Data and text mining

Drug repositioning through integration of prior knowledge and projections of drugs and diseases

Ping Xuan [1], Yangkun Cao [1],Tiangang Zhang [2,*], Xiao Wang [3], Shuxiang Pan [1], Tonghui Shen [1]

[1]School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China, [2]School of Mathematical Science, Heilongjiang University, Harbin 150080, China, [3]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Identifying and developing novel therapeutic effects for existing drugs contributes to reduction of drug development costs. Most of the previous methods focus on integration of the heterogeneous data of drugs and diseases from multiple sources for predicting the candidate drug-disease associations. However, they fail to take the prior knowledge of drugs and diseases and their sparse characteristic into account. It is essential to develop a method that exploits the more useful information to predict the reliable candidate associations.
Results: We present a method based on non-negative matrix factorization, DisDrugPred, to predict the drug-related candidate disease indications. A new type of drug similarity is firstly calculated based on their associated diseases. DisDrugPred completely integrates two types of disease similarities, the associations between drugs and diseases, and the various similarities between drugs from different levels including the chemical structures of drugs, the target proteins of drugs, the diseases associated with drugs, and the side effects of drugs. The prior knowledge of drugs and diseases and the sparse characteristic of drug-disease associations provide a deep biological perspective for capturing the relationships between drugs and diseases. Simultaneously, the possibility that a drug is associated with a disease is also dependant on their projections in the low-dimension feature space. Therefore, DisDrugPred deeply integrates the diverse prior knowledge, the sparse characteristic of associations, and the projections of drugs and diseases. DisDrugPred achieves superior prediction performance than several state-of-the-art methods for drug-disease association prediction. During the validation process, DisDrugPred also can retrieve more actual drug-disease associations in the top part of prediction result which often attracts more attention from the biologists. Moreover, case studies on 5 drugs further confirm DisDrugPred's ability to discover potential candidate disease indications for drugs.
Contact: zhang@hlju.edu.cn

1 Introduction

Developing a new drug is a lengthy, complex, and expensive process which generally takes 10-15 years and 0.8-1.5 billon dollars (Dickson *et al*., 2004; Tamimi and Ellis, 2009; Pushpakom *et al*., 2018). Drug repositioning is to identify novel therapeutic effects for the drugs that have been approved by the regulatory agencies (Padhy and Gupta, 2011; Shahreza et al., 2017). The approved drugs have known and well-characterized bioavailability, safety and pharmacology which can significantly accelerate drug development. Compared to developing a drug de novo, drug repositioning may reduce the drug development period to 6.5 years and the cost for repositioning a drug is $300 million (Nosengo *et al*., 2016; Pritchard *et al*., 2017).

Computational prediction of novel therapeutic indications for approved drugs may screen candidate drug-disease associations for further experimental validation (Hurle *et al*., 2013; Li *et al*., 2015; Chen *et al*., 2016). The previous works can be roughly grouped into two categories. Since the drugs execute their functions by targeting the related genes (Yamanishi *et al*., 2008; Bleakley *et al*., 2009; Fakhraei *et al*., 2014), the drugs and diseases that are associated with each other are usually related to some common genes. Furthermore, the more common genes they are related to, the more likely that they are associated with each other. Thus, several methods of the first category are proposed to infer the association propensity of a drug and a disease based on their related genes or gene expressions (Sirota *et al*., 2011; Wang *et al*., 2014a). Similarly, the association propensity can also be estimated according to the protein complexes shared by the drug and disease (Yu *et al*., 2015) and their common perturbed genes (Peyvandipour *et al*., 2018). However, these methods fail to be applied to the drugs and diseases without common interacted genes or proteins.

The second category takes advantage of the various data that includes the similarities of drugs, diseases, and targets, as well as the interactions and associations between drugs, targets and diseases, for drug repositioning. The similarities of drugs and diseases are integrated by a kernel function to predict drug-disease associations (Wang *et al*., 2013). Several methods infer the candidate drug indications by information flow or random walks on a heterogeneous network composed of drugs, targets and diseases (Wang *et al*., 2014b; Luo *et al*., 2016; Liu *et al*., 2016; Luo *et al*., 2018). A couple of methods exploit the data of drugs and diseases and predict novel drug uses by a logistical regression model, a statistical model, sparse subspace learning, or similarity constrained matrix factorization (Gottlieb *et al*., 2011; Iwata *et al*.,

2015; Liang *et al*., 2017; Zhang *et al*., 2018a). In addition, recent researches indicated that besides proteins, the microRNAs and lncRNAs may also be used as the targets of drugs (Chen *et al*., 2018a; Qu *et al*., 2018; Chen *et al*., 2015). Responses are also one kind of important attributes of drugs (Liu *et al*., 2018; Zhang *et al*., 2018b). Therefore, microRNAs, lncRNAs, and responses related to drugs, are potentially additional information for drug-disease association prediction. However, there are no enough experimentally verified microRNAs, lncRNAs, and responses so far for accurately predicting drug-related diseases. Overall, integrating the heterogeneous data from multiple sources is essential for exploring the drug-disease associations. However, these previous methods ignore the prior knowledge of drugs and diseases and the biological characteristic of drug-disease associations.

In this article, we present DisDrugPred, a novel method for predicting the candidate drug-disease associations. We first calculate a new type of drug similarity based on the diseases that are associated with the drugs. DisDrugPred then completely exploits the similarity and association, as well as interaction data about drugs, diseases, and target proteins of drugs. DisDrugPred deeply integrates not only the diverse prior knowledge of drugs and diseases but also the projections of drugs and diseases in low-dimensional feature space. Integrating the prior knowledge about the case in which two drugs (diseases) will be more similar can capture the relationships between the drug-disease associations and the similarities of drugs (diseases) from the biological perspectives. Projecting the drugs and diseases into a common and low-dimensional feature space contributes to the measurement of the distances between them. These distances between drugs and diseases are also closely related to their association possibilities. Hence a unified model is constructed and an iterative optimization algorithm is developed for solving the model to obtain the association possibilities of drugs and diseases. The experimental results based on cross validation show that DisDrugPred significantly outperforms than several state-of-the-art prediction methods. In particular, when focusing on the top part of prediction result, DisDrugPred successfully retrieves more actual drug-disease associations. Case studies on 5 drugs further confirms that DisDrugPred is able to discover the potential disease indications of drugs.

2 Methods and Materials

Our goal is to predict the potential therapeutic indications, i.e., the candidate diseases, for a given drug of interest. We first calculate a new type of similarity between drugs to exploit the information of their associated diseases. A novel prediction model based on non-negative matrix factorization (Lee *et al*., 2001) is proposed by integrating the multi-source data about drugs and diseases. The drug-disease association scores are able to be obtained by solving the model with an iterative algorithm. A greater association score of drug r_i and disease d_j means that r_i is more likely to be associated with d_j.

2.1 Datasets for drug indication prediction

The associations between drugs and diseases, the chemical substructure profiles of drugs, the domain profiles of target proteins of drugs, the target annotation profiles of drugs, and the disease semantic similarities are obtained from the previous work on prediction of drug-disease associations (Liang *et al*., 2017). The 3051 drug-disease association data is originally extracted from the Unified Medical Language System (Bodenreider, 2004), and it contains the treatment relationships between 763 drugs and 681 diseases. The chemical substructure profile of drugs can be constructed by using the chemical fingerprints which are extracted from the database, PubChem (Kim *et al*., 2015). The domains of drug-related proteins and the gene ontology annotations of these proteins are respectively obtained from the databases, InterPro (Mitchell *et al*., 2015) and UniProt (Consortium, 2018). We extract the side effect indications of drugs from the Database SIDER (Kuhn *et al*., 2015), and 571 ones among 763 drugs have their side effect indications. The disease similarities that incorporated the disease ontology and the disease-related genes are extracted from the DincRNA database (Cheng *et al*., 2018), and 386 ones among 681 diseases have this kind of disease similarity. The disease names come from the U.S. National Library of Medicine (MeSH, http://www.ncbi.nlm.nih.gov/mesh).

2.2 Calculation and representation of multi-source data

Five types of drug similarities. As two drugs, such as r_a and r_b, with more common chemical substructures are usually more similar, the previous work LRSSL (Liang *et al*., 2017) calculates the cosine similarity on their chemical substructure vectors as the first type of similarity of r_a and r_b (Figure 1(a)). Moreover, the drugs with more common domains of target proteins or interacted with more target proteins with similar functions often have relatively higher similarity (Ding *et al*., 2013; Perlman *et al*., 2011). Hence LRSSL also calculated the second and third types of drug similarities based on cosine similarity measure.

Since the drugs associated with similar diseases are also more similar, we additionally calculated the fourth type of similarity. Calculating the drug similarities based on the diseases associated with the drugs is part of the novelty of our work. Inspired by the miRNA similarity measure (Wang *et al*., 2010), we firstly obtain the disease sets related to drugs r_a and r_b, and denote them as $DT_a = \{d_1, d_4\}$ and $DT_b = \{d_2, d_4, d_5\}$ (Figure 1(a)). The similarity between DT_a and DT_b is then calculated as the similarity of r_a and r_b which is denoted as $RS(r_a, r_b)$. $RS(r_a, r_b)$ is defined as,

$$RS(r_a, r_b) = \frac{\sum_{i=1}^{m} \max_{1 \le j \le n} (DS(d_{ai}, d_{bj})) + \sum_{j=1}^{n} \max_{1 \le i \le m} (DS(d_{bj}, d_{ai}))}{m + n}, \quad (1)$$

where $DS(d_{ai}, d_{bj})$ is the semantic similarity of diseases d_{ai} and d_{bj} which belong to DT_a and DT_b respectively. m and n are the numbers of diseases included by DT_a and DT_b. Matrix $R_4 \in \Re^{N_r \times N_r}$ is used to represent this type of drug similarity. $\Re$ represents a set of real numbers and N_r is the number of drugs. $\Re^{N_r \times N_r}$ is a real coordinate space with $N_r \times N_r$ dimensions.

The disease semantic similarities are calculated by using the Wang's method (Wang *et al*., 2010). The method constructs a directed acyclic graph (DAG) for a disease that contains all of the semantic terms related to the disease, such as the DAG of *Breast Neoplasms* in Figure 1 (b). The similarity of two diseases is calculated based on their DAGs. The more their DAGs have common terms, the more similar two diseases are. The values of disease semantic similarity range between 0 and 1. Note that as only the disease semantic similarities cover all the diseases related with our interested drugs, the fourth type of drug similarity is just calculated based on these semantic similarities.

In addition, drugs sharing more similar side effects tend to interact with common target proteins and further have more similar functions (Gottlieb *et al*., 2012; Sridhar *et al*., 2016; Zitnik *et al*., 2018). Thus the fifth type of drug similarity is measured by the cosine similarity based on the side effects related to the drugs. All of the five types of drug similarities are represented by the matrices R_1, R_2, R_3, R_4 and $R_5 \in \Re^{N_r \times N_r}$ where $(R_t)_{ij} (1 \le t \le m_r)$ is the tth type of similarity of drugs r_i and r_j, and m_r is the number of the drug similarity types.

Representation of disease similarities. First, the semantic similarity of two diseases quantifies how similar the disease terms related to them are. Two diseases are generally more similar when they have more common terms. Wang et al. have calculated the disease semantic similarities (Wang *et al*., 2010), and these similarities are widely used by the previous work on drug-disease association prediction (Liang *et al*., 2017; Zhang *et al*., 2018a). Our method also exploit the disease similarities whose values range between 0 and 1. Second, the Disease Ontology (DO) has been developed as a formal ontology for human disease, and it aims to provide an etiological based disease classification (Schriml *et al*., 2018). Simultaneously, the functional similarity of two diseases may be measured by their related genes. Cheng et al. integrated the disease ontology and disease-related genes to obtain another type of disease similarity (Cheng *et al*., 2014). In our study, two types of disease similarities are denoted as the matrices D_1 and $D_2 \in \Re^{N_d \times N_d}$ where $(D_s)_{ij} (1 \le s \le m_d)$ is the sth type of similarity of diseases d_i and d_j, N_d is the numbers of diseases and m_d is the number of the disease similarity types (Figure 1(b)).

Representation of the drug-disease associations. As shown in Figure 1(c), the drug-disease bipartite graph is formed by the known associations between drugs and diseases. According to the graph, matrix $A = (A_{ij}) \in \Re^{N_d \times N_r}$ is constructed to represent the association case between N_d diseases and N_r drugs, where A_{ij} is 1 if disease d_i was observed to be associated with drug r_j or 0 otherwise.

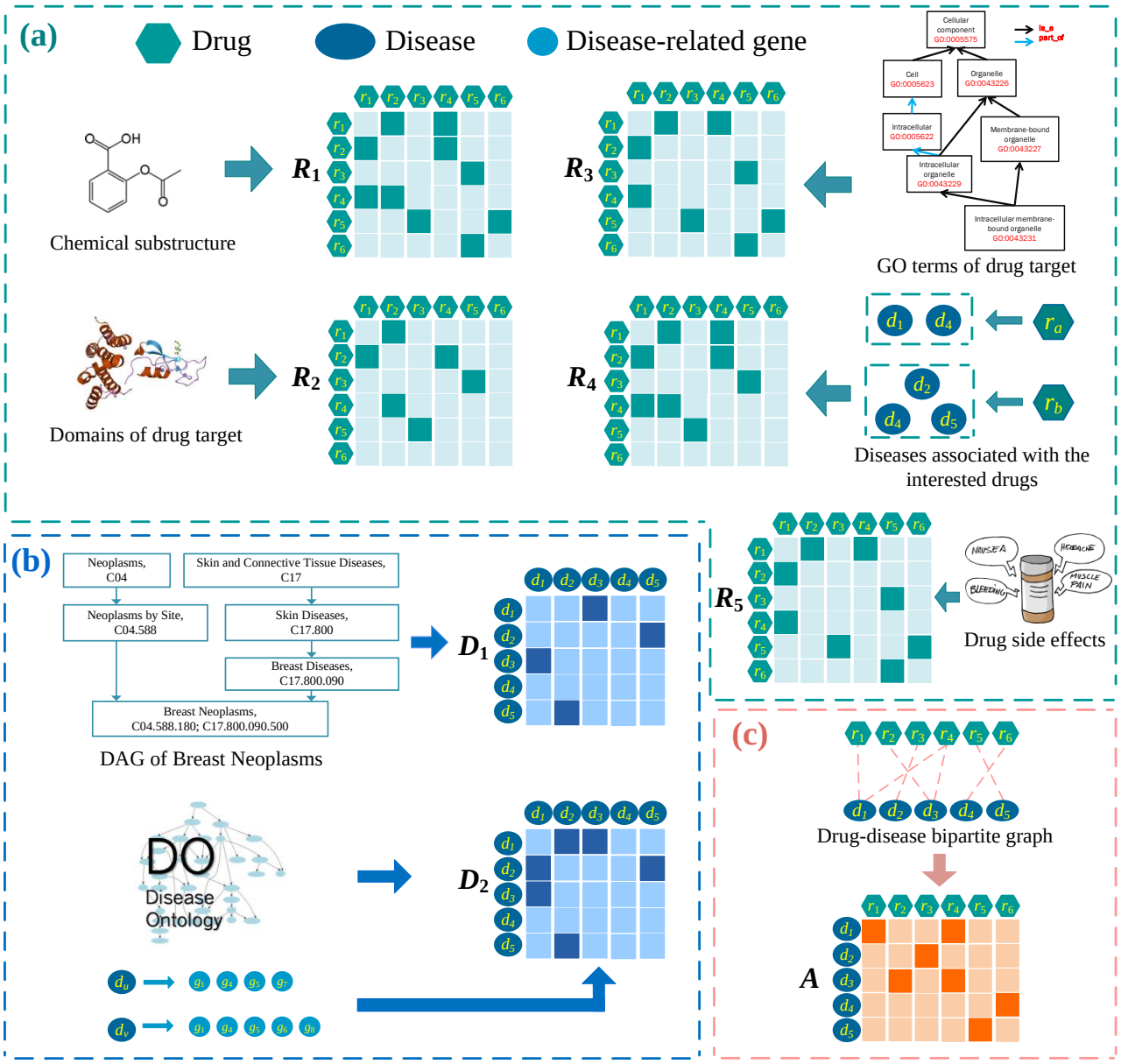

Fig. 1: Representation of the data about drugs and diseases from the multiple sources. **(a)** calculate and represent five types of drug similarities according to their features from different views. **(b)** calculate two types of similarities between diseases and denote them with matrices. **(c)** construct the drug-disease association matrix A according to the known associations between drugs and diseases.

2.3 Drug-disease association prediction model

Modeling the drug-disease association relationships. Let $P = (P_{ij}) \in \Re^{N_d \times N_r}$ be the association score matrix, where $P_{ij} \geq 0$ is a score measuring how probably disease d_i is associated with drug r_j. The observed drug-disease associations and the unobserved ones are represented by 1s and 0s, respectively. Since the non-zero elements of A are very sparse, the optimization item based on matrix factorization is often established based on the observed associations (Natarajan *et al.*, 2014; Chen *et al.*, 2018b; Zhao *et al.*, 2018). Suppose Ω be the set of observed drug-disease associations, and $Y = (Y_{ij}) \in \Re^{N_d \times N_r}$ be the indicator matrix where Y_{ij} is 1 if $(d_i, r_j) \in \Omega$ or 0 otherwise. Obviously, Y is equal to A, and only the known drug-disease associations would contribute to the error term being minimized by using Y. The estimated association cases between drugs and diseases in P should be as consistent with the observed cases in A as possible. As a result, we construct an optimization term as follows,

$$\min_{P \geq 0} \|A \odot (P - A)\|_F^2, \tag{2}$$

where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix and $\odot$ is the Hadamard product.

Modeling the projections of diseases and drugs. When the drugs and diseases are projected the common k-dimension feature space, it is more possible that there is a potential association between a drug and a disease with similar low-dimensional features. Let $W_s \in \Re^{N_d \times k} (1 \leq s \leq m_d)$ be the projection matrix of the sth type of disease similarity, and $H_t \in \Re^{N_r \times k} (1 \leq t \leq m_r)$ is the projection matrix corresponding to the tth type of drug similarity. As the close degrees between the k-dimension features of drugs and diseases $D_s W_s (R_t H_t)^T (1 \leq s \leq m_d, 1 \leq t \leq m_r)$ offer a guidance for estimation of the drug-disease association scores, we add a new optimization term to the objective function,

$$\min_{P, W_s, H_t \geq 0} \|A \odot (P - A)\|_F^2 + \alpha_1 \sum_{s=1}^{m_d} \sum_{t=1}^{m_r} \|P - D_s W_s (R_t H_t)^T\|_F^2, \tag{3}$$

where α_1 is a parameter for making a tradeoff between the first optimal term and the second one.

Modeling the prior knowledge of disease similarities. It is well known that the more two diseases are associated similar drugs, the more similar they are. $(D_s)_{ij}$ in the disease similarity matrix D_s is the sth type of actual similarity between diseases d_i and d_j. The ith row of matrix P, denoted as P_i, contains the possibilities that disease d_i is associated with the various drugs. In the transpose of P, i.e. P^T, its jth column $(P^T)_j$ contains the association possibilities between

disease d_j and all the drugs. The expected similarities between the diseases, PP^T, should be as close to the sth type of actual disease similarity in D_s as possible, which gives rise to the following function,

$$\min_{P,W_s,H_t\geq 0}\|A\odot(P-A)\|_F^2+\alpha_1\sum_{s=1}^{m_d}\sum_{t=1}^{m_r}\|P-D_sW_s(R_tH_t)^T\|_F^2 + \alpha_2\sum_{s=1}^{m_d}\|D_s-PP^T\|_F^2, \tag{4}$$

where α_2 is used to control the contribution of the third term.

Modeling the prior knowledge of drug similarities. As mentioned before, the four types of drug similarities, i.e. R_1, R_2, R_3, R_4 and R_5, reflect the similarities between the drugs from the different perspectives. The prior knowledge about the drug similarities is when two drugs are associated with more similar diseases, they usually have a higher similarity. In the drug-disease association score matrix P, the ith row of the transpose of P, $(P^T)_i$ records the association possibilities between drug r_i and the various diseases. The jth column of P, P_j, is the possibility column vector of drug r_j to be associated with all the diseases. Then $(P^T)_iP_j$ is the expected similarity of r_i and r_j, while $(R_1)_{ij}$ is the actual first type of similarity of these two drugs. The deviation between the expected drug similarities and the actual ones can be introduced as the fourth term to the optimization function,

$$\min_{P,W_s,H_t\geq 0}\|Y\odot(P-A)\|_F^2+\alpha_1\sum_{s=1}^{m_d}\sum_{t=1}^{m_r}\|P-D_sW_s(R_tH_t)^T\|_F^2 + \alpha_2\sum_{s=1}^{m_d}\|D_s-PP^T\|_F^2+\alpha_3\sum_{t=1}^{m_r}\|R_t-P^TP\|_F^2, \tag{5}$$

where α_3 is used to adjust the contribution of the term about the drug similarities.

Modeling smoothness prior. The smoothness prior specifies that a drug and one of its k most similar neighbors are more often associated with two groups of similar diseases. Hence we respectively construct m_r graphs composed of drug nodes according to m_r types of drug similarities and construct a regularization term based these graphs, i.e., graph regularization term (Cai *et al.*, 2011). For the tth ($1\leq t\leq m_r$) type of drug similarity, the adjacency matrix of its corresponding graph is $M_t\in\Re^{N_r\times N_r}$. Its element in the ith row and jth column, $(M_t)_{ij}$, is defined as,

$$(M_t)_{ij}=\begin{cases}1, \text{if the drug } r_j \text{ is one of the } k \text{ most similar neighbors of the drug } r_i\\ 0, \text{otherwise}\end{cases}. \tag{6}$$

Since a drug and its k similar neighbors are more likely to associate with more similar diseases, the following regularization term for smoothness can be constructed with the matrices M_t,

$$\begin{aligned}&\frac{1}{2}\sum_{t=1}^{m_r}\sum_{i,j=1}^{N_r}\|\mathbf{p}_i-\mathbf{p}_j\|^2(M_t)_{ij}\\ &=\sum_{t=1}^{m_r}(Tr(PU_tP^T)-Tr(PM_tP^T))\\ &=\sum_{t=1}^{m_r}Tr(PL_tP^T),\end{aligned} \tag{7}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ denote the ith and jth column vectors of matrix P respectively, and they reflect the cases that the drugs r_i and r_j are potentially associated with all the diseases. $U_t\in\Re^{N_r\times N_r}$ is a diagonal matrix whose elements $(U_t)_{ii}=\sum_{j=1}^{N_r}(M_t)_{ij}$, and $L_t=U_t-M_t$ is the Laplacian matrix of the tth graph. The smoothness term should also be minimized and we have the following function,

$$\begin{aligned}\min_{P,W_s,H_t\geq 0}&\|A\odot(P-A)\|_F^2+\alpha_1\sum_{s=1}^{m_d}\sum_{t=1}^{m_r}\|P-D_sW_s(R_tH_t)^T\|_F^2\\ &+\alpha_2\sum_{s=1}^{m_d}\|D_s-PP^T\|_F^2+\alpha_3\sum_{t=1}^{m_r}\|R_t-P^TP\|_F^2\\ &+\alpha_4\sum_{t=1}^{m_r}Tr(P^TL_tP),\end{aligned} \tag{8}$$

where α_4 regulates the contribution of the smoothness term.

Modeling the biological characteristic of associations. Only a limited number of diseases are associated with a specific drug, so each column of matrix P that records the association scores between the drug and all the diseases should be sparse. The l_1-regularization is imposed on the columns of P for learning the sparse associations. After adding the sparse penalty term, we get the following function

$$\begin{aligned}\min_{P,W_s,H_t\geq 0}&\|A\odot(P-A)\|_F^2+\alpha_1\sum_{s=1}^{m_d}\sum_{t=1}^{m_r}\|P-D_sW_s(R_tH_t)^T\|_F^2\\ &+\alpha_2\sum_{s=1}^{m_d}\|D_s-PP^T\|_F^2+\alpha_3\sum_{t=1}^{m_r}\|R_t-P^TP\|_F^2\\ &+\alpha_4\sum_{t=1}^{m_r}Tr(PL_tP^T)+\alpha_5\sum_{k=1}^{N_r}\|P_k\|_1^2,\end{aligned} \tag{9}$$

where P_k is the kth column of P and N_r is the number of drugs, and α_5 is a parameter that controls the contribution of penalty term.

Introducing regularization term for preventing overfitting. In order to prevent the overfitting in our prediction model, we add the l_2-regularization on the projection matrices, W_s and H_t ($1\leq s\leq m_d, 1\leq t\leq m_r$). We then get the final objective function $L(P,W_s,H_t)$

$$\begin{aligned}\min_{P,W_s,H_t\geq 0}&\|A\odot(P-A)\|_F^2+\alpha_1\sum_{s=1}^{m_d}\sum_{t=1}^{m_r}\|P-D_sW_s(R_tH_t)^T\|_F^2\\ &+\alpha_2\sum_{s=1}^{m_d}\|D_s-PP^T\|_F^2+\alpha_3\sum_{t=1}^{m_r}\|R_t-P^TP\|_F^2\\ &+\alpha_4\sum_{t=1}^{m_r}Tr(PL_tP^T)+\alpha_5\sum_{k=1}^{N_r}\|P_k\|_1^2\\ &+\alpha_6(\sum_{s=1}^{m_d}\|W_s\|_F^2+\|\sum_{t=1}^{m_r}H_t\|_F^2),\end{aligned} \tag{10}$$

where α_6 is a regulation parameter.

2.4 Optimization

As the objective function (10) with the variables P, W_s, and H_t is not convex, it is impractical to get its global minimum. We present an algorithm to find its local minimum by separating the optimization problem into several subproblems and then optimizing them iteratively.

***P*-subproblem:** When updating $\boldsymbol{P}$ with $\boldsymbol{W_s}$ and $\boldsymbol{H_t}$ ($\boldsymbol{1\leq s\leq m_d, 1\leq t\leq m_r}$) fixed, the subproblem for solving $\boldsymbol{P}$ is as follows,

$$\begin{aligned}\min_{\boldsymbol{P\geq 0}}\boldsymbol{L(P)}=&\|\boldsymbol{A\odot(P-A)}\|_{\boldsymbol{F}}^{\boldsymbol{2}}+\boldsymbol{\alpha_1}\sum_{\boldsymbol{s=1}}^{\boldsymbol{m_d}}\sum_{\boldsymbol{t=1}}^{\boldsymbol{m_r}}\|\boldsymbol{P-D_sW_s(R_tH_t)^T}\|_{\boldsymbol{F}}^{\boldsymbol{2}}\\ &+\boldsymbol{\alpha_2}\sum_{\boldsymbol{s=1}}^{\boldsymbol{m_d}}\|\boldsymbol{D_s-PP^T}\|_{\boldsymbol{F}}^{\boldsymbol{2}}+\boldsymbol{\alpha_3}\sum_{\boldsymbol{t=1}}^{\boldsymbol{m_r}}\|\boldsymbol{R_t-P^TP}\|_{\boldsymbol{F}}^{\boldsymbol{2}}\\ &+\boldsymbol{\alpha_4}\sum_{\boldsymbol{t=1}}^{\boldsymbol{m_r}}\boldsymbol{Tr(PL_tP^T)}+\boldsymbol{\alpha_5}\sum_{\boldsymbol{k=1}}^{\boldsymbol{N_r}}\|\boldsymbol{P_k}\|_{\boldsymbol{1}}^{\boldsymbol{2}}.\end{aligned} \tag{11}$$

After the Frobenius norms of matrices have been transformed to their trace norms (Srebro *et al*., 2005), $\boldsymbol{L(P)}$ is able to be rewritten as:

$$\begin{aligned}\boldsymbol{L(P)} =& Tr(\boldsymbol{A} \odot (\boldsymbol{PP^T} - \boldsymbol{PA^T} - \boldsymbol{AP^T} + \boldsymbol{AA^T})) \\ &+ \boldsymbol{\alpha_1} \sum_{1\le s\le m_d} \sum_{1\le t\le m_r} Tr(\boldsymbol{PP^T} - \boldsymbol{PR_tH_tW_s^TD_s^T} - \boldsymbol{D_sW_sH_t^TR_t^TP^T} \\ &\qquad + \boldsymbol{D_sW_sH_t^TR_t^TR_tH_tW_s^TD_s^T}) \\ &+ \boldsymbol{\alpha_2} \sum_{1\le s\le m_d} Tr(\boldsymbol{D_sD_s^T} - \boldsymbol{D_sPP^T} - \boldsymbol{PP^TD_s^T} + \boldsymbol{PP^TPP^T}) \\ &+ \boldsymbol{\alpha_3} \sum_{1\le t\le m_r} Tr(\boldsymbol{R_tR_t^T} - \boldsymbol{R_tP^TP} - \boldsymbol{P^TPR_t^T} + \boldsymbol{P^TPP^TP}) \\ &+ \boldsymbol{\alpha_4} \sum_{1\le t\le m_r} Tr(\boldsymbol{PL_tP^T}) \\ &+ \boldsymbol{\alpha_5}\|\boldsymbol{e_{1\times N_d}P}\|_F^2,\end{aligned} \tag{12}$$

where $\boldsymbol{e_{1\times N_d}}$ is the $\boldsymbol{1 \times N_d}$ vector where all elements are 1. By setting the derivative of $\boldsymbol{L(P)}$ with respect to $\boldsymbol{P}$ to 0, we have

$$\begin{aligned}&\boldsymbol{A} \odot (\boldsymbol{2P - 2A}) + \boldsymbol{\alpha_1} \sum_{1\le s\le m_d} \sum_{1\le t\le m_r} (\boldsymbol{2P - 2D_sW_sH_t^TR_t^T}) \\ &+ \boldsymbol{\alpha_2} \sum_{1\le s\le m_d} (\boldsymbol{-4D_sP + 4PP^TP}) + \boldsymbol{\alpha_3} \sum_{1\le t\le m_r} (\boldsymbol{-4PR_t + 4PP^TP}) \\ &+ \boldsymbol{\alpha_4}(\sum_{1\le t\le m_r} \boldsymbol{2P(U_t - M_t)}) + \boldsymbol{\alpha_5(2e_{1\times N_d}^Te_{1\times N_d}P)} = \boldsymbol{0}.\end{aligned} \tag{13}$$

By multiplying both sides of equation (13) by $\boldsymbol{P_{ij}}$, we get the following equation

$$\begin{aligned}&(\boldsymbol{A} \odot (\boldsymbol{2P - 2A}) + \boldsymbol{\alpha_1} \sum_{1\le s\le m_d} \sum_{1\le t\le m_r} (\boldsymbol{2P - 2D_sW_sH_t^TR_t^T}) \\ &+ \boldsymbol{\alpha_2} \sum_{1\le s\le m_d} (\boldsymbol{-4D_sP + 4PP^TP}) + \boldsymbol{\alpha_3} \sum_{1\le t\le m_r} (\boldsymbol{-4PR_t + 4PP^TP}) \\ &+ \boldsymbol{\alpha_4}(\sum_{1\le t\le m_r} \boldsymbol{2P(U_t - M_t)}) + \boldsymbol{\alpha_5(2e_{1\times N_d}^Te_{1\times N_d}P)})_{ij}\boldsymbol{P_{ij}} = \boldsymbol{0}.\end{aligned} \tag{14}$$

According to the coordinate gradient descent algorithm in (Tan and Fevotte, 2009), $\boldsymbol{P_{ij}}$ can be updated by multiplying it with the ratio of the negative terms to the positive terms in the left side of equation (14),

$$P_{ij}^{new} \leftarrow P_{ij} \cdot \frac{(2\boldsymbol{A} \odot \boldsymbol{A} + 2\alpha_1 \sum_{1\le s\le m_d} \sum_{1\le t\le m_r} \boldsymbol{D_sW_sH_t^TR_t^T} + 4\alpha_2 \sum_{1\le s\le m_d} \boldsymbol{D_sP} + 4\alpha_3 \sum_{1\le t\le m_r} \boldsymbol{PR_t} + 2\alpha_4 \sum_{1\le t\le m_r} \boldsymbol{PM_t})_{ij}}{(2\boldsymbol{A} \odot \boldsymbol{P} + 2m_dm_r\alpha_1\boldsymbol{P} + 4m_d\alpha_2\boldsymbol{PP^TP} + 4m_r\alpha_3\boldsymbol{PP^TP} + 2\alpha_4 \sum_{1\le t\le m_r} \boldsymbol{PU_t} + 2\alpha_5\boldsymbol{e_{1\times N_d}^Te_{1\times N_d}P})_{ij}}. \tag{15}$$

$\boldsymbol{W_s}$**-subproblem**: When $\boldsymbol{P}$ and $\boldsymbol{H_t}$ are fixed, the subproblem for solving $\boldsymbol{W_s}$ $(\boldsymbol{1 \le s \le m_d})$ is:

$$\min_{\boldsymbol{W_s}\ge 0} \boldsymbol{L(W_s)} = \boldsymbol{\alpha_1} \sum_{1\le t\le m_r} \|\boldsymbol{P - D_sW_s(R_tH_t)^T}\|_F^2 + \boldsymbol{\alpha_6}\|\boldsymbol{W_s}\|_F^2. \tag{16}$$

We transform the Frobenius norms of matrices in $\boldsymbol{L(W_s)}$ to their trace norms and rewritten $\boldsymbol{L(W_s)}$ as:

$$\begin{aligned}\boldsymbol{L(W_s)} =& \boldsymbol{\alpha_1} \sum_{1\le t\le m_r} Tr(\boldsymbol{PP^T} - \boldsymbol{PR_tH_tW_s^TD_s^T} - \boldsymbol{D_sW_sH_t^TR_t^TP^T} \\ &+ \boldsymbol{D_sW_sH_t^TR_t^TR_tH_tW_s^TD_s^T}) \\ &+ \boldsymbol{\alpha_6}Tr(\boldsymbol{W_sW_s^T}).\end{aligned} \tag{17}$$

By setting the derivative of $\boldsymbol{L(W_s)}$ with respect to $\boldsymbol{W_s}$ to 0, we get

$$\boldsymbol{\alpha_1} \sum_{1\le t\le m_r} (\boldsymbol{-2D_s^TPR_tH_t + 2D_s^TD_sW_sH_t^TR_t^TR_tH_t}) + \boldsymbol{2\alpha_6W_s} = \boldsymbol{0}. \tag{18}$$

After both sides of (18) are multiplied by $(\boldsymbol{W_s})_{ij}$, we obtain the equation as follows,

$$\begin{aligned}&(\boldsymbol{\alpha_1} \sum_{1\le t\le m_r} (\boldsymbol{-2D_s^TPR_tH_t + 2D_s^TD_sW_sH_t^TR_t^TR_tH_t}) + \boldsymbol{2\alpha_6W_s})_{ij}(\boldsymbol{W_s})_{ij} \\ &= \boldsymbol{0}.\end{aligned} \tag{19}$$

The equation leads to the following $\boldsymbol{W_s}$'s updating rule by applying the coordinate gradient descent algorithm (Tan and Fevotte, 2009),

$$(\boldsymbol{W_s})_{ij}^{new} \leftarrow (\boldsymbol{W_s})_{ij} \cdot \frac{(\boldsymbol{2\alpha_1} \sum_{1\le t\le m_r} \boldsymbol{D_s^TPR_tH_t})_{ij}}{(\boldsymbol{2\alpha_1} \sum_{1\le t\le m_r} \boldsymbol{D_s^TD_sW_sH_t^TR_t^TR_tH_t + 2\alpha_6W_s})_{ij}}. \tag{20}$$

$\boldsymbol{H_t}$**-subproblem:** When updating $\boldsymbol{H_t}$ with $\boldsymbol{P}$ and $\boldsymbol{W_s}$ fixed, we may solve the subproblem of $\boldsymbol{H_t}$ $(\boldsymbol{1 \le t \le m_r})$,

$$\min_{\boldsymbol{H_t}\ge 0} \boldsymbol{L(H_t)} = \boldsymbol{\alpha_1} \sum_{1\le s\le m_d} \|\boldsymbol{P - D_sW_s(R_tH_t)^T}\|_F^2 + \boldsymbol{\alpha_6}\|\boldsymbol{H_t}\|_F^2. \tag{21}$$

Similar to the process of solving the subproblems of $\boldsymbol{P}$ and $\boldsymbol{W_s}$, $\boldsymbol{L(H_t)}$ is transformed firstly according to the characteristic of matrix traces. It is then taken a derivative with respect to $\boldsymbol{H_t}$. Finally, the gradient descent algorithm (Tan and Fevotte, 2009) is applied to get $\boldsymbol{H_t}$'s updating rule,

$$(\boldsymbol{H_t})_{ij}^{new} \leftarrow (\boldsymbol{H_t})_{ij} \cdot \frac{(\boldsymbol{2\alpha_1} \sum_{1\le s\le m_d} \boldsymbol{R_t^TP^TD_sW_s})_{ij}}{(\boldsymbol{2\alpha_1} \sum_{1\le s\le m_d} \boldsymbol{R_t^TR_tH_tW_s^TD_s^TD_sW_s + 2\alpha_6H_t})_{ij}}. \tag{22}$$

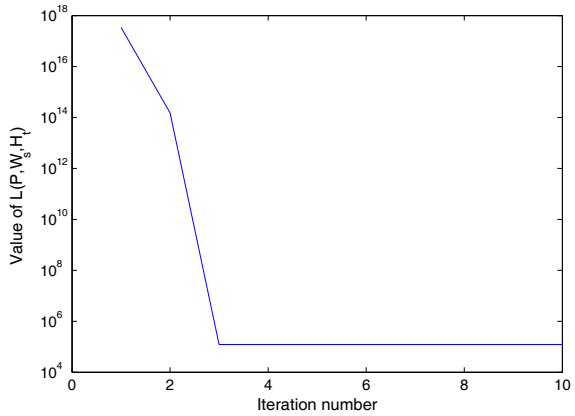

Fig. 2: Convergence of $\boldsymbol{L(P, W_s, H_t)}$

The convergence curve of the objective function $\boldsymbol{L(P, W_s, H_t)}$, confirm that the function can converge to its local minima (Figure 2). $\boldsymbol{L(P, W_s, H_t)}$ is solved by iteratively using the updating rules of $\boldsymbol{P}$, $\boldsymbol{W_s}$, and $\boldsymbol{H_t}$. The iterative process is over when the absolute difference of $\boldsymbol{L(P, W_s, H_t)}$ at two adjacent moments is less than a threshold ($\boldsymbol{\varepsilon = 10^{-6}}$) or the maximum number of iterations, 100, is reached. Finally, $\boldsymbol{P_{ij}}$ is regarded as the estimated association score between disease $\boldsymbol{d_i}$ and drug $\boldsymbol{r_j}$ (Figure 3).

3 Experimental evaluations and discussions

3.1 Evaluation metrics

We perform 5-fold cross-validation for evaluating the performance of a method in predicting drug-disease associations. All known drug-disease associations are randomly divided into 5 equal subsets, four of which are used for training a prediction model, while the remaining subset is used for evaluation. The associations in the remaining subset are added into the testing set and regarded as positive samples. The testing set also contains all the unobserved drug-disease associations which are regarded as negative samples. In the ranking list of associations, the higher the positive samples are ranked, the better the prediction

Algorithm: DisDrugPred algorithm for inferring the potential drug-disease associations
Input: a disease-drug association matrix $\boldsymbol{A} \in \Re^{N_d \times N_r}$, a disease similarity matrix $\boldsymbol{D_1 D_2} \in \Re^{N_d \times N_d}$, and the drug similarity matrices $\boldsymbol{R_1}, \boldsymbol{R_2}, \boldsymbol{R_3}, \boldsymbol{R_4}, \boldsymbol{R_5} \in \Re^{N_r \times N_r}$

Output: The disease-drug association score matrix $\boldsymbol{P}$ where $\boldsymbol{P_{ij}}$ is the association score of disease $\boldsymbol{d_i}$ and drug $\boldsymbol{r_j}$.

Initialize the elements of $\boldsymbol{P}$, $\boldsymbol{W_s}$, and $\boldsymbol{H_t}(\boldsymbol{1 \le s \le m_d, 1 \le t \le m_r})$ with the random values in the range $[\boldsymbol{0, 1}]$
While $\boldsymbol{L(P, W_s, H_t)}$ not converged do
    fix $\boldsymbol{W_s}$ and $\boldsymbol{H_t}$, and update $\boldsymbol{P}$ using the rule:

$$\boldsymbol{P_{ij}^{new} \leftarrow P_{ij} \cdot \frac{(2A \odot A + 2\alpha_1 \sum_{1 \le s \le m_d} \sum_{1 \le t \le m_r} D_s W_s H_t^T R_t^T + 4\alpha_2 \sum_{1 \le s \le m_d} D_s P + 4\alpha_3 \sum_{1 \le t \le m_r} P R_t + 2\alpha_4 \sum_{1 \le t \le m_r} P M_t)_{ij}}{(2A \odot P + 2m_d m_r \alpha_1 P + 4m_d \alpha_2 P P^T P + 4m_r \alpha_3 P P^T P + 2\alpha_4 \sum_{1 \le t \le m_r} P U_t + 2\alpha_5 e_{1 \times N_d}^T e_{1 \times N_d} P)_{ij}}}$$

    For $\boldsymbol{s = 1}$ to $\boldsymbol{m_d}$ do
        fix $\boldsymbol{P}$ and $\boldsymbol{H_t}$, and update $\boldsymbol{W_s}$ using the rule:

$$\boldsymbol{(W_s)_{ij}^{new} \leftarrow (W_s)_{ij} \cdot \frac{(2\alpha_1 \sum_{1 \le t \le m_r} D_s^T P R_t H_t)_{ij}}{(2\alpha_1 \sum_{1 \le t \le m_r} D_s^T D_s W_s H_t^T R_t^T R_t H_t + 2\alpha_6 W_s)_{ij}}}$$

    End For
    For $\boldsymbol{t = 1}$ to $\boldsymbol{m_r}$ do
        fix $\boldsymbol{P}$ and $\boldsymbol{W_s}$, and update $\boldsymbol{H_t}$ using the rule:

$$\boldsymbol{(H_t)_{ij}^{new} \leftarrow (H_t)_{ij} \cdot \frac{(2\alpha_1 \sum_{1 \le s \le m_d} R_t^T P^T D_s W_s)_{ij}}{(2\alpha_1 \sum_{1 \le s \le m_d} R_t^T R_t H_t W_s^T D_s^T D_s W_s + 2\alpha_6 H_t)_{ij}}}$$

    End For
End While

Fig. 3: Iterative algorithm for estimation of the disease-drug association scores

performance is. Note that association dataset is separated to 5 folds for cross-validation, the fourth type of drug similarity is recomputed by only using the drug-disease associations used for training in each cross validation test.

The Receiver Operating Characteristic (ROC, Hajian-Tilaki, 2013) curve can be drawn with the true positive rates (TPRs) and the false positive rates (FPRs) at different ranking cutoffs. TPR is the proportion of positive samples identified correctly among the total positive samples, while FPR is the ratio of misidentified negative samples accounting for all the negative samples. TPR and FPR are defined as follows,

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{TN + FP}, \tag{23}$$

where TP and TN are the numbers of correctly identified positive and negative samples, and FN and FP are the numbers of positive and negative samples that are misidentified. The area under the ROC curve (AUC) is calculated to quantify the overall prediction performance.

There is serious imbalance between the known drug-disease associations (positive samples) and the unobserved ones (negative samples). In such case, the precision-recall (PR) curve is more informative than the ROC curve (Saito T and Rehmsmeier M, 2015). Precision and recall are defined as,

$$Precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}. \tag{24}$$

Precision is the proportion of the correctly identified positive samples among the retrieved samples, and recall is the same as TPR. We also evaluate the performance of association prediction by using PR curve and the area under PR curve (AUPR). In terms of 5-fold cross-validation, the final performance is obtained by using averaging CV. Averaging CV means that we obtain a separate performance (AUC or AUPR) for each of the 5 folds when used as a test set, and the 5 performances are averaged to give the final performance (Pahikkala *et al*., 2015).

Considering the candidates in the top part of ranking list are usually selected by the biologists to further validate with wet-lab experiments, it is better to make the top part contain more positive samples. We thus calculate the recall rate within top part, which is the proportion of positive samples identified correctly in the top k list among the total positive ones, as another evaluation metric.

3.2 Comparison with other methods

To evaluate the performance of the presented method, *DisDrugPred*, we compare it with several state-of-the-art methods for drug-disease association prediction: TL_HGBI (Wang *et al*., 2014b), MBiRW (Luo *et al*., 2016), LRSSL (Liang *et al*., 2017) and SCMFDD (Zhang *et al*., 2018a). We describe these methods in more detail below:

TL_HGBI (Wang *et al*., 2014b): TL_HGBI constructed the disease-drug-target network and incorporated the drug similarities based on their chemical structures, the target similarities, the disease phenotypic similarities, the drug-target interactions, and the disease-drug associations. It inferred the new disease-drug associations based on information flow in the three-layer network. TL_HGBI's prediction model is listed as follows,

$$W_{dr}^{k+1} = \alpha W_{dr}^{k} \times (W_{rr} \times W_{rt}^{k} \times W_{tt} \times W_{rr}^{kT}) + (1 - \alpha) W_{dr}^{0}, \tag{25}$$

$$W_{rt}^{k+1} = \alpha (W_{dr}^{kT} \times W_{dd} \times W_{dr}^{k} \times W_{rr}) \times W_{rt}^{k} + (1 - \alpha) W_{rt}^{0}, \tag{26}$$

where W_{rr}, W_{tt}, and W_{dd} are the weight matrices on the drug-drug links, the target-target links and the disease-disease links, respectively. W_{dr}^{k} and W_{rt}^{k} are the weights of the disease-drug links and the drug-target links at the kth iteration. When the iterative information propagation is converged, W_{dr} contains the association possibilities between diseases and drugs.

MBiRW (Luo *et al*., 2016): MBiRW constructed a drug-disease network by exploiting the drug similarities based on their chemical substructures, the disease semantic similarities, and the drug-disease associations. The association propensities between drugs and diseases are obtained by random walks on the drug network and the disease network, respectively. The prediction model of MBiRW is defined as follows,

$$Rr = \alpha * MR * RD_{t-1} + (1 - \alpha) * A, \tag{27}$$

$$Rd = \alpha * RD_{t-1} * MD + (1 - \alpha) * A, \tag{28}$$

where MR and MD are the transition matrices corresponding to the drug and disease networks, A is the drug-disease association matrix, and RD_{t-1} contains the drug-disease association scores at time $t - 1$. MBiRW combined

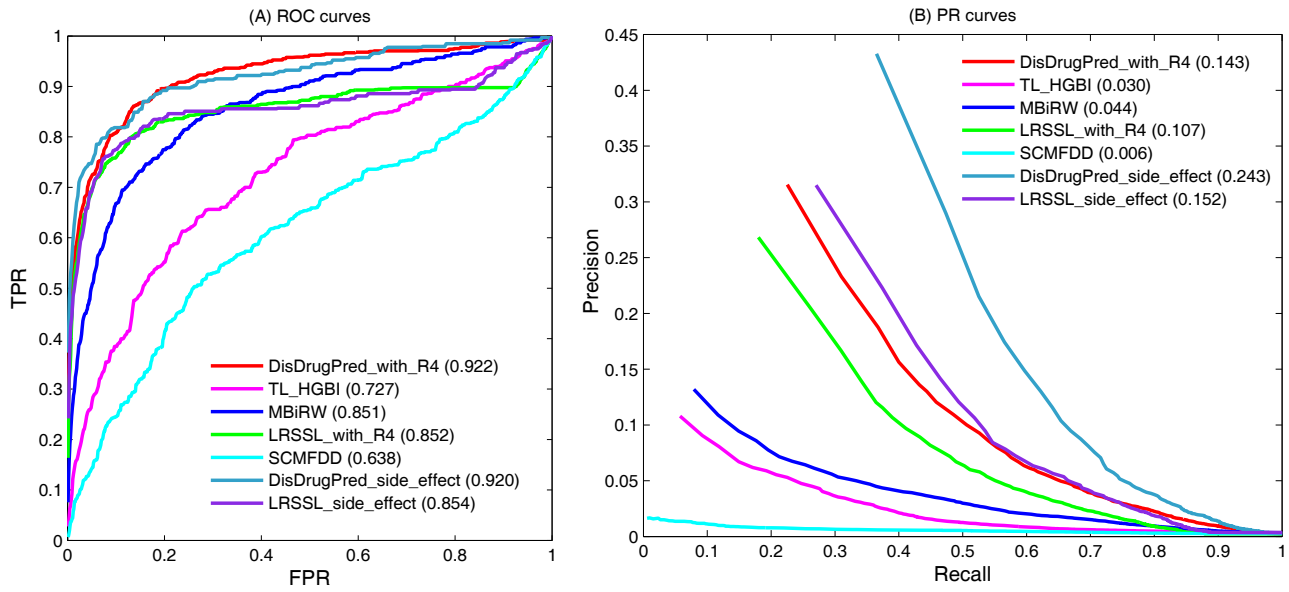

Fig. 4: ROC curves and PR curves of drug-disease association prediction by different methods.

the association propensities in Rr and Rd to get the association scores at time t,

$$RD_t = (rflag * Rr + dflag * Rd)/(rflag + dflag), \tag{29}$$

where $rflag$ and $dflag$ are used to balance the contributions of the drug and disease networks.

LRSSL (Liang *et al.*, 2017): The method uses more data about the drugs in training than TL_HGBI, MBiRW, and SCMFDD. It regards the chemical substructure profiles of the drugs, the target domain profiles, and the target annotation profiles as the drug node attributes, respectively. It also exploits the known drug-disease associations and the local topological structure of graph composed of all the drugs. Its objective function is defined as follows,

$$\min_{F,G_p} \|F-Y\|_F^2 + Tr(F^T LF) + \mu \sum_{p=1}^{m} \|X_p^T G_p - F\|_F^2 + \lambda \sum_{p=1}^{m} \sum_{j=1}^{c} \|G_p(:,j)\|_1^2, \tag{30}$$

where X_p represents the pth type of drug node attribute, Y is the drug-disease association matrix, and L is the Laplacian matrix of drug graph. G_p is used to project each type of drug node attribute to a space whose dimension is equal to the number of diseases, which is also helpful for solving the drug-disease association scores in F.

SCMFDD (Zhang *et al.*, 2018a): SCMFDD focuses on the drug-disease associations, the disease semantic similarities, and the drug similarities based on their substructures. It factorizes the drug-disease association relationships into the low-rank drug and disease feature vectors x_i and y_j as follows,

$$\begin{aligned} \min_{X,Y} L = & \frac{1}{2} \sum_{ij} (a_{ij} - x_i y_j^T)^2 + \frac{\mu}{2} \sum_i \|x_i\|^2 \\ & + \frac{\mu}{2} \sum_j \|y_j\|^2 + \frac{\lambda}{2} \sum_{ij} \|x_i - x_j\|^2 w_{ij}^d \\ & + \frac{\lambda}{2} \sum_{ij} \|y_i - y_j\|^2 w_{ij}^s, \end{aligned} \tag{31}$$

Furthermore, the drug and disease similarities w_{ij}^d and w_{ij}^s are introduced as constraints for learning the drug and disease features, respectively. $x_i y_j^T$ is the estimated association score between the ith disease and the jth drug.

DisDrugPred's hyperparameters, $\alpha_1 \sim \alpha_6$, should be tuned and their values are selected from {0.05, 0.1, 0.2, 0.5, 1, 5, 10, 20, 50}. DisDrugPred yields the best performance when α_1=10, α_2=10, α_3=0.1, α_4=10, α_5=10, and α_6=10, and the optimal set of parameters was obtained by using grid search. To make fair comparisons, the hyperparameters of the other methods are set to their optimal values suggested by their literatures (i.e. $\alpha = 0.4$ and $\beta = 0.3$ for TL_HGBI, $\alpha = 0.3$, $l = 2$ and $r = 2$ for MBiRW, $\mu = 0.01$, $\lambda = 0.01$, $\gamma = 2$, and $k = 10$ for LRSSL, $k = 45\%$, $\mu = 1$ and $\lambda = 4$ for SCMFDD). In addition, the sensitivity coefficients (SC, van Riel *et al.*, 2006) of DisDrugPred's 6 parameters are evaluated by changing one of parameters and fixing the remaining ones. The SC values of $\alpha_1 \sim \alpha_6$ are 5.23e-04, 0.0148, 0.0121, 0.0032, 0.0191, and 7.38e-05, respectively. Hence *DisDrugPred* is not sensitive to the perturbation of α_1, α_4, and α_6, while α_2, α_3, and α_5 have relatively greater impacts on DisDrugPred.

As AUC and AUPR are the better metrics in comparing learning algorithms with probability estimations (Ling *et al.*, 2003, Saito T and Rehmsmeier M, 2015), we use them to evaluate DisDrugPred and the other methods. The ROC curves and their corresponding AUCs obtained by different approaches are given in Figure 4(A). DisDrugPred_with_R4 and LRSSL_with_R4 are the instances of DisDrugPred and LRSSL which exploit four types of drug similarities, i.e., R_1, R_2, R_3, and R_4. DisDrugPred_with_R4 achieves the highest average AUC over all of the 763 drugs (AUC=0.922). It outperforms TL_HGBI by 19.5%, MBiRW by 7.1%, LRSSL with R4 by 7%, and SCMFDD by 28.4%. As shown in Figure 4(B), DisDrugPred_with_R4 also produces the highest average AUPR on 763 drugs (AUPR=0.143). Its' AUPR is 11.3%, 9.9%, 3.6%, and 13.7% better than TL_HGBI, MBiRW, LRSSL_with_R4, and SCMFDD, respectively. LRSSL_with_R4 yields the second best performance. Its' AUC is slightly better than MBiRW while its' AUPR is 6.3% higher than MBiRW. SCMFDD did not perform as well as the other methods as it is very sensitive to the disease and drug similarities. DrugDisPred_with_R4 and LRSSL_with_R4 utilize multiple types of drug similarities, while the other methods focus on only one type of drug similarity. These two methods show the better performances over the other methods, which indicates that integrating more types of drug similarities is essential for improving the prediction accuracy.

In addition, the instances of DisDrugPred and LRSSL are constructed by using R_1, R_2, R_3, R_4, and the fifth type of drug similarity based on their side effects (R_5), and they are referred to as DisDrugPred_side_effect and LRSSL_side_effect. The former still performs better than the latter in terms of both AUC and AUPR, which confirms the superiority of DisDrugPred's algorithm. Since only 571 ones of 763 drugs have their side effects, the subsequent analysis still concentrates on DisDrugPred_with_R4 and LRSSL_with_R4 which cover all of 763 drugs.

For all the prediction results on 763 drugs, we perform a Wilcoxon test to evaluate whether DisDrugPred's performance is significantly better than the other methods. The statistical results (Table 1) indicate that DisDrugPred yields the significantly better performance under the p-value threshold of 0.05 in terms of not only AUCs but AUPRs as well.

Table 1. The statistical result of the paired Wilcoxon test on the AUCs of 763 drugs comparing DisDrugPred and all of four other methods.

p-value between DisDrugPred and another method	TL_HGBI	MBiRW	LRSSL	SCMFDD
p-value of ROC curve	7.2981e-140	4.2955e-55	2.4715e-11	3.1511e-297
p-value of PR curve	2.1728e-41	1.6194e-15	2.5977e-10	8.9884e-229

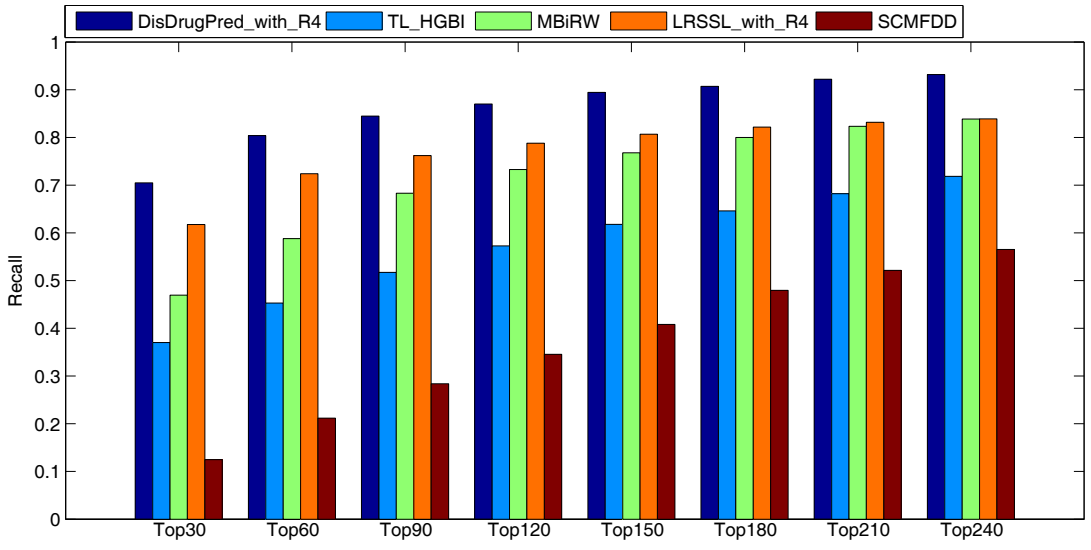

Fig. 5: The recalls across all the tested drugs at different top k cutoffs.

The higher the recall rate on the top k ranked potential drug-disease associations is, the more the real associations are identified successfully. DisDrugPred performs better than the other methods at various k cutoffs (Figure 5), and ranks 70.5% in top 30, 84.5% in top 90, and 89.4% in top 150. Although the AUC of LRSSL is very close to that of MBiRW, all of the recall rates of LRSSL are higher than MBiRW. The former ranks 61.8%, 76.2%, and 80.7% in top 30, 90, and 150, respectively, and the latter ranks 47%, 68.3%, and 76.8%. TL_HGBI is not as good as MBiRW, and it ranks 37%, 51.7%, and 61.8% in top 30, 90, and 150. SCMFDD still did not perform as well as the other methods and the corresponding recall rates are 12.5%, 28.4%, and 40.8%.

3.3 Importance of the drug similarities, the disease similarities, and DisDrugPred's algorithm

To validate the importance of incorporating the fourth type of drug similarity (R_4), two DisDrugPred's instances, DisDrugPred_with_R4 and DisDrugPred_without_R4, are constructed. The former is trained with R_4, while the latter is trained without R_4. At the same time, since LRSSL is able to be extended to exploit more types of drug similarities, we also construct two LRSSL's instances that are trained with $R4$ and without $R4$, LRSSL_with_R4 and LRSSL_without_R4, respectively.

First, the instances of DisDrugPred and LRSSL with R_4 perform better than the ones without R_4, respectively (Supplementary Figure 1). DisDrugPred_with_R4's AUC and AUPR are 1.9% and 1.9% higher than DisDrugPred_without_R4. LRSSL_with_R4's AUC and AUPR also increase by 5.1% and 1.9% compared with LRSSL_without_R4. It shows the importance of incorporating the drug similarities R_4 for improving prediction performance.

Second, the performances of DisDrugPred's instances are better than LRSSL's instances whenever their models are trained by using R_4 or not (Supplementary Figure 1). DisDrugPred_with_R4 achieves 7% and 3.6% higher AUC and AUPR than LRSSL_with_R4. DisDrugPred_without_R4's AUC and AUPR also increase by 10.2% and 3.6% compared with LRSSL_without_R4. It confirms that the algorithm of DisDrugPred also help with the improvement of prediction performance.

In addition, to evaluate the effect of exploiting multiple types of disease similarities, an instance of DisDrugPred is constructed by using the first and second type of disease similarities (D_1 and D_2), and is referred to as DisDrugPred_with_D2. Another DisDrugPred's instance is trained without D_2, and is named DisDrugPred_without_D2. Since the other methods just exploit D_1 and they are not available for using both D_1 and D_2, we only estimate DisDrugPred's performance. As shown in Supplementary Figure 1, DisDrugPred_with_D2's AUPR is a little bit higher than DisDrugPred_without_D2 and it increases by 0.3%, while DisDrugPred_with_D2's AUC is equal to DisDrugPred_without_D2's one. It indicates D_2 has a slight effect on the prediction performances.

3.4 Case studies on 5 drugs

To further demonstrate DisDrugPred's ability to discover the potential drug-disease associations, case studies on 5 drugs, *ciprofloxacin*, *clonidine*, *ampicillin*, *etoposide*, and *cefotaxime*, are conducted. For each of these 5 drugs, the candidate drug-disease associations are prioritized by their association scores, and the top 10 candidates are collected, 50 candidates in total (Table 2).

First, the Comparative Toxicogenomics Database (CTD) provides the key information about the drugs and their effects on human diseases which were manually curated from the published literatures (Davis *et al*., 2016). DrugBank is also a database that captures the clinical trial information of drugs including the drug and the disease for which the trial was conducted (Wishart *et al*., 2017). The repoDB database records the approved and failed drugs and their respective indications (Brown *et al*., 2017). As shown in Table 2, 29 candidates are contained by CTD and they are supported by the direct evidences, 13 candidates are included by DrugBank, and 1 candidate is recorded by repoDB. It indicates these candidate diseases are indeed associated with the corresponding drugs.

Next, ClinicalTrials.gov (https://clinicaltrials.gov/) is an online resource provided by the U.S. National Library of Medicine, and it includes a great many clinical trials about various drugs and the corresponding diseases. PubChem is an open chemistry database supported by the National Institutes of Health (https://pubchem.ncbi.nlm.nih.gov/), and it provides information on chemical substances which include the drugs and their biological activities (Kim *et al*., 2015). There are 4 candidates included by ClinicalTrials.gov and 2 candidates contained by PubChem, indicating these drug-disease associations are supported by the clinical trials. In addition, the 4 candidates labeled by "literature" are supported by the literatures, and the drugs are confirmed to have effects on the corresponding diseases.

Besides the manually curated drug-disease associations, the CTD database also contains the potential associations inferred by the literatures. There is 1 *etoposide*-related candidate disease, *Urinary Tract Infections*, contained by the inferred part of CTD. Hence *etoposide* is more likely to be associated with

Table 2. The top 10 candidates related to the drugs ciprofloxacin, clonidine, ampicillin, etoposide, and cefotaxime, respectively. (1) 'CTD' means a drug-disease association is included by the comparative toxicogenomics database (CTD) and the association is curated manually. (2) 'ClinicalTrials' means that a drug-disease association has been recorded in the online database ClinicalTrials.gov. (3) 'DrugBank' means that the drug-disease association is contained by the DrugBank database that captures the drug trial information. (4) 'repoDB' means that the drug-disease association is included by the repoDB database that records the approved and failed drugs and their indications. (5) 'PubChem' means that the PubChem database has recorded the toxicological information about the drug and disease. (6) 'literature' means that there is a published literature to support the drug-disease association. (7) 'inferred candidate' means that the drug-disease association is the potential one inferred by the literatures and included by CTD. (8) 'unconfirmed' means that there is no evidence to confirm the drug-disease association.

Drug ID	Rank	Disease name	Description	Rank	Disease name	Description
ciprofloxacin	1	Gram-Negative Bacterial Infections	CTD	6	Pneumonia, Bacterial	CTD
	2	Streptococcal Infections	DrugBank	7	Soft Tissue Infections	CTD
	3	Bacterial Infections	CTD	8	Serratia Infections	PubChem
	4	Enterobacteriaceae Infections	CTD	9	Chlamydia Infections	CTD
	5	Salmonella Infections	CTD	10	Helicobacter Infections	CTD
clonidine	1	Pain	CTD	6	Sleep Disorders	ClinicalTrials
	2	Neurologic Manifestations	unconfirmed	7	Nausea	CTD
	3	Depressive Disorder	CTD	8	Edema	CTD
	4	Vomiting	CTD	9	Facial Pain	literature (Yoon *et al*., 2015)
	5	Muscle Cramp	PubChem	10	Muscle Rigidity	unconfirmed
ampicillin	1	Streptococcal Infections	CTD	6	Septicemia	DrugBank, repoDB
	2	Proteus Infections	CTD	7	Gram-Positive Bacterial Infections	CTD
	3	Bacterial Infections	CTD	8	Enterobacteriaceae Infections	DrugBank
	4	Pneumonia, Bacterial	CTD, ClinicalTrials	9	Wound Infection	CTD
	5	Gram-Negative Bacterial Infections	CTD	10	Staphylococcal Skin Infections	DrugBank
etoposide	1	Breast	CTD	6	Lymphoma	CTD
	2	Sarcoma	CTD	7	Urinary Tract Infections	inferred candidate by 1 literature
	3	Leukemiad	DrugBank	8	Ovarian Neoplasms	literature (Bozkaya, 2017)
	4	Hodgkin Disease	CTD	9	Melanoma	DrugBank
	5	Lymphoma, Non-Hodgkin	CTD	10	Head and Neck Neoplasms	DrugBank
cefotaxime	1	Bacterial Infections	CTD, ClinicalTrials	6	Gram-Positive Bacterial Infections	CTD, DrugBank
	2	Enterobacteriaceae Infections	DrugBank	7	Helicobacter Infections	literature (van der Voort *et al*., 2000)
	3	Gram-Negative Bacterial Infections	CTD, DrugBank	8	Eye Infections, Bacterial	literature (Kramann *et al*., 2001)
	4	Pseudomonas Infections	DrugBank	9	Staphylococcal Skin Infections	DrugBank
	5	Respiratory Tract Infections	CTD, ClinicalTrials	10	Septicemia	DrugBank

Urinary Tract Infections. In the total 50 candidates, 2 of them are not confirmed by the observed evidences and they are labeled with "unconfirmed". All the case studies indicate that DisDrugPred is indeed capable of discovering potential candidate drug-disease associations.

3.5 Prediction of novel drug-disease associations

After having evaluated its prediction performance by cross validation and case studies, we applied DisDrugPred to predict the novel drug-disease associations. All of the known drug-disease associations were utilized to train DisDrugPred's prediction model. The potential candidate associations were then obtained by using the model and listed in supplementary table ST1. In addition, the fourth type of drug similarity based on their associated diseases is shown in supplementary table ST2.

4 Conclusions

A method based on non-negative matrix factorization, DisDrugPred, is developed for predicting the potential drug-disease associations. On the basis of calculating the fourth type of drug similarity, DisDrugPred captures the various intra-relationships of drugs and diseases, i.e., the 5 types of drug similarities and 2 types of disease similarities. Meanwhile, it also captures the inter-relationships among drugs and diseases, i.e., the known drug-disease associations. Moreover, the various prior knowledge and the projections of drugs and diseases are deeply integrated to enhance reasoning on the drug-disease associations. In addition, the experimental results confirm DisDrugPred's algorithm also contributes to its' superior performance. An iterative algorithm is developed to obtain the estimated drug-disease association scores, and these scores can be used for ranking the candidate diseases for each of the drugs. In our experiments, we find DisDrugPred consistently outperforms than the other methods tested here in terms of not only AUCs but also AUPRs. In particular, DisDrugPred is more useful for the biologists as its top ranking list contains more real drug-disease associations. Case studies on five drugs demonstrate DisDrugPred's ability in discovering the potential disease indications. DisDrugPred can serve as a prioritization tool to generate the reliable candidates for subsequent identification of actual drug-disease associations with the wet-lab experiments.

Acknowledgements

The work was supported by the Natural Science Foundation of China (61702296, 61302139), the Heilongjiang Postdoctoral Scientific Research Staring Foundation (BHL-Q18104), the Natural Science Foundation of Heilongjiang Province (FLHPY2019329), the Fundamental Research Foundation of Universities in Heilongjiang Province for Technology Innovation (KJCX201805), the Fundamental Research Foundation of Universities in Heilongjiang Province for Youth Innovation Team (RCYJTD201805), the Young Innovative Talent Research Foundation of Harbin Science and Technology Bureau (2016RQQXJ135), and the Foundation of Graduate Innovative Research (YJSCX2018-140HLJU, YJSCX2018-047HLJU).

References

Bleakley, K. and Yamanishi, Y. (2009) Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, **25(18)**, 2397-2403.

Brown, A.S. and Patel, C.J. (2017) A standard database for drug repositioning. *Scientific data*, **4**, 170029.

Bodenreider, O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, **32(suppl_1)**, D267-D270.

Bozkaya, Y., Dogan, M., Umut Erdem, G., Tulunay, G., Uncu, H., Arik, Z., et al. (2017) Effectiveness of low-dose oral etoposide treatment in patients with recurrent and platinum-resistant epithelial ovarian cancer. *Journal of Obstetrics and Gynaecology*, **37(5)**, 649-654.

Cai, D., He, X., Han, J. and Huang, T.S. (2011) Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**, 1548-1560.

Chen, X., Yan, C.C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2015) Drug-target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics*, **17(4)**, 696-712.

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L. and Yan, G. (2016) NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS computational biology*, **12(7)**, e1004975.

Chen, X., Guan, N.N., Sun, Y.Z., Li, J.Q. and Qu, J. (2018) MicroRNA-small molecule association identification: from experimental results to computational models. *Published online October* . **16**, 2018.

Chen, X., Wang, L., Qu, J., Guan, N.N., Li, J.Q. and Berger, B. (2018) Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics*. doi: 10.1093/bioinformatics/bty503

Cheng, L., Li, J., Ju, P., Peng, J. and Wang, Y. (2014) SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PloS one*, **9(6)**, e99415.

Cheng, L., Hu, Y., Sun, J., Zhou, M. and Jiang, Q. (2018) DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics*, **34(11)**, 1953-1956.

Consortium, U. (2018) UniProt: the universal protein knowledgebase. *Nucleic acids research*, **46(5)**, 2699.

Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., King, B.L., McMorran, R., et al. (2016) The comparative toxicogenomics database: update 2017. *Nucleic acids research*, **45(D1)**, D972-D978.

Dickson, M. and Gagnon, J.P. (2004) Key factors in the rising cost of new drug discovery and development. *Nature reviews Drug discovery*, **3(5)**, 417.

Ding, H., Takigawa, I., Mamitsuka, H. and Zhu, S. (2013) Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in bioinformatics* , **15(5)**, 734-747.

Fakhraei, S., Huang, B., Raschid, L. and Getoor, L. (2014) Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **11(5)**, 775-787.

Gottlieb, A., Stein, G.Y., Ruppin, E. and Sharan, R. (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, **7(1)**, 496.

Gottlieb, A., Stein, G.Y., Ruppin, E. and Sharan, R. (2012) INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular systems biology*, **8(1)**, 592.

Hajian-Tilaki, K. (2013) Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, **4(2)**, 627.

Hurle, M., Yang, L., Xie, Q., Rajpal, D., Sanseau, P. and Agarwal, P. (2013) Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, **93(4)**, 335-341.

Iwata, H., Sawada, R., Mizutani, S. and Yamanishi, Y. (2015) Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *Journal of chemical information and modeling*, **55(2)**, 446-459.

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., et al. (2015) PubChem substance and compound databases. *Nucleic acids research*, **44(D1)**, D1202-D1213.

Kramann, C., Pitz, S., Schwenn, O., Haber, M., Hommel, G. and Pfeiffer, N. (2001) Effects of intraocular cefotaxime on the human corneal endothelium1. *Journal of Cataract & Refractive Surgery* , **27(2)**, 250-255.

Kuhn, M., Letunic, I., Jensen, L.J. and Bork, P. (2015) The SIDER database of drugs and side effects. *Nucleic acids research*, **44(D1)**, D1075-D1079.

Lee, D.D. and Seung, H.S. (2001) Algorithms for non-negative matrix factorization. In, *Advances in neural information processing systems*, 556-562.

Li, J., Zheng, S., Chen, B., Butte, A.J., Swamidass, S.J. and Lu, Z. (2015) A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, **17(1)**, 2-12.

Liang, X., Zhang, P., Yan, L., Fu, Y., Peng, F., Qu, L., et al. (2017) LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning. *Bioinformatics*, **33(8)**, 1187-1196.

Ling, C.X., Huang, J. and Zhang, H. (2003)AUC: a better measure than accuracy in comparing learning algorithms. *Conference of the canadian society for computational studies of intelligence.*, Springer, p. 329-341.

Liu, H., Song, Y., Guan, J., Luo, L. and Zhuang, Z. (2016) Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC bioinformatics*, **17(17)**, 539.

Liu, H., Zhao, Y., Zhang, L. and Chen, X. (2016) Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Molecular Therapy-Nucleic Acids*, **13**, 303-311.

Lotfi Shahreza, M., Ghadiri, N., Mousavi, S.R., Varshosaz, J. and Green, J.R. A (2017) review of network-based approaches to drug repositioning. *Briefings in bioinformatics*, **19(5)**, 878-892.

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.X., et al. (2016) Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics*, 32(17), 2664-2671.

Luo, H., Li, M., Wang, S., Liu, Q., Li, Y. and Wang, J. (2018) Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, **34(11)**, 1904-1912.

Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic acids research*, **43(D1)**, D213-D221.

Natarajan, N. and Dhillon, I.S. (2014) Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, **30(12)**, i60-i68.

Nosengo, N. (2016) Can you teach old drugs new tricks? *Nature News*, **534(7607)**, 314.

Padhy, B. and Gupta, Y. (2011) Drug repositioning: re-investigating existing drugs for new therapeutic indications. *Journal of postgraduate medicine*, **57(2)**, 153-160.

Pahikkala, T., Airola, A., Pietila, S., Shakyawar, S., Szwajda, A., Tang, J., et al. (2015) Toward more realistic drug-target interaction predictions. *Briefings in bioinformatics*, **16(2)**, 325-337.

Perlman, L., Gottlieb, A., Atias, N., Ruppin, E. and Sharan, R. (2011)Combining drug and gene similarity measures for drug-target elucidation. *Journal of computational biology* , **18(2)**, 133-145.

Peyvandipour, A., Saberian, N., Shafi, A., Donato, M., Draghici, S. and Valencia, A. (2018) A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, **34(16)**, 2817-2825.

Pritchard, J.L.E., O'Mara, T.A. and Glubb, D.M. (2017) Enhancing the promise of drug repositioning through genetics. *Frontiers in pharmacology*, **8**, 896.

Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., et al. (2018) Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, **18(1)**, 41.

Qu, J., Chen, X., Sun, Y.Z., Li, J.Q. and Ming, Z. (2018) Inferring potential small molecule-miRNA association based on triple layer heterogeneous network. *Journal of cheminformatics*, **10(1)**, 30.

Saito, T. and Rehmsmeier, M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, **10(3)**, e0118432.

Schriml, L.M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., et al. (2015) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research* , **47(D1)**, D995-D962.

Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., Sweet-Cordero, A., et al. (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, **3(96)**, 96ra77.

Srebro, N. and Shraibman, A. (2005) Rank, trace-norm and max-norm. *International Conference on Computational Learning Theory*, Springer; p. 545-560.

Sridhar, D., Fakhraei, S. and Getoor, L. (2016) A probabilistic approach for collective similarity-based drug-drug interaction prediction. *Bioinformatics* , **32(20)**, 3175-3182.

Tamimi, N.A. and Ellis, P. (2009) Drug development: from concept to marketing! *Nephron Clinical Practice*, **113(3)**, c125-c131.

Tan,V.Y. and Févotte,C. (2009) Automatic relevance determination in nonnegative matrix factorization, *In SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*.

van der Voort, P.H., van der Hulst, R.W., Zandstra, D.F., van der Ende, A., Geraedts, A.A. and Tytgat, G.N. (2000) In vitro susceptibility of Helicobacter pylori to, and in vivo suppression by, antimicrobials used in selective decontamination of the digestive tract. *Journal of Antimicrobial Chemotherapy 2000*, **46(5)**, 803-805.

van Riel, N.A. (2006) Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Briefings in bioinformatics*, **7(4)**, 364-374.

Wang, D., Wang, J., Lu, M., Song, F. and Cui, Q. (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26(13)**, 1644-1650.

Wang, Y., Chen, S., Deng, N. and Wang, Y. (2013) Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PloS one*, **8(11)**, e78518.

Wang, L., Wang, Y., Hu, Q. and Li, S. (2014) Systematic Analysis of New Drug Indications by Drug-Gene-Disease Coherent Subnetworks. *CPT: pharmacometrics & systems pharmacology*, **3(11)**, 1-9.

Wang, W., Yang, S., Zhang, X. and Li, J. (2014) Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, **30(20)**, 2923-2930.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., et al. (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, **46(D1)**, D1074-D1082.

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. and Kanehisa, M. (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics s*, **24(13)**, i232-i240.

Yoon, S.Y., Kang, S.Y., Kim, H.W., Kim, H.C. and Roh, D.H. (2015)Clonidine reduces nociceptive responses in mouse orofacial formalin model: potentiation by sigma-1 receptor antagonist BD1047 without impaired motor coordination. *Biological and Pharmaceutical Bulletin*, **38(9)**, 1320-1327.

Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y. and Gao, L. (2015) Inferring drug-disease associations based on known protein complexes. *BMC medical genomics*, **8(2)**, S2.

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018) Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics* , **19(1)**, 233.

Zhang, L., Chen, X., Guan, N.N., Liu, H. and Li, J.Q. (2018) A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction. *Frontiers in Pharmacology*, **9**.

Zhao, Y., Chen, X. and Yin, J. (2018) A novel computational method for the identification of potential miRNA-disease association based on symmetric non-negative matrix factorization and Kronecker regularized least square. *Frontiers in genetics*, **9**, 324.

Zitnik, M., Agrawal, M. and Leskovec, J. (2018) Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **34(13)**, i457-i466.