


METHODOLOGY ARTICLE

Open Access



Drug-target interactions prediction using marginalized denoising model on heterogeneous networks

Chunyan Tang^{1,2*} , Cheng Zhong^{2*}, Danyang Chen² and Jianyi Wang³

* Correspondence: tangchunyan@gxu.edu.cn; chzhong@gxu.edu.cn

¹School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

²School of Computer, Electronics and Information, Guangxi University, Nanning, China

Full list of author information is available at the end of the article

Abstract

Background: Drugs achieve pharmacological functions by acting on target proteins. Identifying interactions between drugs and target proteins is an essential task in old drug repositioning and new drug discovery. To recommend new drug candidates and reposition existing drugs, computational approaches are commonly adopted. Compared with the wet-lab experiments, the computational approaches have lower cost for drug discovery and provides effective guidance in the subsequent experimental verification. How to integrate different types of biological data and handle the sparsity of drug-target interaction data are still great challenges.

Results: In this paper, we propose a novel drug-target interactions (DTIs) prediction method incorporating marginalized denoising model on heterogeneous networks with association index kernel matrix and latent global association. The experimental results on benchmark datasets and new compiled datasets indicate that compared to other existing methods, our method achieves higher scores of *AUC* (area under curve of receiver operating characteristic) and larger values of *AUPR* (area under precision-recall curve).

Conclusions: The performance improvement in our method depends on the association index kernel matrix and the latent global association. The association index kernel matrix calculates the sharing relationship between drugs and targets. The latent global associations address the false positive issue caused by network link sparsity. Our method can provide a useful approach to recommend new drug candidates and reposition existing drugs.

Keywords: Drug-target interaction, Marginalized denoising model, Drug discovery prediction, Drug repositioning prediction



Background

Identifying drug-target interactions (DTIs) is a critical work in drug discovery and drug repositioning. Although high-throughput screening and other biological assays are becoming available, the experimental methods for DTIs identification remain to be extremely costly. Different computational methods for predicting potential DTIs were proposed in the past decade [1–5].

In 2008, Yamanishi et al. [6] proposed a bipartite network model by integrating the chemical and genomic spaces to predict DTIs for four classes of target proteins, which are enzymes, ion channels (IC), G protein-coupled receptors (GPCR), and nuclear receptors (NR). Based on the four datasets, several methods to improve the accuracy of DTIs prediction were proposed. In early studies, the DTI prediction problem was treated as a binary classification problem. Some classical classifiers such as support vector machines (SVM) and regularized least squares (RLS) were used to predict drug-target interactions. A supervised bipartite local model (BLM) using SVM classifier was proposed to predict drug and target sets respectively [7]. To solve the problem of selecting negative samples, a semi-supervised learning method called Laplacian Regularized Least Squares (LapRLS) was proposed [8]. To analyze the relevance between the network topological information and DTIs prediction, a Gaussian interaction profile (GIP) kernel was defined to capture the topological information in DTIs network. And a Regularized Least Squares (RLS) classifier was employed with GIP kernel to predict DTIs [9]. The methods mentioned above focus on existing drug-target interaction pairs and mainly deal with the old drug reposition problem.

To predict new drugs or targets, a bipartite local model with neighbor-based interaction profile inferring (BLM-NII) was proposed [10]. A weighted nearest neighbor (WNN) profile and the GIP kernel were incorporated to handle new drug compounds [11]. A robust model against the overfitting problem of traditional statistical methods was proposed based on the Random Forest (RF) method [12]. Matrix factorization (MF) method is a feature extraction method widely used in recommendation system [13]. The MF method was used to identify latent features of drugs and targets to handle new drug discovery problem [14–19]. Zheng et al. [15] used collaborative matrix factorization (CMF) to predict potential DTIs. Liu et al. [17] used the logistic matrix factorization and neighborhood information of drugs and targets to predict DTIs. There are also some methods which form the final kernel matrix using the linear combination of two or more kernel matrices [7–9, 15, 17]. Hao et al. [18] combined different kernel matrix with nonlinear kernel diffusion, and employed the diffused kernel matrix with RLS classifier to predict DTIs. Hao et al. [19] integrated logistic matrix factorization and kernel diffusion to improve the accuracy of DTIs prediction. The model based on diffused kernel matrix outperforms the model based on the linearly weighted kernel matrix for DTIs prediction.

To predict more realistic drug-target interactions, some researchers used drug-target binding affinity. Binding affinity indicates the strength of interactions between drug-target pairs. Binding affinity is usually measured by the dissociation constant (K_d), inhibition constant (K_i), or the half maximal inhibitory concentration (IC50). `Krocker_ri` [9, 20] is a method to predict drug-target binding affinity [21]. For more accurate prediction on continuous drug-target binding affinity data, a non-linear method

called SimBoost was proposed by using the gradient boosting regression trees as the learning model [22].

With rapidly development of deep learning, some deep learning frameworks have been applied in the field of drug discovery [23–27]. Stacked auto-encoder was used to construct deep representation of drug-target pairs [24]. Hu et al. [25] used convolutional Neural Network (CNN) to predict DTIs. A new compound-protein interaction (CPIs) prediction approach was developed by combining graphical neural network (GNN) for compounds and convolutional neural network (CNN) for proteins [26]. Based on deep neural network (DNN), Tian et al. [27] proposed a method called DL_CPI to predict large-scale compound-protein interactions. Deep learning method has the advantages in dealing with growing compounds data. But analyzing deep learning models is difficult due to their black-box nature, more effective models are needed to improve the accuracy of DTIs prediction.

From the perspective of networks, the DTIs prediction problem can be treated as a network link prediction problem [28]. Chen et al. [29] developed a model of network-based random walk with restart on heterogeneous networks (NRWRH) to predict potential DTIs. Lan et al. [30] used the models of random walk with restart, k nearest neighbors (k NN), and heat kernel diffusion to label unknown DTIs to predict potential DTIs. Recently, Chen et al. used marginalized denoising model (MDM) to predict hidden or missing links in a given relational matrix by transforming a network link prediction problem to a matrix denoising problem [31]. The MDM-based method can predict new protein-protein interaction in the PPI network better than the MF-based methods. But the MDM-based predicting method has not been applied to the heterogeneous network such as drug-target interactions.

To further improve the accuracy of DTIs prediction, this paper proposes an integrated method using the marginalized denoising model on heterogeneous networks, association index and kernels fusion. We transform the DTIs prediction problem to a noise reduction problem on heterogeneous networks. The heterogeneous network is constructed by combining drug and target kernel matrices and the existing DTIs network. To construct the kernel matrix, we introduce the association index kernel matrix to measure the sharing interaction relationship between drugs and the sharing interaction relationship between targets. The sharing interaction relationship is derived from the common targets between drugs and the common drugs between targets. Furthermore, we not only use the information of associations of the nearest neighbors to perform DTIs prediction, but also incorporate the global association between drugs and targets to reduce the sparsity of DTIs network and improve prediction accuracy.

The rest of this paper is organized as follows. The experimental results are reported in section 2. The discussion of experimental results is given in section 3. The conclusion is given in section 4. The source of the benchmark dataset selected and new compiled dataset, construction of the matrices of similarity between drugs and similarity between targets, MDM model and our proposed prediction method are described in section 5.

Results

Similar to the previous studies [11, 15, 17, 19], we conducted the experiments by five trials of 10-fold cross-validation (CV). We employed the area under curve of receiver

operating characteristic (*AUC*) and area under precision-recall curve (*AUPR*) as the evaluation metrics. To valid our prediction method in drug reposition, in completely new drug discovery, and in completely new targets discovery respectively, we conducted the cross-validation under the following three settings:

- (1) CVP (cross-validation based on the drug-target interaction pairs): Validating for drug reposition. 90% of the drug-target interaction pairs in drug-target interaction network Y were randomly selected as training data, and the left 10% of the drug-target interaction pairs were selected as testing data. The CVP setting is used to verify the performance of the prediction method in drug reposition.
- (2) CVD (cross-validation based on the drugs): Validating for new drug in known targets. 90% of rows (drugs) in Y were randomly selected as training data, and the left 10% of rows (drugs) were selected as testing data. The CVD setting is used to verify the performance of the prediction method in new drug discovery.
- (3) CVT (cross-validation based on the targets): Validating for new target in known drugs. 90% of columns (targets) in Y were randomly selected as training data, and the left 10% of columns (targets) were selected as testing data. The CVT setting is used to verify the performance of the prediction method in new target discovery.

We evaluated our method DTIP_MDHN and three existing DTIs prediction methods BLM-NII [10], RLS-WNN [11], NRLMF [17], and DNILMF [19] on the benchmark datasets and the new dataset1. BLM-NII method integrated BLM model with neighbor-based interaction profile to handle the new drugs/targets problem. RLS-WNN is a GIP-based prediction method with a weighted nearest neighbor profile for predicting new drug compounds. NRLMF and DNILMF are two MF-based prediction methods. We implemented algorithm DTIP_MDHN in MATLAB. The experiment was conducted at the high-performance computing center of Guangxi University.¹

We first evaluated our method DTIP_MDHN and other four methods BLM-NII, RLS-WNN, NRLMF and DNILMF in terms of *AUC* and *AUPR* on benchmark data. To verify the performance of the prediction methods in drug reposition, we conducted the experiment under CVP setting. The experimental results are shown in Table 1. In addition, to verify the performance of prediction methods in new drug/target discovery, we conducted the experiment under CVR and CVC settings. The experimental results are shown in Tables 2 and 3, respectively.

We can see from Tables 1, 2 and 3 that compared with other four methods, our method DTIP_MDHN achieves better results for *AUC* metric on the benchmark datasets under CVP and CVD settings, and obtains higher scores on IC, GPCR and NR datasets under CVT setting. For *AUPR* metric, DTIP_MDHN outperforms other four methods on all datasets under CVP and CVD settings, and achieves higher scores on IC, GPCR and NR datasets under CVT setting. The GPCR and NR datasets are sparser than Enzyme and IC datasets, so the prediction accuracy on GPCR and NR datasets is always lower in previous study. In our method DTIP_MDHN, the Jaccard index is introduced to measure the sharing interaction relationship between drugs and targets, the indirect interactions are introduced by global association to solve the data sparsity

¹<http://hpc.gxu.edu.cn>

Table 1 AUC and AUPR scores of five methods under CVP setting

Dataset	Method	AUPR	AUC
Enzyme	BLM-NII	0.7560	0.9792
	RLS-WNN	0.7160	0.9640
	NRLMF	0.8920	0.9870
	DNILMF	0.9220	0.9890
	DTIP_MDHN	0.9609	0.9970
Ion Channel (IC)	BLM-NII	0.8256	0.9810
	RLS-WNN	0.7170	0.9590
	NRLMF	0.9060	0.9890
	DNILMF	0.9380	0.9900
	DTIP_MDHN	0.9744	0.9976
GPCR	BLM-NII	0.5420	0.9550
	RLS-WNN	0.5200	0.9440
	NRLMF	0.7490	0.9690
	DNILMF	0.8120	0.9750
	DTIP_MDHN	0.9540	0.9957
Nuclear Receptor (NR)	BLM-NII	0.6740	0.9153
	RLS-WNN	0.5890	0.9010
	NRLMF	0.7280	0.9500
	DNILMF	0.7510	0.9550
	DTIP_MDHN	0.8626	0.9913

The best results in each column are in **bold**

Table 2 AUC and AUPR scores of five methods under CVD setting

Dataset	Methods	AUPR	AUC
Enzyme	BLM-NII	0.2568	0.8230
	RLS-WNN	0.2780	0.8820
	NRLMF	0.3580	0.8710
	DNILMF	0.7960	0.9640
	DTIP_MDHN	0.8378	0.9834
Ion Channel (IC)	BLM-NII	0.3310	0.7973
	RLS-WNN	0.2580	0.7970
	NRLMF	0.3440	0.8130
	DNILMF	0.8220	0.9610
	DTIP_MDHN	0.8587	0.9845
GPCR	BLM-NII	0.3250	0.8315
	RLS-WNN	0.2950	0.8910
	NRLMF	0.3640	0.8950
	DNILMF	0.7810	0.9670
	DTIP_MDHN	0.8487	0.9864
Nuclear Receptor (NR)	BLM-NII	0.4389	0.8010
	RLS-WNN	0.5040	0.8900
	NRLMF	0.5450	0.9000
	DNILMF	0.7760	0.9560
	DTIP_MDHN	0.8463	0.9917

The best results in each column are in **bold**

Table 3 *AUC* and *AUPR* scores of five methods under CVT setting

Dataset	Methods	<i>AUPR</i>	<i>AUC</i>
Enzyme	BLM-NII	0.7376	0.9190
	RLS-WNN	0.5660	0.9470
	NRLMF	0.8120	0.9660
	DNILMF	0.8890	0.9780
	DTIP_MDHN	0.8848	0.9463
Ion Channel (IC)	BLM-NII	0.7658	0.9153
	RLS-WNN	0.6960	0.9500
	NRLMF	0.7850	0.9640
	DNILMF	0.8870	0.9700
	DTIP_MDHN	0.9092	0.9705
GPCR	BLM-NII	0.3532	0.7781
	RLS-WNN	0.5500	0.9260
	NRLMF	0.5560	0.9300
	DNILMF	0.6840	0.9330
	DTIP_MDHN	0.8865	0.9593
Nuclear Receptor (NR)	BLM-NII	0.4523	0.5430
	RLS-WNN	0.5310	0.9350
	NRLMF	0.4490	0.8510
	DNILMF	0.4830	0.8560
	DTIP_MDHN	0.8113	0.9823

The best results in each column are in **bold**

problem. The experimental result indicates that our method can improve the prediction accuracy, and it is more suitable for predicting DTIs on more sparse datasets such as GPCR and NR under CVT setting.

Constructing the final kernel matrices of drugs and targets is a key step to predict latent DTIs. To compare the effects of different final kernel matrices on the DTIs prediction results, we evaluated our constructed final kernel matrices *KFJD* and *KFJT* with other two final kernel matrices in GIP [9] and DNILMF [19], in terms of *AUC* and *AUPR* on benchmark data. In Table 4, we denote the final kernel matrices of drugs and targets constructed from GIP [9] as *KGD* and *KGT* respectively, the final kernel matrices of drugs and targets constructed from DNILMF [19] as *KFD* and *KFT* respectively. Table 4 shows the scores of *AUC* and *AUPR* of DTIP_MDHN using these kernel matrices under CVP setting.

The experimental results in Table 4 indicate that our constructed final kernel matrices of drugs and targets *KFJD* and *KFJT* indeed leads to more accurate predictions in our method DTIP_MDHN than the final kernel matrices in GIP [9] and DNILMF [19].

Next, we evaluated our proposed prediction model with other machine learning models, such as supervised learning models SVM and RF, and Matrix Factorization (MF) model. We extracted our constructed final kernel matrices *KFJD/KFJT* as the features of drug-target pairs, drug-target interaction matrix *Y* as the classification labels for supervised learning prediction models. We used BLM [7] as the SVM-based method, DDR [12] as the RF-based method, and DNILMF [19] as the MF-based method. While the scores of

Table 4 AUC and AUPR of DTIP_MDHN using 3 kernel matrices under CVP setting

Dataset	final kernel matrices	AUPR	AUC
Enzyme	KGD/KGT	0.8540	0.9831
	KFD/KFT	0.9480	0.9867
	KFJD/KFJT	0.9738	0.9995
Ion Channel (IC)	KGD/KGT	0.8735	0.9904
	KFD/KFT	0.9482	0.9917
	KFJD/KFJT	0.9700	0.9994
GPCR	KGD/KGT	0.8660	0.9812
	KFD/KFT	0.9480	0.9973
	KFJD/KFJT	0.9651	0.9990
Nuclear Receptor (NR)	KGD/KGT	0.7483	0.9867
	KFD/KFT	0.8086	0.9856
	KFJD/KFJT	0.8315	0.9988

The best results in each column are in **bold**

AUPR and AUC were calculated under CVP setting. Table 5 shows the scores of AUC and AUPR for four prediction models.

The experimental results in Table 5 indicate that our proposed prediction model achieves higher scores of AUC and AUPR than SVM, RF, and MF models in DTIs prediction.

In addition to the final kernel matrices of drugs and targets, there are two key parameters in DTIP_MDHN. One is the noise value (*noise*), and another one is the dimension of latent layer (*k*). We evaluated how the values of *noise* and *k* affect the scores of AUC and AUPR for DTIP_MDHN on the benchmark datasets respectively. The *noise* is set to 0.65, 0.75, 0.85 and 0.95, and *k* is set to the value in range [10, 150] according to the

Table 5 AUC and AUPR scores of four prediction models

Dataset	Method	AUPR	AUC
Enzyme	BLM	0.9552	0.9890
	DDR	0.9457	0.9849
	DNILMF	0.9367	0.9939
	DTIP_MDHN	0.9609	0.9970
Ion Channel (IC)	BLM	0.8814	0.9891
	DDR	0.9535	0.9914
	DNILMF	0.9499	0.9926
	DTIP_MDHN	0.9744	0.9976
GPCR	BLM	0.8344	0.9716
	DDR	0.8224	0.9841
	DNILMF	0.8353	0.9804
	DTIP_MDHN	0.9543	0.9957
Nuclear Receptor (NR)	BLM	0.5949	0.8489
	DDR	0.8302	0.9431
	DNILMF	0.7993	0.9727
	DTIP_MDHN	0.8626	0.9913

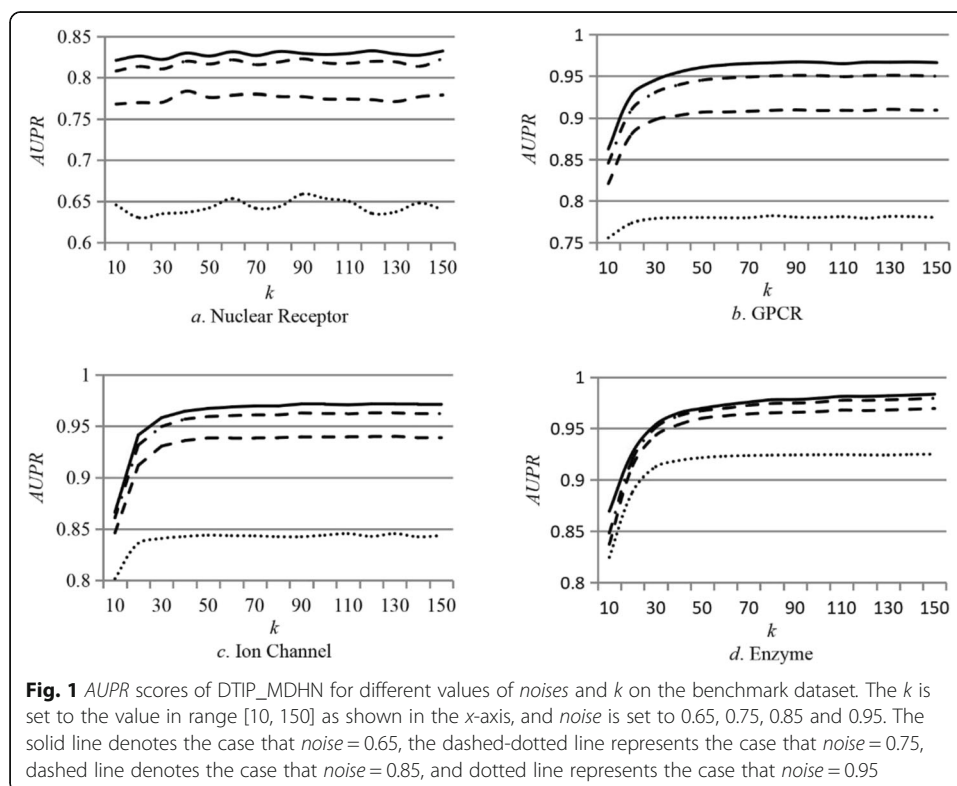
The best results in each column are in **bold**

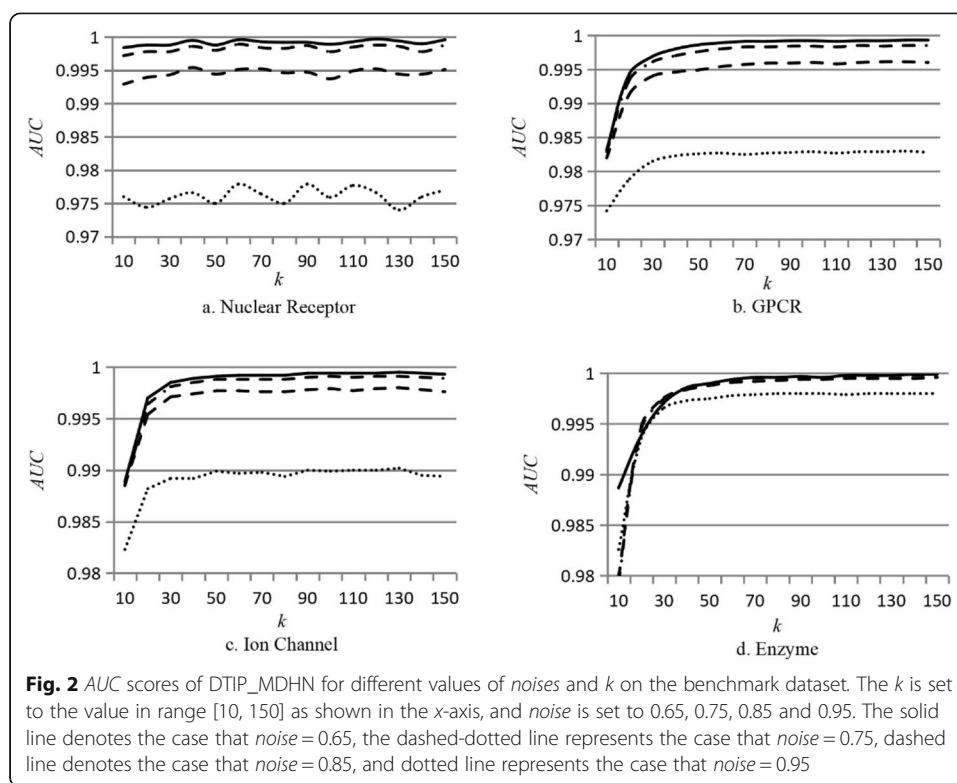
setting in MDM [31]. The experimental results on four datasets are shown in Figs. 1 and 2 respectively, where the solid line denotes the case that $noise = 0.65$, the dashed-dotted line represents the case that $noise = 0.75$, dashed line denotes the case that $noise = 0.85$, and dotted line represents the case that $noise = 0.95$.

From Figs. 1 and 2 we can see that DTIP_MDHN obtains the highest scores of AUC and $AUPR$ on the four datasets when $noise = 0.65$. The dimension of latent layer (k) indicates the degree of dimensionality reduction in the Auto-Encode (AE). The key information will lose from original data if k is too small. The non-critical and redundant information still exists if the value of k is too large. In general, the choice of value of k depends on the dimension of different datasets. By analyzing the results in Figs. 1 and 2, we set the value of k according to the number of drugs for different datasets. Table 6 shows the values of k and $noise$ on the benchmark datasets.

To verify the validity of DTIP_MDHN method, we sort the new drug-target interaction pairs predicted by DTIP_MDHN in descending order of the prediction scores and obtain top 5 of the scores for Enzyme, IC, GPCR and NR respectively. If a new drug-target interaction is validated in the current version of KEGG [32], SuperTarget [33], DRUGBANK [34], and ChEMBL [35], the “Validated” item is labeled by “yes”; otherwise it is labeled by “No”. Table 7 shows the top 5 of new drug-target interactions predicted by DTIP_MDHN on the benchmark datasets.

As shown in Table 7, the top 5 of new drug-target interactions for Enzyme dataset are validated in current databases. 3 of the top 5 new drug-target interactions for IC and GPCR datasets are validated in current databases respectively. 2 of the top 5 new drug-target interactions for NR dataset are validated in current databases. The statistics for the “Validated” item in Table 10 shows that, the hit rate of prediction for all the





four datasets is about 75%. In fact, the NR dataset is the most challenging dataset for DTIs prediction because it is the sparsest dataset among benchmark datasets [6, 8, 18].

We further analyze the no-validated DTI pairs in NR dataset. From Table 7 we can see that the top one of predicted items in NR dataset is a DTI pair between D00316 (Etretinate) and hsa6096 (ROR β). The study in [36] indicated that several retinoids bind to ROR β (hsa6096) to provide a novel pathway for retinoid action. As Etretinate is an aromatic retinoid, a second-generation retinoid, there is a high probability of interaction between Etretinate and ROR β . For the fifth item in NR dataset, D01115 (Eplere-none) is predicted to interact with a Glucocorticoid receptor (hsa2908). Although the interaction between D01115 and hsa2908 has not been found in the current version of KEGG, DRUGBANK, ChEMBL and SuperTarget, an antagonist activity assay confirms this interaction result in PubChem BioAssay ID: AID 761383 from ChEMBL [37].

The benchmark datasets were generated in 2008. Many new interactions are appended to the current version of the KEGG [32], SuperTarget [33], DrugBank [34], and BRENDA [38] nowadays. To enhance the diversity of experimental dataset and inspect the performance of our proposed method on the new database, we used the new dataset1 from KEGG to perform DTIs prediction. Following the category in KEGG, the

Table 6 Values of *k* and *noise* on the benchmark datasets

Dataset	Number of drugs	<i>k</i>	<i>noise</i>
Enzyme	445	100	0.65
Ion Channel	210	60	0.65
GPCR	223	60	0.65
Nuclear Receptor	54	20	0.65

Table 7 Top 5 Interactions predicted by DTIP_MDHN on the benchmark datasets

Dataset	KEGG Drug ID	Drug name	KEGG Has ID	Uniport ID	Gene name	Validated
Enzyme	D00542	Halothane	has:1571	P05181	CYP2E1	Yes
	D00139	Methoxsalen	has:1543	P04798	CYP1A1	Yes
	D00437	Nifedipine	has:1559	P11712	CYP2C9	Yes
	D00410	Metrapone	has:1543	P04798	CYP1A1	Yes
	D00574	Aminoglutethimide	has:1589	P08686	CYP21A2	Yes
Ion Channel	D03365	Nicotine	has:1137	P43681	CHRNA4	Yes
	D00640	Propafenone hydrochloride	has:6336	Q9Y5Y9	SCN10A	Yes
	D02098	Proparacaine hydrochloride	has:8645	O95279	KCNK5	No
	D02356	Verapamil	has:2893	P48058	GRIA4	No
	D00552	Benzocaine	has:6331	Q14524	SCN5A	Yes
GPCR	D00683	Albuterol sulfate	has:153	P08588	ADRB1	Yes
	D02359	Ritodrine	has:153	P08588	ADRB1	No
	D02147	Albuterol	has:153	P08588	ADRB1	Yes
	D01386	Ephedrine hydrochloride	has:153	P08588	ADRB1	Yes
	D00604	Clonidine hydrochloride	has:148	P35348	ADRA1A	No
Nuclear Receptor	D00316	Etretinate	has:6096	Q58EY0	ROR β	No
	D01132	Tazarotene	has:6097	P51449	ROR γ	No
	D00182	Norethindrone	has:2099	P03372	ESR1	Yes
	D00348	Isotretinoin	has: 5915	P10826	RARB	Yes
	D01115	Eplerenone	has:2908	P04150	NR3C1	No

target proteins can be divided into 8 datasets. In addition to the datasets of Enzyme, IC, GPCR and NR, the 4 new datasets are protein kinase (PK), transporter (TR), cell surface molecule and ligand (CSM), cytokine and cytokine receptor (CR). After deleting the redundant and invalid data, we compiled the new datasets with 11,912 known interactions linking 4495 unique drugs and 959 unique targets. We conducted the experiment to evaluate our method DTIP_MDHN and the newest MF-based method DNILMF. Some drugs may act on two or more different types of targets. For example, Cocaine (D00110) can act on SCN9A (hsa6335) which belongs to Ion channels, and can act on SLC6A2 (hsa6530) which is belongs to Transporters. So, we added a dataset containing all 8 classes of target proteins on KEGG as input in the experiment. This dataset is denoted as “ALL”.

Table 8 shows the *AUC* and *AUPR* scores for two prediction methods on the new datasets1 under CVP setting, in which DNILMF used the optimized parameters ($numLatent = 90$, $c = 20$, $thisAlpha = 0.7$, $\lambda_u = 10$, $\lambda_v = 10$, $K = 2$) for Enzyme and “ALL” datasets, used the parameters ($numLatent = 90$, $c = 6$, $thisAlpha = 0.4$, $\lambda_u = 2$, $\lambda_v = 2$, $K = 2$) for the other datasets, and DTIP_MDHN used the parameters $noise = 0.65$ and the value of k in Table 6.

From Table 8, we can see that for the new dataset1 of Enzyme, IC, GPCR, and NR, the scores of *AUC* and *AUPR* computed by DNILMF and DTIP_MDHN are basically the same as that for the benchmark datasets. For the datasets of protein kinase, transporter, cell surface molecule and ligand, cytokine and cytokine receptor, the scores of *AUPR* and *AUC* are mostly about 0.9 and 0.99 respectively. For the “ALL” dataset, the

Table 8 AUC and AUPR for DNILMF and DTIP_MDHN on new dataset1 under CVP setting

Dataset	Method	AUPR	AUC
Enzyme ^a	DNILMF	0.9245	0.9950
	DTIP_MDHN	0.9071	0.9911
Ion Channel (IC)	DNILMF	0.9921	0.9991
	DTIP_MDHN	0.9968	0.9998
GPCR	DNILMF	0.9239	0.9935
	DTIP_MDHN	0.9615	0.9963
Nuclear Receptor (NR)	DNILMF	0.9341	0.9897
	DTIP_MDHN	0.9610	0.9910
protein kinase	DNILMF	0.8713	0.9875
	DTIP_MDHN	0.9408	0.9959
transporter	DNILMF	0.8852	0.9907
	DTIP_MDHN	0.9523	0.9978
cytokine and cytokine receptor	DNILMF	0.8166	0.9827
	DTIP_MDHN	0.8630	0.9842
cell surface molecule and ligand	DNILMF	0.8557	0.9817
	DTIP_MDHN	0.9076	0.9887
ALL ^a	DNILMF	0.7578	0.9813
	DTIP_MDHN	0.9743	0.9978

^aoptimized parameters ($numLatent = 90$, $c = 20$, $thisAlpha = 0.7$, $\lambda_u = 10$, $\lambda_v = 10$, $K = 2$) were used in DNILMF

score of AUPR is about 0.97 for DTIP_MDHN, and only about 0.63 for DNILMF. The “ALL” dataset is much sparser than any single dataset because the “ALL” dataset treats the interaction information on the above 8 datasets as a whole one. In DNILMF method, only the local neighborhood information is used to measure the similarity between drugs and targets. In DTIP_MDHN method, the global association is exploited to represent the indirect association relationship between drugs and targets, the influence of link sparsity is reduced, and the prediction accuracy is improved.

To verify the availability of our proposed method in binding affinity prediction, we evaluated our method DTIP_MDHN and two existing binding affinity prediction methods Kronecker_rls [9, 20] and SimBoost [22] on the David and Metz datasets, in terms of AUPR, AUC and concordance index (CI) [21, 39]. Kronecker_rls [9, 20] is a DTIs prediction method that first be used to predict binding affinity [21]. SimBoost is a supervised learning model and selects the gradient boosting regression trees to predict continuous binding affinity.

To measure with AUPR and AUC, the quantitative datasets were binarized by using relatively stringent cut-off thresholds ($K_d < 30$ nM and $K_i < 28.18$ nM) [21]. It means that if $K_d < 30$ nM or $K_i < 28.18$ nM, the affinity value is set to 1, otherwise, the affinity value is set to 0. To measure with the continuous values of K_d and K_i , the concordance index (CI) was used as an evaluation metric [21, 39].

For the continuous values of K_d and K_i , we use Pearson Correlation Coefficient (PCC) instead of the Jaccard index to calculate the association index kernel because Jaccard kernel matrix works well on binary interaction matrix, while PCC is originally developed to measure the relationship between two continuous variables [40] and can be calculated with the function *corrcoef()* in MATLAB. The values of CI calculated by

DTIP_MDHN with PCC kernel are denoted as CI_{PCC} in the following Tables 9, 10, 11, 12, 13 and 14.

Table 9 shows the scores of $AUPR$, AUC , and CI for three prediction methods Kronecker_rls [21], SimBoost [22], and our method DTIP_MDHN. Tables 10 and 11 shows the scores of $AUPR$, AUC , and CI for Kronecker_rls and our method DTIP_MDHN under CVD and CVT setting, respectively. Because the supervised learning methods do not distinguish CVP, CVD and CVT setting, the comparison with SimBoost method was not included in Tables 10 and 11. These results in Tables 9, 10 and 11 were based on the protein target normalized SW sequence similarity and compound drug 2-dimensional structural similarity. We used the default parameters $noise = 0.65$ and $k = 60$ for $AUPR$, AUC and CI , and used parameters $noise = 0.95$ and $k = 60$ for CI_{PCC} in our method DTIP_MDHN.

From Tables 9, 10 and 11, we can see that our method DTIP_MDHN has higher scores of $AUPR$ and AUC than Kronecker_rls and SimBoost. We can also see that SimBoost has higher score of CI than DTIP_MDHN and Kron_rls on David dataset, and Kron_rls has higher score of CI than DTIP_MDHN and SimBoost on Metz dataset under CVP setting in Table 9. Kron_rls has higher score of CI than DTIP_MDHN on David dataset, but DTIP_MDHN has higher score of CI than Kron_rls on Metz dataset under CVD and CVT setting in Tables 10 and 11 respectively. For DTIP_MDHN, the scores of CI_{PCC} are higher than that of CI on David and Metz datasets under CVP and CVT setting and on David dataset under CVD setting. This illustrates that PCC kernel matrix achieves higher accuracy than Jaccard kernel matrix in predicting drug-target binding affinity.

Next, we evaluate the prediction accuracy of DTIP_MDHN and Kronecker_rls with different chemical structure and sequence similarity kernels on David dataset in terms of $AUPR$, AUC and CI under different CV setting. We denote the two-dimensional Tanimoto coefficients similarity kernel matrix as 2D, denote three-dimensional Tanimoto coefficients similarity kernel matrix as 3D, and denote the extended-connectivity fingerprint ECFP4 similarity kernel matrix as ECFP4 in Tables 12, 13 and 14. The value of CI calculated by DTIP_MDHN with PCC kernel labeled as CI_{PCC} . Kronecker_rls uses the default parameters and DTIP_MDHN uses $noise = 0.65$ and $k = 60$ for $AUPR$, AUC , and CI , and DTIP_MDHN uses $noise = 0.95$ and $k = 60$ for CI_{PCC} .

From the experimental results shown in Tables 12, 13 and 14, we can see that in terms of $AUPR$ and AUC , our method DTIP_MDHN outperforms over Kronecker_rls method for all three similarity kernels under all three CV setting. In terms of CI , DTIP_MDHN with PCC kernel achieves better prediction accuracy than DTIP_MDHN

Table 9 $AUPR$, AUC , and CI for three binding affinity prediction methods under CVP setting

Dataset	Method	$AUPR$	AUC	CI	CI_{PCC}
David	Kronecker_rls	0.6586	0.9388	0.8740	–
	SimBoost	0.7580	0.9560	0.8840	–
	DTIP_MDHN	0.7706	0.9671	0.8623	0.8740
Metz	Kronecker_rls	0.5720	0.9340	0.9340	–
	SimBoost	0.6290	0.9580	0.8510	–
	DTIP_MDHN	0.8303	0.9960	0.8702	0.8812

The best results in each column are in **bold**. CI_{PCC} is the values of CI calculated by DTIP_MDHN with PCC kernel

Table 10 *AUPR*, *AUC*, and *CI* for two binding affinity prediction methods under CVD setting

Dataset	Method	<i>AUPR</i>	<i>AUC</i>	<i>CI</i>	<i>CI_PCC</i>
David	Kronecker_rls	0.2203	0.7055	0.6981	–
	DTIP_MDHN	0.6896	0.8892	0.5489	0.8014
Metz	Kronecker_rls	0.4301	0.8596	0.7244	–
	DTIP_MDHN	0.7648	0.9947	0.8563	0.8043

The best results in each column are in **bold**. *CI_PCC* is the values of *CI* calculated by DTIP_MDHN with *PCC* kernel

with Jaccard kernel in most cases. This illustrates that *PCC* is more suitable for continuous variable correlation comparison by different similarity kernels. DTIP_MDHN with *PCC* kernel gains better result than Kronecker_rls with 2-dimensional and 3-dimensional drug similarity kernels. Kronecker_rls with ECFP4 fingerprint drug similarity kernels gain better result than DTIP_MDHN with *PCC* kernel.

To inspect our method DTIP_MDHN for large-scale compound-protein interaction (CPIs) prediction, we evaluated DTIP_MDHN with BLM [7] and a deep learning model combining GNN and CNN [26] on new database 2. The new database 2 contains CPIs of *Homo sapiens* retrieved from STITCH database (Version 5.0) [41]. STITCH database contains a comprehensive resource for both known and predicted interactions of compounds and proteins. In order to ensure the accuracy of CPIs data, we extracted the CPIs interactions with combined scores greater than 900 from CPIs of *Homo sapiens* interactions. It means the CPIs interactions that we used in our experiment have the interaction probability greater than 90%. The experimental data contains 13,286 drugs, 5313 targets, and 116,199 interactions. The detailed compound protein interaction information can be referred to Additional file 1. Table 15 shows the values of *AUC* and *AUPR* for DTIP_MDHN, BLM, and GNN&CNN on STITCH dataset under CVP setting.

From Table 15, we can see that DTIP_MDHN obtains higher score of *AUC* than BLM and GNN&CNN in large-scale CPIs prediction, which indicates that our method DTIP_MDHN can identify true negatives from the testing data more accurate than BLM and GNN&CNN methods. On the other hand, we can also see that GNN&CNN achieves higher score of *AUPR* than BLM and DTIP_MDHN. This is because GNN&CNN has high sensitivity with reliable negative samples.

Discussion

In this paper, we propose a novel drug-target interactions (DTIs) prediction method incorporating marginalized denoising model on heterogeneous networks with association index kernel matrix and latent global association. We combine the chemical structure similarity matrix of drugs, the sequence similarity matrix of targets with the GIP kernel

Table 11 *AUPR*, *AUC*, and *CI* for two binding affinity prediction methods under CVT setting

Dataset	Method	<i>AUPR</i>	<i>AUC</i>	<i>CI</i>	<i>CI_PCC</i>
David	Kronecker_rls	0.5012	0.8912	0.8037	–
	DTIP_MDHN	0.7919	0.9701	0.7875	0.8293
Metz	Kronecker_rls	0.2729	0.8355	0.6292	–
	DTIP_MDHN	0.7473	0.9541	0.7803	0.8045

The best results in each column are in **bold**. *CI_PCC* is the values of *CI* calculated by DTIP_MDHN with *PCC* kernel

Table 12 AUPR, AUC, and CI for different kernels under CVP setting

Kernel	Method	AUPR	AUC	CI	CI_PCC
2D	Kronecker_rls	0.6586	0.9388	0.8740	–
	DTIP_MDHN	0.7706	0.9471	0.8623	0.8740
3D	Kronecker_rls	0.6642	0.9419	0.8778	–
	DTIP_MDHN	0.7712	0.9474	0.8821	0.8919
ECFP4	Kronecker_rls	0.6654	0.9444	0.8793	–
	DTIP_MDHN	0.7654	0.9457	0.8856	0.9020

The best results in each column are in **bold**. CI_PCC is the values of CI calculated by DTIP_MDHN with PCC kernel

matrix and the association index kernel matrix to construct final kernel matrix. We use the association index kernel matrix to enhance the relevance between drugs and targets by calculating the sharing association between drugs and targets. In the building model step, we build a heterogeneous network with drug kernel matrix, target kernel matrix, and existing drug-target interaction network to construct global links for drugs, targets and known drug-target interactions, and further extract latent global associations from the heterogeneous network. The latent global associations between drugs and targets are important to reduce the data sparsity.

The experimental results on benchmark dataset show that our proposed prediction method outperforms the existing binary classification predicting methods and MF-based predicting methods in term of AUC and AUPR. Specifically, for the sparser datasets such as GPCR and NR, the prediction accuracy of our method is increased of 10% ~ 20% than other comparative methods. To compare the effects of different final kernel matrices on the DTIs prediction results, we evaluated our constructed final kernel matrices with other two final kernel matrices in GIP [9] and DNILMF [19]. The experimental results indicate that our constructed final kernel matrices of drugs and targets indeed leads to more accurate predictions than the final kernel matrices in GIP [9] and DNILMF [19]. To evaluate our proposed prediction model with supervised learning models SVM, RF, and Matrix Factorization (MF) model DNILMF, we extracted our constructed final kernel matrices *KFJD/KFJT* as the features of drug-target pairs, drug-target interaction matrix *Y* as the classification labels. The experimental results show that our proposed prediction model achieves higher predictions accuracy than SVM, RF, and DNILMF in DTIs prediction. We also evaluated the key parameters *noise* and *k* within a certain value range to optimize the prediction accuracy. The results show that DTIP_MDHN obtains higher predictions accuracy on the four datasets when *noise* = 0.65, and the optimized value of *k* vary with the number of drugs for different datasets.

Table 13 AUPR, AUC, and CI for different kernels under CVD setting

Kernel	Method	AUPR	AUC	CI	CI_PCC
2D	Kronecker_rls	0.2203	0.7055	0.6981	–
	DTIP_MDHN	0.6896	0.8892	0.5489	0.8014
3D	Kronecker_rls	0.3308	0.7700	0.7441	–
	DTIP_MDHN	0.7044	0.8970	0.5547	0.7470
ECFP4	Kronecker_rls	0.3117	0.7487	0.7504	–
	DTIP_MDHN	0.7138	0.9028	0.6990	0.7350

The best results in each column are in **bold**. CI_PCC is the values of CI calculated by DTIP_MDHN with PCC kernel

Table 14 *AUPR*, *AUC*, and *CI* for different kernels under CVT setting

Kernel	Method	<i>AUPR</i>	<i>AUC</i>	<i>CI</i>	<i>CI_PCC</i>
2D	Kronecker_rls	0.5010	0.8912	0.8037	–
	DTIP_MDHN	0.7919	0.9701	0.7875	0.8293
3D	Kronecker_rls	0.5494	0.9045	0.8192	–
	DTIP_MDHN	0.7903	0.9708	0.7854	0.8344
ECFP4	Kronecker_rls	0.6024	0.9201	0.8408	–
	DTIP_MDHN	0.7942	0.9700	0.8313	0.8354

The best results in each column are in **bold**. *CI_PCC* is the values of *CI* calculated by DTIP_MDHN with PCC kernel

To enhance the diversity of experiment data and inspect the performance of our proposed method on the new database, we evaluated our method DTIP_MDHN and the method DNLMF for the 8 classes of target proteins extracted from the current KEGG BRITE database. The experimental results also show that the scores of *AUC* and *AUPR* of DTIP_MDHN are higher than that of DNLMF on the compiled new DTIs database.

The experimental results on Davis and Metz datasets show that our method also can improve the accuracy for predicting drug-target binding affinity. For the continuous values of K_d and K_i , we evaluated our method with two association index method, Pearson Correlation Coefficient (PCC) and Jaccard index, respectively. The experimental results show that PCC is more suitable to measure the relationship between two continuous variables, while Jaccard kernel matrix works well on binary interaction matrix.

To inspect our method DTIP_MDHN for large-scale compound-protein interaction (CPIs) prediction, we evaluated DTIP_MDHN with BLM [7] and a deep learning model combining GNN and CNN [26] on our new dataset 2. The experimental dataset contains 13,286 drugs, 5313 targets, and 116,199 interactions. This dataset is much sparser than the benchmark dataset and new dataset 1. The experimental results indicate that our method DTIP_MDHN can identify true negatives from the sparse dataset more accurately than other comparative methods.

Conclusion

The performance improvement in our method depends on the association index kernel matrix and latent global association. The association index kernel matrix calculates the sharing relationship between drugs and targets. The latent global association addresses the false positive issue caused by network link sparsity. Our method can provide a useful approach to recommend new drug candidates and reposition existing drugs.

The features of a drug-target pair can be characterized more accurately by the biologic physicochemical properties. One future research direction is to use the key biologic physicochemical properties with feature selection method to improve similarity

Table 15 *AUC* and *AUPR* on STITCH dataset under CVP setting

Method	<i>AUPR</i>	<i>AUC</i>
BLM	0.4856	0.9078
DTIP_MDHN	0.8125	0.9850
GNN&CNN	0.8367	0.9460

The best results in each column are in **bold**

measurement in pharmacology, and extend our method to predict potential interaction relationship in other biologic interaction networks that play a part in pharmacology. Meanwhile, with application of deep learning in the field of drug discovery [23–27], it is also another future research direction for predicting drug-target interactions using deep learning framework on multiple information including biologic physicochemical properties.

Methods

Problem description

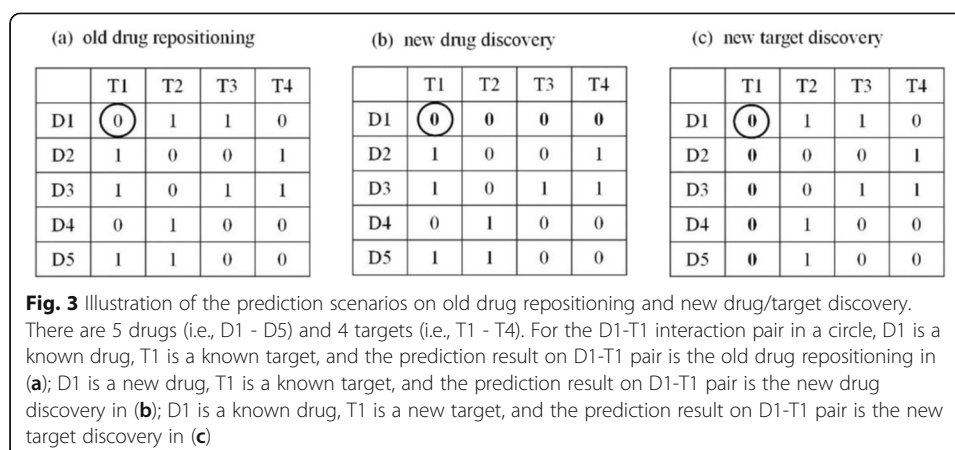
Given a set of drugs $D = \{d_1, d_2, \dots, d_n\}$ and a set of target proteins $T = \{t_1, t_2, \dots, t_m\}$, a drug similarity matrix $SD \in \mathbb{R}^{n \times n}$, a target similarity matrix $ST \in \mathbb{R}^{m \times m}$, and a matrix of known interactions $Y \in \mathbb{R}^{n \times m}$ between drugs and targets are defined, where n is the number of drugs, m is the number of target proteins, and each item $Y_{ij} \in \{0, 1\}$, $i = 1, 2, \dots, n$, and $j = 1, 2, \dots, m$. If drug d_i has a known interaction with target t_j , the value of Y_{ij} is 1, otherwise is 0, $i = 1, 2, \dots, n$, and $j = 1, 2, \dots, m$. The goal of drug-target interactions (DTIs) prediction is to recommend new drug- target pairs using above three matrices and other source of information.

The prediction of old drug repositioning is to predict the interaction probability of drug and target when drug and target are known but drug has no known interaction with target. The prediction of new drug/target discovery is to predict the interaction probability of drug and target when drug is newly developed and target is a known protein or a protein target is newly identified and drug is a known compound.

We illustrated the prediction scenarios on old drug repositioning, new drug/target discovery in Fig. 3. There are 5 drugs (i.e., D1 - D5) and 4 targets (i.e., T1 - T4) in Fig. 3. For the D1-T1 interaction pair in a circle, D1 is a known drug, T1 is a known target, and the prediction result on D1-T1 pair is the old drug repositioning in Fig. 3a; D1 is a new drug, T1 is a known target, and the prediction result on D1-T1 pair is the new drug discovery in Fig. 3b; D1 is a known drug, T1 is a new target, and the prediction result on D1-T1 pair is the new target discovery in Fig. 3c [19].

Datasets

The benchmark datasets were originally provided by Yamanishi et al. [6]. The datasets are publicly available at <http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>. Protein sequences of targets were obtained from the KEGG GENES database [32]. The target similarity matrix is composed of the sequence similarity score between proteins, and it is computed by a normalized version of Smith-Waterman score [42]. Chemical compounds were obtained from the KEGG DRUG and COMPOUND databases [32]. The drug similarity matrix is composed of the chemical structure similarity score between drugs, and it is computed by the SIMCOMP tool [43]. The drug-target interaction matrix is composed of the known drug-target interaction pairs retrieved from databases of KEGG BRITE [32], SuperTarget [33], DrugBank [34], and BRENDA [38]. The benchmark datasets contain four datasets. The first one is enzymes containing 445 drugs and 664 targets. The second one is ion channels (IC) containing 210 drugs and 204 targets. The third one is G-protein coupled receptors (GPCR) containing 223 drugs and 95



targets. And the last one is nuclear receptors (NR) containing 54 drugs and 26 targets. Table 16 lists the statistics for the benchmark datasets [6].

In the past decade, an exponential growth of chemical biology data available in the public databases, such as KEGG [32], SuperTarget [33], Drugbank [34], ChEMBL [35], and STITCH [41]. To enhance the diversity of experimental datasets and inspect our proposed predicting method for the latest database, we extracted two new DTIs datasets from KEGG and STITCH respectively.

For new dataset 1, we obtained the classification information of drugs based on the “target-based classification of drugs” in the KEGG BRITE database,² including 8 datasets which are enzymes, ion channels (IC), G protein-coupled receptors (GPCR), nuclear receptors (NR), Cytokines and receptors (CR), Cell surface molecules and ligands (CSM), Protein kinases (PK), and Transporters (TR). The chemical structure similarity matrix of drugs is computed by the SIMCOMP2 tool.³ Protein sequence similarity matrix of targets is composed of the scores derived from KEGG SSDB Paralog database. After deleting the redundant and invalid data of drugs, targets, and drug-target interaction pairs, we obtained a total of 8 new datasets containing 11,912 known interactions, 4495 unique drugs, and 959 unique targets. The statistics for new dataset 1 are listed in Table 17. The detailed drug target interaction information can be referred to Additional file 2.

As shown in Table 17, the amounts of drugs and targets in enzymes, ion channels (IC), G protein-coupled receptors (GPCR), and nuclear receptors (NR) are significantly different from that of the corresponding datasets in benchmark datasets. These datasets are important supplement to benchmark datasets in the experimental verification.

To inspect our proposed method for predicting large-scale compound-protein interactions (CPIs), we retrieved CPIs of *Homo sapiens* from STITCH database (Version 5.0) [41] as new dataset 2.⁴ The compound similarity matrix is derived from the scores of chemical_chemical links in STITCH database.⁵ Similarly, the protein sequence similarity matrix is obtained as new dataset 1. After deleting the redundant and invalid data

²https://www.kegg.jp/kegg-bin/get_htext?br08310.keg

³<https://www.genome.jp/tools/simcomp2/>

⁴http://stitch.embl.de/download/protein_chemical.links.v5.0/9606.protein_chemical.links.v5.0.tsv.gz

⁵http://stitch.embl.de/download/chemical_chemical.links.v5.0.tsv.gz

Table 16 Statistics for the benchmark datasets [6]

Dataset	Number of drugs	Number of targets	Number of drug-target Interactions	Average degree of drugs	Average degree of targets
Enzymes	445	664	2926	6.57	4.40
IC	210	204	1476	7.02	7.23
GPCR	223	95	635	2.84	6.68
NR	54	26	90	1.66	3.46

of drugs, targets, and drug-target interaction pairs, we obtained 5,979,099 interactions between 15,324 unique proteins in *Homo sapiens* and 224,203 unique compounds.

To validate our proposed method for predicting drug-target binding affinity, we selected two kinase datasets from the studies by Davis et al. [44] and Metz et al. [45] respectively. These two datasets are available at <http://staff.cs.utu.fi/~aatapa/data/DrugTarget/>. In Davis dataset [44], the target similarity matrix is computed by a normalized version of Smith-Waterman score [42]. There are 3 drug similarity matrices in Davis dataset, two-dimensional and three-dimensional Tanimoto coefficients similarity matrices, and the extended-connectivity fingerprint ECFP4 [46] similarity matrix. The drug-target interaction affinity matrix used kinase disassociation constant (K_d). There are 68 drugs, 442 targets, and 1527 interactions in Davis dataset.

In Metz dataset [45], the target similarity matrix is computed by a normalized version of Smith-Waterman score [42]. The drug similarity matrix is a two-dimensional Tanimoto coefficients similarity matrix. The drug-target interaction affinity matrix used kinase inhibition constant (K_i). There are 1421 drugs, 156 targets, and 3200 interactions in Metz dataset.

The statistics for these two kinase datasets are listed in Table 18.

Method

We propose a new method to learn drug kernel matrix and target kernel matrix. We integrate drug kernel matrix, target kernel matrix, and drug-target interaction network to build a heterogeneous network. We apply the marginalized denoising model on heterogeneous network to improve the accuracy of drug-target interaction prediction. Our proposed prediction method consists of the following four steps:

Step 1: Calculate drug kernel matrix $KFJD$ by combining drug similarity matrix SD , Gaussian interaction profile kernel matrix for drugs KGD , and association index kernel matrix for drugs KJD , where KGD and KJD are constructed from drug-target interaction network Y .

Step 2: Calculate target kernel matrix $KFJT$ by combining target similarity matrix ST , Gaussian interaction profile kernel matrix for targets KGT , and association index kernel matrix for targets KJT , where KGT and KJT are constructed from Y' which is the transpose of drug-target interaction network Y .

Table 17 Statistics for the new dataset 1

Dataset	Number of drugs	Number of targets	Number of drug-target Interactions	Average degree of drugs	Average degree of targets
Enzymes	1178	370	2705	2.30	7.31
IC	462	127	3629	7.85	28.57
GPCR	1582	128	3472	2.19	27.13
NR	422	19	558	1.32	29.37
CR	199	101	283	1.42	2.80
CSM	102	78	234	2.29	3.00
PK	280	95	625	2.23	6.58
TR	270	41	406	1.50	9.9

Step 3: Construct a heterogeneous network M by drug kernel matrix $KFJD$, target kernel matrix $KFJT$, and drug-target interaction network Y .

Step 4: Create a marginalized denoising model (MDM) on the constructed heterogeneous network M with local and global associations between nodes (targets and drugs) to predict latent drug-target interaction pairs.

The procedure of our proposed prediction method is shown in Fig. 4.

Constructing final kernel matrix

The final kernel matrix combines different kernels with drug similarity matrix and target similarity matrix for potential DTIs prediction. Based on kernel fusion [18, 19], we calculate drug kernel matrix by combining drug similarity matrix with GIP kernel matrix and Jaccard kernel matrix, and calculate target kernel matrix by combining target protein similarity matrix with GIP kernel matrix and Jaccard kernel matrix.

The final drug kernel matrix $KFJD$ and final target kernel matrix $KFJT$ are calculated according to the following steps.

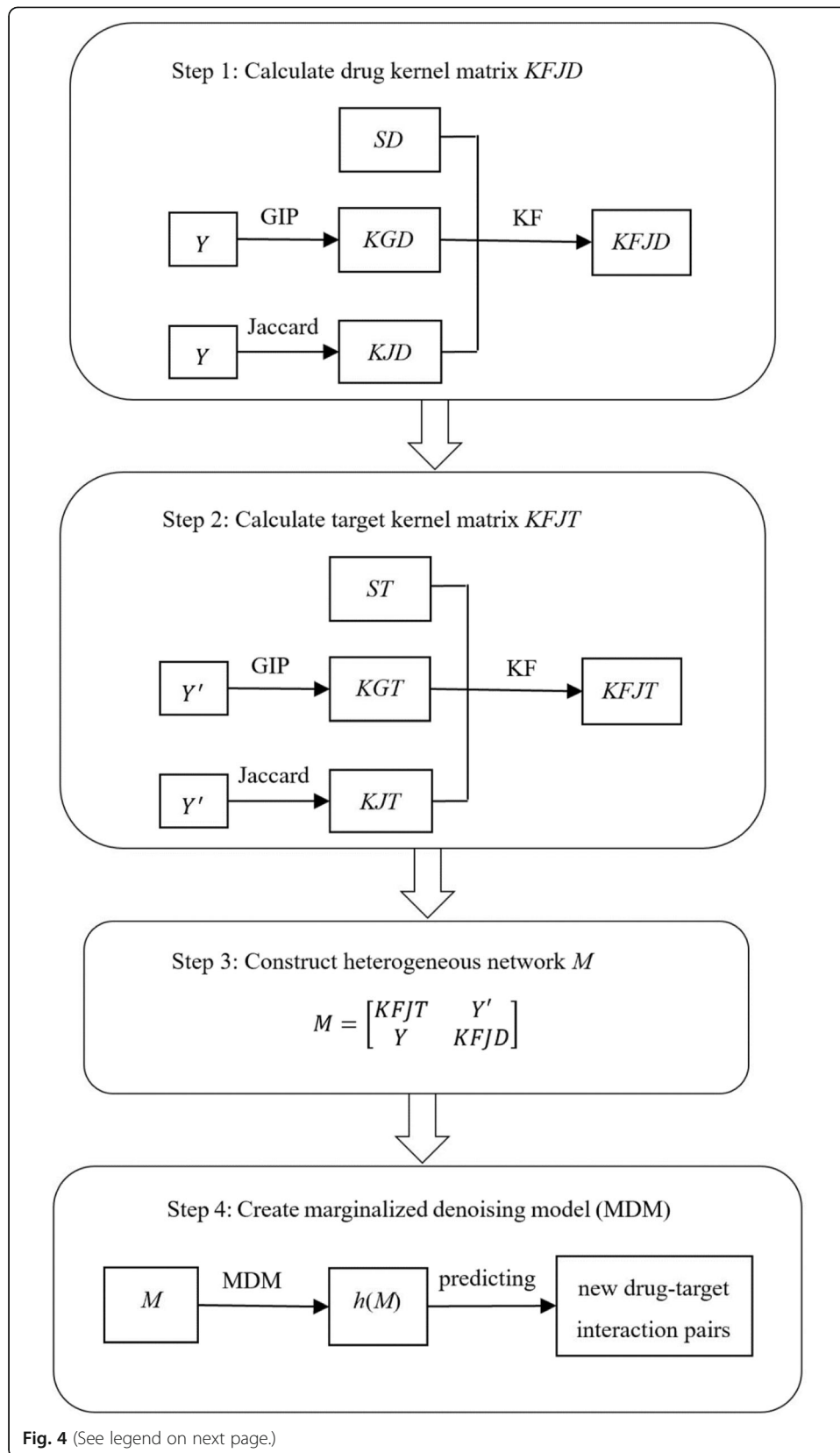
Firstly, GIP kernel matrix for drugs KGD and GIP kernel matrix for targets KGT are calculated respectively [9]:

$$\begin{aligned}
 KGD_{d_i, d_j} &= \exp\left(-\gamma_d \|y_{d_i} - y_{d_j}\|^2\right), 1 \leq i, j \leq n \\
 KGT_{t_i, t_j} &= \exp\left(-\gamma_t \|y_{t_i} - y_{t_j}\|^2\right), 1 \leq i, j \leq m
 \end{aligned} \tag{1}$$

where y_{d_i} and y_{d_j} are interaction profiles of drugs d_i and d_j respectively, which are represented by binary vectors encoding presence or absence of interaction with every target in interaction matrix Y . Similarly, y_{t_i} and y_{t_j} are interaction profiles of targets t_i and t_j respectively, which are represented by binary vectors encoding presence or

Table 18 Statistics for the kinase datasets

Dataset	Number of drugs	Number of targets	Number of drug-target Interactions	Average degree of drugs	Average degree of targets
David	68	442	1527	22.46	3.45
Ketz	1421	156	3200	2.25	20.51



(See figure on previous page.)

Fig. 4 Procedure of our proposed predicting method. Drug kernel matrix $KFJD$ was calculated by combining drug similarity matrix SD , GIP kernel matrix for drugs KGD , and association index kernel matrix for drugs KJD , where KGD and KJD are constructed from drug-target interaction network Y (seen in step 1). target kernel matrix $KFJT$ was calculated by combining target similarity matrix ST , GIP kernel matrix for targets KGT , and association index kernel matrix for targets KJT , where KGT and KJT are constructed from Y' which is the transpose of Y (seen in step 2). Next, a heterogeneous network M was constructed by drug kernel matrix $KFJD$, target kernel matrix $KFJT$, and drug-target interaction network Y (seen in step 3). Finally, a marginalized denoising model (MDM) was created on the heterogeneous network M with local and global associations between nodes (targets and drugs) to predict latent drug-target interaction pairs (seen in step 4)

absence of interaction with every drug in interaction matrix Y . Parameters γ_d and γ_t are used to control kernel bandwidth and are defined as follows [9]:

$$\begin{aligned}\gamma_d &= 1 / \left(\frac{1}{n_d} \sum_{i=1}^{n_d} |y_{d_i}|^2 \right) \\ \gamma_t &= 1 / \left(\frac{1}{n_t} \sum_{j=1}^{n_t} |y_{t_j}|^2 \right)\end{aligned}\quad (2)$$

Secondly, Jaccard profile kernel matrix for drugs and Jaccard profile kernel matrix for targets are calculated respectively.

Jaccard index [47] is commonly used in association index. Compared with cosine, Pearson correlation coefficient, and other association index, Jaccard index is more suitable for binary data with high sparsity, and Jaccard index is used to measure the degree of sharing association between two nodes in biological interaction network [40]. Hence, we use Jaccard index to construct an association index kernel matrix between drugs and an association index kernel matrix between targets in DTIs network respectively. Next, we discuss how to calculate Jaccard kernel matrix for drug KJD and Jaccard kernel matrix for target KJT .

The value of Jaccard index kernel for drugs d_i and d_j in DTIs network, KJD_{d_i, d_j} , is calculated as follows [40]:

$$KJD_{d_i, d_j} = \frac{D_{11}}{D_{01} + D_{10} + D_{11}}, 1 \leq i, j \leq n \quad (3)$$

where D_{01} , D_{10} , and D_{11} are three parameters to measure the sharing relationship between d_i and d_j . D_{01} is total number of targets when the value of $Y(d_i, t_k)$ is 0 and the value of $Y(d_j, t_k)$ is 1, D_{10} denotes total number of targets when the value of $Y(d_i, t_k)$ is 1 and the value of $Y(d_j, t_k)$ is 0, D_{11} represents total number of targets when the value of $Y(d_i, t_k)$ is 1 and the value of $Y(d_j, t_k)$ also is 1, where Y is the target-drug interaction matrix, and t_k is a target contained in Y , $i, j = 1, 2, \dots, n$, and $k = 1, 2, \dots, m$.

Similarly, the value of Jaccard index kernel for targets t_i and t_j in DTIs network, KJT_{t_i, t_j} , is computed as follows [40]:

$$KJT_{t_i, t_j} = \frac{T_{11}}{T_{01} + T_{10} + T_{11}}, 1 \leq i, j \leq m \quad (4)$$

where T_{01} , T_{10} , and T_{11} are three parameters to measure the sharing relationship between t_i and t_j . T_{01} is total number of drugs when the value of $Y(d_k, t_i)$ is 0 and the value of $Y(d_k, t_j)$ is 1, T_{10} denotes total number of drugs when the value of $Y(d_k, t_i)$ is 1 and the value of $Y(d_k, t_j)$ is 0, T_{11} represents total number of drugs when the value of

$Y(d_k, t_i)$ is 1 and the value of $Y(d_k, t_j)$ also is 1, where Y is the target-drug interaction matrix, and d_k is a drug contained in Y , $i, j = 1, 2, \dots, m$, and $k = 1, 2, \dots, n$.

Thirdly, based on the nonlinear kernel fusion technique [17, 18], the final drug kernel matrix $KFJD$ is calculated according to three matrices SD , KGD and KJD , and the final target kernel matrix $KFJT$ is calculated according to three matrices ST , KGT and KJT .

The calculation for $KFJD$ is described as follows.

The three kernel matrices SD , KGD , and KJD are first normalized according to Hao's method [18]. The normalized matrices are denoted by $PD1$, $PD2$, and $PD3$ respectively [18]:

$$\begin{aligned}
 PD1(d_i, d_j) &= \begin{cases} \frac{SD(d_i, d_j)}{2\sum_{k \neq i} SD(d_i, d_k)}, & j \neq i, 1 \leq i, j \leq n \\ 1/2, & j = i \end{cases} \\
 PD2(d_i, d_j) &= \begin{cases} \frac{KGD(d_i, d_j)}{2\sum_{k \neq i} KGD(d_i, d_k)}, & j \neq i, 1 \leq i, j \leq n \\ 1/2, & j = i \end{cases} \\
 PD3(d_i, d_j) &= \begin{cases} \frac{KJD(d_i, d_j)}{2\sum_{k \neq i} KJD(d_i, d_k)}, & j \neq i, 1 \leq i, j \leq n \\ 1/2, & j = i \end{cases}
 \end{aligned} \tag{5}$$

Then, we apply the k nearest neighbors (k NN) algorithm to compute local similarity matrices $LD1$, $LD2$, and $LD3$ for $PD1$, $PD2$, and $PD3$ respectively [18]:

$$\begin{aligned}
 LD1(d_i, d_j) &= \begin{cases} \frac{PD1(d_i, d_j)}{\sum_{d_k \in N_i} PD1(d_i, d_k)}, & d_j \in N_i, 1 \leq i, j \leq n \\ 0, & d_j \notin N_i \end{cases} \\
 LD2(d_i, d_j) &= \begin{cases} \frac{PD2(d_i, d_j)}{\sum_{d_k \in N_i} PD2(d_i, d_k)}, & d_j \in N_i, 1 \leq i, j \leq n \\ 0, & d_j \notin N_i \end{cases} \\
 LD3(d_i, d_j) &= \begin{cases} \frac{PD3(d_i, d_j)}{\sum_{d_k \in N_i} PD3(d_i, d_k)}, & d_j \in N_i, 1 \leq i, j \leq n \\ 0, & d_j \notin N_i \end{cases}
 \end{aligned} \tag{6}$$

where N_i denotes the k nearest neighbors of drug d_i , $i = 1, 2, \dots, n$. In formula (6), the similarity between any two non-nearest neighbors is set to zero to reduce the influence on prediction results from the non-nearest drug-target interaction pairs.

The key step of fusion operation is an iterative calculation [18]:

$$\begin{aligned}
 PD_{t+1}^1 &= LD1 \times \frac{(PD_t^2 + PD_t^3)}{2} \times LD1' \\
 PD_{t+1}^2 &= LD2 \times \frac{(PD_t^1 + PD_t^3)}{2} \times LD2' \\
 PD_{t+1}^3 &= LD3 \times \frac{(PD_t^1 + PD_t^2)}{2} \times LD3'
 \end{aligned} \tag{7}$$

where PD_{t+1}^1 , PD_{t+1}^2 , and PD_{t+1}^3 are the results of $PD1$, $PD2$, and $PD3$ after t iterations respectively, and $LD1'$, $LD2'$, and $LD3'$ are the transposes of $LD1$, $LD2$, and $LD3$ respectively.

During each iteration, the values of PD_{t+1}^1 , PD_{t+1}^2 , and PD_{t+1}^3 are further updated by $PD_{t+1}^1 = (PD_{t+1}^1 + PD_{t+1}^1')/2 + I$, $PD_{t+1}^2 = (PD_{t+1}^2 + PD_{t+1}^2')/2 + I$, and $PD_{t+1}^3 = (PD_{t+1}^3 + PD_{t+1}^3')/2 + I$ respectively, where I is an identity matrix.

After t iterations, the final drug kernel matrix $KFJD$ can be obtained by [17]:

$$KFJD = (PD_t^1 + PD_t^2 + PD_t^3)/3 \quad (8)$$

Similarly, the final target kernel matrix $KFJT$ can be obtained as follows:

$$KFJT = (PT_t^1 + PT_t^2 + PT_t^3)/3 \quad (9)$$

More detailed description about the kernel fusion can be seen in [18, 19, 48].

Marginalized denoising model

Our method treats DTIs prediction problem as network link prediction problem. We use Marginalized denoising model (MDM) [31] on heterogeneous network composed of the final drug and target kernel matrices and the known drug-target interaction matrix to predict potential DTIs. Marginalized denoising model [31] is inspired by the idea of marginalized denoising auto-encoders [49].

Auto-Encoder (AE) is a type of artificial neural networks, which is used to learn efficient data coding in an unsupervised manner [50, 51]. The AE encodes original input dataset x with weight w into latent representation h and decodes h into output y , where $h = f(x)$ and $y = g(h)$. The AE is trained to minimize reconstruction error $\mathcal{L}(x, g(f(x)))$ to guarantee that output y closely matches original data x . The AE is widely used to extract features and reduce dimensionality. The AE can also be used to learn new features.

Denoising Auto-Encoder (DAE) [52] transforms original input dataset x into partially corrupted input \tilde{x} and trains \tilde{x} to recover undistorted original input x . To train an auto-encoder to denoised data, a preliminary stochastic mapping $x \rightarrow \tilde{x}$ is performed to corrupt the data, and \tilde{x} with weight w is used as an input for normal auto-encoder. The loss function of DAE is represented by $\mathcal{L}(x, g(f(\tilde{x})))$ instead of $\mathcal{L}(x, g(f(x)))$. The corrupted input \tilde{x} can be constructed by randomly setting original input x to zero with given probability p , where $0 < p < 1$. The original noises in original input dataset x are removed during the corrupting process. To a certain extent, the training data are close to the testing data after the training data are denoised, and the robustness of weight w is enforced after training.

Marginalized denoising auto-encoder (mDA) [49] is a variant of DAE. The mDA is used to solve the problem with high computational cost of the DAE. "Marginalized" means that the loss function $\mathcal{L}(x, g(f(\tilde{x})))$ is approximated by the expected value $\mathbb{E} \|\mathcal{L}(x, g(f(\tilde{x})))\|_{p(\tilde{x}|x)}$ of loss function with conditional distribution $p(\tilde{x}|x)$ based on the weak law of large number [53].

Our prediction method

The latent drug-target interactions are impacted by the existing drug-target interaction pairs in the drug-target interaction network. The probability of predicting drug-target interactions may also be influenced by the matrix of similarities between drugs and the matrix of similarities between targets [19].

We treat the drug-target interactions (DTIs) prediction problem as network link prediction problem. To improve the prediction accuracy, we propose a DTIs prediction method using marginalized denoising model on heterogeneous network. The

heterogeneous network can be represented by matrix $M = \begin{bmatrix} KFJT & Y' \\ Y & KFJD \end{bmatrix}$ of size $(m+n) \times (m+n)$, where $KFJT \in \mathbb{R}^{m \times m}$ is the target kernel matrix, $KFJD \in \mathbb{R}^{n \times n}$ is the drug kernel matrix, $Y \in \mathbb{R}^{n \times m}$ is the drug-target interaction network, Y' is the transpose of Y , m is the number of targets, and n is the number of drugs.

To generate the training data, we inject random noise to original input matrix M to construct the corrupted matrix \tilde{M} . The set of corrupted matrices $\tilde{\mathcal{M}} = \{\tilde{M}_1, \tilde{M}_2, \dots, \tilde{M}_c\}$ is the training data. Then, we train the mapping function $h(\tilde{M})$ such that the final output M^* closely matches the original matrix M . That is to minimize the loss function $\mathcal{L}(h(\tilde{M}))$:

$$\mathcal{L}(h(\tilde{M})) = \sum_{\tilde{M} \in \tilde{\mathcal{M}}} \|M - h(\tilde{M})\|_F^2 \quad (10)$$

$$M^* = h(\tilde{M}) = \sum_{l=1}^{m+n} L_{il} \tilde{M}_{lj} + \sum_{l=1}^{m+n} \sum_{k=1}^{m+n} \tilde{M}_{il} G_{lk} \tilde{M}_{jk} + b_i, 1 \leq i, j \leq m+n \quad (11)$$

where the mapping function $h(\tilde{M})$ consists of the latent local and global associations between any two drug or target nodes in M , $\|\cdot\|_F^2$ denotes the Frobenius norm of matrix, \tilde{M} s in corrupted matrices set $\tilde{\mathcal{M}}$ are constructed by randomly setting the value of elements in M to zero with given probability p , where $0 < p < 1$, b_i is a bias value, L is local association weighted matrix, $\sum_{l=1}^{m+n} L_{il} \tilde{M}_{lj}$ is latent local interaction between nodes i and j via node l , G is global association weighted matrix and $\sum_{l=1}^{m+n} \sum_{k=1}^{m+n} \tilde{M}_{il} G_{lk} \tilde{M}_{jk}$ is latent global association between nodes i and j via nodes l and k , $1 \leq i, j \leq m+n$.

We illustrate an example of latent global association in Fig. 5. The solid line shows the existing association, and the dashed line shows latent global association.

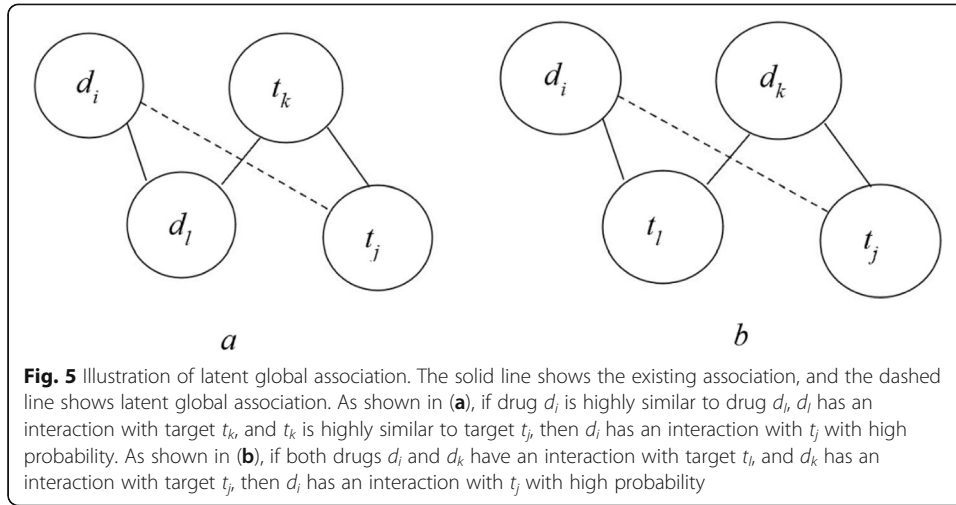
As shown in Fig. 5a, if drug d_i is highly similar to drug d_l , d_l has an interaction with target t_k , and t_k is highly similar to target t_j , then d_i has an interaction with t_j with high probability. We can also see from Fig. 5b that, if both drugs d_i and d_k have an interaction with target t_l , and d_k has an interaction with target t_j , then d_i has an interaction with t_j with high probability. The latent global association represents the weighted value of indirect drug-target interaction. The iterative training with latent local and global associations will obtain a more precise drug-target interaction prediction result M^* .

To prevent loss function $\mathcal{L}(h)$ from overfitting and enhance the learning performance, we construct a new objective function $\mathcal{L}(L, G, b)$ by Tikhonov regularization terms:

$$\mathcal{L}(L, G, b) = \sum_{\tilde{M} \in \tilde{\mathcal{M}}} \|M - L\tilde{M} - \tilde{M}G\tilde{M}^T - (b * \mathbf{1}_n)^T\|_F^2 + \frac{\lambda_1}{2} (\|L\|_F^2) + \frac{\lambda_2}{2} (\|G\|_F^2) \quad (12)$$

where L and G represent latent local and global association weighted matrices respectively, b is a bias vector, $\mathbf{1}_n$ denotes an all-one column vector of size n , and λ_1 and λ_2 are the regularization coefficients. Tikhonov regularization is used to ensure the smoothness of fitting curves of L and G [54].

In the denoising auto-encoder, the more the training data used, the more accurate the prediction results are. Ideally, we use infinite training data to compute weight



matrices L and G . However, when the size of set $\tilde{\mathcal{M}}$ is increased, the computation cost becomes more expensive. According to the weak law of large number [53], when the size of set $\tilde{\mathcal{M}}$ becomes very large, we can rewrite the sum part of formula (12) into the expectation form as follows:

$$\mathcal{L}(L, G, b) = \mathbb{E}_{p(\tilde{\mathcal{M}}|M)} \left[\left\| M - L\tilde{M} - \tilde{M}G\tilde{M}^T - b\mathbf{1}_n^T \right\|_F^2 \right] + \frac{\lambda_1}{2} (\|L\|_F^2) + \frac{\lambda_2}{2} (\|G\|_F^2) \quad (13)$$

where $p(\tilde{\mathcal{M}}|M)$ is a conditional distribution, and the expectation is with respect to the random variable $\tilde{\mathcal{M}}$.

To apply formula (13) in large data matrix, low rank approximation is used [31]. Formula (13) is rewritten with respect to $L = UU^T$ and $G = VV^T$ as follows:

$$\begin{aligned} \mathcal{L}(U, V, b) = & 0.5 * \text{tr}(M^T M) - \text{tr} \left(U^T * \tilde{M} M^T * U + V^T * \tilde{M}^T M^T \tilde{M} * V + M^T b \mathbf{1}_n^T \right) \\ & + 0.5 * \text{tr} \left(U^T * U U^T \tilde{M} \tilde{M}^T * U + V^T * \tilde{M}^T \tilde{M} V V^T \tilde{M}^T \tilde{M} * V + b^T * b \mathbf{1}_n^T \right) \\ & + \text{tr} \left(U^T * \tilde{M} V V^T \tilde{M}^T \tilde{M}^T * U + U^T * b \mathbf{1}_n^T \tilde{M}^T * U + V^T * \tilde{M}^T b \mathbf{1}_n^T \tilde{M} * V \right) \\ & + 0.5 * \text{tr}(U^T * \lambda_1 I * U) + 0.5 * \text{tr}(V^T * \lambda_2 I * V) \end{aligned} \quad (14)$$

where $U, V \in \mathbb{R}^{(m+n) \times k}$, k is the dimension of latent variables U and V , $\text{tr}(\cdot)$ represents the trace of matrices, and I is the identity matrix.

To minimize the norm function $\mathcal{L}(U, V, b)$, the partial gradient of formula (14) is calculated with respect to U , V and b as follows:

$$\frac{\partial \mathcal{L}}{\partial U} = \mathbb{E} \left[\left(U U^T \tilde{M} \tilde{M}^T + \tilde{M} \tilde{M}^T U U^T + \tilde{M} V V^T \tilde{M}^T \tilde{M}^T + \tilde{M} \tilde{M} V V^T \tilde{M}^T + b \mathbf{1}_n^T \tilde{M}^T + \tilde{M} b^T - M \tilde{M}^T - \tilde{M} M^T \right) U + \lambda_1 U \right] \quad (15)$$

$$\frac{\partial \mathcal{L}}{\partial V} = \mathbb{E} \left[\tilde{M}^T \left(U U^T \tilde{M} + \tilde{M}^T U U^T + \tilde{M} V V^T \tilde{M}^T + \tilde{M} V V^T \tilde{M}^T + b \mathbf{1}_n^T + b^T - M - M^T \right) \tilde{M} \right] V + \lambda_2 V \quad (16)$$

$$\frac{\partial \mathcal{L}}{\partial b} = \mathbb{E} \left[\left(UU^T \tilde{M} + \tilde{M} V V^T \tilde{M}^T + b \mathbf{1}_n^T - M \right) \mathbf{1}_n \right] \quad (17)$$

Given q as the residual probability for \tilde{M} , $q = 1 - p$, we label a constant matrix containing no \tilde{M} as C , and calculate the gradients for different terms of \tilde{M} . For a term containing only one \tilde{M} , $\mathbb{E}[C\tilde{M}] = C\mathbb{E}[\tilde{M}] = qCM$. For a term containing two \tilde{M} s, we need to analyze the cases that the two \tilde{M} s are the same or not, e.g., if the two \tilde{M} s are the same, $\mathbb{E}[\tilde{M}^T C \tilde{M}] = q^2 M^T C M$, otherwise $\mathbb{E}[\tilde{M}^T C \tilde{M}] = q(1 - q) \text{diag}(M^T * \text{diag}(C))$. The term containing two \tilde{M} s, $\mathbb{E}[\tilde{M}^T C \tilde{M}]$, is given in formula (18) [31]:

$$\mathbb{E}[\tilde{M}^T C \tilde{M}] = q^2 M^T C M + q(1 - q) \text{diag}(M^T * \text{diag}(C)) \quad (18)$$

For the term containing three or more \tilde{M} s, we need to analyze the cases that all the \tilde{M} s are the same or any two \tilde{M} s are the same or all the \tilde{M} s are different. The term containing three \tilde{M} s, $\mathbb{E}[\tilde{M} C \tilde{M}^T \tilde{M}^T]$, is given as follows [31]:

$$\begin{aligned} \mathbb{E}[\tilde{M} C \tilde{M}^T \tilde{M}^T] &= q^3 M C M^T M^T + q^2(1 - q)(\text{diag}(M * \text{diag}(C)) M^T + M * C * \text{diag}(\text{diag}(M)) + \text{diag}(M) * \text{sum}(C * M, 2)) \\ &\quad + q(1 - 2q)(1 - q) \text{diag}(\text{diag}(M) \circ (\text{diag}(C))) \end{aligned} \quad (19)$$

where the function diag^* outputs the diagonal elements of a matrix, the operator \circ denotes the Hadamard product (element-wise product), and the function $\text{sum}(*, 2)$ outputs the sum by rows of a matrix.

We use the L-BFGS (Limited-memory BFGS) [55] to optimize the objective functions with respect to latent variables U , V , and b . The L-BFGS [55] is an optimization algorithm in the family of quasi-Newton methods. The Newton's method is an iterative optimization using Taylor's second-order expansion. The Newton's method finds extrema for loss function by computing Hessian matrix. It is too expensive to compute Hessian matrix for every iteration. The L-BFGS algorithm optimizes the calculation of Newton's method and simplifies the calculation of Hessian matrix. L-BFGS has the feature of fast convergence and no storage of Hessian matrix.

Finally, we calculate the final matrix $M^* = UU^T M + MVV^T M^T + b$, and compute the evaluation metrics AUC (area under curve of receiver operating characteristic) and $AUPR$ (area under precision-recall curve) by comparing M with M^* .

Based on the above steps, we propose a drug-target interaction prediction algorithm using marginalized denoising model on heterogeneous network called DTIP_MDHN, in which its input files *SDFFile*, *STFile* and *YFile* are derived from <http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>, *noise* is the noise value, and k is the dimension of latent layer. Algorithm DTIP_MDHN is described in algorithm 1.

Algorithm 1: DTIP_MDHNInput: *SDFFile*, *STFile*, *YFile*, *noise*, *k*Output: *AUC*, *AUPR*

Begin

1. Generate drug chemical structure similarity matrix *SD*, target sequence similarity matrix *ST*, and known drug-target interaction matrix *Y* from *SDFFile*, *STFile*, and *YFile* respectively;
2. Compute GIP kernel matrices *KGD* and *KGT* for drugs and targets according to Eq. (1) and (2), respectively;
3. Compute Jaccard kernel matrices *KJD* and *KJT* for drugs and targets according to Eq. (3) and (4), respectively;
4. Compute the final kernel matrix *KFJD* by *SD*, *KGD*, and *KJD* according to Eq. (5)-(8);
5. Compute the final kernel matrix *KFJT* by *ST*, *KGT*, and *KJT* according to target version of Eq. (5)-(8);
6. Construct matrix $M = \begin{bmatrix} KFJT & Y' \\ Y & KFJD \end{bmatrix}$;
7. Initialize *U*, *V*, and *b* randomly;
8. Compute objective function $\mathcal{L}(U, V, b)$, $grad_U = \frac{\partial \mathcal{L}}{\partial U}$, $grad_V = \frac{\partial \mathcal{L}}{\partial V}$, and $grad_b = \frac{\partial \mathcal{L}}{\partial b}$ by Eq. (14)-(17) respectively;
9. Optimize objective function $\mathcal{L}(U, V, b)$ with $grad_U$, $grad_V$, and $grad_b$ by L-BFGS;
10. Calculate the final matrix $M^* = UU'M + MVV'M' + b$ using the optimal output *U*, *V*, and *b*;
11. Calculate the scores of *AUC* and *AUPR* by comparing *M* with *M**.

End.

Our method DTIP_MDHN can obtain more accurate prediction than other existing methods because it introduces Jaccard index kernel matrix to measure the sharing interaction relationship between drugs and targets, and uses both local and global associations to reduce the sparsity of DTIs network.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12859-020-03662-8>.

Additional file 1. Compound-Protein Interaction pairs selected for large-scale CPIs prediction. This file records the detailed compound-protein interaction pairs selected for large-scale CPIs prediction. These data were extracted from STITCH database (Version 5.0) with combined scores greater than 900 in *Homo sapiens* and contains 13,286 drugs, 5313 targets, and 116,199 interactions. In this file, compounds are derived from PubChem with the prefix "CID", proteins are derived from Ensembl with the prefix "ENSP", scores are the combined scores in STITCH database. The scores indicate the interaction probability of corresponding compound protein interaction pair. All detailed information about these interactions can be found in STITCH database.

Additional file 2. Drug-Target Interaction pairs in the new Dataset 1. This file records the detailed drug-target interaction pairs on enzymes, ion channels, GPCRs, nuclear receptors, Cytokines and receptors, Cell surface molecules and ligands, Protein kinases, and Transporters of the new Dataset 1. The new database 1 was extracted from KEGG database and contains 4495 drugs, 959 targets, and 11,912 known interactions.

Abbreviations

DTI: Drug-target interaction; AUC: Area under curve of receiver operating characteristic; AUPR: Area under precision-recall curve; DTIP_MDHN: A drug-target interaction prediction method using marginalized denoising model on heterogeneous network; GIP: Gaussian interaction profile; MF: Matrix factorization; MDM: Marginalized denoising model; CPI: Compound-protein interaction; PPI: Protein-protein interaction; ECFP: The extended-connectivity fingerprint; DAE: Denoising Auto-Encoder; CI: The concordance index

Acknowledgments

The authors thank the editor and anonymous reviewers for their constructive comments and suggestions, which greatly help us improve our manuscript.

Authors' contributions

C.T. and C.Z. designed the methodology. C.T. implemented the analysis. C.T., C.Z., and J.W. performed the experiments and analyzed the results. C.T., C.Z., D.C., and J.W. wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is financially supported by the National Natural Science Foundation of China under Grant No. 61962004, and Natural Science Foundation of Guangxi under Grant No. 2014GXNSFAA118396. The funders did not play any role in this study.

Availability of data and materials

The benchmark datasets were publicly available at <http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>. Algorithm DTI_MDHN is implemented in MATLAB. The software suite of our method is available at <https://doi.org/10.6084/m9.figshare.11980161>.

Ethics approval and consent to participate

No ethics approval was required for the study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. ²School of Computer, Electronics and Information, Guangxi University, Nanning, China. ³Medical College, Guangxi University, Nanning, China.

Received: 18 March 2020 Accepted: 14 July 2020

Published online: 23 July 2020

References

1. Csérmelyi P, Korcsmáros T, Kiss HJM, et al. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther*. 2012;138(3):333–408. <https://doi.org/10.1016/j.pharmthera.2013.01.016>.
2. Ding H, Takigawa I, Mamitsuka H, et al. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief Bioinform*. 2014;15(5):734–47. <https://doi.org/10.1093/bib/bbt056>.
3. Chen X, Yan CC, Zhang X, et al. Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform*. 2016;17(4):696–712. <https://doi.org/10.1093/bib/bbv066>.
4. Cheng T, Hao M, Takeda T, et al. Large-scale prediction of drug-target interaction: a data-centric review. *AAPS J*. 2017;19:1264–75. <https://doi.org/10.1208/s12248-017-0092-6>.
5. Ezzat A, Wu M, Li XL, et al. Computational prediction of drug–target interactions using chemo-genomic approaches: an empirical survey. *Brief Bioinform*. 2019;20(4):1337–57. <https://doi.org/10.1093/bib/bby002>.
6. Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*. 2008;24(13):i232–40. <https://doi.org/10.1093/bioinformatics/btn162>.
7. Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*. 2009;25(18):2397–403. <https://doi.org/10.1093/bioinformatics/btp433>.
8. Xia Z, Wu LY, Zhou X, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst Biol*. 2010;4(Suppl 2):S6. <https://doi.org/10.1186/1752-0509-4-S2-S6>.
9. Laarhoven TV, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*. 2011;27(21):3036–43. <https://doi.org/10.1093/bioinformatics/btr500>.
10. Mei J-P, Kwok C-K, Yang P, et al. Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics*. 2013;29(2):238–45. <https://doi.org/10.1093/bioinformatics/bts670>.
11. Twan VL, Elena M, Peter C. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One*. 2013;8(6):e66952. <https://doi.org/10.1371/journal.pone.0066952>.
12. Olayan RS, Haitham A, Bajic VB. DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches. *Bioinformatics*. 2018;34(7):1164–73. <https://doi.org/10.1093/bioinformatics/btx731>.
13. Koren Y, Bell RM, Volinsky C. Matrix factorization techniques for recommender systems. *IEEE Computer*. 2009;42(8):30–7. <https://doi.org/10.1109/MC.2009.263>.
14. Cobanoglu MC, Liu C, Hu F, et al. Predicting drug–target interactions using probabilistic matrix factorization. *J Chem Inf Model*. 2013;53(12):3399–409. <https://doi.org/10.1021/ci400219z>.
15. 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*. Chicago; 2013. p. 1025–33. <https://doi.org/10.1145/2487575.2487670>.
16. Ezzat A, Zhao P, Wu M, et al. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans Comput Biol Bioinform*. 2017;14(3):646–56. <https://doi.org/10.1109/TCBB.2016.2530062>.
17. Liu Y, Wu M, Miao C, et al. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput Biol*. 2016;12(2):e1004760. <https://doi.org/10.1371/journal.pcbi.1004760>.
18. Hao M, Wang Y, Bryant SH. Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique. *Anal Chim Acta*. 2016;909:41–50. <https://doi.org/10.1016/j.aca.2016.01.014>.
19. Hao M, Bryant SH, Wang Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Sci Rep*. 2017;7:40376. <https://doi.org/10.1038/srep40376>.
20. Pahikkala T, Airola A, Stock M, et al. Efficient regularized least-squares algorithms for conditional ranking on relational data. *Mach Learn*. 2013;93:321–56.
21. Pahikkala T, Airola A, Pietilä S, et al. Toward more realistic drug-target interaction predictions. *Brief Bioinform*. 2015;16(2):325–37. <https://doi.org/10.1093/bib/bbu010>.

22. He T, Heidemeyer M, Ban F, et al. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *J Cheminform*. 2017;9:24. <https://doi.org/10.1186/s13321-017-0209-z>.
23. Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23(6):1241–50. <https://doi.org/10.1016/j.drudis.2018.01.039>.
24. Hu PW, Chan KCC, You ZH. Large-scale prediction of drug-target interactions from deep representations. In: Proceedings of the 2016 International joint conference on neural networks (IJCNN), Vancouver, British Columbia, Canada, July 24–29; 2016. <https://doi.org/10.1109/IJCNN.2016.7727339>.
25. Hu SS, Zhang C, Chen P, et al. Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. *BMC Bioinformatics*. 2019;20(Suppl 25):689. <https://doi.org/10.1186/s12859-019-3263-x>.
26. Masashi T, Kentaro T, Jun S. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*. 2019;35(2):309–18. <https://doi.org/10.1093/bioinformatics/bty53546>.
27. Tian K, Shao M, Zhou S, et al. Boosting compound-protein interaction prediction by deep learning. *Methods*. 2016;110:64–72. <https://doi.org/10.1016/j.ymeth.2016.06.024>.
28. Yamanishi Y. Chemogenomic approaches to infer drug-target interaction networks. *Data Min Syst Biol*. 2013;939:97–113. https://doi.org/10.1007/978-1-62703-107-3_9.
29. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol BioSyst*. 2012;8(7):1970–8. <https://doi.org/10.1039/C2MB00002D>.
30. Lan W, Wang J, Li M, et al. Predicting drug–target interaction using positive-unlabeled learning. *Neurocomputing*. 2016; 206:50–7. <https://doi.org/10.1016/j.neucom.2016.03.080>.
31. Chen Z, Zhang W. A marginalized denoising method for link prediction in relational data. In: Proceedings of the 2014 SIAM international conference on data mining, Philadelphia, Pennsylvania, USA, April 24–26; 2014. p. 298–306. <https://doi.org/10.1137/1.9781611973440.34>.
32. Kanehisa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*. 2006;34(1):D354–7. <https://doi.org/10.1093/nar/gkj102>.
33. Günther S, Kuhn D, Dunkel M, et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res*. 2008;36(Database issue):D919–22. <https://doi.org/10.1093/nar/gkm862>.
34. Wishart, D. S, Knox C, Guo A C, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;36(Database issue): D901-D906. doi: <https://doi.org/10.1093/nar/gkm958>.
35. Bento AP, Gaulton A, Hersey A, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res*. 2014;42(Database issue):1083–90. <https://doi.org/10.1093/nar/gkt1031>.
36. Stehlin-Gaon C, Willmann D, Zeyer D, et al. All-trans retinoic acid is a ligand for the orphan nuclear receptor ROR β . *Nat Struct Biol*. 2003;10(10):820–5. <https://doi.org/10.1038/nsb979>.
37. Yang C, Shen HC, Wu Z, et al. Discovery of novel oxazolindione derivatives as potent and selective mineralocorticoid receptor antagonists. *Bioorg Med Chem Lett*. 2013;23(15):4388–92. <https://doi.org/10.1016/j.bmcl.2013.05.077>.
38. Schomburg I. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*. 2004;32(suppl_1):D431–3. <https://doi.org/10.1093/nar/gkh081>.
39. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005;92(4):965–70. <https://doi.org/10.1093/biomet/92.4.965>.
40. Bass JF, Diallo A, Nelson J, et al. Using networks to measure similarity between genes: association index selection. *Nat Methods*. 2013;10(12):1169–76. <https://doi.org/10.1038/nmeth.2728>.
41. Szklarczyk D, Santos A, von Mering C, et al. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res*. 2016;44(D1):D380–4. <https://doi.org/10.1093/nar/gkv1277>.
42. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147(1):195–7.
43. Hattori M, Okuno Y, Goto S, et al. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc*. 2003;125(39):11853–65. <https://doi.org/10.1021/ja036030u>.
44. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol*. 2011;29(11): 1046–51. <https://doi.org/10.1038/nbt.1990>.
45. Metz JT, Johnson EF, Soni NB, et al. Navigating the kinome. *Nat Chem Biol*. 2011;7(4):200–2. <https://doi.org/10.1038/nchembio.530>.
46. Rogers D, Brown RD, Hahn M. Using extended connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen*. 2005;10:682–6. <https://doi.org/10.1177/1087057105281365>.
47. Jaccard P. The distribution of the Flora in the Alpine zone. *New Phytol*. 1912;11(2):37–50.
48. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333–7. <https://doi.org/10.1038/nmeth.2810>.
49. Chen M, Xu Z, Weinberger KQ, et al. Marginalized denoising autoencoders for domain adaptation. In: Proceeding of the 29th international conference on machine learning, Edinburgh, Scotland, UK; 2012. arXiv preprint arXiv: 1206.4683.
50. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323:533–6. <https://doi.org/10.1038/323533a0>.
51. Baldi P, Hornik K. Neural networks and principal component analysis: learning from examples without local minima. *Neural Netw*. 1989;2(89):53–8. [https://doi.org/10.1016/0893-6080\(89\)90014-2](https://doi.org/10.1016/0893-6080(89)90014-2).
52. Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning, ACM; 2008. p. 1096–103. <https://doi.org/10.1145/1390156.1390294>.
53. Govindarajulu Z. On weak laws of large numbers. *Proc Math Sci*. 1970;71(6):266–74.
54. Guan N, Tao Z, Luo Z, et al. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Trans Image Process*. 2011;20(7):2030–48. <https://doi.org/10.1109/TIP.2011.2105496>.
55. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Math Program*. 1989;45(1–3):503–28.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.