

Data and text mining

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

Ping Xuan<sup>1</sup>, Yangkun Cao<sup>1</sup>, Tiangang Zhang<sup>2,\*</sup>, Xiao Wang<sup>3</sup>,  
Shuxiang Pan<sup>1</sup> and Tonghui Shen<sup>1</sup>

<sup>1</sup>School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China, <sup>2</sup>School of Mathematical Science, Heilongjiang University, Harbin 150080, China and <sup>3</sup>School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 7, 2018; revised on February 24, 2019; editorial decision on March 10, 2019; accepted on March 12, 2019

## 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 five drugs further confirm DisDrugPred’s ability to discover potential candidate disease indications for drugs.

**Availability and implementation:** The fourth type of drug similarity and the predicted candidates for all the drugs are available at <https://github.com/pingxuan-hlju/DisDrugPred>.

**Contact:** zhang@hlju.edu.cn

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Developing a new drug is a lengthy, complex and expensive process which generally takes 10–15 years and 0.8–1.5 billion dollars (Dickson and Gagnon, 2004; Pushpakom *et al.*, 2018; Tamimi and Ellis, 2009). Drug repositioning is to identify novel therapeutic effects for the drugs that have been approved by the regulatory agencies (Lotfi Shahreza *et al.*, 2018; Padhy and Gupta, 2011). 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, 2016; Pritchard *et al.*, 2017).

Computational prediction of novel therapeutic indications for approved drugs may screen candidate drug–disease associations for further experimental validation (Chen *et al.*, 2016; Hurle *et al.*, 2013; Li *et al.*, 2015). The previous works can be roughly grouped into two categories. Since the drugs execute their functions by targeting the related genes (Bleakley and Yamanishi, 2009; Fakhraei *et al.*, 2014; Yamanishi *et al.*, 2008), 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 (Liu *et al.*, 2016; Luo *et al.*, 2016, 2018; Wang *et al.*, 2014b). 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.*, 2015, 2018a; Qu *et al.*, 2018). 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 not 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 five drugs further confirm that DisDrugPred is able to discover the potential disease indications of drugs.

## 2 Materials and methods

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 and Seung, 2001) is proposed by integrating the multisource 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  $d_j$ .

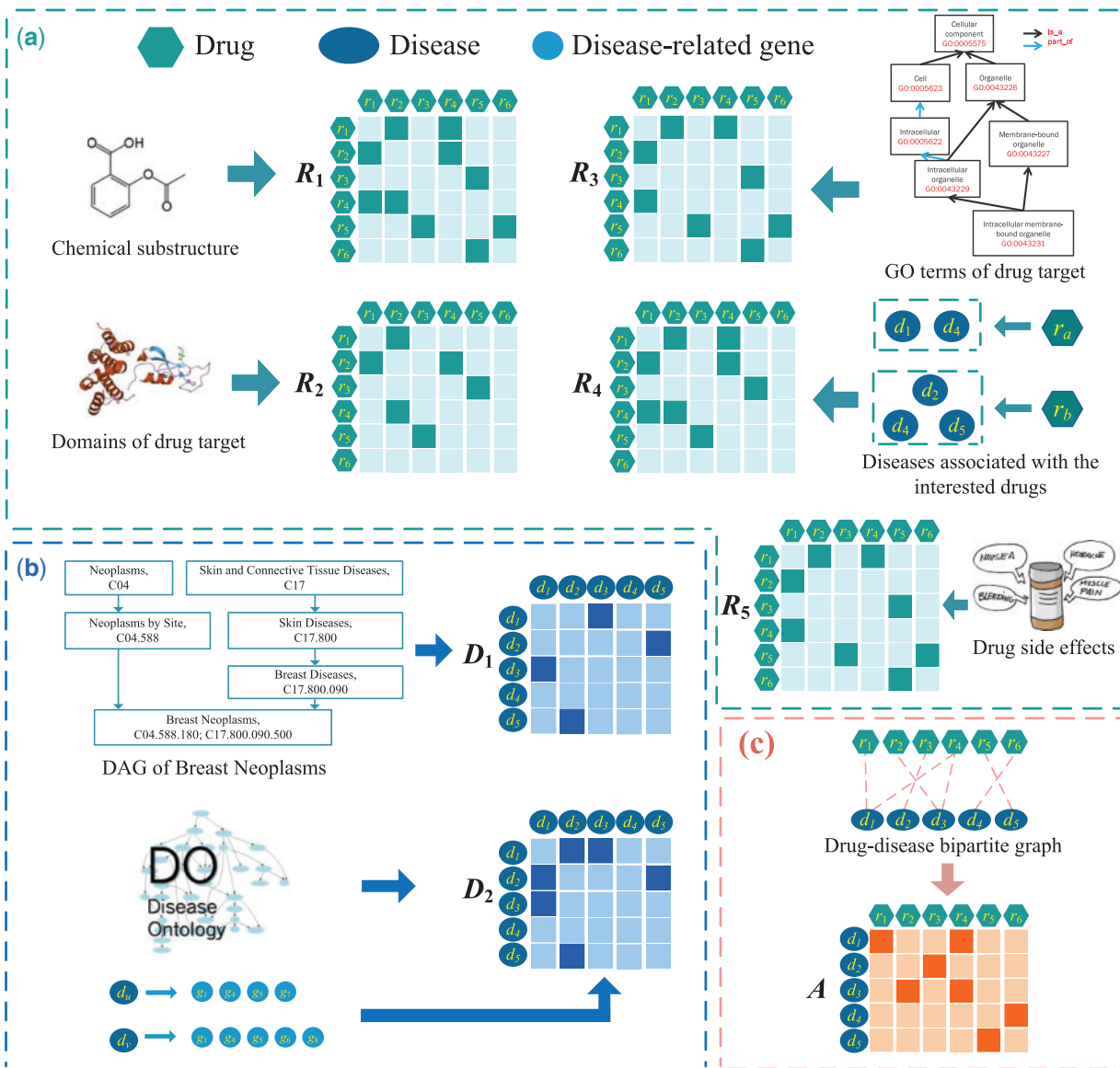
### 2.1 Datasets for drug indication prediction

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 associations are 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.*, 2016), and 571 ones among 763 drugs have their side effect indications. The disease similarities that incorporated the disease ontology (DO) 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 US National Library of Medicine (MeSH, <http://www.ncbi.nlm.nih.gov/mesh>).

### 2.2 Calculation and representation of multisource data

#### 2.2.1 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  $r_b$  (Fig. 1a). Moreover, the drugs with more common domains of target proteins



**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

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  $DT_b = \{d_2, d_4, d_5\}$  (Fig. 1a). 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,

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

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  $R_4 \in \mathbb{R}^{N_r \times N_r}$  is used to represent this type of drug similarity.  $\mathbb{R}$  represents a set of real numbers and  $N_r$  is the number of drugs.  $\mathbb{R}^{N_r \times N_r}$  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 to 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  $R_5 \in \mathbb{R}^{N_r \times N_r}$ , where  $(R_t)_{ij} (1 \leq t \leq m_r)$  is the  $t$ th type of similarity of drugs  $r_i$  and  $r_j$ , and  $m_r$  is the number of the drug similarity types.

### 2.2.2 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 exploits the disease similarities whose values range between 0 and 1. Second, the 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 DO 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  $D_2 \in \mathbb{R}^{N_d \times N_d}$ , where  $(D_s)_{ij} (1 \leq s \leq m_d)$  is the  $s$ th 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 (Fig. 1b).

### 2.2.3 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  $A = (A_{ij}) \in \mathbb{R}^{N_d \times N_r}$  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.

## 2.3 Drug-disease association prediction model

### 2.3.1 Modeling the drug-disease association relationships

Let  $P = (P_{ij}) \in \mathbb{R}^{N_d \times N_r}$  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 (Chen *et al.*, 2018b; Natarajan and Dhillon, 2014; Zhao *et al.*, 2018). Suppose  $\Omega$  be the set of observed drug-disease associations, and  $Y = (Y_{ij}) \in \mathbb{R}^{N_d \times N_r}$  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,

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

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

### 2.3.2 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  $W_s \in \mathbb{R}^{N_d \times k} (1 \leq s \leq m_d)$  be the projection matrix of the  $s$ th type of disease similarity, and  $H_t \in \mathbb{R}^{N_r \times k} (1 \leq t \leq m_r)$  is the

projection matrix corresponding to the  $t$ th 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,

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

where  $\alpha_1$  is a parameter for making a tradeoff between the first optimal term and the second one.

### 2.3.3 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  $s$ th type of actual similarity between diseases  $d_i$  and  $d_j$ . The  $i$ th 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  $j$ th 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  $s$ th type of actual disease similarity in  $D_s$  as possible, which gives rise to the following function,

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

where  $\alpha_2$  is used to control the contribution of the third term.

### 2.3.4 Modeling the prior knowledge of drug similarities

As mentioned before, the five 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  $i$ th row of the transpose of  $P$ ,  $(P^T)_i$  records the association possibilities between drug  $r_i$  and the various diseases. The  $j$ th column of  $P$ ,  $P_j$ , is the possibility column vector of drug  $r_j$  to be associated with all the diseases. Then  $(P^T)_i P_j$  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,

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

where  $\alpha_3$  is used to adjust the contribution of the term about the drug similarities.

### 2.3.5 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  $t$ th  $(1 \leq t \leq m_r)$  type of drug similarity, the adjacency matrix of its corresponding graph is  $M_t \in \mathbb{R}^{N_r \times N_r}$ . Its element in the  $i$ th row and  $j$ th column,  $(M_t)_{ij}$ , is defined as,

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

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  $M_t$ ,

$$\begin{aligned} & \frac{1}{2} \sum_{t=1}^{m_r} \sum_{i,j=1}^{N_r} \|p_i - p_j\|^2 (M_t)_{ij} \\ &= \sum_{t=1}^{m_r} (Tr(PU_t P^T) - Tr(PM_t P^T)) \\ &= \sum_{t=1}^{m_r} Tr(PL_t P^T), \end{aligned} \quad (7)$$

where  $p_i$  and  $p_j$  denote the  $i$ th and  $j$ th 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.  $U_t \in \mathbb{R}^{N_r \times N_r}$  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  $t$ th graph. The smoothness term should also be minimized and we have the following function,

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

where  $\alpha_4$  regulates the contribution of the smoothness term.

### 2.3.6 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

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

where  $P_k$  is the  $k$ th column of  $P$  and  $N_r$  is the number of drugs, and  $\alpha_5$  is a parameter that controls the contribution of penalty term.

### 2.3.7 Introducing regularization term for preventing overfitting

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  $L(P, W_s, H_t)$

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

where  $\alpha_6$  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  $P$  with  $W_s$  and  $H_t$  ( $1 \leq s \leq m_d$ ,  $1 \leq t \leq m_r$ ) fixed, the subproblem for solving  $P$  is as follows,

$$\begin{aligned} \min_{P \geq 0} L(P) &= \|A \odot (P - A)\|_F^2 + \alpha_1 \sum_{s=1}^{m_d} \sum_{t=1}^{m_r} \|P - D_s W_s (R_t H_t)^T\|_F^2 \\ &+ \alpha_2 \sum_{s=1}^{m_d} \|D_s - PP^T\|_F^2 + \alpha_3 \sum_{t=1}^{m_r} \|R_t - P^T P\|_F^2 \\ &+ \alpha_4 \sum_{t=1}^{m_r} Tr(PL_t P^T) + \alpha_5 \sum_{k=1}^{N_r} \|P_k\|_1^2. \end{aligned} \quad (11)$$

After the Frobenius norms of matrices have been transformed to their trace norms (Srebro and Shraibman, 2005),  $L(P)$  is able to be rewritten as:

$$\begin{aligned} L(P) &= Tr(A \odot (PP^T - PA^T - AP^T + AA^T)) \\ &+ \alpha_1 \sum_{1 \leq s \leq m_d} \sum_{1 \leq t \leq m_r} Tr(PP^T - PR_t H_t W_s^T D_s^T - D_s W_s H_t^T R_t^T P^T \\ &\quad + D_s W_s H_t^T R_t^T R_t H_t W_s^T D_s^T) \\ &+ \alpha_2 \sum_{1 \leq s \leq m_d} Tr(D_s D_s^T - D_s PP^T - PP^T D_s^T + PP^T PP^T) \\ &+ \alpha_3 \sum_{1 \leq t \leq m_r} Tr(R_t R_t^T - R_t P^T P - P^T PR_t^T + P^T PP^T P) \\ &+ \alpha_4 \sum_{1 \leq t \leq m_r} Tr(PL_t P^T) \\ &+ \alpha_5 \|e_{1 \times N_d} P\|_F^2, \end{aligned} \quad (12)$$

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

$$\begin{aligned} & A \odot (2P - 2A) + \alpha_1 \sum_{1 \leq s \leq m_d} \sum_{1 \leq t \leq m_r} (2P - 2D_s W_s H_t^T R_t^T) \\ &+ \alpha_2 \sum_{1 \leq s \leq m_d} (-4D_s P + 4PP^T P) + \alpha_3 \sum_{1 \leq t \leq m_r} (-4PR_t + 4PP^T P) \\ &+ \alpha_4 \left( \sum_{1 \leq t \leq m_r} 2P(U_t - M_t) \right) + \alpha_5 (2e_{1 \times N_d}^T e_{1 \times N_d} P) = 0. \end{aligned} \quad (13)$$

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

$$\begin{aligned} & (A \odot (2P - 2A) + \alpha_1 \sum_{1 \leq s \leq m_d} \sum_{1 \leq t \leq m_r} (2P - 2D_s W_s H_t^T R_t^T) \\ &+ \alpha_2 \sum_{1 \leq s \leq m_d} (-4D_s P + 4PP^T P) + \alpha_3 \sum_{1 \leq t \leq m_r} (-4PR_t + 4PP^T P) \\ &+ \alpha_4 \left( \sum_{1 \leq t \leq m_r} 2P(U_t - M_t) \right) + \alpha_5 (2e_{1 \times N_d}^T e_{1 \times N_d} P))_{ij} P_{ij} = 0. \end{aligned} \quad (14)$$

According to the coordinate gradient descent algorithm in Tan and Févotte (2009),  $P_{ij}$  can be updated by multiplying it with the



ratio of the negative terms to the positive terms in the left side of Equation (14),

$$P_{ij}^{\text{new}} \leftarrow P_{ij} \cdot \frac{(2A \odot A + 2\alpha_1 \sum_{1 \leq s \leq m_d} \sum_{1 \leq t \leq m_r} D_s W_s H_t^T R_t^T + 4\alpha_2 \sum_{1 \leq s \leq m_d} D_s P + 4\alpha_3 \sum_{1 \leq t \leq m_r} P R_t + 2\alpha_4 \sum_{1 \leq t \leq m_r} P M_t)_{ij}}{(2A \odot P + 2m_d m_r \alpha_1 P + 4m_d \alpha_2 P P^T P + 4m_r \alpha_3 P P^T P + 2\alpha_4 \sum_{1 \leq t \leq m_r} P U_t + 2\alpha_5 e_{1 \times N_d}^T e_{1 \times N_d} P)_{ij}}. \quad (15)$$

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

$$\min_{W_s \geq 0} L(W_s) = \alpha_1 \sum_{1 \leq t \leq m_r} \|P - D_s W_s (R_t H_t)^T\|_F^2 + \alpha_6 \|W_s\|_F^2. \quad (16)$$

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

$$L(W_s) = \alpha_1 \sum_{1 \leq t \leq m_r} \text{Tr}(P P^T - P R_t H_t W_s^T D_s^T - D_s W_s H_t^T R_t^T P^T + D_s W_s H_t^T R_t^T H_t W_s^T D_s^T) + \alpha_6 \text{Tr}(W_s W_s^T). \quad (17)$$

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

$$\alpha_1 \sum_{1 \leq t \leq m_r} (-2D_s^T P R_t H_t + 2D_s^T D_s W_s H_t^T R_t^T R_t H_t) + 2\alpha_6 W_s = 0. \quad (18)$$

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

$$\left( \alpha_1 \sum_{1 \leq t \leq m_r} (-2D_s^T P R_t H_t + 2D_s^T D_s W_s H_t^T R_t^T R_t H_t) + 2\alpha_6 W_s \right)_{ij} (W_s)_{ij} = 0. \quad (19)$$

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

$$(W_s)_{ij}^{\text{new}} \leftarrow (W_s)_{ij} \cdot \frac{\left( 2\alpha_1 \sum_{1 \leq t \leq m_r} D_s^T P R_t H_t \right)_{ij}}{\left( 2\alpha_1 \sum_{1 \leq t \leq m_r} D_s^T D_s W_s H_t^T R_t^T R_t H_t + 2\alpha_6 W_s \right)_{ij}}. \quad (20)$$

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

$$\min_{H_t \geq 0} L(H_t) = \alpha_1 \sum_{1 \leq s \leq m_d} \|P - D_s W_s (R_t H_t)^T\|_F^2 + \alpha_6 \|H_t\|_F^2. \quad (21)$$

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

$$(H_t)_{ij}^{\text{new}} \leftarrow (H_t)_{ij} \cdot \frac{\left( 2\alpha_1 \sum_{1 \leq s \leq m_d} R_t^T P^T D_s W_s \right)_{ij}}{\left( 2\alpha_1 \sum_{1 \leq s \leq m_d} R_t^T R_t H_t W_s^T D_s^T D_s W_s + 2\alpha_6 H_t \right)_{ij}}. \quad (22)$$

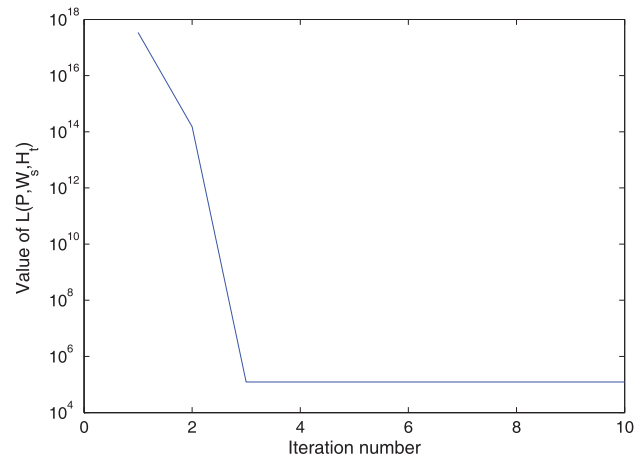


Fig. 2. Convergence of  $L(P, W_s, H_t)$

The convergence curve of the objective function  $L(P, W_s, H_t)$ , confirm that the function can converge to its local minima (Fig. 2).  $L(P, W_s, H_t)$  is solved by iteratively using the updating rules of  $P$ ,  $W_s$  and  $H_t$ . The iterative process is over when the absolute difference of  $L(P, W_s, H_t)$  at two adjacent moments is less than a threshold ( $\varepsilon = 10^{-6}$ ) or the maximum number of iterations, 100, is reached. Finally,  $P_{ij}$  is regarded as the estimated association score between disease  $d_i$  and drug  $r_j$  (Fig. 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 five 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 performance is. Note that association dataset is separated to 5-fold 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,

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

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

**Algorithm:** DisDrugPred algorithm for inferring the potential drug-disease associations

**Input:** a disease-drug association matrix  $A \in \mathbb{R}^{N_d \times N_r}$ , a disease similarity matrix  $D_1 D_2 \in \mathbb{R}^{N_d \times N_d}$ , and the drug similarity matrices  $R_1, R_2, R_3, R_4, R_5 \in \mathbb{R}^{N_r \times N_r}$

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

1 Initialize the elements of  $P, W_s$ , and  $H_t$  ( $1 \leq s \leq m_d, 1 \leq t \leq m_r$ ) with the random values in the range  $[0, 1]$

2 While  $L(P, W_s, H_t)$  not converged do

3 fix  $W_s$  and  $H_t$ , and update  $P$  using the rule:

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

4 For  $s = 1$  to  $m_d$  do

5 fix  $P$  and  $H_t$ , and update  $W_s$  using the rule:

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

6 End For

7 For  $t = 1$  to  $m_r$  do

8 fix  $P$  and  $W_s$ , and update  $H_t$  using the rule:

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

9 End For

10 End While

**Fig. 3.** Iterative algorithm for estimation of the disease–drug association scores

samples). In such case, the precision–recall (PR) curve is more informative than the ROC curve (Saito and Rehmsmeier, 2015). Precision and recall are defined as,

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

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-fold when used as a test set, and the five 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,

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

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

where  $W_{rr}$ ,  $W_{rt}$  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  $k$ th 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,

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

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

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

$$RD_t = (rflag * Rr + dflag * Rd) / (rflag + dflag), \quad (29)$$

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,

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

where  $X_p$  represents the  $p$ th 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,

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

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  $i$ th disease and the  $j$ th drug.

*DisDrugPred*'s hyperparameters,  $\alpha_1$ – $\alpha_6$ , 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  $\alpha_1 = 10$ ,  $\alpha_2 = 10$ ,  $\alpha_3 = 0.1$ ,  $\alpha_4 = 10$ ,  $\alpha_5 = 10$  and  $\alpha_6 = 10$ , 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, 2006) of *DisDrugPred*'s six parameters are evaluated by changing one of parameters and fixing the remaining ones. The SC values of  $\alpha_1$ – $\alpha_6$  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  $\alpha_1$ ,  $\alpha_4$  and  $\alpha_6$ , while  $\alpha_2$ ,  $\alpha_3$  and  $\alpha_5$  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 and Rehmsmeier, 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  $R_4$  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.

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 (Fig. 5), and ranks 70.5% in top 30, 84.5% in top

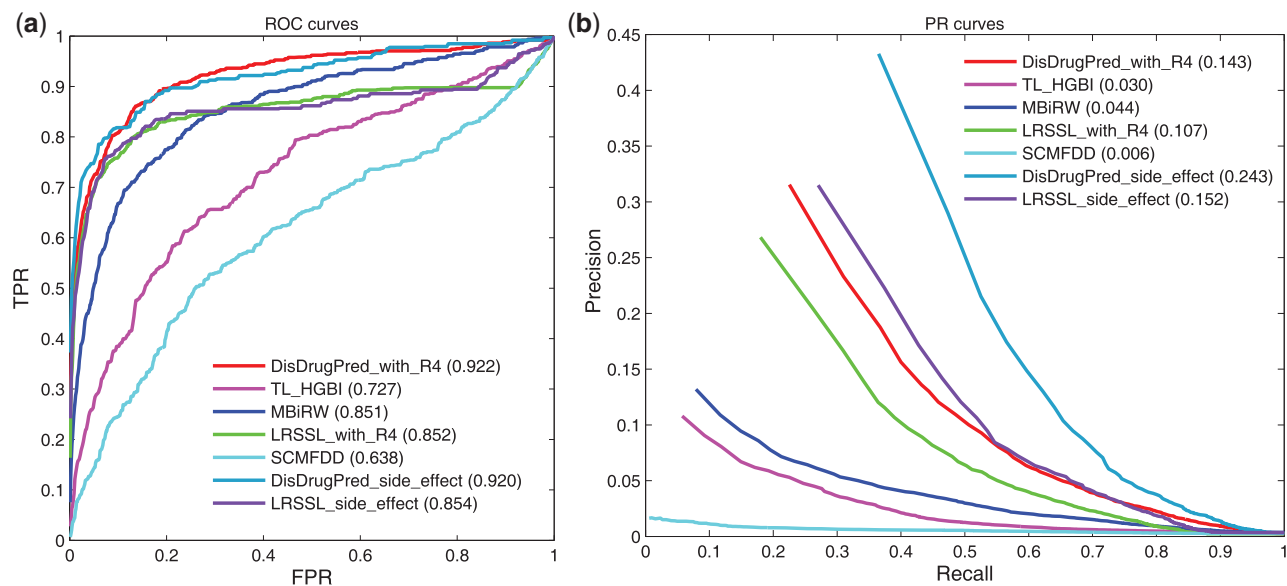
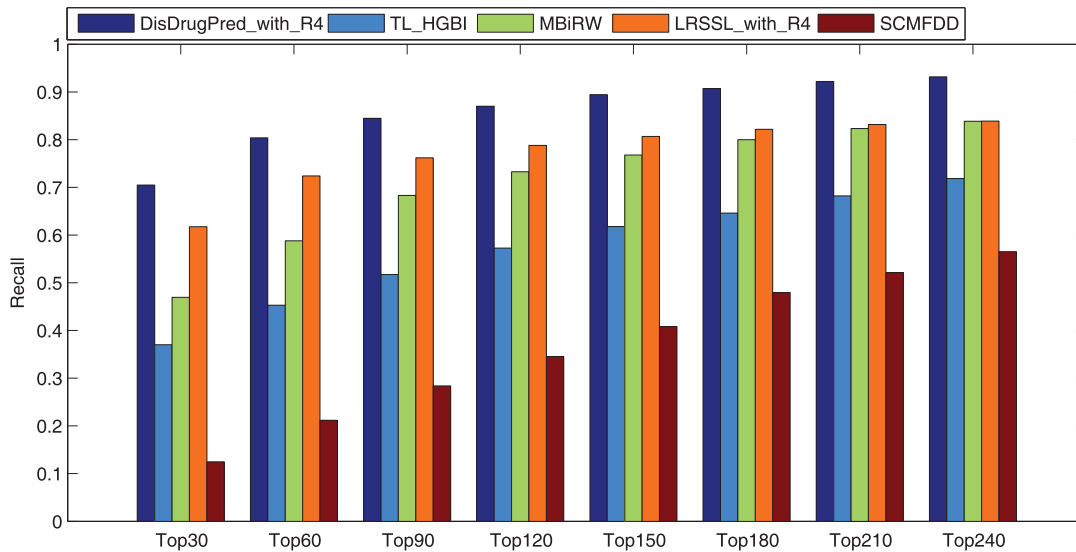


Fig. 4. (a) ROC curves and (b) PR curves of drug–disease association prediction by different methods



**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

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\_ $R_4$  and DisDrugPred\_without\_ $R_4$ , 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  $R_4$  and without  $R_4$ , LRSSL\_with\_ $R_4$  and LRSSL\_without\_ $R_4$ , respectively.

First, the instances of DisDrugPred and LRSSL with  $R_4$  perform better than the ones without  $R_4$ , respectively (Supplementary Fig. S1). DisDrugPred\_with\_ $R_4$ 's AUC and AUPR are 1.9% and 1.9% higher than DisDrugPred\_without\_ $R_4$ . LRSSL\_with\_ $R_4$ 's AUC and AUPR also increase by 5.1% and 1.9% compared with LRSSL\_without\_ $R_4$ . 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 Fig. S1). DisDrugPred\_with\_ $R_4$  achieves 7% and 3.6% higher AUC and AUPR than LRSSL\_with\_ $R_4$ . DisDrugPred\_without\_ $R_4$ 's AUC and AUPR also increase by 10.2% and 3.6% compared with LRSSL\_without\_ $R_4$ . 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 types of disease similarities ( $D_1$  and  $D_2$ ), and is referred to as DisDrugPred\_with\_ $D_2$ . Another DisDrugPred's instance is trained without  $D_2$ , and is named DisDrugPred\_without\_ $D_2$ . 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 S1, DisDrugPred\_with\_ $D_2$ 's AUPR is a little bit higher than DisDrugPred\_without\_ $D_2$  and it increases by 0.3%, while DisDrugPred\_with\_ $D_2$ 's AUC is equal to DisDrugPred\_without\_ $D_2$ 's one. It indicates  $D_2$  has a slight effect on the prediction performances.

### 3.4 Case studies on five drugs

To further demonstrate DisDrugPred's ability to discover the potential drug-disease associations, case studies on five drugs, *ciprofloxacin*, *clonidine*, *ampicillin*, *etoposide* and *cefotaxime*, are conducted. For each of these five 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 and Patel, 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

**Table 2.** The top 10 candidates related to the drugs *ciprofloxacin*, *clonidine*, *ampicillin*, *etoposide* and *cefotaxime*, respectively

Drug ID	Rank	Disease name	Description	Rank	Disease name	Description
<i>Ciprofloxacin</i>	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
<i>Clonidine</i>	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 <i>et al.</i> , 2015)
	5	Muscle cramp	PubChem	10	Muscle rigidity	Unconfirmed
<i>Ampicillin</i>	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
<i>Etoposide</i>	1	Breast	CTD	6	Lymphoma	CTD
	2	Sarcoma	CTD	7	Urinary tract infections	Inferred candidate by 1 literature
	3	Leukemia	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
<i>Cefotaxime</i>	1	Bacterial infections	CTD, ClinicalTrials	6	Gram-positive bacterial infections	CTD, DrugBank
	2	Enterobacteriaceae infections	DrugBank	7	Helicobacter infections	Literature (van der Voort <i>et al.</i> , 2000)
	3	Gram-negative bacterial infections	CTD, DrugBank	8	Eye infections, bacterial	Literature (Kramann <i>et al.</i> , 2001)
	4	Pseudomonas infections	DrugBank	9	Staphylococcal skin infections	DrugBank
	5	Respiratory tract infections	CTD, ClinicalTrials	10	Septicemia	DrugBank

Note: (1) 'CTD' means a drug-disease association is included by the 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.

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 US 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 four candidates included by ClinicalTrials.gov and two candidates contained by PubChem, indicating these drug-disease associations are supported by the clinical trials. In addition, the four 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 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 *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 S1](#). In addition, the fourth type of drug similarity based on their associated diseases is shown in [Supplementary Table S2](#).

## 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 five types of drug similarities and two 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.

## Funding

The work was supported by the Natural Science Foundation of China [61702296, 61302139], the Heilongjiang Postdoctoral Scientific Research Starting 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 [2016RQXXJ135] and the Foundation of Graduate Innovative Research [YJSCX2018-140HLJU, YJSCX2018-047HLJU].

*Conflict of Interest:* none declared.

## References

- Bleakley, K. and Yamanishi, Y. (2009) Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, **25**, 2397–2403.
- Bodenreider, O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.
- Bozkaya, Y. et al. (2017) Effectiveness of low-dose oral etoposide treatment in patients with recurrent and platinum-resistant epithelial ovarian cancer. *J. Obstet. Gynaecol.*, **37**, 649–654.
- Brown, A.S. and Patel, C.J. (2017) A standard database for drug repositioning. *Sci. Data*, **4**, 170029.
- Cai, D. et al. (2011) Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **8**, 1548–1560.
- Chen, X. et al. (2015) Drug–target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.*, **17**, 696–712.
- Chen, X. et al. (2016) NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput. Biol.*, **12**, e1004975.
- Chen, X. et al. (2018a) MicroRNA-small molecule association identification: from experimental results to computational models. *Brief. Bioinform.*, **16**, 2018.
- Chen, X. et al. (2018b) Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics*, **34**, 4256–4265.
- Cheng, L. et al. (2014) SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One*, **9**, e99415.
- Cheng, L. et al. (2018) DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics*, **34**, 1953–1956.
- Consortium, U. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
- Davis, A.P. et al. (2016) The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.*, **45**, D972–D978.
- Dickson, M. and Gagnon, J.P. (2004) Key factors in the rising cost of new drug discovery and development. *Nat. Rev. Drug Discov.*, **3**, 417.
- Ding, H. et al. (2013) Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Brief. Bioinform.*, **15**, 734–747.
- Fakhraei, S. et al. (2014) Network-based drug–target interaction prediction with probabilistic soft logic. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **11**, 775–787.
- Gottlieb, A. et al. (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.
- Gottlieb, A. et al. (2012) INDI: a computational framework for inferring drug interactions and their associated recommendations. *Mol. Syst. Biol.*, **8**, 592.
- Hajian-Tilaki, K. (2013) Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.*, **4**, 627.
- Hurle, M. et al. (2013) Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.*, **93**, 335–341.
- Iwata, H. et al. (2015) Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *J. Chem. Inform. Model.*, **55**, 446–459.
- Kim, S. et al. (2015) PubChem substance and compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.
- Kramann, C. et al. (2001) Effects of intraocular cefotaxime on the human corneal endothelium 1. *J. Cataract Refract. Surg.*, **27**, 250–255.
- Kuhn, M. et al. (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res.*, **44**, D1075–1079.
- Lee, D.D. and Seung, H.S. (2001) Algorithms for non-negative matrix factorization. *Adv. Neural Inform. Process. Syst.*, **556**, 562.
- Li, J. et al. (2015) A survey of current trends in computational drug repositioning. *Brief. Bioinform.*, **17**, 2–12.
- Liang, X. et al. (2017) LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning. *Bioinformatics*, **33**, 1187–1196.
- Ling, C.X. et al. (2003) AUC: a better measure than accuracy in comparing learning algorithms. In: *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, Heidelberg, Berlin, Germany, pp. 329–341.
- Liu, H. et al. (2016) Inferring new indications for approved drugs via random walk on drug–disease heterogeneous networks. *BMC Bioinform.*, **17**, 539.
- Liu, H. et al. (2018) Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Mol. Ther. Nucleic Acids*, **13**, 303–311.
- Lotfi Shahreza, M. et al. (2018) Review of network-based approaches to drug repositioning. *Brief. Bioinform.*, **19**, 878–892.
- Luo, H. et al. (2018) Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, **34**, 1904–1912.
- Luo, H. et al. (2016) Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics*, **32**, 2664–2671.
- Mitchell, A. et al. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
- Natarajan, N. and Dhillon, I.S. (2014) Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, **30**, i60–68.
- Nosengo, N. (2016) Can you teach old drugs new tricks? *Nat. News*, **534**, 314.
- Padhy, B. and Gupta, Y. (2011) Drug repositioning: re-investigating existing drugs for new therapeutic indications. *J. Postgrad. Med.*, **57**, 153–160.
- Pahikkala, T. et al. (2015) Toward more realistic drug–target interaction predictions. *Brief. Bioinform.*, **16**, 325–337.
- Perlman, L. et al. (2011) Combining drug and gene similarity measures for drug–target elucidation. *J. Comput. Biol.*, **18**, 133–145.
- Peyvandipour, A. et al. (2018) A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, **34**, 2817–2825.
- Pritchard, J.L.E. et al. (2017) Enhancing the promise of drug repositioning through genetics. *Front. Pharmacol.*, **8**, 896.
- Pushpakom, S. et al. (2018) Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.*, **18**, 41.
- Qu, J. et al. (2018) Inferring potential small molecule-miRNA association based on triple layer heterogeneous network. *J. Cheminform.*, **10**, 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**, e0118432.
- Schriml, L.M. et al. (2018) Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.

- Sirota, M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, **3**, 96ra77.
- Srebro, N. and Shraibman, A. (2005) Rank, trace-norm and max-norm. In: *International Conference on Computational Learning Theory*. Springer, Heidelberg, Berlin, Germany, pp. 545–560.
- Sridhar, D. *et al.* (2016) A probabilistic approach for collective similarity-based drug–drug interaction prediction. *Bioinformatics*, **32**, 3175–3182.
- Tamimi, N.A. and Ellis, P. (2009) Drug development: from concept to marketing!. *Nephron Clin. Pract.*, **113**, c125–c131.
- Tan, V.Y. and Févotte, C. (2009) Automatic relevance determination in non-negative matrix factorization. In: *SPARS'09-signal Processing with Adaptive Sparse Structured Representations*. Saint Malo, UK.
- van der Voort, P.H. *et al.* (2000) *In vitro* susceptibility of *Helicobacter pylori* to, and *in vivo* suppression by, antimicrobials used in selective decontamination of the digestive tract. *J. Antimicrob. Chemother.*, **46**, 803–805.
- van Riel, N.A. (2006) Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief. Bioinform.*, **7**, 364–374.
- Wang, D. *et al.* (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26**, 1644–1650.
- Wang, L. *et al.* (2014a) Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks. *CPT Pharmacometrics Syst. Pharmacol.*, **3**, 1–9.
- Wang, W. *et al.* (2014b) Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, **30**, 2923–2930.
- Wang, Y. *et al.* (2013) Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One*, **8**, e78518.
- Wishart, D.S. *et al.* (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
- Yamanishi, Y. *et al.* (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.
- Yoon, S.Y. *et al.* (2015) Clonidine reduces nociceptive responses in mouse orofacial formalin model: potentiation by sigma-1 receptor antagonist BD1047 without impaired motor coordination. *Biol. Pharm. Bull.*, **38**, 1320–1327.
- Yu, L. *et al.* (2015) Inferring drug–disease associations based on known protein complexes. *BMC Med. Genomics*, **8**, S2.
- Zhang, L. *et al.* (2018a) A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction. *Front. Pharmacol.*, **9**.
- Zhang, W. *et al.* (2018b) Predicting drug–disease associations by using similarity constrained matrix factorization. *BMC Bioinform.*, **19**, 233.
- Zhao, Y. *et al.* (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. *Front. Genet.*, **9**, 324.
- Zitnik, M. *et al.* (2018) Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **34**, i457–i466.