Briefings in Bioinformatics, 22(4), 2021, 1–14

https://doi.org/10.1093/bib/bbaa267
Problem Solving Protocol

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

Computational drug repositioning based on multi-similarities bilinear matrix factorization

Mengyun Yang, Gaoyan Wu, Qichang Zhao, Yaohang Li and Jianxin Wang

Corresponding author: Jianxin Wang, Tel: +86-731-88820212; Fax: +86-731-88877936; E-mail: jxwang@mail.csu.edu.cn

Abstract

With the development of high-throughput technology and the accumulation of biomedical data, the prior information of biological entity can be calculated from different aspects. Specifically, drug–drug similarities can be measured from target profiles, drug–drug interaction and side effects. Similarly, different methods and data sources to calculate disease ontology can result in multiple measures of pairwise disease similarities. Therefore, in computational drug repositioning, developing a dynamic method to optimize the fusion process of multiple similarities is a crucial and challenging task. In this study, we propose a multi-similarities bilinear matrix factorization (MSBMF) method to predict promising drug-associated indications for existing and novel drugs. Instead of fusing multiple similarities into a single similarity matrix, we concatenate these similarity matrices of drug and disease, respectively. Applying matrix factorization methods, we decompose the drug–disease association matrix into a drug-feature matrix and a disease-feature matrix. At the same time, using these feature matrices as basis, we extract effective latent features representing the drug and disease similarity matrices to infer missing drug–disease associations. Moreover, these two factored matrices are constrained by non-negative factorization to ensure that the completed drug–disease association matrix is biologically interpretable. In addition, we numerically solve the MSBMF model by an efficient alternating direction method of multipliers algorithm. The computational experiment results show that MSBMF obtains higher prediction accuracy than the state-of-the-art drug repositioning methods in cross-validation experiments. Case studies also demonstrate the effectiveness of our proposed method in practical applications. **Availability:** The data and code of MSBMF are freely available at https://github.com/BioinformaticsCSU/MSBMF. Corresponding author: Jianxin Wang, School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, P. R. China. E-mail: jxwang@mail.csu.edu.cn **Supplementary Data:** Supplementary data are available online at https://academic.oup.com/bib.

Key words: drug repositioning; matrix factorization; drug–disease associations; multi-similarities; association prediction; ADMM

Mengyun Yang is a PhD candidate in the School of Computer Science and Engineering, Central South University, China, and with Provincial Key Laboratory of Informational Service for Rural Area of Southwestern Hunan, Shaoyang University, China. His current research interests include machine learning, deep learning and bioinformatics.
Gaoyan Wu is a graduate student in School of Computer Science and Engineering, Central South University, China. Her main research interests include matrix completion and drug–target interaction prediction.
Qichang Zhao is a PhD candidate in School of Computer Science and Engineering, Central South University, China. His current research interests include machine learning, deep learning and bioinformatics.
Yaohang Li is an associate professor in computer science at Old Dominion University, USA. His research interests are in protein structure modelling, computational biology, bioinformatics, Monte Carlo methods, big data algorithms and parallel and distributive computing.
Jianxin Wang is the dean and a professor in the School of Computer Science and Engineering, Central South University, China and with Hunan Provincial Key Lab of Bioinformatics, Central South University, Changsha, 410083, China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics and computer network.
Submitted: 20 July 2020; **Received (in revised form):** 31 August 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

The procedure of traditional drug discovery is time consuming, expensive and high risk [1, 2]. Although the spending on drug research continues to grow every year, the number of new drugs approved for production annually has stagnated. The process of identifying new indications for existing drugs, known as drug repositioning, is an effective way to improve the efficiency of drug development [3]. The approved drugs have safety, efficacy and toleration data available after preliminary tests and clinical trials, which has the potential to save a large amount of money and time for developing new treatments [4, 5]. With the development of high-throughput technology, a large number of multi-omic data have been generated and developed, which provide a variety of biological sources for computational drug repositioning. These omics data can be roughly divided into three categories, including drug-, disease- and protein/gene-related databases. More specifically, some drug-related databases are popularly used, such as DrugBank [6], PubChem [7], CTD [8] and SIDER [9]. Some disease-related databases, e.g. Disease Ontology (DO) [10], MalaCards [11], Online Mendelian Inheritance in Man (OMIM) [12] and DisGeNET [13], provide disease terms or genetic information. The protein/gene-related databases, such as UniProtKB [14], BioGrid [15], HPRD [16] and PDB [17], provide extended information for drug research.

Recently, quite a few computational methods have been proposed for drug repositioning. Network-based and machine learning-based methods are among the most popularly used ones. Based on the assumption that similar drugs share the similar molecular pathways to treat similar diseases, both classes of methods focus on extracting accurate and reliable similarity information between drugs as well as diseases to predict new drug–disease associations.

The fundamental idea of network-based methods is to construct homogeneous or heterogeneous biological networks, such as disease–disease, drug–drug, drug–target or disease–gene networks to infer the similarity relationships between biological entities. The similarity information extracted from these biological networks plays an important role in predicting the potential drug–disease associations. For example, Luo *et al.* [18] proposed a comprehensive similarity scheme to measure the similarities of drug and disease and then used bi-random walk algorithm in a heterogeneous drug–disease association network to predict potential indications for approved and novel drugs. Wang *et al.* [19] constructed a heterogeneous drug–target–disease network, which included drug similarity, protein sequence similarity, disease phenotype similarity, drug–target interactions and drug–disease associations. A triple layer heterogeneous graph-based inference (TL-HGBI) method was then developed to compute the potential drug–disease association probabilities on the heterogeneous drug–target–disease network. Martínez *et al.* [20] developed a network-based drug–disease prioritization method named DrugNet to predict new candidate indications for drugs and new therapeutic drugs for diseases. Yu *et al.* [21] constructed a drug–complex–disease tripartite networks to infer drug–disease associations, which is not like other network-based algorithm that need to obtain drug and disease similarities. Protein complexes were considered as the bridge between drug–complex network and complex–disease network and then indirect drug–disease score was computed using symmetric probability model.

The machine learning-based methods typically follow the protocol of training classifiers with the prior information of diseases and drugs and use them to predict new drug–disease associations. Based on the multi-class support vector machine classifier, Napolitano *et al.* [22] proposed a drug-centered computational method, which calculates the average drug similarities based on three drug-related information sources (chemical structures, drug–target interactions and gene expressions), which serve as the classification features to identify a therapeutic class for the given drugs. Liang *et al.* [23] proposed a sparse subspace learning method (LRSSL) integrating drug chemical data, target domain data and target annotation data to predict indications for new and approved drugs. Integrating drug-exposure gene expression profiles, disease–gene expression profiles and the existing drug–disease relationship, Saberian *et al.* [24] used a distance metric learning technique to infer approved drugs for a specified indication. It is worth mentioning that matrix factorization and matrix completion methods have been demonstrated promising success for computational drug repositioning in recent years [25–31]. These methods address the drug repositioning problem as a recommendation problem. Luo *et al.* [25] proposed a recommendation system named drug repositioning recommendation system (DRRS) based on the singular value thresholding (SVT) model [32] to fill out the incomplete drug–disease matrix, which is derived from a heterogeneous drug–disease network by integrating drug–drug and disease–disease similarities as sub-networks. Assuming noisy similarity calculations and hence a noisy drug–disease association matrix, Yang *et al.* [26] enhanced DRRS algorithm by a bounded nuclear norm regularization (BNNR) method. BNNR is not only able to tolerate the potential noise from calculating drug and disease similarities but also restricts the predicted scores in the range of $[0, 1]$. Dai *et al.* [28] proposed a matrix factorization model with a genomic space to predict new drug indications. The feature vectors of drug and disease are extracted from the genomic space and are utilized to predict the missing drug–disease pairs. Zhang *et al.* [31] proposed a similarity constrained matrix factorization (SCMFDD) method for predicting drug–disease associations. SCMFDD only used single drug feature-based similarity and disease semantic similarity. The authors collected five types of drug features, such as substructures, targets, pathways, enzymes and drug–drug interactions. The experiment results show that substructure- and drug interaction-based similarities are more effective in SCMFDD.

Most of the network- and machine learning-based drug repositioning methods utilize only one single measure to evaluate the similarity between pairwise diseases as well as between pairwise drugs. The disease similarity is often calculated based on disease phenotype, while the drug similarity is evaluated according to their chemical structures. In fact, the similarities of pairwise drugs and/or pairwise diseases are not only noisy but also multi-modal, which can be measured from different aspects. Fusing the multiple similarity measures can effectively tolerate the noise in individual similarity computation, extract effective features and hence provide more precise quantification of the drug–drug and the disease–disease relationships. The popular way of similarity fusion is to integrate multiple similarity measures into a comprehensive similarity either by using certain operators (e.g. maximum, minimum or average) or by the network-based fusion method. These similarity fusion approaches are typically carried out in the preprocessing stage where the fusion parameters are determined and then fixed afterwards. These fusion parameters can no longer be changed during the optimization process in drug–disease association training. Moreover, fusing multiple similarity measures into a single comprehensive similarity value may lead to information loss. All of these inspire us to develop a dynamic approach

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

to optimize the fusion process and the drug–disease association training process simultaneously so as to make fully use of the multiple similarity measures between pairwise drugs and pairwise diseases.

In this study, we propose a multi-similarities bilinear matrix factorization (MSBMF) method by merging multiple similarity measures of pairwise drugs and pairwise diseases to improve the reliability of predicted indications for approved and novel drugs. First of all, five measures of drug similarities and two measures of disease similarities are computed using publicly available software packages. Then, two concatenated drug and disease similarity matrices are constructed based on the above similarity calculations. Finally, we integrate these two concatenated similarity matrices and the drug–disease association matrix into our MSBMF model, where the fusion of the drug and disease similarity information and the optimization of the drug–disease associations are carried out at the same time. The overall workflow of our method is illustrated in Figure 1. The main contributions of our MSBMF algorithm include the following:

- MSBMF provides an effective scheme to dynamically integrate multiple similarities between pairwise drugs and pairwise diseases into drug–disease association training.
- MSBMF model incorporates two non-negative constraints to ensure that the entries of predicted association matrix are non-negative, which can be interpreted.
- An iterative computational method based on the alternating direction method of multipliers (ADMM) is developed to effectively solve the MSBMF model.

Related works

In this section, we provide a detail review of five state-of-the-art drug repositioning approaches, including DRRS [25], BNNR [26], MBiRW [18], DrugNet [20] and HGBI [33].

DRRS: considering drug repositioning as a recommendation system problem, Luo *et al.* [25] proposed a DRRS algorithm. A global drug–disease adjacency matrix was constructed by integrating drug similarity matrix, disease similarity matrix and drug–disease association matrix. To obtain a low-rank approximation of the adjacency matrix, Luo *et al.* used an SVT algorithm to complete the missing entries. The SVT model can be formulated as follows:

$$\begin{aligned} &\min_X \tau \|X\|_* + \frac{1}{2}\|X\|_F^2 \\ &\text{s.t.} \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M), \end{aligned} \tag{1}$$

where M is the adjacency matrix, Ω is a set of index pairs containing all known entries, $\mathcal{P}_\Omega$ is a projection operator on Ω and $\|X\|_*$ represents the nuclear norm of X, which can lead to the low-rank approximation for X. Actually, there are some noise existing in similarity matrices since the similarity values are computationally estimated instead of experimentally validated. The objective function of the matrix completion model should take the noise of matrix M into account. In addition, the predicted values using DRRS algorithm may lead to negative values, which bring difficulty in biological interpretation.

BNNR: In order to address the above issues in DRRS algorithm, Yang *et al.* [26] proposed a BNNR method for computational drug repositioning. The model of BNNR is described as follows:

$$\begin{aligned} &\min_X \|X\|_* + \frac{\alpha}{2}\left\|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(M)\right\|_F^2 \\ &\text{s.t. } 0 \le X \le 1. \end{aligned} \tag{2}$$

BNNR model not only introduces a soft regularization term to tolerate noise data but also adds a bounded constraint to ensure the predict values are with in the range $[0, 1]$. However, singular value decomposition is required in the process of optimizing the nuclear norm, which makes BNNR algorithm time consuming for large-scale problems.

MBiRW: considering the effect of known drug–disease association information to similarity measures, Luo *et al.* [18] proposed a comprehensive similarity scheme to enhance drug and disease similarities using a logistic function [34] and a graph clustering method named ClusterONE [35]. On the updated heterogeneous drug–disease network, the bi-random walk is performed to rank candidate indications for drugs, which is called MBiRW algorithm. In fact, based on the idea of similarity scheme, the more prior biological information can be also integrated to further improve similarity measures, such as drug–target interactions and disease–gene associations. As shown in [25], the process of random walk is equivalent to that of approximating the eigenvector associated with the largest eigenvalue of its transition matrix in the heterogeneous drug–disease network, which is equivalent to the matrix completion model when the largest eigenvalue is dominating in the transition matrix.

DrugNet: based on a heterogeneous network prioritization, Martínez *et al.* [20] developed a method called DrugNet for drug repositioning, which can integrate different types of data from the similarity network and the association network. DrugNet implements drug–disease and disease–drug prioritization on the heterogeneous drug–disease network by a propagation flow algorithm named ProphNet [36]. In practice, for a given drug, if a large number of disease associations are available, the prediction can be quite accurate. However, DrugNet suffers from the cold start problem, which may not have much advantage in identifying potential indications for novel drugs without any known associated diseases.

HGBI: based on the guilt-by-association principle [37, 38], Wang *et al.* [33] proposed an HGBI algorithm to predict drug–target interactions, which had been extended to drug repositioning [19]. HGBI algorithm has fast convergence, which is suitable for large-scale problems. Due to small similarity values providing little information in association inference [39], HGBI involves a process of removing similarity values lower than a certain threshold. It makes the updated similarity measures rough. We suggest that designing an accurate similarity measure to further improve the prediction performance of HGBI.

Materials

Datasets

We obtain the validated drug–disease association data from [40], which is recognized as the gold standard dataset in drug repositioning. The total number of drug–disease associations is 1933, and the numbers of related drugs and diseases are 593 and 313, respectively. We construct the corresponding drug–disease association matrix, where the known associations are represented as 1s, while the unknowns are denoted as 0s.

Drug similarity measures

The drug similarities used in MSBMF are calculated based on their chemical structures, anatomical therapeutic chemical (ATC) codes, side effects, drug–drug interactions and target profiles.

Chemical structure similarity R_{chem}. The chemical structure similarity measures the similarity of the chemical compounds in

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

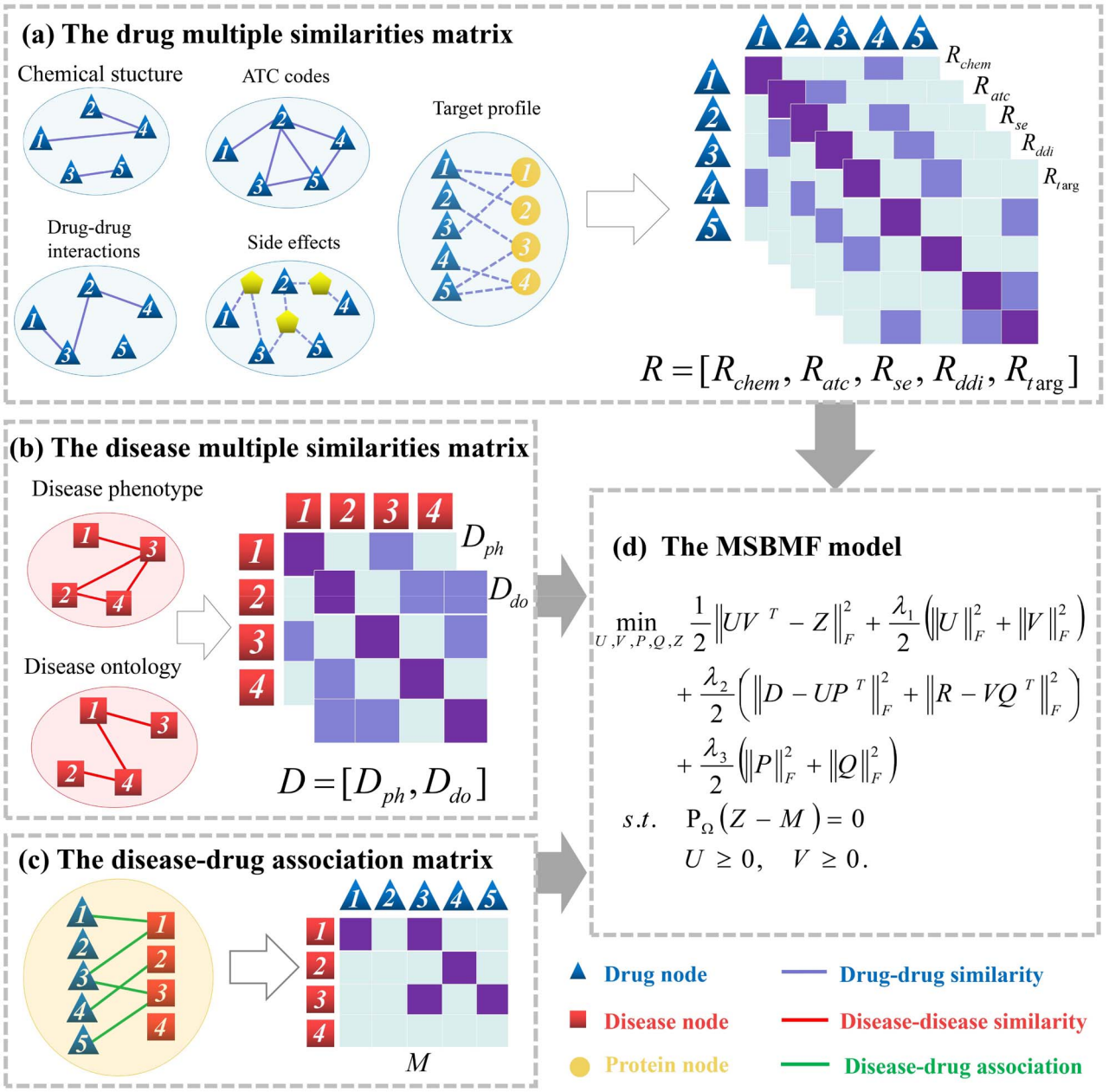

Figure 1. The overall workflow of MSBMF. (A) The drug multiple similarities matrix. (B) The disease multiple similarities matrix. (C) The drug–disease association matrix. (D) The model of MSBMF.

drugs with respect to their structural qualities, which are calculated using the chemical development kit (CDK) [41]. The procedure of deriving chemical structure similarity is described as follows. First of all, the Canonical SMILES [42] files of the related drugs are downloaded from DrugBank [6]. Then, we use the CDK tool to calculate the hashed fingerprints for all drugs with default parameters. Finally, the chemical structure similarity is reported by the Tanimoto similarity score.

ATC codes similarity R_{atc}. The World Health Organization ATC classification system is employed to represent drugs. This hierarchical classification system categorizes drugs according to their therapeutic effects and chemical characteristics on organs or systems. All ATC codes in this study are extracted from DrugBank. The more ATC codes two drugs share, we consider the more similar they are. A semantic similarity algorithm [43] is applied to compute the similarity score between ATC codes.

Side effects similarity R_{se}. Side effects of drugs are extracted from SIDER database [9]. A given drug is represented by a profile listing all known side effects. The similarity is computed using the Jaccard similarity coefficient [44] between the side effect profiles of a given pair of drugs. Specifically, we use the following formula to measure the side effects similarity between drugs i and j:

$$R_{se}(i,j) = \frac{|\,SE_i \cap SE_j\,|}{|\,SE_i \cup SE_j\,|},$$

where SE_i denotes the set of side-effects of drug i.

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

Drug–drug interactions similarity R_{ddi}. Drug–drug interactions are extracted from DrugBank. Each drug is represented as an interaction profile consisting of all drugs that are known to interact with the specific drug. Drug–drug interaction similarity is then computed according to the Jaccard scores of their drug–drug interactions profiles. Denoting DDI_i and DDI_j to represent the drug–drug interactions profiles of drug i and drug j, respectively, the drug–drug interactions similarity is defined as follows:

$$R_{ddi}(i,j) = \frac{|\ DDI_i \cap DDI_j\ |}{|\ DDI_i \cup DDI_j\ |}.$$

Target profile similarity R_{targ}. The drug–target information is extracted from DrugBank. Given a drug, its target profile includes all of its known associated targets. Similar to side effects similarity and drug–drug interactions similarity, the similarities based on drug–target interactions are calculated using the Jaccard scores of their target profiles. The target profile similarity between drug i and drug j is measured by

$$R_{targ}(i,j) = \frac{|\ T_i \cap T_j\ |}{|\ T_i \cup T_j\ |},$$

where T_i and T_j represent the target profiles of drugs i and j, respectively.

Disease similarity measures

The similarities between pairwise diseases used in MSBMF include phenotype similarity and ontology similarity.

Disease phenotype similarity D_{ph}. The disease phenotype similarities are collected from MimMiner [45] and then normalized to the range of $[0, 1]$. The similarity score between two diseases is computed according to the frequencies of the medical subject headings vocabulary in their medical descriptions, which are obtained from the OMIM database [12].

DO similarity D_{do}. DO is an important annotation in human genes, which can be used to describe relationships between diseases. The DO terms are organized in a directed acyclic graph and then the semantic similarity between two diseases is measured by their relative positions in the DAG. According to the structure of DO terms, we compute the disease ontology similarity using the gene ontology-based algorithm [46].

Methods

Matrix factorization

Matrix factorization is an effective numerical method used in applications of data representation, recommendation systems and document clustering [47–50], which intends to calculate an optimal approximation to the target matrix by decomposing it into two low-rank matrices. Generally, the mathematical model of matrix factorization is formulated as

$$\min_{U,V} \left\| UV^T - M \right\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $M \in \mathbb{R}^{n\times m}$ is the given matrix, $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$ are the latent feature matrices of M and r is the subspace dimensionality ($r \leq min(n, m)$). There are quite a few of optimization algorithms that can be used to solve the above matrix factorization model, including the alternating least squares (ALS) [51], multiplicative updating algorithm (Mult) [52] and the ADMM [53–55]. Because of their remarkable simplicity and less parameters to tune, ALS and Mult algorithms have been popularly applied to a wide variety of applications. However, compared with ALS and Mult, the classic ADMM algorithm is superior in terms of solution quality [56]. In this study, we adopt ADMM to solve our proposed matrix factorization model.

MSBMF for drug repositioning

Based on the assumption that similar drugs share the similar molecular pathways to treat similar diseases, the underlying latent factors determining drug–disease associations are highly correlated. In other words, the drug–disease matrix M is low rank. M can be split into two low-dimensional feature matrices, i.e. disease feature U and drug feature V. In practical applications, since the knowledge of negative associations are usually not available, M is extremely sparse, typically with less than 1% known associations, while the rest of the elements are unknown. So the error term is only evaluated on the entries with known associations. Meanwhile, Tikhonov regularization terms are often used to avoid overfitting. As a result, the baseline matrix factorization model in drug repositioning is formulated as

$$\min_{U,V} \frac{1}{2} \left\| \mathcal{P}_\Omega(UV^T - M) \right\|_F^2 + \frac{\lambda_1}{2}(\|U\|_F^2 + \|V\|_F^2), \tag{2}$$

where λ_1 is the harmonic parameter balancing the error term and the regularization terms, Ω is an index set of known associations in matrix M and $\mathcal{P}_\Omega$ is a projection operator onto Ω defined as

$$(\mathcal{P}_\Omega(X))_{ij} = \begin{cases} X_{ij}, & (i,j) \in \Omega \\ 0, & (i,j) \notin \Omega. \end{cases}$$

However, the objective function of the above baseline model does not involve a large amount of prior information of diseases and drugs, such as disease similarities and drug similarities. Given a disease similarity matrix D and a drug similarity matrix R, due to the fact that U and V can be considered as the matrices containing the latent feature vectors of diseases and drugs, UU^T and VV^T are expected to match D and R, respectively [29, 57]. Therefore, model (2) is extended to

$$\begin{aligned} \min_{U,V} \frac{1}{2} \left\| \mathcal{P}_\Omega(UV^T - M) \right\|_F^2 + \frac{\lambda_1}{2}(\|U\|_F^2 + \|V\|_F^2) \\ + \frac{\lambda_2}{2}(\left\| D - UU^T \right\|_F^2 + \left\| R - VV^T \right\|_F^2), \end{aligned} \tag{3}$$

by adding the disease and drug similarity terms. Model (3) deals with a single disease and drug similarity measure. Here, in order to incorporate multiple similarity measures, we propose a MSBMF model for drug repositioning, which is formulated as follows:

$$\begin{aligned} \min_{U,V,P,Q,Z} \frac{1}{2} \left\| UV^T - Z \right\|_F^2 + \frac{\lambda_1}{2}(\|U\|_F^2 + \|V\|_F^2) \\ + \frac{\lambda_2}{2}(\left\| D_m - UP^T \right\|_F^2 + \left\| R_m - VQ^T \right\|_F^2) \\ + \frac{\lambda_3}{2}(\|P\|_F^2 + \|Q\|_F^2) \\ \text{s.t.} \quad \mathcal{P}_\Omega(Z) = \mathcal{P}_\Omega(M) \\ U \geq 0, \ \ V \geq 0, \end{aligned} \tag{4}$$

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

where D_m and R_m are matrices concatenating multi-similarities measures of diseases and drugs, respectively, and λ_1, λ_2 and λ_3 are balancing parameters. In this work, we set $D_m = [D_{ph}, D_{do}]$ and $R_m = [R_{chem}, R_{atc}, R_{se}, R_{ddi}, R_{targ}]$. The approximations of the similarity matrix D_m and R_m are constructed using the feature matrices U and V as basis, where P and Q are matrices including latent features representing disease similarity and drug similarity, respectively. Z is an auxiliary matrix for facilitating mathematical optimization.

MSBMF is able to deal with multi-similarities matrices of drug and disease simultaneously during the optimization process, rather than fusing them into a comprehensive similarity matrix. It incorporates the multi-similarities and drug–disease associations together into the objective function. Moreover, MSBMF has two non-negative constraints, ensuring that the predicted entries of association matrix are non-negative for the biological interpretability. There is no specific requirement for ordering the similarities matrices.

We use the ADMM framework to solve model (4). By introducing two splitting matrices X and Y, (4) is transformed into

$$\begin{aligned}\min_{U,V,P,Q,X,Y,Z} & \frac{1}{2}\left\|UV^T - Z\right\|_F^2 + \frac{\lambda_1}{2}(\|U\|_F^2 + \|V\|_F^2) \\ & + \frac{\lambda_2}{2}(\left\|D_m - UP^T\right\|_F^2 + \left\|R_m - VQ^T\right\|_F^2) \\ & + \frac{\lambda_3}{2}(\|P\|_F^2 + \|Q\|_F^2) \\ s.t. \quad & \mathcal{P}_\Omega(Z) = \mathcal{P}_\Omega(M) \\ & U = X, V = Y \\ & X \geq 0, \quad Y \geq 0.\end{aligned} \tag{5}$$

The augmented Lagrangian function is given by

$$\begin{aligned}\mathcal{L}(U,V,P,Q,X,Y,Z) = & \frac{1}{2}\left\|UV^T - Z\right\|_F^2 + \frac{\lambda_1}{2}(\|U\|_F^2 + \|V\|_F^2) \\ & + \frac{\lambda_2}{2}(\left\|D_m - UP^T\right\|_F^2 + \left\|R_m - VQ^T\right\|_F^2) + \frac{\lambda_3}{2}(\|P\|_F^2 + \|Q\|_F^2) \\ & + <\Phi, U - X> + <\Psi, V - Y> + \frac{\mu}{2}(\|U - X\|_F^2 + \|V - Y\|_F^2),\end{aligned} \tag{6}$$

where Φ and Ψ are the Lagrange multipliers and μ is the penalty parameter. We minimize $\mathcal{L}(U, V, P, Q, X, Y, Z)$ by alternatively optimizing one variable while fixing the others. Therefore, at the kth iteration, U_{k+1}, V_{k+1}, P_{k+1}, Q_{k+1}, X_{k+1}, Y_{k+1} and Z_{k+1} need to be computed alternatively such that

$$\begin{aligned}U_{k+1} &= \arg\min_U \mathcal{L}(U, V_k, P_k, Q_k, X_k, Y_k, Z_k), \\ V_{k+1} &= \arg\min_V \mathcal{L}(U_{k+1}, V, P_k, Q_k, X_k, Y_k, Z_k), \\ P_{k+1} &= \arg\min_P \mathcal{L}(U_{k+1}, V_{k+1}, P, Q_k, X_k, Y_k, Z_k), \\ Q_{k+1} &= \arg\min_Q \mathcal{L}(U_{k+1}, V_{k+1}, P_{k+1}, Q, X_k, Y_k, Z_k), \\ X_{k+1} &= \arg\min_{X:X\geq 0} \mathcal{L}(U_{k+1}, V_{k+1}, P_{k+1}, Q_{k+1}, X, Y_k, Z_k), \\ Y_{k+1} &= \arg\min_{Y:Y\geq 0} \mathcal{L}(U_{k+1}, V_{k+1}, P_{k+1}, Q_{k+1}, X_{k+1}, Y, Z_k), \\ Z_{k+1} &= \arg\min_{Z:P_\Omega(Z)=P_\Omega(M)} \mathcal{L}(U_{k+1}, V_{k+1}, P_{k+1}, Q_{k+1}, X_{k+1}, Y_{k+1}, Z).\end{aligned} \tag{7}$$

The above optimization problems are regularized least squares problems, whose corresponding closed-form solutions are listed as follows:

$$\begin{aligned}U_{k+1} = (Z_k V_k + \lambda_2 D_m P_k - \Phi_k + \mu_k X_k)(V_k^T V_k \\ + \lambda_2 P_k^T P_k + (\lambda_1 + \mu_k)I)^{-1},\end{aligned} \tag{8a}$$

$$\begin{aligned}V_{k+1} = (Z_k^T U_{k+1} + \lambda_2 R_m Q_k - \Psi_k + \mu_k Y_k)(U_{k+1}^T U_{k+1} \\ + \lambda_2 Q_k^T Q_k + (\lambda_1 + \mu_k)I)^{-1},\end{aligned} \tag{8b}$$

$$P_{k+1} = \lambda_2 D_m^T U_{k+1}(\lambda_2 U_{k+1}^T U_{k+1} + \lambda_3 I)^{-1}, \tag{8c}$$

$$Q_{k+1} = \lambda_2 R_m^T V_{k+1}(\lambda_2 V_{k+1}^T V_{k+1} + \lambda_3 I)^{-1}, \tag{8d}$$

$$X_{k+1} = \mathcal{Q}_+(U_{k+1} + \frac{1}{\mu_k}\Phi_k), \tag{8e}$$

$$Y_{k+1} = \mathcal{Q}_+(V_{k+1} + \frac{1}{\mu_k}\Psi_k), \tag{8f}$$

$$Z_{k+1} = \mathcal{P}_\Omega(M) + \mathcal{P}_{\bar{\Omega}}(U_{k+1}V_{k+1}^T), \tag{8g}$$

where I denotes the identity matrix, $\mathcal{Q}_+$ is a non-negative projection defined as

$$(\mathcal{Q}_+(X))_{ij} = \begin{cases} X_{ij}, & X_{ij} > 0 \\ 0, & otherwise, \end{cases}$$

$\bar{\Omega}$ denotes a complement to Ω and μ_k is the learning rate at iteration k.

In summary, the overall iterative scheme of optimizing MSBMF model is presented in Algorithm 1. We adopt a scheme with gradually increasing learning rate to achieve fast convergence [58]. After performing the MSBMF algorithm, the to-be-complete disease–drug association matrix M becomes a non-negative matrix M^* with missing elements filled out as predicted scores. According to these predicted scores, the potential drug–disease pairs can be inferred.

Results and discussion

Performance evaluation

To evaluate the performance of MSBMF, we carry out two kinds of computational experiments, including a 10-fold cross-validation and *de novo* tests, to identifying potential indications for approved and novel drugs. For the 10-fold cross-validation, we randomly divide the existing drug–disease associations into ten mutually exclusive parts, which have approximately the same size. Each part is considered as the testing set in turn, whereas the remaining nine parts are used as the training set. We repeat the 10-fold cross-validation 10 times and report the average values as the final results. For *de novo* test, we select drugs with only one known disease association as the target

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

Algorithm 1: MSBMF Algorithm
Input: The disease–drug association matrix $M \in \mathbb{R}^{n\times m}$, the multiply similarities of disease matrices: $D_m \in \mathbb{R}^{n\times 2n}$, the multiply similarities of drug matrices: $R_m \in \mathbb{R}^{m\times 5m}$, subspace dimensionality r, parameters λ_1, λ_2 and λ_3.
Output: Predicted association matrix M^*.
initialize randomly four non-negative matrices: $U_0 \in \mathbb{R}^{n\times r}$, $V_0 \in \mathbb{R}^{m\times r}$, $P_0 \in \mathbb{R}^{2n\times r}$ and $Q_0 \in \mathbb{R}^{5m\times r}$; $X_0 = U_0$, $Y_0 = V_0$, $Z_0 = M$, $\Phi_0 = \mathbf{0}$, $\Psi_0 = \mathbf{0}$, μ_0, μ_{max} and rate changing factor $\rho > 1$;
$k \leftarrow 0$;
repeat
compute $U_{k+1}, V_{k+1}, P_{k+1}, Q_{k+1}, X_{k+1}, Y_{k+1}$ and Z_{k+1} by (8);
update the multipliers by
$\Phi_{k+1} \leftarrow \Phi_k + \mu_k(U_{k+1} - X_{k+1})$;
$\Psi_{k+1} \leftarrow \Psi_k + \mu_k(V_{k+1} - Y_{k+1})$;
update μ_{k+1} by $\mu_{k+1} \leftarrow min\{\rho\mu_k, \mu_{max}\}$;
$k \leftarrow k + 1$;
until convergence
$M^* \leftarrow X_{k+1}Y_{k+1}^T$;
return M^*.

drugs and then remove the known drug–disease association of each target drug in turn. As a result, these drugs are treated as novel drugs in *de novo* experiment, where the capabilities of the drug repositioning algorithms in handling the cold-start cases are tested. In the 10-fold cross-validation, the five drug similarities are employed. However, in *de novo*, we remove R_{se} and R_{ddi} from the prior drug similarities. Because a new drug may not have complete data (e.g. no side effect information). According to the predicted scores of the unknown drug–disease pairs, we rank all candidate indications associated with the test drug in a descending order. We use three evaluation metrics to assess the performance of our MSBMF and other methods in comparison, including the area under the receiver operating characteristic (ROC) curve (AUC), the area under the precision-recall curve (AUPR) and precision.

Parameters tuning and setting

In MSBMF algorithm, the tunable parameters include the latent dimension r and the three coefficient λ_1, λ_2 and λ_3. We set $r = [\tau \min(m, n)]$, where $\tau \in [0, 1]$ and $[\cdot]$ denotes the rounding function. For preventing overfitting due to many parameters, we set λ_2 and λ_3 to the same value and remove one parameter. Because they are used to penalize the related terms of P and Q in model (4). Finally, three parameters need to be determined, including τ, λ_1 and λ_2.

While exhaustively searching the parameter space is computationally costly, we adopt a 'fixing one and determining the others' strategy. For the gold standard dataset, we first set τ to 0.5 and then pick the values of λ_1 and λ_2 from $\{0.001, 0.01, 0.1, 1\}$ by cross-validation. Then, we fix the determined values of the coefficients and select τ from $\{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. The computational results for determining the coefficients λ_1 and λ_2 are listed in Table 1. One can find that the sum of AUC and AUPR values reaches maximum when $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$. As shown in Table 2, MSBMF consistently yields approximately the same good performance when $\tau \geq 0.7$, indicating a low-rank matrix can effectively approximate the original association matrix. Accordingly, we empirically set $\tau = 0.7$.

Table 1. The sum of AUC and AUPR values using different λ_1 and λ_2 values in the 10-fold cross-validation

λ_1 λ_2	0.001	0.01	0.1	1
0.001	1.070	1.070	0.891	0.649
0.01	1.173	1.221	1.114	0.741
0.1	1.315	**1.367**	1.273	0.940
1	1.208	1.211	1.209	1.153

Table 2. The sum of AUC and AUPR values using different τ values with fixing $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$

τ	0.1	0.3	0.5	0.7	0.9	1
AUC+AUPR	1.153	1.316	1.367	1.393	1.400	1.398

The stopping criteria of MSBMF algorithm are

$$f_k \leq tol_1 \quad \text{and} \quad \frac{|f_{k+1} - f_k|}{max\{1, |f_k|\}} \leq tol_2,$$

where $f_k = \frac{\|X_{k+1}Y_{k+1} - X_kY_k\|_F}{\|X_kY_k\|_F}$, and tol_1 and tol_2 are the given tolerances. Here, we set $tol_1 = 2 \times 10^{-3}$ and $tol_2 = 10^{-4}$.

Comparison with state-of-the-art drug repositioning methods

We compare MSBMF with five state-of-the-art drug repositioning approaches, including BNNR [26], DRRS [25], MBiRW [18], DrugNet [20] and HGBI [33]. The parameters in these methods are set to either the optimal values by the grid searching (BNNR: α and β are chosen from $\{0.01, 0.1, 1, 10\}$; DrugNet: α is picked from $\{0.1, 0.2, ..., 0.9\}$) or the recommended values by the authors (DRRS: τ and δ are adaptive parameters; MBiRW: $\alpha = 0.3, l = r = 2$; HGBI: $\alpha = 0.4$). According to the prior similarity information mentioned in their own literature, the drug chemical structure similarity and the disease phenotype similarity are used to feed these above compared models.

Based on the multi-view learning framework, Wang *et al.* [59] proposed a similarity network fusion (SNF) method. SNF uses K nearest neighbors and cross-diffusion process to merge various similarity networks into a single fused similarity network, which can capture the common feature information from different similarities. For fair and comprehensive comparison on multiple similarities, we also use the SNF algorithm to integrate the five drug similarities for drug measures and the two disease similarities for disease measures in the above methods for comparison. These corresponding methods are named BNNR-SNF, DRRS-SNF, MBiRW-SNF, DrugNet-SNF and HGBI-SNF, correspondingly. The hyperparameters in SNF algorithm (i.e. K nearest neighbours and iteration steps t) are set to 10 empirically.

We evaluate the performance of all methods in 10-fold cross-validation and *de novo* tests. Table 3 reports AUC, AUPR and precision values of all compared methods in these two experiments on the gold standard dataset. As shown in Table 3, MSBMF outperforms the other approaches in 10-fold cross-validation in terms of AUC, AUPR and precision values. In the 10-fold cross-validation, MSBMF obtains the best AUC, AUPR and precision values of 0.941, 0.421 and 0.455, which are 0.967%, 4.726% and 3.409% higher than BNNR, the method with second best performance, respectively. Compared to BNNR in *de novo*, MSBMF achieves AUC, AUPR and precision values of 0.875, 0.301 and

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

Table 3. AUC, AUPR and precision values of all compared methods in 10-fold cross-validation and *de novo* tests on the gold standard dataset

Tests	Metrics	**MSBMF**	BNNR(-SNF)	DRRS(-SNF)	MBiRW(-SNF)	DrugNet(-SNF)	HGBI(-SNF)
10-fold CV	AUC	**0.941**	0.932(0.919)	0.930 (0.916)	0.917 (0.881)	0.868 (0.908)	0.829 (0.923)
	AUPR	**0.421**	0.402(0.385)	0.341 (0.361)	0.264 (0.240)	0.155 (0.166)	0.102 (0.305)
	Precision	**0.455**	0.440(0.425)	0.375 (0.400)	0.304 (0.290)	0.192 (0.200)	0.130 (0.348)
de novo	AUC	0.875	0.830(0.831)	0.824(0.813)	0.818(0.795)	0.782(**0.881**)	0.746(0.873)
	AUPR	0.301	0.199(0.291)	0.197(0.318)	0.189(0.149)	0.102(0.281)	0.075(**0.336**)
	Precision	0.368	0.251(0.345)	0.269(0.386)	0.234(0.170)	0.135(0.339)	0.099(**0.398**)

The best results are highlighted in bold and the second best results are underlined.

0.368, respectively. They are 5.421%, 51.256% and 46.614% higher than BNNR. The improvement is actually significant. Actually, DrugNet-SNF and MSBMF obtain the best and second best AUC values in *de novo*, but the AUPR and precision values of MSBMF are 7.117% and 8.555% higher than DrugNet-SNF.

In general, when the SNF algorithm is used to fuse multiple similarities of drugs and diseases, most prediction results from the compared methods are improved compared to the versions without using fused similarities. This indicates the effectiveness of using multiple similarities compared to individual similarities. The only exception is MBiRW, where the performance of MBiRW-SNF downgrades. This is because the comprehensive similarity scheme in MBiRW changes the structure of the fused network, and thus the random walk algorithm has difficulty to correctly infer some potential associations when the fused similarity matrices are employed. We applied SNF technique to the MBiRW algorithm without the similarity scheme, denoted as BiRW-SNF. The 10-fold cross-validation and *de novo* tests are conducted by BiRW-SNF algorithm. The corresponding AUC, AUPR and precision are 0.923, 0.306 and 0.348 in 10-fold cross-validation, while 0.857, 0.353 and 0.415 in *de novo*, respectively. Comparing with the prediction results of MBiRW and MBiRW-SNF in Table 3, we find the performance of BiRW-SNF is better than those of MBiRW and MBiRW-SNF. It shows that SNF can improve the MBiRW algorithm without the similarity preprocessing, which confirms our assertion.

After all, MSBMF outperforms the methods using SNF-fused similarity matrices in the 10-fold cross-validation and has an excellent performance in *de novo*, showing that dynamically incorporating multiple drug and disease similarity information into the optimization process yields significantly better effectiveness than fusing multiple similarity information into a single similarity matrix in preprocessing. In addition, the compared methods only do well in one of the two experiments. Specifically, BNNR and DRRS have good performance in the 10-fold cross-validation, and the other two approaches, DrugNet-SNF and HGBI-SNF, are suitable for *de novo* experiment. However, our proposed MSBMF can perform excellent under the two experimental settings. Figures 2 and 3 show the ROC and PR curves of the 10-fold cross-validation and *de novo* tests, respectively.

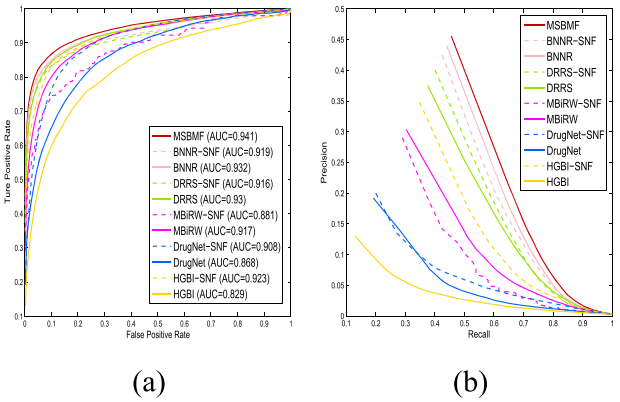

Figure 2. Prediction results of all approaches in the 10-fold cross-validation for the gold standard dataset. (A) ROC curves and AUC values. (B) Precision-recall curves.

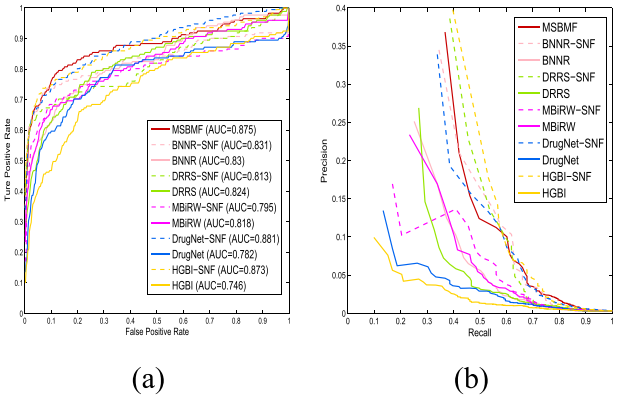

Figure 3. Prediction results of all approaches in *de novo* for the gold standard dataset. (A) ROC curves and AUC values. (B) Precision-recall curves.

The sensitivity analysis of parameters

In this section, we focus on the sensitivity analysis for three parameters, i.e. λ_1, λ_2 and τ, in the 10-fold cross-validation. When $\lambda_1 = 0.1$, $\lambda_2 = 0.01$ and $\tau = 0.7$, MSBMF can yield excellent performance. We vary one parameter while keeping the rest two parameters fixed to study how the parameter benefits the AUC and AUPR values.

Figure 4 shows the performance trend of MSBMF with different values of λ_1. We can find the AUC and AUPR can achieve the best values on $\lambda_1 = 0.1$. When the value of λ_1 reduces from 0.1, the values of two measures drop rapidly. It indicates that the regularization terms (i.e. $\|U\|_F$ and $\|V\|_F$) of model (4) should not be weakened too much. Similarly, Figure 5 shows the performance trend of MSBMF with different values of λ_2. As shown in Figure 5, the best AUC and AUPR present on $\lambda_2 = 0.01$. When λ_2 increases to 1 from 0.01, the values of AUC and AUPR keep going down. It demonstrates that the error terms and the regularization terms of P and Q in model (4) should not be strengthened.

Finally, the effect of parameter τ on the prediction accuracy is discussed. Figure 6 shows the AUC and AUPR trends of MSBMF with various τ. As the value of τ increases, these two indications gradually increase. When $\tau > 0.7$, the trends of AUC and

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

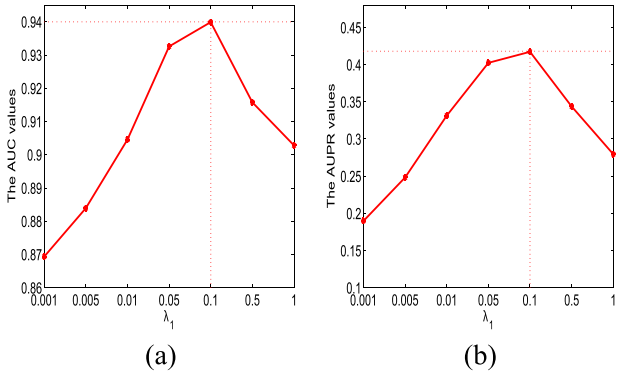

Figure 4. (A) Variation of the AUC values with the different settings of λ_1. (B) Variation of the AUPR values with the different settings of λ_1.

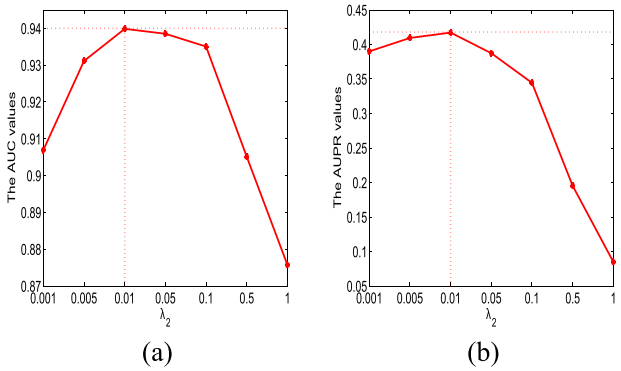

Figure 5. (A) Variation of the AUC values with the different settings of λ_2. (B) Variation of the AUPR values with the different settings of λ_2.

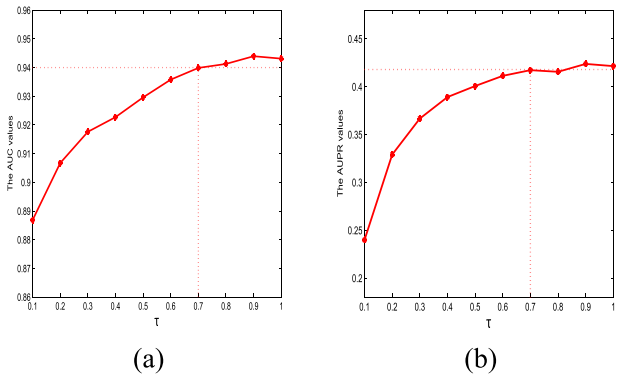

Figure 6. (A) Variation of the AUC values with the different settings of τ. (B) Variation of the AUPR values with the different settings of τ.

AUPR are levelling off. Actually, it demonstrates that the low-rank structure of drug–disease matrix can be approximated well when $\tau = 0.7$. If τ continue to increase to 0.9 or 1, it will not only generate overfitting but also increase the computational complexity.

Analysis of single similarity and combined multi-similarities in MSBMF

In order to analyze the effects among individual similarities and integrated multi-similarities in MSBMF model clearly, we focus on analyzing the drug multiple similarities while fixing the two disease similarities. For a single drug similarity, we iteratively set

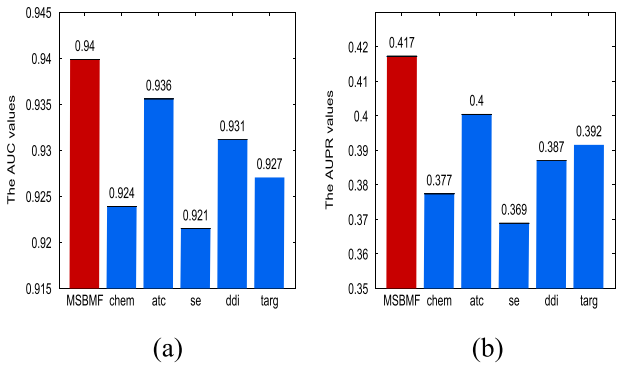

Figure 7. The performance of single similarity and multi-similarities in 10-fold cross-validation. (A) The AUC comparison. (B) The AUPR comparison. Integrating multiple similarities in MSBMF yields more accurate predictions than the models using single similarity.

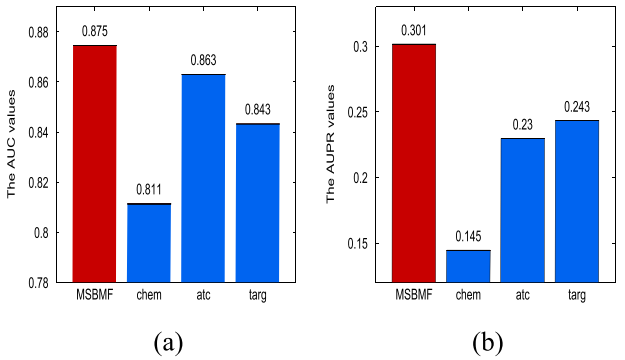

Figure 8. The performance of single similarity and multi-similarities in *de novo* experiment. (A) The AUC comparison. (B) The AUPR comparison. Integrating multiple similarities in MSBMF yields more accurate predictions than the models using single similarity.

the matrix R_m to one of the drug similarity matrices (R_{chem}, R_{atc}, R_{se}, R_{ddi} and R_{targ}) in Algorithm 1.

We first compare the performance of MSBMF model incorporating all 5 drug similarities with those using single similarity in the 10-fold cross-validation. As shown in Figure 7, for individual similarities, R_{atc} is most effective in the AUC and AUPR values. Meanwhile, in *de novo* experiment, we only concatenate R_{chem}, R_{atc} and R_{targ} as a comprehensive similarity matrix. As shown in Figure 8, R_{targ} is most effective in the AUPR values. This is because when no disease association for a novel drug available, the interaction information in R_{targ} starts to play an more important role than the others in predictions. Nevertheless, MSBMF with multi-similarities yields higher AUC and AUPR values than MSBMF with single drug similarity. This confirms that integrating multiple similarities can effectively improve prediction accuracy in our proposed method.

Moreover, we further analyze the effect of the number of concatenated similarity matrices to the prediction capability of MSBMF. For multi-similarities experiments, we consider all possible permutations of these drug similarities. For instance, MSBMF with 2-similarities (denoted MSBMF-2) include 20 permutations such as $[R_{chem}, R_{atc}]$, $[R_{atc}, R_{chem}]$, $[R_{chem}, R_{se}]$, $[R_{se}, R_{chem}]$..., etc. Similarly, MSBMF using 3, 4 and 5 drug similarities have 60, 120 and 120 permutations, respectively. The box plots in Figure 9 display the AUC and AUPR results of MSBMF-1, MSBMF-2, MSBMF-3, MSBMF-4 and MSBMF-5 in the 10-fold cross-validation, where the blue star denotes the average value

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

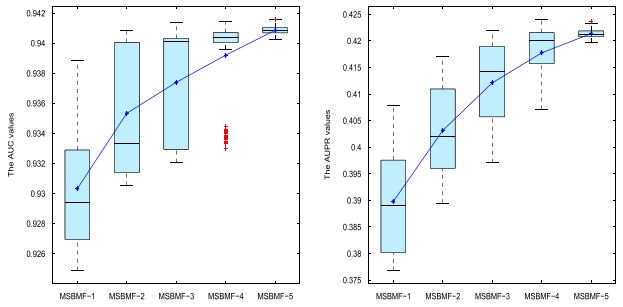

Figure 9. The box plots of all combined multi-similarities in the 10-fold cross-validation. (A) The statistics of AUC results. (B) The statistics of AUPR results. More similarity matrices continuously improve the prediction accuracy of MSBMF.

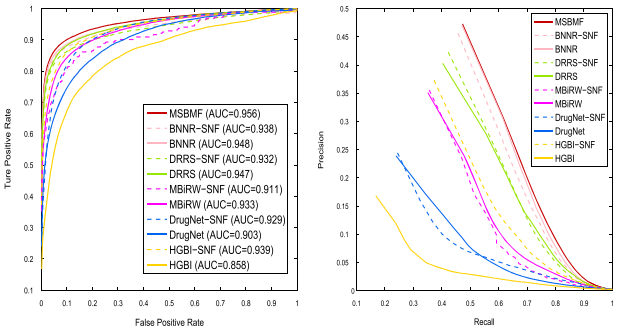

Figure 10. The prediction results of all approaches in the 10-fold cross-validation for Cdataset. (A) ROC curves and AUC values. (B) Precision-recall curves.

of each combined multi-similarities and the black line in blue box represents the corresponding median value. When more similarity matrices are involved, the AUC and AUPR values continue to improve, indicating that the additional similarity measures, even though not performing as well individually, contribute. This is due to the fact that, during training, MSBMF adaptively learns effective latent features from multiple similarities. When additional similarity measures are incorporated, MSBMF is capable of extracting useful features that can contribute to optimizing the drug–disease association matrix. Figure 9 also indicates that the order of similarity matrices in concatenation has little impact on the prediction results.

Case studies

To validate the capability of MSBMF in practical applications, we conduct case studies in identifying novel indications for approved drugs. On the gold standard dataset, we use MSBMF to predict the unknown drug–disease associations with all existing associations and multiple similarities. For each drug, the candidate indications are ranked according to the predicted scores in a descending order. In recent years, the development of drugs to treat neoplastic and psychotic diseases has attracted great attention from researchers. Here, we select four common antineoplastic drugs (doxorubicin, gemcitabine, vincristine and methotrexate) and an antipsychotic drug (risperidone) to conduct case studies and retrieve the evidences for the candidate indications in CTD database. As shown in Table 4, the top 10 candidate indications for these five drugs by MSBMF algorithm are listed and the confirmed ones by CTD database are highlighted in bold. Besides, we have also used five other methods to predict the corresponding top 10 candidate indications, including BNNR, DRRS, DrugNet, MBiRW and HGBI algorithms. These prediction results are shown in Supplementary Table S1–S5. Comparing Table 4 and Supplementary Tables S1–S5, for all indications for these five drugs, we count the retrieved total number of MSBMF, BNNR, DRRS, DrugNet, MBiRW and HGBI methods in CTD database, which are 24, 19, 19, 19, 10 and 3, respectively. More specifically, taking doxorubicin as an example, we observe that MSBMF retrieves five indications, the other methods identify only 0–3 indications. In particular, Kaposi Sarcoma, Susceptibility to (148000) and Hepatocellular Carcinoma (114550) are not retrieved by the other methods. The results in this case study indicate that MSBMF is more effective in predicting potential indications for marked drugs than the other methods in comparison.

Moreover, risperidone is an antipsychotic drug, which mainly used to treat schizophrenia, schizoaffective disorder and encephalopathy. Our algorithm identifies three related diseases in top 10 candidates, including Alcohol Dependence (103780), Insensitivity to Pain with Hyperplastic Myelinopathy (147530) and Migraine with or without Aura, Susceptibility to, 1 (157300). They have not been confirmed and may be promising indications and worth further research. In addition, we computed and collected the drug–disease pairs with scores greater than 0.95 in Supplementary Table S6, which are unconfirmed in CTD database. We believe them will give more useful references for medical researchers.

Experiments on the other datasets

To demonstrate the flexibility and reliability of MSBMF method, we carry out MSBMF on the other two datasets, i.e. Cdatatset [18] and a new larger dataset. Cdataset contains 663 drugs obtained from DrugBank, 409 diseases collected in OMIM database and 2352 known drug–disease associations. Integrating Cdataset and the CTD database (February 2020 release), we collects a new larger drug–disease dataset, named Ydataset. It includes 1478 drugs, 655 diseases and 8448 validated drug–disease associations. The multiple drug and disease similarities are calculated in the same way as described in Section 3.2 and Section 3.3. We evaluate the performance of MSBMF on Cdataset and Ydataset by conducting 10-fold cross-validation and *de novo* tests.

Table 5 shows AUC, AUPR and precision values of all compared methods in 10-fold cross-validation and *de novo* tests on Cdataset. As shown in Table 5, MSBMF achieves the best performance in terms of AUC, AUPR and precision compared with the other methods. Specifically, MSBMF obtains the AUC, AUPR and precision values of 0.956, 0.446 and 0.473 in the 10-fold cross-validation and 0.883, 0.298 and 0.367 in *de novo*, respectively. Compared with MSBMF, BNNR is the 2nd best method in the 10-fold cross-validation, but in *de novo*, the AUC, AUPR and precision of BNNR is lower 8.744%, 54.404% and 44.488% than that of our method. Moreover, HGBI-SNF is the second best approach in *de novo*, but in the 10-fold cross-validation, the AUC, AUPR and precision of HGBI-SNF is lower 1.810%, 34.743% and 26.810% than that of MSBMF. In addition, the ROC and precision-recall curves of these two experiments are shown in Figures 10 and 11.

Table 6 gives the comparison results of all methods on Ydataset. As shown in Table 6, again, MSBMF achieves the best

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

Table 4. Top 10 candidate indications for doxorubicin, gemcitabine, vincristine, methotrexate and risperidone

Drugs (DrugBank IDs)	Top 10 candidate diseases (OMIM IDs)
Doxorubicin (DB00997)	**Esophageal Cancer (133239)**; Reticulum Cell Sarcoma (267730); Testicular Germ Cell Tumor (273300); Leukemia, Chronic Lymphocytic, Susceptibility to, 2 (109543); Small Cell Cancer of The Lung (182280); **Kaposi Sarcoma, Susceptibility to (148000)**; **Colorectal Cancer (114500)**; **Hepatocellular Carcinoma (114550)**; Dohle Bodies and Leukemia (223350); **Renal Cell Carcinoma, Nonpapillary (144700)**.
Gemcitabine (DB00441)	Mismatch Repair Cancer Syndrome (276300); **Prostate Cancer (176807)**; Gastric Cancer, Hereditary Diffuse (137215); **Multiple Myeloma (254500)**; **Colorectal Cancer (114500)**; **Hepatocellular Carcinoma (114550)**; Melanoma, Cutaneous Malignant, Susceptibility to, 1 (155600); **Classic Hodgkin Lymphoma (236000)**; Thrombocythemia 1 (187950); Miller-Dieker Lissencephaly Syndrome (247200).
Vincristine (DB00541)	Testicular Germ Cell Tumor (273300); **Small Cell Cancer of The Lung (182280)**; **Breast Cancer (114480)**; **Kaposi Sarcoma, Susceptibility to (148000)**; Leukemia, Chronic Lymphocytic (151400); **Bladder Cancer (109800)**; **Osteogenic Sarcoma (259500)**; **Colorectal Cancer (114500)**; Gastric Cancer, Hereditary Diffuse (137215); **Lung Cancer (211980)**.
Methotrexate (DB00563)	Neuroblastoma (256700); Thrombocytopenic Purpura, Autoimmune (188030); **Multiple Myeloma (254500)**; **Prostate Cancer (176807)**; Wilms Tumor 1 (194070); Sarcoidosis, Susceptibility to, 1 (181000); **Lung Cancer (211980)**; **Leukemia, Chronic Lymphocytic (151400)**; Spastic Paraplegia and Evans Syndrome (601608); Multiple Sclerosis, Susceptibility to (126200).
Risperidone (DB00734)	**Obsessive-Compulsive Disorder (164230)**; Hyperthermia, Cutaneous, with Headaches and Nausea (145590); Alcohol Dependence (103780); Dementia, Lewy Body (127750); Camurati-Engelmann Disease (131300); **Panic Disorder 1 (167870)**; **Attention Deficit-Hyperactivity Disorder (143465)**; Insensitivity to Pain with Hyperplastic Myelinopathy (147530); Migraine with or without Aura, Susceptibility to, 1 (157300); Restless Legs Syndrome, Susceptibility to, 1 (102300).

The predicted indications in bold have been confirmed by CTD database.

Table 5. AUC, AUPR and precision values of all compared methods in 10-fold cross-validation and *de novo* tests on Cdataset

Tests	Metrics	**MSBMF**	BNNR(-SNF)	DRRS(-SNF)	MBiRW(-SNF)	DrugNet(-SNF)	HGBI(-SNF)
10-fold CV	AUC	**0.956**	0.948(0.939)	0.947(0.932)	0.933(0.911)	0.903(0.929)	0.858(0.939)
	AUPR	**0.446**	0.441(0.424)	0.378(0.390)	0.310(0.306)	0.201(0.202)	0.129(0.331)
	Precision	**0.473**	0.471(0.457)	0.403(0.423)	0.351(0.356)	0.239(0.245)	0.168(0.373)
de novo	AUC	**0.883**	0.812(0.841)	0.819(0.792)	0.804(0.796)	0.785(**0.883**)	0.732(0.872)
	AUPR	**0.298**	0.193(0.260)	0.181(0.222)	0.174(0.106)	0.106(0.242)	0.075(0.262)
	Precision	**0.367**	0.254(0.299)	0.243(0.260)	0.232(0.119)	0.147(0.288)	0.107(0.328)

The best results are highlighted in bold and the second best results are underlined.

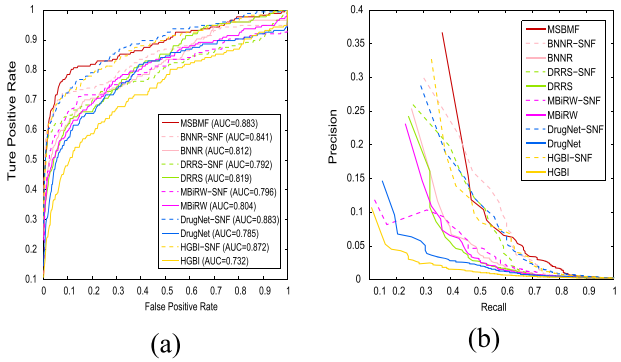

Figure 11. The prediction results of all approaches in *de novo* for Cdataset. (A) ROC curves and the corresponding AUC values. (B) Precision-recall curves.

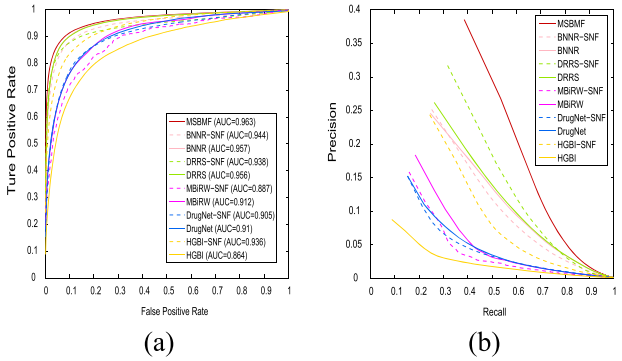

Figure 12. The prediction results of all approaches in the 10-fold cross-validation for Ydataset. (A) ROC curves and AUC values. (B) Precision-recall curves.

results in terms of AUC, AUPR and precision in the 10-fold cross-validation, which are 0.963, 0.360 and 0.386, respectively. In *de novo*, MSBMF achieves the best AUC value compared to other methods, but the AUPR and precision values of MSBMF slightly inferior to HGBI-SNF. The ROC and precision-recall curves of these two experiments are shown in Figures 12 and 13. In summary, the prediction results on Cdataset and Ydataset illustrate that our method exhibits good generality on the other datasets.

Conclusions

In this article, we have presented MSBMF, a new computational method of integrating multiple similarity measures of drugs

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

Table 6. AUC, AUPR and precision values of all compared methods in 10-fold cross-validation and *de novo* tests on Ydataset

Tests	Metrics	**MSBMF**	BNNR(-SNF)	DRRS(-SNF)	MBiRW(-SNF)	DrugNet(-SNF)	HGBI(-SNF)
10-fold CV	AUC	**0.963**	0.957(0.944)	0.956(0.938)	0.912(0.887)	0.910(0.905)	0.864(0.936)
	AUPR	**0.360**	0.229(0.225)	0.241(0.287)	0.145(0.121)	0.122(0.120)	0.069(0.206)
	Precision	**0.386**	0.246(0.252)	0.262(0.317)	0.184(0.159)	0.153(0.153)	0.088(0.244)
de novo	AUC	**0.927**	0.897(0.906)	0.887(0.706)	0.878(0.854)	0.845(0.887)	0.733(0.906)
	AUPR	0.167	0.107(0.132)	0.103(0.124)	0.105(0.037)	0.072(0.131)	0.047(**0.172**)
	Precision	0.203	0.129(0.167)	0.121(0.169)	0.146(0.048)	0.088(0.165)	0.069(**0.223**)

The best results are highlighted in bold and the second best results are underlined.

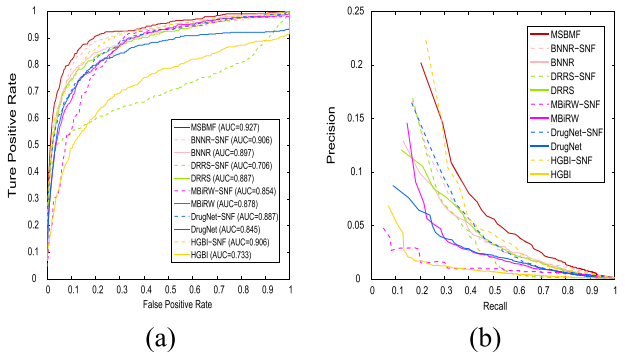

Figure 13. The prediction results of all approaches in *de novo* for Ydataset. (A) ROC curves and the corresponding AUC values. (B) Precision-recall curves.

and diseases for drug repositioning. MSBMF provides an effective scheme for dynamically integrating multiple similarities and extracting useful features to infer potential drug–disease associations. The fusion of multiple similarities and completing the drug–disease association matrix are carried out at the same time in MSBMF training. The non-negative constraint in MSBMF also ensures that the predicted scores of associations are non-negative. Our computational results have shown that MSBMF yields better performance in both cross-validation and *de novo* tests compared with the other state-of-the-art methods. When more similarity measures for drugs or diseases calculated from different aspects are involved, the performance of MSBMF continues to improve. Moreover, our case studies demonstrate that MSBMF is a powerful tool that can be effectively applied to practical applications. However, our proposed method has two potential limitations. First, MSBMF involves non-convex optimization, which often leads to the local optimal solutions instead of the global optimal solution. Second, in the process of matrix factorization, the rank of decomposition matrix must be artificially estimated in advance, which cannot be set adaptively according to the characteristics of real data. In the future, we plan to design a comprehensive algorithm integrating multiple similarities and multi-layers networks for predicting potential drug–disease associations.

Key Points

- We propose a multi-similarities bilinear matrix factorization (MSBMF) method to dynamically integrate multiple similarities between pairwise drugs and pairwise diseases into drug–disease association training.
- MSBMF model incorporates two non-negative constraints to ensure that the entries of predicted association matrix are non-negative, which can be interpreted.
- An iterative computational method based on the alternating direction method of multipliers is developed to effectively solve the MSBMF model.

Funding

National Natural Science Foundation of China (grant no. 61972423); the Graduate Research Innovation Project of Hunan (grant no. CX20190125); Hunan Provincial Science and Technology Program (no. 2018wk4001); 111Project (no. B18059).

References

1. Chong CR, Sullivan DJ. New uses for old drugs. *Nature* 2007;**448**(7154):645–6.
2. Tamimi NAM, Ellis P. Drug development: from concept to marketing! *Nephron Clin Pract* 2009;**113**(3):c125–31.
3. Novac N. Challenges and opportunities of drug repositioning. *Trends Pharmacol Sci* 2013;**34**(5):267–72.
4. Pushpakom S, Iorio F, Eyers PA, *et al*. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;**18**:41–58.
5. Luo H, Li M, Yang M, *et al*. Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief Bioinform* 2020. doi: 10.1093/bib/bbz176.
6. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**(Database issue):D668–72.
7. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem Substance and Compound Databases. *Nucleic Acids Res* 2016;**44**(D1): D1202–13.
8. Davis AP, Grondin CJ, Johnson RJ, *et al*. The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res* 2013;**41**(Database issue):D1104–14.
9. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**(D1):D1075–9.
10. Kibbe WA, Arze C, Felix V, *et al*. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015;**43**(D1):D1071–8.
11. Rappaport N, Twik M, Plaschkes I, *et al*. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* 2017;**45**(D1):D877–87.
12. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

genes and genetic disorders. *Nucleic Acids Res* 2002;**30**(1): 52–5.
13. Piñero J, Bravo À, Queralt-Rosinach N, *et al*. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**(D1):D833–9.
14. Apweiler R, Bairoch A, Wu CH, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**(Database issue):D115–9.
15. Stark C, Breitkreutz BJ, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006; **34**(Database issue):D535–9.
16. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database–2009 update. *Nucleic Acids Res* 2009;**37**(Database issue):D767–72.
17. Rose PW, Prlić A, Altunkaya A, *et al*. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* 2017;**45**(D1):D271–81.
18. Luo H, Wang J, Li M, *et al*. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 2016;**32**(17):2664–71.
19. Wang W, Yang S, Zhang X, Li J. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 2014;**30**(20):2923–30.
20. Martínez V, Navarro C, Cano C, *et al*. DrugNet: network-based drug-disease prioritization by integrating heterogeneous data. *Artif Intell Med* 2015;**63**(1):41–9.
21. Yu L, Huang J, Ma Z, *et al*. Inferring drug-disease associations based on known protein complexes. *BMC Med Genomics* 2015;**8**(Suppl 2):S2.
22. Napolitano F, Zhao Y, Moreira VM, *et al*. Drug repositioning: a machine-learning approach through data integration. *J Chem* 2013;**5**(1):30.
23. Liang X, Zhang P, Yan L, *et al*. LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning. *Bioinformatics* 2017;**33**(8):1187–96.
24. Saberian N, Peyvandipour A, Donato M, *et al*. A new computational drug repurposing method using established disease-drug pair knowledge. *Bioinformatics* 2019;**35**(19):3672–8.
25. Luo H, Li M, Wang S, *et al*. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018;**34**(11):1904–12.
26. Yang M, Luo H, Li Y, Wang J. Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics (ISMB/ECCB 2019)* 2019;**35**(14):i455–63.
27. Yang M, Luo H, Li Y, *et al*. Overlap matrix completion for predicting drug-associated indications. *PLoS Comput Biol* 2019;**15**(12):e1007541.
28. Dai W, Liu X, Gao Y, *et al*. (2015) Matrix factorization-based prediction of novel drug indications by integrating genomic space *Comput Math Methods Med*, **2015**(2015), 275045.
29. Cui Z, Gao Y-L, Liu J-X, *et al*. The computational prediction of drug-disease interactions using the dual-network $L_{2,1}$-CMF method. *BMC Bioinformatic* 2019;**20**(1):5.
30. Xuan P, Cao Y, Zhang T, *et al*. Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics* 2019;**35**(20):4108–19.
31. Zhang W, Yue X, Lin W, *et al*. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* 2018;**19**(1):233.
32. Cai JF, Candes EJ, Shen Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 2008;**20**(4):1956–82.
33. Wang W, Yang S, Li J. Drug target predictions based on heterogeneous graph inference. *Pac Symp Biocomput* 2013;**18**:53–64.
34. Vanunu O, Magger O, Ruppin E, *et al*. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010;**6**(1):e1000641.
35. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 2012;**9**(5):471–2.
36. Martínez V, Cano C, Blanco A. ProphNet: a generic prioritization method through propagation of information. *BMC Bioinformatics* 2014;**15**(Suppl 1):S5.
37. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin Pharmacol Ther* 2009;**86**(5):507–10.
38. Barabási AL, Gulbahce N, Loscalzo J, *et al*. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**(1):56–68.
39. Chen Y, Jiang T, Jiang R. Uncover disease genes by maximizing information flow in the phenome-interactome network. *Bioinformatics* 2011;**27**(13):i167–76.
40. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**(1):496.
41. Steinbeck C, Han Y, Kuhn S, *et al*. The chemistry development kit (CDK): an open-source java library for chemo-and bioinformatics. *Chem* 2003;**34**(21):493–500.
42. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**(1):31–6.
43. Resnik P. (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI '95: Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada.
44. Jaccard P. Nouvelles recheres Sur la distribution florale. *Bull Soc Vaud Sci Nat* 1908;**44**:223–70.
45. van Driel MA, Bruggeman J, Vriend G, *et al*. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**(5):535–42.
46. Wang JZ, Du Z, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;**23**(10):1274–81.
47. Huang DS, Zheng CH. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 2006;**22**(15):1855–62.
48. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. *Computer* 2009;**42**(8):30–7.
49. Hosoda K, Watanabe M, Wersing H, *et al*. A model for learning topographically organized parts-based representations of objects in visual cortex: topographic nonnegative matrix factorization. *Neural Comput* 2009;**21**(9):2605–33.
50. Huang X, Zheng X, Yuan W, *et al*. Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization. *Inform Sci* 2011;**181**(11):2293–302.
51. Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environ* 1994;**5**(2):111–26.
52. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. *Adv Neural Inf Process Syst* 2001;**13**:556–62.
53. Boyd S, Parikh N, Chu E, *et al*. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn* 2010;**3**(1):1–122.

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022

54. Yang J, Yuan X. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math Comput* 2012;**82**(281):301–29.
55. Xu Y, Yin W, Wen Z, Zhang Y. An alternating direction algorithm for matrix completion with nonnegative factors. *Front Math China* 2012;**7**(2):365–84.
56. Zhang Y. An alternating direction algorithm for nonnegative matrix factorization. *CAAM Technical Reports* 2010. https://hdl.handle.net/1911/102146.
57. Zheng X, Hao D, Hiroshi M, *et al*. (2013) Collaborative Matrix Factorization with Multiple Similarities for Predicting Drug-Target Interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA.
58. Shang F, Cheng J, Liu Y, *et al*. Bilinear factor matrix norm minimization for robust PCA: algorithms and applications. *IEEE Trans Pattern Anal Mach Intell* 2018;**40**(9):2066–80.
59. Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**(3):333–7.

Downloaded from https://academic.oup.com/bib/article/22/4/bbaa267/5956157 by Technische Universitaet Muenchen user on 01 December 2022