PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

iDrug: Integration of drug repositioning and drug-target prediction via cross-network embedding

Huiyuan Chen[1], **Feixiong Cheng**[2,3,4], **Jing Li**[1,4]*

1 Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America, **2** Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, United States of America, **3** Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, Ohio, United States of America, **4** Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio, United States of America

* jingli@cwru.edu

Citation: Chen H, Cheng F, Li J (2020) iDrug: Integration of drug repositioning and drug-target prediction via cross-network embedding. PLoS Comput Biol 16(7): e1008040. https://doi.org/10.1371/journal.pcbi.1008040

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: November 22, 2019

Accepted: June 10, 2020

Published: July 15, 2020

Data Availability Statement: Our code and dataset are available at: https://github.com/Case-esaC/iDrug.

Funding: JL was supported in part by NSF CCF1815139 and a faculty investment fund from CWRU. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Computational drug repositioning and drug-target prediction have become essential tasks in the early stage of drug discovery. In previous studies, these two tasks have often been considered separately. However, the entities studied in these two tasks (i.e., drugs, targets, and diseases) are inherently related. On one hand, drugs interact with targets in cells to modulate target activities, which in turn alter biological pathways to promote healthy functions and to treat diseases. On the other hand, both drug repositioning and drug-target prediction involve the same drug feature space, which naturally connects these two problems and the two domains (diseases and targets). By using the *wisdom of the crowds*, it is possible to transfer knowledge from one of the domains to the other. The existence of relationships among drug-target-disease motivates us to jointly consider drug repositioning and drug-target prediction in drug discovery. In this paper, we present a novel approach called iDrug, which seamlessly integrates drug repositioning and drug-target prediction into one coherent model via cross-network embedding. In particular, we provide a principled way to transfer knowledge from these two domains and to enhance prediction performance for both tasks. Using real-world datasets, we demonstrate that iDrug achieves superior performance on both learning tasks compared to several state-of-the-art approaches. Our code and datasets are available at: https://github.com/Case-esaC/iDrug.

Author summary

Traditional high-throughput techniques for drug discovery are often expensive, time-consuming, and with high failure rates. Computational drug repositioning and drug-target prediction have thus become essential tasks in the early stage drug discovery. The emergence of large-scale heterogeneous biological networks has offered unprecedented opportunities for developing machine learning approaches to identify novel drug-disease or drug-target interactions. However, most existing works focused either on the drug-disease

network or on the drug-target network, thus failed to capture the inherent dependencies between these two networks. These two biological networks are naturally connected since they involve the same drug feature space. In our opinion, ignoring this rich source of information is a major shortcoming of some existing works. In this paper, we present a novel approach called iDrug, which seamlessly integrates the drug-disease network and the drug-target network into one coherent model via cross-network embedding. As a result, iDrug is able to take full usage of the knowledge within these two biological networks to better exploit new biomedical insights of drug-target-disease. Therefore, iDrug has broad applications in drug discovery.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Targeted therapies and personalized treatments are the most promising strategies to treat complex human diseases, especially for cancer. Accurate identification of drug mechanism of actions (MoAs) is thus of great importance in drug discovery process. Two important tasks, drug repositioning (also known as drug-disease prediction) [1] and drug-target prediction [2], have been actively investigated to better understand the drugs' MoAs. Drug repositioning (as well as drug-target prediction), aiming to identify new indications (new targets) for existing drugs, had gained increasing interests over the last decades. Traditional *in vitro* and *in vivo* prediction of interactions between drug-disease or drug-target are desirable in many cases, but with an expensive and protracted process of experimentation and testing [3, 4]. On the other hand, computational approaches provide an alternative tool to efficiently predict potential candidates with certain reasonable accuracy, thus narrowing down the search space to be investigated by the follow-up wet-lab experiments [5].

Many computational approaches have been developed for each task independently. Two excellent surveys on drug repositioning [6] and drug-target prediction [7] contain a very detailed overview of different machine learning techniques for each domain. Among many different methods, one of the most popular approaches is the network-based inference model [8, 9], which formulates drug-disease (or drug-target) prediction as a missing link prediction problem on a heterogeneous network. Advances in this direction are also essential for identifying biological significance of new disease genes, and for uncovering drug targets and biomarkers for complex diseases.

For drug repositioning, Wu et al. proposed a weighted bipartite network to identify the connected communities of drugs and diseases using a network clustering approach [10]. Two network topology-based methods, ProbS and HeatS, were also introduced to predict new indications for different diseases by considering the disease pathway and phenotype features [11]. MBiRW further used a bi-random walk algorithm on a two-layer network to identify potential novel indications for a given drug [12]. Zhang et al. introduced a matrix factorization method to predict novel drug-disease associations by integrating multiple drug and disease similarities [13]. Similarly, Chen et al. further applied multiple kernel learning to incorporate multiple heterogeneous data sources of drug and disease into the prediction framework [14]. Recently, Zeng et al. introduced a network-based deep learning method for *in silico* drug

repositioning, which could learn nonlinear features of drugs from the heterogeneous networks by a multi-modal deep autoencoder [15].

For drug-target prediction, Bleakley et al. applied support vector machine to predict novel targets based on a bipartite local model [16]. Chen et al. presented a random walk with restart on a bipartite network to predict potential drug-target interactions on a large scale [17]. Ezzat et al. proposed two matrix factorization methods for drug-target prediction and constantly boost the accuracy via graph regularization. In addition, Zheng et al. proposed a coupled matrix factorization model, which projected drugs and targets into a common low-rank feature space [18]. Nascimento et al. also integrated multiple heterogeneous information sources for both drugs and targets by using multiple kernel learning [19]. Chen et al. developed several effective computational models to predict potential drug–target interactions from heterogeneous biological data, which could provide better understanding of various interactions [20]. Luo et al. presented a network integration pipeline for drug-target prediction via low-rank matrix factorization, which integrated diverse information from heterogeneous data sources [21]. Recently, Lee et al. constructed a novel drug-target prediction model to extract local residue patterns of target protein sequences using a CNN-based deep learning approach, which exhibited better performance than previous shallow models [22].

Computational frameworks from the two domains share many common characteristics. First, most of them, for both domains, obey the *guilt by association* principle, which assumes that similar drugs tend to treat (bind to) similar diseases (targets) with high probability and vice versa [6, 7]. Second, they have exactly the same network structures, i.e., a bipartite network consisting of one layer of drugs and the other layer of diseases (targets). Moreover, they adopt many common features of drugs, targets, and diseases, such as drug's chemical structure, target's sequence, and disease phenotype. Based on those commonalities, it is not surprising that same machine learning methods can be adopted or directly applied to two prediction tasks interchangeably with barely no compromise on accuracy. For example, random walk with restart [17], couple matrix factorization [18], and multiple kernel learning [19] were first successfully proposed to predict the new drug-target interactions. Subsequently, they can be perfectly adaptive to solve drug repositioning problem as well [12–14].

It is generally a challenging task to compare different approaches in either domain for a couple of reasons. First, many methods are not only different in the computational approaches that they used, most of the time, they are also very different in the data that they analyzed. Sometimes, it is just impractical to separate data from computational approaches. Second, many of the approaches are based on the global structure of the network in predicting missing links. Although they normally give better results comparing to methods based on local structure, they may not provide intuitive explanations of the predicted results. Third, most computational approaches cannot afford experimental validations. Coupled with the issue of different types of input data required, it is hard to compare different methods objectively. One possible solution for this problem is through crowdsourcing projects such as the Dream Challenges [23]. Regardless these challenges, there is still room to improve computational approaches. In particular, most of the existing studies considered drug-disease and drug-target prediction as two isolated tasks and the relationships between these two domains—namely, cross-domain knowledge—are typically ignored. Some recent studies have shown that such cross-domain knowledge is very useful in improving the success rate of drug development [24]. Indeed, therapeutic effect of drugs on a disease is through their abilities to modulate the biological targets within the disease pathways, which in turn promotes healthy functioning of the metabolic system and cure the disease. In other words, targets can provide evidence to understand drugs' MoAs, which could serve as a useful bridge in drug discovery. Therefore, it is reasonable to integrate drug repositioning and drug-target prediction together to better exploit different

domain-specific knowledge. Wang et al. presented a three-layer heterogeneous network model named TL_HGBI to learn the potential relationships among drug-target-disease in a unified model [25]. However, it can only be applied to a small-scale dataset since its drug layer required that the drugs must interact with known targets and diseases, leading to the issue of data sparsity. In real-life situation, due to the nature of data collection (i.e., data were generated by different labs in different time), it is very unlikely that the triple relationships <drug, target, disease> are always available while integrating drug-disease and drug-target domain. Moreover, the random walk with restart algorithm adopted by TL_HGBI assumed that the next step of the random walker only depended on the current node, which might suffer from the bias induced by noise. The limitation motivates us to seek answers to a natural question: can we still leverage all the available data provided in two domains to alleviate the data sparsity issue, generate better performance, and extend to large-scale dataset in drug discovery?

In this work, we propose a novel framework—iDrug, which not only jointly performs drug repositioning and drug-target prediction at the same time, but also integrates diverse information from heterogeneous data sources. The key idea of iDrug is mainly inspired by cross-network embedding [26–28], which aims to borrow information from some related domains to achieve better performance in the domain of interest. Compared with single network embedding, cross-network embedding simultaneously considers at least two types of networks from different domains [27]. To be specific, two types of relationships are considered for each node in the networks: (i) *within-network relationship*, which preserves the specific structural feature of a node in its own domain; (ii) *cross-network dependency*, which describes the associations between nodes across different networks/domains. Fig 1 shows an example of two heterogeneous networks corresponding to two domains $\mathcal{D}_1$ and $\mathcal{D}_2$. In each domain, the edges connecting those nodes in the same layer (e.g., the disease layer) are defined based on their similarities (e.g., disease-disease similarity). The edges across two different layers in the same domain (e.g., drug-disease links) are labeled based on known associations. The goal is to predict novel cross links in the same domain, which solves the drug repositioning or drug-target prediction problem. Note that several drug nodes such as {②, ③, ⑤, ⑦, ⑧, ⑩} in Fig 1 also have another type of links, which connects the same drug nodes between $\mathcal{D}_1$ and $\mathcal{D}_2$. We call these links as

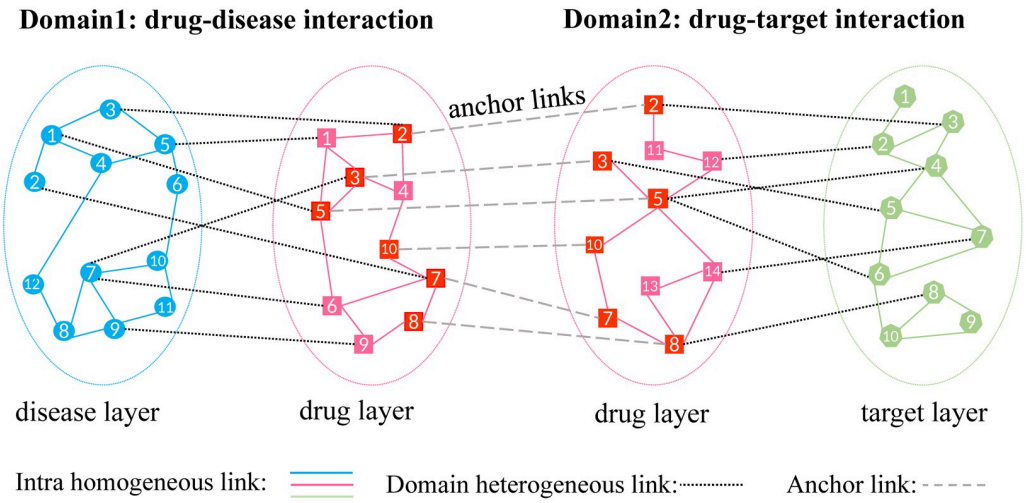

Fig 1. An overview of iDrug. An illustration of the cross-network framework across two domains: drug-disease and drug-target networks. iDrug requires only partially overlapped drug nodes between these two domains. The anchor links among drug nodes are used to transfer domain knowledge across the networks.

https://doi.org/10.1371/journal.pcbi.1008040.g001

anchor links, which have been shown to play a central role in multi-layered network mining tasks [29, 30]. We thus regard these anchor links as bridges to fully transfer domain-specific knowledge to benefit each other during the learning process. iDrug has several advantages over existing single-domain methods. First, unlike single-domain approaches, iDrug is able to jointly perform two tasks, drug repositioning and drug-target prediction in one unified model, which has broader applicability in real-life drug discovery. In addition, by transferring knowledge across different domains, iDrug can substantially alleviate data sparsity issue due to complementary property of the two related domains and thus mutually enhance the performance. Moreover, unlike some previous methods that requires totally overlap of the drug layer [25, 31], iDrug only requires partial overlap of drugs from the two domains. Therefore, it will be able to include more data from both domains. Overall, iDrug provides an alternative opportunity for us to better gain new biomedical insights of drug-target-disease relationships.

Materials and methods

In this section, we first formulate the drug repositioning and drug-target prediction problem as a matrix completion problem. We then provide the details of multiple heterogeneous data sources, the framework of iDrug, and the learning algorithm.

Problem definition

Notation. Following the convention, we use bold upper-case for matrices (e.g., $\mathbf{A}$), the (i, j)-th element of matrix $\mathbf{A}$ denotes as $\mathbf{A}(i, j)$. $\mathbf{A}(i,:)$ or $\mathbf{A}(:, i)$ denotes its i-th row or column and $\mathbf{A}^T$ denotes the transpose of matrix $\mathbf{A}$.

Domain 1: Drug-disease prediction. In domain $\mathcal{D}_1$, we try to predict new indications of drugs using a drug-disease bipartite network, drug-drug similarity, and disease-disease similarity information. We start by representing the bipartite network as a sparse $n_1 \times m_1$ matrix $\mathbf{X}^{(1)}$, where n_1 is the number of drugs and m_1 is the number of diseases. $\mathbf{X}^{(1)}(i, j) = 1$ if i-th drug and j-th disease are known to interact and $\mathbf{X}^{(1)}(i, j) = 0$ otherwise. The drug-drug similarity can be encoded into a $n_1 \times n_1$ square matrix $\mathbf{A}_u^{(1)}$, with $\mathbf{A}_u^{(1)}(i, j)$ representing the similarity score between i-th and j-th drugs. Analogously, the disease-disease similarity can be represented by a $m_1 \times m_1$ square matrix $\mathbf{A}_v^{(1)}$. We can then regard this problem as an analog of user-item preferences problem in recommender system [32], in which users and items denote drugs and diseases, respectively. The goal is to predict the new interactions between drugs and diseases by completing the matrix $\mathbf{X}^{(1)}$.

Domain 2: Drug-target prediction. Similarly, we denote the drug-target bipartite network as a sparse matrix $\mathbf{X}^{(2)} \in \mathbb{R}^{n_2 \times m_2}$ in domain $\mathcal{D}_2$, where n_2 and m_2 are the numbers of drugs and targets, respectively. The matrices $\mathbf{A}_u^{(2)} \in \mathbb{R}^{n_2 \times n_2}$ and $\mathbf{A}_v^{(2)} \in \mathbb{R}^{m_2 \times m_2}$ denote the drug-drug similarity and target-target similarity, respectively. Note that both $\mathbf{A}_u^{(1)}$ and $\mathbf{A}_u^{(2)}$ represent drug-drug similarity but in different domains, resulting in different sizes. Here we only require partial overlap of drugs between $\mathcal{D}_1$ and $\mathcal{D}_2$. Novel drug-target interactions can then be inferred by completing the matrix $\mathbf{X}^{(2)}$.

Construction of drug-disease and drug-target network

Cross-network construction. To construct the networks, we consider two different datasets. For the first one, the initial drug-disease interactions in domain $\mathcal{D}_1$ are obtained from the Comparative Toxicogenomics Database (CTD): http://ctdbase.org/. For the second dataset, drug-disease relationships are obtained from Gottlieb et al. [33], which has been frequently used in many previous studies [12, 34] and is regraded as a gold standard dataset.

Table 1. Statistics of drug-disease network (Domain 1) from CTD database and drug-target network (Domain 2) from DrugBank database. 469 drugs are overlapped between two networks in total.

Domain	|Drug|	|Target|	|Disease|	|Interaction|
Domain 1: drug-disease	1,321	-	3,966	111,481
Domain 2: drug-target	946	3,610	-	10,234

https://doi.org/10.1371/journal.pcbi.1008040.t001

The CTD dataset contains 1, 048, 547 drug-disease associations. We only focus on those diseases that have OMIM https://www.omim.org/ identifiers for conveniently computing disease similarity scores later. We thus collect total of 1, 321 drugs, 3, 966 disease as well as 111, 481 drug-disease interactions.

The initial drug-target associations in domain $\mathcal{D}_2$ can be directly obtained from DrugBank https://www.drugbank.ca/. We mainly focus on the approved small molecule compounds and require each drug to have at least two targets, resulting in 946 drugs, 3, 610 targets, and 10, 234 drug-target interactions. The requirement to have at least two targets for each drug is motivated by the notion that a drug can be used to treat a different disease most likely due to its off target activities. Across the two domains $\mathcal{D}_1$ and $\mathcal{D}_2$, we have 469 common drugs in the two networks. The statistics of these two networks are shown in Table 1.

Dataset two is much smaller and only contains 1, 933 known drug-disease associations involving 593 drugs and 313 diseases. The drug-disease relationships in this dataset are human curated and are believed to be more reliable than the ones in the first dataset. We further collect the targets of those 593 drugs from the DrugBank database and construct the drug-target network that consists of 1, 011 targets and 3, 427 known drug-target interactions.

In addition to the cross links of the two heterogeneous networks, we also construct homogeneous edges and their weights based on the following similarity measures.

Drug-drug similarity. Although there are a number of measurements developed for computing drug-drug similarities, a recent study showed that Tanimoto coefficient similarity is highly efficient for fingerprint-based similarity measurement [35]. Chemical structures of drugs in Canonical SMILES form are directly downloaded from DrugBank. The Chemical Development Kit is then applied to compute the Tanimoto similarity score of any two drugs using their corresponding 2D chemical fingerprints [36]. Briefly, two drugs have a higher similarity score if they have more similar of chemical structures.

Target-target similarity. Protein targets consist of long chains of amino acid sequences, which perform a vast array of functions within organisms. Target-target similarity scores are thus calculated using the Smith-Waterman algorithm (e.g., 11 for gap open penaltyand 1 for its extension, BLOSUM62 for the scoring matrix.) based on their amino acid sequences. The similarity scores are then normalized into [0, 1] using the same method proposed in a previous work [16].

Disease-disease similarity. A recent study shows that a disease similarity defined based on the semantic similarity between MeSH terms describing the diseases is an accurate measure for heritable diseases at the molecular level [37]. The measure is defined based on the concept of information content of a MeSH term in an ontology, defined as the negative logarithm of the probability. Therefore, the disease similarity measure is unbounded, non-negative real number. We directly download their similarities from the Disimweb server http://www.paccanarolab.org/disimweb.

For the CTD dataset, the intra-similarity distributions of drug-drug and disease-disease are shown in Fig 2. The intra-similarity distributions of drug-drug and target-target are shown in Fig 3. The drug-drug similarity distributions from the two domains follow similar patterns.

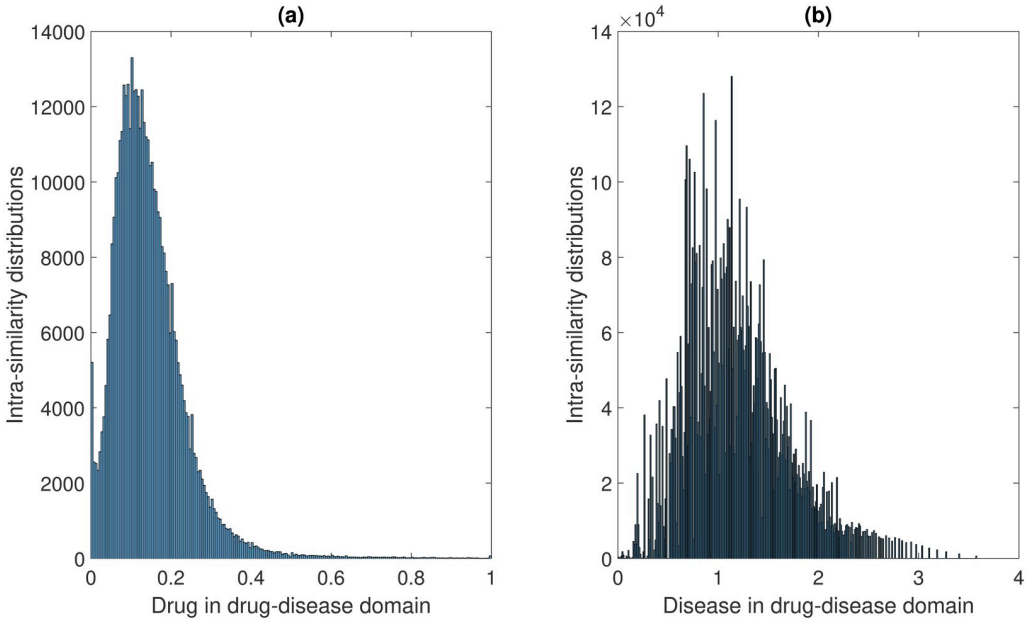

Fig 2. The intra-similarity distributions in drug-disease domain. (a) The intra-similarity distributions of drug pairs, the drug-drug similarities are calculated based on Tanimoto Score. (b) The intra-similarity distributions of disease pairs, the disease-disease similarities are computed based on the semantic similarity of MeSH terms. Note that all the self-similarity values of drugs and diseases have already been excluded in the histograms.

https://doi.org/10.1371/journal.pcbi.1008040.g002

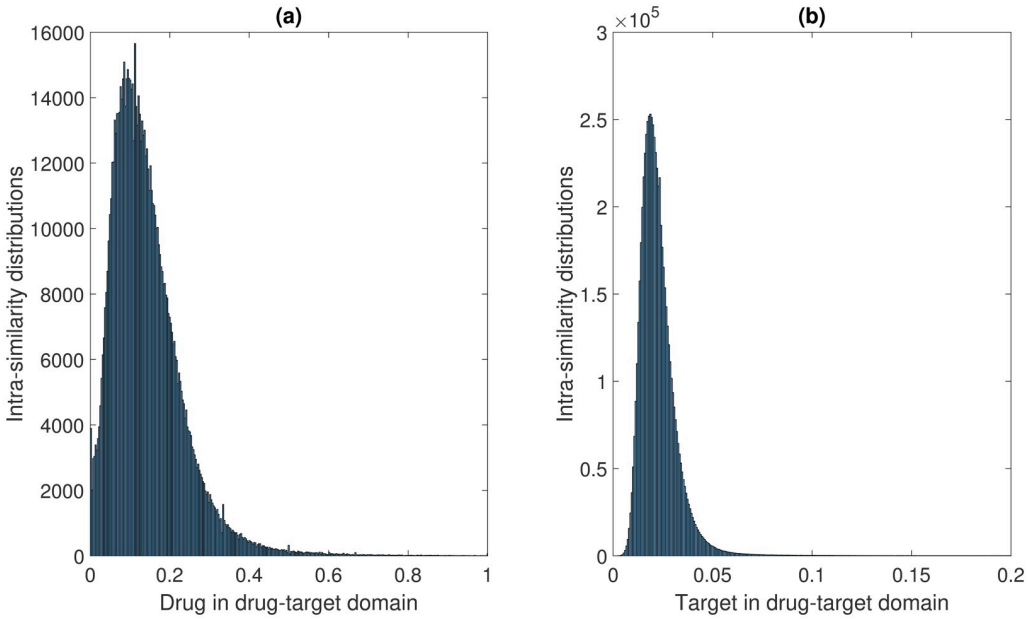

Fig 3. The intra-similarity distributions in drug-target domain. (a) The intra-similarity distributions of drug pairs, the drug-drug similarities are calculated based on Tanimoto Score. (b) The intra-similarity distributions of target pairs, the target-target similarities are calculated using the Smith-Waterman algorithm on target sequences. Note that all the self-similarity values of drugs and targets have already been excluded in the histograms. For target-target similarities, we only show the similarity values within [0, 0.2] since most of them are located in this range.

https://doi.org/10.1371/journal.pcbi.1008040.g003

Overall, most drug pairs have similarities smaller than 0.4, which is not surprising given that most drugs may not be related. The disease similarities are not normalized. For two diseases, the lower their shared MeSH terms on the ontology, the higher the information content and their similarities. The target-target similarities are generally very small because of the diverse set of targets.

The iDrug model

The key idea behind iDrug is to treat the problem as a cross-network embedding problem [38] by considering both within-network and cross-network relationships. We next provide more details for within-network factorization and cross-network consistency, and propose the unified model.

1) Within-network factorization: For a single-domain such as drug-disease prediction, we adopt the basic idea of graph regularized non-negative matrix factorization [38–40], which decomposes the drug-disease interaction matrix $\mathbf{X}^{(1)} \in \mathbb{R}^{n_1 \times m_1}$ into two r_1-rank feature matrices $\mathbf{U}^{(1)} \in \mathbb{R}^{n_1 \times r_1}$ (i.e., drugs' feature space) and $\mathbf{V}^{(1)} \in \mathbb{R}^{m_1 \times r_1}$ (i.e., diseases' feature space) by minimizing the following objective function:

$$\min_{\mathbf{U}^{(1)} \geq 0, \mathbf{V}^{(1)} \geq 0} \| \mathbf{W}^{(1)} \odot (\mathbf{X}^{(1)} - \mathbf{U}^{(1)} \mathbf{V}^{(1)^T}) \|_F^2 + \alpha \cdot (\mathrm{Tr}(\mathbf{U}^{(1)^T} \mathbf{L}_u^{(1)} \mathbf{U}^{(1)}) + \mathrm{Tr}(\mathbf{V}^{(1)^T} \mathbf{L}_v^{(1)} \mathbf{V}^{(1)})) \quad (1)$$

where $\mathbf{L}_u^{(1)} = \mathbf{D}_u^{(1)} - \mathbf{A}_u^{(1)}$, $\mathbf{L}_v^{(1)} = \mathbf{D}_v^{(1)} - \mathbf{A}_v^{(1)}$ and $\odot$ is the Hadamard product, $\mathrm{Tr}(\cdot)$ is the Trace operator, $\|\cdot\|_F$ is the matrix Frobenius norm and α is the regularization parameter. $\mathbf{D}_u^{(1)}$ and $\mathbf{D}_v^{(1)}$ are the diagonal degree matrices of $\mathbf{A}_u^{(1)}$ and $\mathbf{A}_v^{(1)}$, such as $\mathbf{D}_u^{(1)}(i,i) = \sum_{j=1}^{n_1} \mathbf{A}_u^{(1)}(i,j)$; $\mathbf{W}^{(1)} \in \mathbb{R}^{n_1 \times m_1}$ is weight matrix indicating the weight of entities on $\mathbf{X}^{(1)}$. Typically, a smaller weight is assigned to unobserved samples. Based on the strategy for one-class collaborative filtering [38, 40], we set $\mathbf{W}^{(1)}(i,j) = 1$ if $\mathbf{X}^{(1)}(i,j)$ is observed and $\mathbf{W}^{(1)}(i,j) = w \in [0, 1)$ otherwise. We use trace optimization (i.e., the second term) to achieve within-network smoothness, which ensures that the low-rank representations of nodes i and j in the same layer will be close to each other. Finally, the non-negativity constraints on the factor matrices $\mathbf{U}^{(1)}$ and $\mathbf{V}^{(1)}$ lead to more interpretable results [41]. Similarly, for drug-target domain, we can decompose matrix $\mathbf{X}^{(2)} \in \mathbb{R}^{n_2 \times m_2}$ into two r_2-rank feature matrices $\mathbf{U}^{(2)} \in \mathbb{R}^{n_2 \times r_2}$ (drugs' feature space) and $\mathbf{V}^{(2)} \in \mathbb{R}^{m_2 \times r_2}$ (targets' feature space) in a similar way as the Eq (1).

2) Cross-network consistency: iDrug further captures cross-network relationships by making the hypotheses that the partial overlap of drugs are consistent with each other across the two domains because they all represent the same drugs. To achieve this goal, we introduce a drug mapping matrix $\mathbf{S}^{(1,2)} \in \mathbb{R}^{n_2 \times n_1}$ to represent the anchor links cross domain $\mathcal{D}_1$ and $\mathcal{D}_2$. To be specific, $\mathbf{S}^{(1,2)}(i,j) = 1$ if the i-th row of $\mathbf{U}^{(2)}$ and j-th row of $\mathbf{U}^{(1)}$ represent the same drug; $\mathbf{S}^{(1,2)}(i,j) = 0$ otherwise. Note that at most one element in each row of $\mathbf{S}^{(1,2)}$ can be 1 because of the one-to-one constraint of anchor links across the two domains. We then observe that $\mathbf{S}^{(1,2)}$ $\mathbf{U}^{(1)}$ in fact project the partial overlap of drug feature space from domain $\mathcal{D}_1$ to domain $\mathcal{D}_2$. In addition, if two drugs are similar in domain $\mathcal{D}_1$, they should also be similar after projecting to domain $\mathcal{D}_2$. The cross-network consistency of the partial overlap of drugs can thus be achieved by minimizing the following disagreement [42]:

$$D(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \| \mathbf{S}^{(1,2)} \mathbf{U}^{(1)} (\mathbf{S}^{(1,2)} \mathbf{U}^{(1)})^T - \mathbf{U}^{(2)} \mathbf{U}^{(2)^T} \|_F^2 \quad (2)$$

In other words, the feature space $\mathbf{U}^{(1)}$ in the drug domain $\mathcal{D}_1$ should match the feature space $\mathbf{U}^{(2)}$ in the drug domain $\mathcal{D}_2$ as much as possible because of their overlapped drugs.

3) The unified model: Finally, we can integrate the domain-specific within-network objective with respect to drug-target and drug-disease in Eq (1), with the cross-network consistency Eq (2) into a unified objective function as follows:

$$\begin{aligned}\min \quad & \mathcal{J} = \underbrace{\sum_{i=1}^{2} \| \mathbf{W}^{(i)} \odot (\mathbf{X}^{(i)} - \mathbf{U}^{(i)}\mathbf{V}^{(i)^T}) \|_F^2}_{\text{domain factorization}} + \underbrace{\beta \| \mathbf{S}^{(1,2)}\mathbf{U}^{(1)}(\mathbf{S}^{(1,2)}\mathbf{U}^{(1)})^T - \mathbf{U}^{(2)}\mathbf{U}^{(2)^T} \|_F^2}_{\text{cross-network consistency}} \\ + \quad & \underbrace{\alpha \sum_{i=1}^{2} (\mathrm{Tr}(\mathbf{U}^{(i)^T}(\mathbf{D}_u^{(i)} - \mathbf{A}_u^{(i)})\mathbf{U}^{(i)}) + \mathrm{Tr}(\mathbf{V}^{(i)^T}(\mathbf{D}_v^{(i)} - \mathbf{A}_v^{(i)})\mathbf{V}^{(i)}))}_{\text{within-network smoothness}} + \underbrace{\gamma \sum_{i=1}^{2} (\| \mathbf{U}^{(i)} \|_1 + \| \mathbf{V}^{(i)} \|_1)}_{\text{regularization}} \\ \text{s.t.} \quad & \mathbf{U}^{(i)} \geq 0, \mathbf{V}^{(i)} \geq 0, \text{ for } i = 1, 2\end{aligned} \tag{3}$$

where the l_1-norm penalty on $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ (e.g., $\|\mathbf{A}\|_1 = \Sigma_{ij}\, \mathbf{A}_{ij}$) can achieve a more sparse solution [43]. The regularization parameters α, β, and γ can adjust the relative importance of within-network smoothness, cross-network consistency, and sparseness of optimal solutions, which will be studied later. For convenience, all symbols in Eq (3) are summarized in Table 2.

Learning algorithm

In this section, we provide rigorous theoretical analysis of iDrug in terms of its correctness, convergence, and complexity.

The objective function in Eq (3) is non-convex when considering all variables. We can optimize it by using the multiplicative update minimization approach [41], i.e., the objective function is alternately minimized with respect to one variable while fixing others. This procedure repeats until convergence, i.e., $\| \mathcal{J}^{(t+1)} - \mathcal{J}^{(t)} \| \leq \delta$, where δ is a small constant. The optimization procedure is summarized in S1 Fig. The details of the correctness and convergence of our algorithm can be found in the supplementary materials (S1 Appendix). Here we only include

Table 2. The symbols used in the objective function Eq (3) and their descriptions.

Symbol	Definition and Description
$\mathbf{X}^{(1)}$,$\mathbf{W}^{(1)}$	The adjacency matrix and the weight matrix of the known drug-disease interactions
$\mathbf{X}^{(2)}$,$\mathbf{W}^{(2)}$	The adjacency matrix and the weight matrix of the known drug-target interactions.
$\mathbf{U}^{(1)}$,$\mathbf{V}^{(1)}$	The low-rank representations of drugs and diseases in the drug-disease domain.
$\mathbf{U}^{(2)}$,$\mathbf{V}^{(2)}$	The low-rank representations of drugs and targets in the drug-target domain.
$\mathbf{S}^{(1,2)}$	The drug mapping matrix to denote anchor links across the two domains.
$\mathbf{A}_u^{(1)}$, $\mathbf{D}_u^{(1)}$	The drug-drug similarity matrix and its degree matrix in the drug-disease domain.
$\mathbf{A}_u^{(2)}$, $\mathbf{D}_u^{(2)}$	The drug-drug similarity matrix and its degree matrix in the drug-target domain.
$\mathbf{A}_v^{(1)}$, $\mathbf{D}_v^{(1)}$	The target-target similarity matrix and its degree matrix in the drug-disease domain.
$\mathbf{A}_v^{(2)}$, $\mathbf{D}_v^{(2)}$	The disease-disease similarity matrix and its degree matrix in the drug-target domain.
n_1,m_1	The number of drugs and diseases in the drug-disease domain.
n_2,m_2	The number of drugs and targets in the drug-target domain.
r_1,r_2	The ranks of matrices $\{\mathbf{U}^{(1)}, \mathbf{V}^{(1)}\}$ and $\{\mathbf{U}^{(2)}, \mathbf{V}^{(2)}\}$.

https://doi.org/10.1371/journal.pcbi.1008040.t002

the updating formula of each variable as follows:

$$\mathbf{U}^{(1)} \leftarrow \mathbf{U}^{(1)} \odot \sqrt{\frac{\mathbf{X}^{(1)}\mathbf{V}^{(1)} + \alpha\mathbf{A}_u^{(1)}\mathbf{U}^{(1)} + 2\beta\Delta}{\mathbf{T}^{(1)} + \alpha\mathbf{D}_u^{(1)}\mathbf{U}^{(1)} + 2\beta\Theta + 0.5\gamma}} \tag{4}$$

$$\mathbf{U}^{(2)} \leftarrow \mathbf{U}^{(2)} \odot \sqrt{\frac{\mathbf{X}^{(2)}\mathbf{V}^{(2)} + \alpha\mathbf{A}_u^{(2)}\mathbf{U}^{(2)} + 2\beta\Xi}{\mathbf{T}^{(2)} + \alpha\mathbf{D}_u^{(2)}\mathbf{U}^{(2)} + 2\beta\mathbf{U}^{(2)}{\mathbf{U}^{(2)}}^T\mathbf{U}^{(2)} + 0.5\gamma}} \tag{5}$$

$$\mathbf{V}^{(i)} \leftarrow \mathbf{V}^{(i)} \odot \sqrt{\frac{{\mathbf{X}^{(i)}}^T\mathbf{U}^{(i)} + \alpha\mathbf{A}_v^{(i)}\mathbf{V}^{(i)}}{((\mathbf{R}^{(i)^T} + w^2\mathbf{V}^{(i)}{\mathbf{U}^{(i)}}^T)\mathbf{U}^{(i)} + \alpha\mathbf{D}_v^{(i)}\mathbf{V}^{(i)} + 0.5\gamma}} \tag{6}$$

where

$$\begin{aligned}
\Delta &= {\mathbf{S}^{(1,2)}}^T\mathbf{U}^{(2)}{\mathbf{U}^{(2)}}^T\mathbf{S}^{(1,2)}\mathbf{U}^{(1)} \\
\Theta &= {\mathbf{S}^{(1,2)}}^T\mathbf{S}^{(1,2)}\mathbf{U}^{(1)}{\mathbf{U}^{(1)}}^T{\mathbf{S}^{(1,2)}}^T\mathbf{S}^{(1,2)}\mathbf{U}^{(1)} \\
\Xi &= \mathbf{S}^{(1,2)}\mathbf{U}^{(1)}{\mathbf{U}^{(1)}}^T{\mathbf{S}^{(1,2)}}^T\mathbf{U}^{(2)} \\
\mathbf{R}^{(i)} &= (1 - w^2)\mathbf{I}^{(i)} \odot (\mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T) \\
\mathbf{T}^{(i)} &= (\mathbf{R}^{(i)} + w^2\mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T)\mathbf{V}^{(i)}
\end{aligned}$$

and $\mathbf{I}^{(i)}$ is an indicator matrix for the observed elements in $\mathbf{X}^{(i)}$, i.e., $\mathbf{I}^{(i)}(u, v) = 1$ if $\mathbf{X}^{(i)}(u, v) > 0$, and $\mathbf{I}^{(i)}(u, v) = 0$ otherwise.

Once we solve Eq (3), for a give drug i and disease j in the drug repositioning domain, we can infer their potential association by $\tilde{\mathbf{X}}^{(1)}(i,j) = \mathbf{U}^{(1)}(i,:)\mathbf{V}^{(1)}(j,:)^T$. Similarly, for a give drug i and target j in the drug-target domain, the novel drug-target associations can be inferred by computing $\tilde{\mathbf{X}}^{(2)}(i,j) = \mathbf{U}^{(2)}(i,:)\mathbf{V}^{(2)}(j,:)^T$.

Complexity. According to the updating rules (Eqs (4) to (6)), the time complexity of our optimization algorithm is $\mathcal{O}(k \cdot (nmr + n^2r + sr) + n^3 + nr^2)$, where k is the number of iterations, $n = \max\{n_1, n_2\}$, $m = \max\{m_1, m_2\}$, $r = \max\{r_1, r_2\}$, and $s = \max\{\text{nnz}(\mathbf{X}^{(1)}), \text{nnz}(\mathbf{X}^{(2)})\}$, where $\text{nnz}(\mathbf{X})$ is the number of non-zero elements in $\mathbf{X}$. In practice, $r \ll \min\{m, n\}$ and s is usually very small due to the sparsity of networks. Moreover, the $\mathcal{O}(n^3)$ term is from the matrix multiplication $\mathbf{S}^{(1,2)^T}\ \mathbf{S}^{(1,2)}$ in Eq (4). Since $\mathbf{S}^{(1,2)}$ is the very sparse mapping matrix and is unchanged in each iteration, we can thus cache $\mathbf{S}^{(1,2)^T}\ \mathbf{S}^{(1,2)}$ in advance to reduce the time complexity. The overall complexity of our algorithm can be denoted as $\mathcal{O}(k \cdot (nmr + n^2r))$ in total. Although iDrug performs cross-domain learning using two biological networks, the computational complexity remains the same as the state-of-the-art matrix factorization algorithms in a single domain, such as GRMF [34] for drug-target prediction.

Results

In this section, we conduct several experiments to evaluate the performance of our proposed iDrug on the two domains, respectively. Specifically, we perform the drug repositioning by fully using the knowledge containing in drug-target domain and vice versa in the cross-validation experiments. Many approaches exist for both problems. However, for some approaches, it is extremely challenging to compare their performance because these approaches are closely tied to the data that they are using. For example, the model structures are usually determined by the available data they have [15, 33]. In this study, we only compare four network-based

methods that can easily take drug/target/disease similarities as inputs. The baseline methods are as follows:

- RLS-Kron [44], a kernel-based classifier that combines chemical and genomic similarity matrices for drug-target prediction.
- TL_HGBI [25]: a random-walk based algorithm for a three layers drug-target-disease network to predict the new interactions between drugs and diseases (targets).
- MBiRW [12]: a bi-random walk algorithm on a bipartite network to identify potential indications by further adjusting the clustering of drugs and diseases.
- GRMF [34]: a matrix factorization method that uses graph regularization to learn low-rank representations for drugs and targets.

Although some of them are originally designed for drug repositioning (e.g., TL_HGBI and MBiRW), all of above methods can be directly applied to the two domains as discussed before. The parameters of these algorithms are first initialized as those in the original paper and then tuned for the optimal performance. For RLS-Kron, the regularization parameter $\sigma = 1$ and the kernel bandwidths $\gamma = 1$. For random-walk based algorithm TL_HGBI and MBiRW, we set their thresholds the same as the optimal settings in their original papers. For GRMF, we tune the regularization parameter by using grid search algorithm [34], and $\lambda_l = 0.5$, $\lambda_d = \lambda_t = 10^{-3}$ are chosen for the best performance. For iDrug, we set rank $r_1 = 90$ and $r_2 = 70$, weight in Eq (6) $w = 0.3$ and regularization parameters $\alpha = \beta = \gamma = 0.01$. The impact of regularization parameters are presented in the Sensitivity Analysis subsection.

Cross-validation for drug-disease prediction

In order to provide a complete picture of the performance for each approach, we conduct five-fold cross-validation experiments under the following three scenarios for drug-disease prediction [19, 34, 45]:

1. 'pair prediction' scenario, which predicts unknown interactions between known drugs and diseases. All known drug-disease associations are split into five folds, in which four folds are used for training and one fold for testing.
2. 'new drug' scenario, which predicts diseases for new drugs. In this scenario, the drugs are divided into five disjoint subsets. The drug-disease associations of four folds of drugs are used for training and the rest pairs are used for testing.
3. 'new disease' scenario, which predicts drugs for new diseases. The setting is similar to scenario two, but data are separated based on diseases.

For scenario 1, the original drug-disease associations are very sparse with a large fraction of unknown interactions. We choose one fold of known pair interactions as positive samples and further randomly select an equal number of unknown interactions as negative samples as test set. The four folds of known interactions and the rest of unknown pairs are used to train the model [19, 21]. By varying the rank threshold, we can calculate various true positive rate (TPR), false positive rate (FPR), Precision and Recall values. We then use Area Under the Receiver Operating Characteristic curve (AUROC) and Area Under the Precision Recall curve (AUPR) to assess the performance of different models [19, 34, 45].

For scenarios 2 and 3, we evaluate different methods mainly based on the precision of top-k metric because we are more interested in the top-k ranked candidates for 'new drug' or 'new disease' in the drug discovery process [6, 7]. Note that the entire drug-target associations are

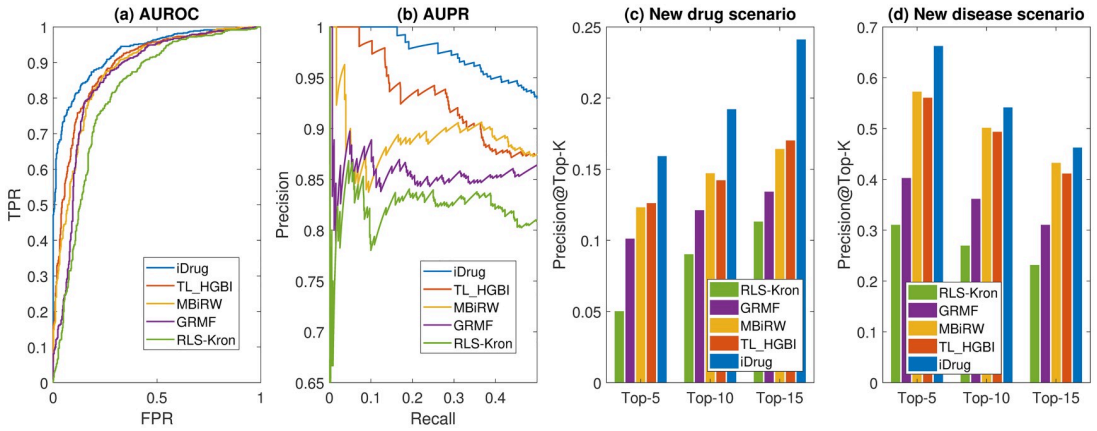

Fig 4. Comparison on the performance of different methods on drug repositioning. (a) The AUROC curves for the 'pair prediction' scenario. (b) The AUPR curves for the 'pair prediction' scenarios. (c) Precision of the top-*k* candidates for the 'new drug' scenario. (d) Precision of the top-*k* candidates for the 'new disease' scenario.

https://doi.org/10.1371/journal.pcbi.1008040.g004

all incorporated when performing the task of drug repositioning. For each method, five-fold cross-validation experiments are repeated 10 times independently and the average performance is reported.

Experimental results. Our results (Fig 4) clearly show that the proposed iDrug consistently outperforms all the other approaches for all three scenarios. Fig 4(a) shows that iDrug achieves an AUROC value of 0.9213, which is better than that of the other methods in the same experimental scenario (e.g., TL_HGBI: 0.886, MBiRW: 0.879, GRMF: 0.863, and RLS-Kron: 0.844). Meanwhile, iDrug achieves an AUPR of 0.938, which are higher than all the other approaches (TL_HGBI: 0.881, MBiRW: 0.876, GRMF: 0.847, and RLS-Kron: 0.813). Several interesting observations can be made based on the results in Fig 4. First, iDrug and TL_HGBI, both of which integrate target information, perform better than the rest methods, indicating the contributions of target information for drug-disease prediction. In a previous study [25], it was shown that the performance gradually decreased when some of the observed drug-target links in the network were removed randomly. Our results here further confirm that it is preferable to jointly model drug-target-disease relationships to better understand drug's MoAs. Second, comparing iDrug with TL_HGBI, TL_HGBI requires a common set of drug nodes across the two domains, therefore can be viewed as a sub-network of iDrug and is more prone to the issue of data sparsity. For instance, for a new compound added to the network, its similarity with existing drugs can be calculated. However, its interactions with diseases and targets might be completely unknown. TL_HGBI cannot address such a *cold-start* problem for novel drug discovery. In contrast, iDrug overcomes this issue by jointly learning on larger drug-disease and drug-target networks, and transferring domain-specific knowledge through anchor links cross domains [27]. Larger networks tend to contain more information thus ease the issue of data sparsity. Therefore, iDrug is expected to perform better as shown here. In comparing iDrug with MBiRW, both approaches employ the concept of drug community/cluster to improve their performance, but in different ways. The community of drugs in MBiRW was constructed based on common drug indications, which highly depended on the known drug-disease associations and might suffer from bias from data. Different from MBiRW, iDrug restricts the community of drugs by imposing consistency constraints among drugs from the two domains, i.e., $D(\mathbf{U}^{(1)}, \mathbf{U}^{(2)})$. It can therefore obtain more reliable community of drugs for both domains. iDrug achieves higher AUPR score than GRMF, presumably due to the fact that GRMF only incorporates information of drugs and diseases. In fact, GRMF

can be viewed as a degraded version of iDrug for single-domain prediction since they both try to learn the low-rank representations of drugs and diseases via matrix factorization with graph regularization. iDrug improves the accuracy by simply transferring knowledge from multiple domains. Finally, the kernel-based RLS-Kron model performs the worst among all the approaches since it is often hard to choose an appropriate kernel function, which usually requires more specific domain knowledge from the experts.

For scenarios 2 and 3 (Fig 4(c) and 4(d)), iDrug also perform better than the other approaches in term of the precision of top-5, top-10 and top-15 metrics. It is interesting to notice the difference in the two scenarios. The precision of top-*k* candidates In the 'new drug' scenario is much lower than the precision in the 'new disease' scenario. We suspect that one of the main reasons is that the number of overlapped drugs connecting the two domains is decreased when splitting the drugs into five disjoint sets in the 'new drug' scenario. The impact of anchor links is thus reduced and weak domain-specific knowledge can be transferred cross domains in the cross-validation experiments. In the 'new disease' scenario, the anchor links are preserved and rich domain-specific knowledge are transferred cross domains, resulting in a higher precision.

Cross-validation for drug-target prediction

In this section, we test and compare iDrug's performance with other approaches for the task of drug-target prediction. Similar to drug-disease prediction, we conduct five-fold cross-validation experiments under three scenarios: 'pair prediction', 'new drug', and 'new target' scenarios. The entire drug-disease network is preserved during the learning process. The experiments are also repeated 10 times independently and the average scores are reported. For the task of drug-target prediction, the performance is obtained by setting $\alpha = \beta = 0.01$ and $\gamma = 0.001$.

Results show that iDrug consistently outperforms all other methods in all three scenarios for drug-target prediction for all the measures (Fig 5). For instance, iDrug achieves a 0.897 AUPR score, much higher than that of the other approaches. The next two closest competitors are TH_HGBI (AUPR: 0.856) and MBiRW (AUPR: 0.849). In terms of the precision of top-*k* metric, iDrug is also able to better predict candidates for novel drugs and novel targets, for the same reason as we discussed in drug-disease prediction experiments. The superior

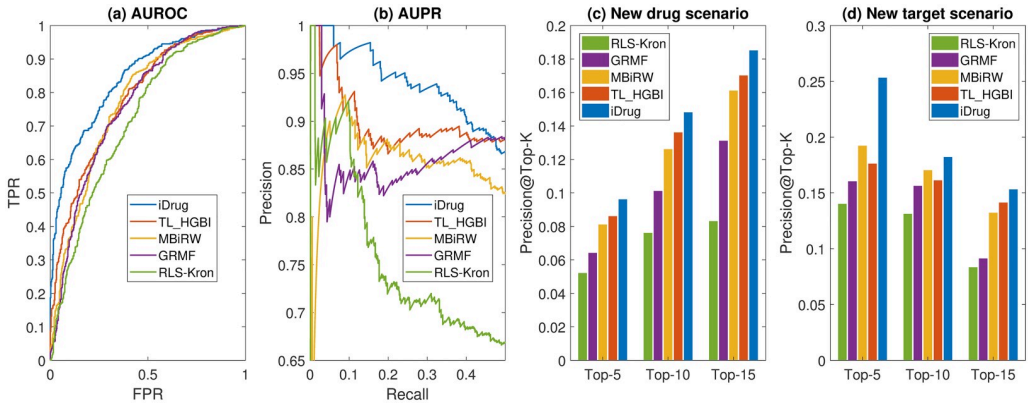

Fig 5. Performance comparison of different methods on drug-target prediction. (a) The AUROC curves for the 'pair prediction' scenario. (b) The AUPR curves for the 'pair prediction' scenario. (c) Precision of the top-*k* candidates for the 'new drug' scenario. (d) Precision of the top-*k* candidates for the 'new target' scenario.

https://doi.org/10.1371/journal.pcbi.1008040.g005

performance of iDrug demonstrates its potential on transferring knowledge across two related domains, thus serving as a promising tool for drug-target prediction.

Sensitivity analysis

There are six hyper parameters in our proposed framework: r_1, r_2, α, β, γ and w. For the ranks of latent matrices (r_1 and r_2), intuitively, the greater the ranks are, the more latent information can be captured, while the higher the computational costs are. In the experiments, we find that $r_i \leq 0.1 \min\{n_i, m_i\}$ can achieve a good trade-off between accuracy and running time. We thus set $r_1 = 90$ and $r_2 = 70$ in the sensitivity analysis of other parameters. Here we present the results of the impacts of these parameters for the task of drug repositioning using the AUPR measurement in scenario 1. The impacts of the regularization parameters for different scenarios (e.g., using precision of top-k candidates) in different domain (e.g., drug-target prediction) can be conducted in a similar fashion. We omit the details here.

The impact of α, β, and γ. As discussed earlier, α, β, and γ represent the contributions of within-domain smoothness, cross-network consistency, and the sparseness of solutions, respectively. We fix one of three parameters, and study the impacts of the remaining two on the inference results while fixing all the other parameters ($w = 0.3$, $r_1 = 90$ and $r_2 = 70$). For example, we first fix $\gamma = 0.01$ and vary α and β within $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$ by using the grid-search technique. Results in S2 Fig show that AUPR is very stable across the wide range for both α and β in the task of drug-disease prediction: all AUPR scores are over 0.910. Similar patterns are observed in studying the impacts of (α, γ) and (β, γ) pairs (S3 and S4 Figs, respectively). In general, a relatively high AUPR score can be achieved when $\alpha = \beta = \gamma \approx 0.01$.

The impact of w. The parameter $w \in (0, 1)$ in the weight matrix $\mathbf{W}$ denotes the cost assigned to unobserved samples, which is very useful for imbalance datasets [38, 40]. To perform sensitivity analysis, we fix $r_1 = 90$, $r_2 = 70$ and $\alpha = \beta = \gamma = 0.01$, and vary w within {0.1, 0.2, 0.3, . . ., 0.9}. The AUPR measure for drug-disease prediction is used to evaluate the performance. S5 Fig shows that iDrug is very robust to the regularization parameter w within (0, 1).

Finally, S6 Fig shows the rate of convergence of our optimization algorithm in the experiments. The values of the objective function in Eq (3) steadily decrease with more iterations and less than 100 iterations are sufficient for convergence.

Experimental results on a gold standard dataset

We further investigate the performance of various methods on a human curated dataset initially studied by Gottlieb et al. [33], which has been commonly used in many previous studies [12, 34]. This dataset only contains 1, 933 known drug-disease associations involving 593 drugs and 313 diseases. We further collect all the 1, 011 targets of those 593 drugs from DrugBank database, which consists of 3, 427 drug-target interactions. We evaluate different methods for the 'pair prediction' scenario for both drug-disease and drug-target predictions.

Results in Fig 6 show that iDrug performs better than the rest of the algorithms for both tasks using the gold standard dataset. For example, for drug-disease prediction, iDrug achieves 0.917 for AUROC and 0.926 for AUPR, which are higher than the two measures from any other methods and are consistent with the results based on the CTD dataset. It is also observed that comparing with other approaches, iDrug achieves the greatest improvement in terms of AUPR. The next two closest competitors are TH_HGBI (AUPR: 0.883) and MBiRW (AUPR:0.846). An explanation is that AUPR punishes highly ranked false positives much more than AUROC does, especially for sparse dataset [46]. Similar trends can be observed for the task of drug-target prediction.

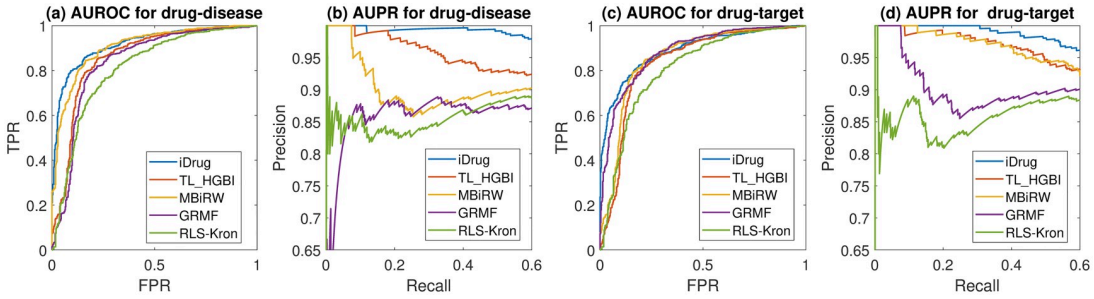

Fig 6. Performance comparison of different methods for the 'pair prediction' scenario on the gold standard dataset. (a) The AUROC curves for drug-disease prediction. (b) The AUPR curves for drug-disease prediction. (c) The AUROC curves for drug-target prediction. (d) The AUPR curves for drug-target prediction.

https://doi.org/10.1371/journal.pcbi.1008040.g006

Notice that the dataset from Gottlieb et al. [33] was collected almost a decade ago. All the drugs, diseases, and drug-disease interactions in the dataset can be found in the CTD dataset. Therefore, we can actually compare our novel prediction results based on the small dataset with the ones actually exist in the CTD dataset but not in the small dataset, which can be viewed as an indication of how iDrug might work in reality in predicting new drug-disease relationships. Towards that end, we collect the top 20 drugs for all the 313 diseases, and identify 514 new drug-disease interactions that can be found in the CTD dataset. Based on the hypergeometric distribution (all possible drug-disease pairs in CTD is 5239086, the number of positive pairs in CTD is 111481, the total number of drug-disease pairs considered 6260 and the number of positive pairs is 514), the result is extremely significant (p-value is 6.19×10^{-144}) and shows that the top ranked predictions by iDrug are greatly enriched in the CTD dataset.

Discussion

Drug repositioning and drug-target prediction have been widely recognized as promising tasks to better understand drug's MoAs. Existing machine learning methods often consider them as two isolated tasks and ignore the potential dependencies between these two related domains. In this paper, we propose a new learning framework, iDrug, which seamlessly integrates drug repositioning and drug-target prediction into a unified model. iDrug treats the problem as a cross-network embedding problem by considering both of within-network and cross-network relationships. We also develop the optimization algorithm to solve the problem and provide rigorous theoretical analysis concerning its correctness, convergence and complexity. Experimental results on both tasks of drug repositioning and drug-target prediction demonstrate that the proposed iDrug outperforms existing algorithms in all cases for both drug-disease and drug-target prediction. The efficiency and effectiveness of iDrug allows us to better understand new biomedical insights of drug-target-disease in drug discovery.

Our approach can be further improved in several directions. For example, although our model considers rich bioinformatics and cheminformatics data from publicly available databases, data quality can not be guaranteed and the network data may be incomplete and contain noise. Even with existing data, data representations, for example using binary fingerprints to represent drug chemical structures, can have significant impact on the prediction performance [33]. To alleviate the problem, one direction of future work is to incorporate more heterogeneous data sources describing drugs, targets, and diseases so that multiple data sources may provide complementary information to allow missing data imputation and noise removal. Other heterogeneous data sources, such as drug's side-effect, drug's ligand binding site information, target's Gene Ontology annotations, disease pathways, and diseases' Human Phenotype

Ontology, can also be integrated into the heterogeneous network for the two tasks [14, 18, 19, 21, 24, 47]. One possible extension is to use coupled matrix factorization to jointly capture the low-rank representations of the network and multiple data sources simultaneously.

Additionally, iDrug is built upon the matrix factorization framework, which approximates unobserved values using linear combinations of latent features. It is therefore can not capture more complex and non-linear drug-disease or drug-target interactions in the latent space. To overcome the linearity of iDrug, we will investigate the application of deep learning techniques, which have shown some initial success in capturing more complex and non-linear feature interactions in medicine and biology [48, 49]. One possible strategy is to train an autoencoder in an unsupervised way to capture the nonlinear feature representations of drugs, targets, and diseases. These nonlinear features can be integrated to identify novel drug-disease and drug-target interactions by using deep cross-network embedding techniques. Another inherent limitation for all factorization based approaches is the interpretability. Some existing works use heterogeneous knowledge graphs that incorporate many different types of data and identify paths from the graphs to interpret identified relationships [24]. However, it is a challenge task how one can combine these two strategies into one unified framework and is worthy further investigations.

More recently, a growing number of studies have suggested that non-coding RNAs (ncRNAs), especially microRNAs (miRNAs), play a significant role in affecting gene expressions and in disease progressions, making them a new class of drug targets [47, 50, 51]. Therefore, it becomes important to understand the relationship between drugs and miRNA targets. Theoretically, the framework proposed here can be applied to miRNA targets by defining a proper miRNA-miRNA similarity. One challenge is that knowledge about existing drug-miRNA interactions is limited. It is an interesting question to explore how to adopt existing approaches including iDrug to identify novel drug-miRNA interactions [50, 52].

Finally, most existing work including iDrug, implicitly assuming the monotherapy strategy in investigating drug-target-disease relationships, cannot easily incorporate polypharmacology or polytherapy strategy. On the other hand, polypharmacology and polytherapy, offer many advantages compared to monotherapy [53–56], including better efficacy, lower individual dosage, and reduced adverse effects. As an example of potential polytherapy, pentamidine and chlorpromazine show no anti-tumor activities when being administrated individually, but their combination inhibits tumor growth more effectively than paclitaxel, an anticancer chemotherapy drug. Furthermore, drug combinations often use existing drugs that have been approved by the Food and Drug Administration (FDA). Therefore, their toxic properties and side effects are usually well studied, and their combination could be directly used safely by patients [53–56]. For future work, we aim to extend our drug layers to contain drug pairs to study drug synergy for complex diseases. Eventually, wet lab experimental testing is a necessary step to validate outcomes of any computational approaches, which cannot be done without collaborations with investigators with expertise in biochemistry and drug development.

Supporting information

S1 Appendix. Optimization algorithm. The details of the optimization algorithm for solving $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$, as well as its correctness and convergence results can be found here.
(PDF)

S1 Fig. The pseudocode of the proposed iDrug. The algorithm to solve the objective function in Eq (3).
(TIF)

S2 Fig. The impact of α and β. Grid-based search method to study the impact of α and β with respect to the AUPR measurement for the task of drug repositioning, while γ is fixed to be 0.01.
(TIF)

S3 Fig. The impact of α and γ. Grid-based search method to study the impact of α and γ with respect to the AUPR measurement for the task of drug repositioning, while β is fixed to be 0.01.
(TIF)

S4 Fig. The impact of β and γ. Grid-based search method to study the impact of β and γ with respect to the AUPR measurement for the task of drug repositioning, while α is fixed to be 0.01.
(TIF)

S5 Fig. The impact of w. The impact of w with respect to the AUPR measurement for the task of drug repositioning.
(TIF)

S6 Fig. Convergence on empirical data. The convergence of iDrug on empirical data for the task of drug repositioning.
(TIF)

Author Contributions

Conceptualization: Huiyuan Chen, Jing Li.

Data curation: Huiyuan Chen, Feixiong Cheng, Jing Li.

Formal analysis: Huiyuan Chen, Jing Li.

Funding acquisition: Jing Li.

Investigation: Huiyuan Chen, Feixiong Cheng, Jing Li.

Methodology: Huiyuan Chen, Jing Li.

Project administration: Jing Li.

Resources: Huiyuan Chen, Jing Li.

Software: Huiyuan Chen.

Supervision: Feixiong Cheng, Jing Li.

Validation: Huiyuan Chen, Jing Li.

Visualization: Huiyuan Chen, Jing Li.

Writing – original draft: Huiyuan Chen, Jing Li.

Writing – review & editing: Huiyuan Chen, Jing Li.

References

1. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery. 2004; 3(8):673–683. https://doi.org/10.1038/nrd1468 PMID: 15286734
2. Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nature Chemical Biology. 2013; 9(4):232–240. https://doi.org/10.1038/nchembio.1199 PMID: 23508189

3. Zerbini LF, Bhasin MK, de Vasconcellos JF, Paccez JD, Gu X, Kung AL, et al. Computational repositioning and preclinical validation of pentamidine for renal cell cancer. Molecular Cancer Therapeutics. 2014; 13(7):1929–1941. https://doi.org/10.1158/1535-7163.MCT-13-0750 PMID: 24785412

4. Whitebread S, Hamon J, Bojanic D, Urban L. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. Drug Discovery Today. 2005; 10(21):1421–1433. https://doi.org/10.1016/S1359-6446(05)03632-9 PMID: 16243262

5. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational methods in drug discovery. Pharmacological Reviews. 2014; 66(1):334–395. https://doi.org/10.1124/pr.112.007336 PMID: 24381236

6. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Briefings in Bioinformatics. 2015; 17(1):2–12. https://doi.org/10.1093/bib/bbv020 PMID: 25832646

7. Ezzat A, Wu M, Li XL, Kwoh CK. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Briefings in Bioinformatics. 2019; 20(4):1337–1357. https://doi.org/10.1093/bib/bby002 PMID: 29377981

8. AY M, Goh KI, Cusick ME, Barabasi AL, Vidal M, et al. Drug–target network. Nature Biotechnology. 2007; 25(10):1119–1127. https://doi.org/10.1038/nbt1338

9. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011; 12(1):56–68. https://doi.org/10.1038/nrg2918 PMID: 21164525

10. Wu C, Gudivada RC, Aronow BJ, Jegga AG. Computational drug repositioning through heterogeneous network clustering. BMC Systems Biology. 2013; 7(5):S6. https://doi.org/10.1186/1752-0509-7-S5-S6

11. Chen H, Zhang H, Zhang Z, Cao Y, Tang W. Network-based inference methods for drug repositioning. Computational and Mathematical Methods in Medicine. 2015; 2015:130620. https://doi.org/10.1155/2015/130620 PMID: 25969690

12. Luo H, Wang J, Li M, Luo J, Peng X, Wu FX, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics. 2016; 32(17):2664–2671. https://doi.org/10.1093/bioinformatics/btw228 PMID: 27153662

13. Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In: AMIA Annual Symposium Proceedings. AMIA;2014;2014:1258-1267.

14. Chen H, Li J. A flexible and robust multi-source learning algorithm for drug repositioning. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM;2017. p. 510–515.

15. Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019; 35(24):5191–5198. https://doi.org/10.1093/bioinformatics/btz418 PMID: 31116390

16. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics. 2009; 25(18):2397–2403. https://doi.org/10.1093/bioinformatics/btp433 PMID: 19605421

17. Chen X, Liu MX, Yan GY. Drug–target interaction prediction by random walk on the heterogeneous network. Molecular BioSystems. 2012; 8(7):1970–1978. https://doi.org/10.1039/c2mb00002d PMID: 22538619

18. 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. ACM; 2013. p. 1025–1033.

19. Nascimento AC, Prudêncio RB, Costa IG. A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinformatics. 2016; 17(1):46. https://doi.org/10.1186/s12859-016-0890-3 PMID: 26801218

20. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, et al. Drug–target interaction prediction: databases, web servers and computational models. Briefings in Bioinformatics. 2016; 17(4):696–712. https://doi.org/10.1093/bib/bbv066 PMID: 26283676

21. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature Communications. 2017; 8(1):1–13. https://doi.org/10.1038/s41467-017-00680-8

22. Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Computational Biology. 2019; 15(6):e1007129. https://doi.org/10.1371/journal.pcbi.1007129 PMID: 31199797

23. AstraZeneca-Sanger Drug Combination DREAM Consortium. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nature Communications. 2019; 10(1):1–17.

24. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017; 6:e26726. https://doi.org/10.7554/eLife.26726 PMID: 28936969

25. Wang W, Yang S, Zhang X, Li J. Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics. 2014; 30(20):2923–2930. https://doi.org/10.1093/bioinformatics/btu403 PMID: 24974205

26. Tang J, Wu S, Sun J, Su H. Cross-domain collaboration recommendation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM;2012. p. 1285–1293.

27. Pan SJ, Yang Q. A survey on transfer learning. In: IEEE Transactions on Knowledge and Data Engineering. 2009; 22(10):1345–1359.

28. Chen H, Li J. Learning multiple similarities of users and items in recommender systems. In: 2017 IEEE International Conference on Data Mining. IEEE; 2017. p. 811–816.

29. Kong X, Zhang J, Yu PS. Inferring anchor links across multiple heterogeneous social networks. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM; 2013. p. 179–188.

30. Chen H, Li J. Exploiting structural and temporal evolution in dynamic link prediction. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management; 2018. p. 427–436.

31. Chen H, Li J. Modeling relational drug-target-disease interactions via tensor factorization with multiple web sources. In: The World Wide Web conference. ACM; 2019. p. 218–227.

32. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer. 2009;(8):30–37. https://doi.org/10.1109/MC.2009.263

33. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular Systems Biology. 2011; 7(1):496. https://doi.org/10.1038/msb.2011.26 PMID: 21654673

34. Ezzat A, Zhao P, Wu M, Li XL, Kwoh CK. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2017; 14(3):646–656. https://doi.org/10.1109/TCBB.2016.2530062

35. Lim H, Poleksic A, Yao Y, Tong H, He D, Zhuang L, et al. Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing. PLoS Computational Biology. 2016; 12(10):e1005135. https://doi.org/10.1371/journal.pcbi.1005135 PMID: 27716836

36. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. Journal of Chemical Information and Computer Sciences. 2003; 43(2):493–500. https://doi.org/10.1021/ci025584y PMID: 12653513

37. Caniza H, Romero AE, Paccanaro A. A network medicine approach to quantify distance between hereditary disease modules on the interactome. Scientific Reports. 2015; 5:17658. https://doi.org/10.1038/srep17658 PMID: 26631976

38. Chen C, Tong H, Xie L, Ying L, He Q. FASCINATE: fast cross-layer dependency inference on multi-layered networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 765–774.

39. Cai D, He X, Han J, Huang TS. Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010; 33(8):1548–1560. PMID: 21173440

40. Pan R, Zhou Y, Cao B, Liu NN, Lukose R, Scholz M, et al. One-class collaborative filtering. In: 2008 Eighth IEEE International Conference on Data Mining. IEEE; 2008. p. 502–511.

41. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems; 2001. p. 556–562.

42. Cheng W, Zhang X, Guo Z, Wu Y, Sullivan PF, Wang W. Flexible and robust co-regularized multi-domain graph clustering. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2013. p. 320–328.

43. Hoyer PO. Non-negative sparse coding. In: Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing. IEEE; 2002. p. 557–565.

44. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics. 2011; 27(21):3036–3043. https://doi.org/10.1093/bioinformatics/btr500 PMID: 21893517

45. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, et al. Toward more realistic drug–target interaction predictions. Briefings in Bioinformatics. 2014; 16(2):325–337. https://doi.org/10.1093/bib/bbu010 PMID: 24723570

46. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning. ACM; 2006. p. 233–240.

47. Chen X, Guan NN, Sun YZ, Li JQ, Qu J. MicroRNA-small molecule association identification: from experimental results to computational models. Briefings in Bioinformatics. 2020. 21(1):47–61.

48. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019; 25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7 PMID: 30617339

49. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular Systems Biology. 2016; 12(7):878. https://doi.org/10.15252/msb.20156651 PMID: 27474269

50. Qu J, Chen X, Sun YZ, Zhao Y, Cai SB, Ming Z, et al. In Silico prediction of small molecule-miRNA associations based on the HeteSim algorithm. Molecular Therapy-Nucleic Acids. 2019; 14:274–286. https://doi.org/10.1016/j.omtn.2018.12.002 PMID: 30654189

51. Christopher AF, Kaur RP, Kaur G, Kaur A, Gupta V, Bansal P. MicroRNA therapeutics: discovering novel targets and developing specific therapy. Perspectives in Clinical Research. 2016; 7(2):68–74. https://doi.org/10.4103/2229-3485.179431 PMID: 27141472

52. Zhao Y, Chen X, Yin J, Qu J. SNMFSMMA: using symmetric nonnegative matrix factorization and Kronecker regularized least squares to predict potential small molecule-microRNA association. RNA Biology. 2020. 17(2):281–291. https://doi.org/10.1080/15476286.2019.1694732 PMID: 31739716

53. Chen H, Li J. DrugCom: synergistic discovery of drug combinations using tensor decomposition. In: 2018 IEEE International Conference on Data Mining. IEEE; 2018. p. 899–904.

54. Cheng F, Kovács IA, Barabási AL. Network-based prediction of drug combinations. Nature Communications. 2019; 10(1):1197. https://doi.org/10.1038/s41467-019-09186-x PMID: 30867426

55. Chen H, Iyengar SK, Li J. Large-scale analysis of drug combinations by integrating multiple heterogeneous information networks. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; 2019. p. 67–76.

56. Chen X, Ren B, Chen M, Wang Q, Zhang L, Yan G. NLLSS: predicting synergistic drug combinations based on semi-supervised learning. PLoS Computational Biology. 2016; 12(7):e1004975. https://doi.org/10.1371/journal.pcbi.1004975 PMID: 27415801