frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 09 April 2020
doi: 10.3389/fbioe.2020.00218

# Predicting Drug-Disease Associations via Multi-Task Learning Based on Collective Matrix Factorization

*Feng Huang*[1], *Yang Qiu*[1], *Qiaojun Li*[1,2], *Shichao Liu*[1,3]* *and Fuchuan Ni*[1,3]*

[1] College of Informatics, Huazhong Agricultural University, Wuhan, China, [2] School of Electronic and Information Engineering, Henan Polytechnic Institute, Henan Nanyang, China, [3] Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China

OPEN ACCESS

**Edited by:**
*Shikui Tu,*
*Shanghai Jiao Tong University, China*

**Reviewed by:**
*Xiaofei Zhang,*
*Central China Normal University, China*
*Chen Zheng,*
*Michigan State University,*
*United States*
*Minzhu Xie,*
*Hunan Normal University, China*

***Correspondence:**
*Shichao Liu*
*scliu@mail.hzau.edu.cn*
*Fuchuan Ni*
*fcni_cn@mail.hzau.edu.cn*

**Specialty section:**
*This article was submitted to*
*Bioinformatics and Computational*
*Biology,*
*a section of the journal*
*Frontiers in Bioengineering and*
*Biotechnology*

**Received:** *20 December 2019*
**Accepted:** *04 March 2020*
**Published:** *09 April 2020*

**Citation:**
*Huang F, Qiu Y, Li Q, Liu S and Ni F*
*(2020) Predicting Drug-Disease*
*Associations via Multi-Task Learning*
*Based on Collective Matrix*
*Factorization.*
*Front. Bioeng. Biotechnol. 8:218.*
*doi: 10.3389/fbioe.2020.00218*

Identifying drug-disease associations is integral to drug development. Computationally prioritizing candidate drug-disease associations has attracted growing attention due to its contribution to reducing the cost of laboratory screening. Drug-disease associations involve different association types, such as drug indications and drug side effects. However, the existing models for predicting drug-disease associations merely concentrate on independent tasks: recommending novel indications to benefit drug repositioning, predicting potential side effects to prevent drug-induced risk, or only determining the existence of drug-disease association. They ignore crucial prior knowledge of the correlations between different association types. Since the Comparative Toxicogenomics Database (CTD) annotates the drug-disease associations as therapeutic or marker/mechanism, we consider predicting the two types of association. To this end, we propose a collective matrix factorization-based multi-task learning method (CMFMTL) in this paper. CMFMTL handles the problem as multi-task learning where each task is to predict one type of association, and two tasks complement and improve each other by capturing the relatedness between them. First, drug-disease associations are represented as a bipartite network with two types of links representing therapeutic effects and non-therapeutic effects. Then, CMFMTL, respectively, approximates the association matrix regarding each link type by matrix tri-factorization, and shares the low-dimensional latent representations for drugs and diseases in the two related tasks for the goal of collective learning. Finally, CMFMTL puts the two tasks into a unified framework and an efficient algorithm is developed to solve our proposed optimization problem. In the computational experiments, CMFMTL outperforms several state-of-the-art methods both in the two tasks. Moreover, case studies show that CMFMTL helps to find out novel drug-disease associations that are not included in CTD, and simultaneously predicts their association types.

**Keywords: drug-disease association, predicting association type, similarity, collective matrix factorization, multi-task learning**

## INTRODUCTION

Drugs are chemicals used to treat, cure, prevent, or diagnose diseases. The development of a new drug has three steps: discovery stage, preclinical stage, and clinical stage (Wilson, 2006), which takes about 15 years (Dimasi, 2001) and costs about 1,000 million U.S. dollars (Adams and Brantner, 2006). Such a costly and time-consuming process remains at huge risk. After marketing, approved drugs will be surveilled to reassess their safety and some side effects may be reported (Liang et al., 2019). If adverse drug reactions cause serious consequences then the drugs are taken off the shelves and approval is withdrawn, bringing enormous economic loss to pharmaceutical companies. Therefore, identifying drug-disease associations is of significant importance. On one hand, finding novel indications for drugs can be helpful for more effective drug development. On the other hand, screening potential side effects for drugs can reduce the risk of medicine use. But traditional wet-lab experiments are expensive and laborious. In light of these challenges, computational methods which associate drugs with diseases have attracted growing attention from the biomedical community.

Recently, a large number of computational methods have been proposed for the drug-disease association prediction. Gottlieb et al. (2011) constructed a drug-disease association classifier based on the integration of drug molecular structures, drug molecular activities, and disease semantic information. Pauwels et al. (2011) put chemical structures of drugs in four machine-learning models to train classifiers. Huang et al. (2013) used the random walk to infer the unobserved links in a heterogeneous network merging drugs, genomic information, and disease phenotypes. Cheng et al. (2013) adopted a resource allocation-based approach to infer unobserved side effects for existing drugs. Oh et al. (2014) extracted features representing drug-disease associations by using similarity-based features and module-distance-based features, and then, respectively, adopted decision tree, multi-layer perception, and random forest to build prediction models. Wang et al. (2014) designed a computational framework based on a three-layer heterogeneous network model (TL-HGBI). Zhang et al. proposed the multi-label learning method (Zhang et al., 2015), and the linear neighborhood similarity-based method (Zhang et al., 2016a, 2017c) for side effect prediction. Moghadam et al. (2016) adopted the kernel fusion technique to combine different drug features and disease features, and then built SVM models. Liang et al. (2017) proposed a Laplacian regularized sparse subspace learning method (LRSSL) which integrated drug chemical structures, drug target domains, and target ontology. Zhang et al. (2016b) defined this task as the recommender problem, and introduced restricted Boltzmann machine and collaborative filtering to predict unobserved side effects. Luo et al. (2018) designed a drug repositioning recommendation system (DRRS) and used a matrix completion algorithm to fill out the unknown entries in drug-disease associations. Zhang et al. (2017b) presented a novel bipartite network-based method, which only used known drug-disease associations to predict unobserved associations. Zhang et al. (2018c) proposed a similarly constrained matrix factorization method, which utilized known drug-disease associations, drug features, and disease semantic information. Xuan et al. (2019) proposed a computational drug repositioning method through the integration of multiple drug similarity and disease similarity.

The existing models for predicting drug-disease associations only focus on indication prediction or side effect prediction, but ignore the relatedness of the two tasks, which is vital for knowledge of drug-disease associations. Despite the fact that some studies (Yang and Agarwal, 2011; Wang et al., 2013) considered drug side effects as auxiliary information for indication prediction, they failed to comprehensively make use of prior knowledge. According to the Comparative Toxicogenomics Database (CTD) (Davis et al., 2013, 2017) some drugs have therapeutic effects on diseases, e.g., sorafenib is usually used to treat leukemia (Auclair et al., 2007). Some drugs play a role in the etiology of diseases which can be regarded as side effects, biomarkers or other effects, e.g., increased sediment in the brain of amyloid beta-protein may correlate with Alzheimer's disease (Yamada et al., 2008), continued exposure to nicotine may cause lung cancer, and over-dose ingestion of caffeine may lead to a headache. Almost all drugs exert only one type of effect on a certain disease. In general, if a drug can be used to treat a disease then one can know that the drug is much less likely to exert other effects on the disease. Hence, predicting two types of drug-disease associations requires multi-task learning with two closely related tasks, where each task is meant to predict one type of association. It is a natural foresight that addressing the two tasks in one uniform framework can make them complement each other. To this end, we devise a model for capturing the relatedness between the tasks and retaining the individuality of each of them.

In this paper, we propose a collective matrix factorization-based multi-task learning method (abbreviated as "CMFMTL") to predict two types of drug-disease associations. From the CTD database, we collect drug-disease associations annotated as therapeutic or marker/mechanism (non-therapeutic), and then construct a drug-disease network with two types of links representing therapeutic effects and non-therapeutic effects. CMFMTL, respectively, approximates the association matrix regarding each link type by matrix tri-factorization, and shares the low-dimensional latent representations for drugs and diseases in the two related tasks for the goal of collective learning. We also develop an efficient algorithm to solve our proposed model. In the computational experiments, CMFMTL outperforms several state-of-the-art methods in both tasks. Moreover, case studies show that CMFMTL helps to find out novel drug-disease associations that are not included in CTD, and simultaneously predicts their association types.

## MATERIALS AND METHODS

### Dataset

The Comparative Toxicogenomics Database (CTD) (Davis et al., 2013, 2017) is a publicly available database that intends to advance understanding about how environmental exposures affect human health. Zhang et al. (2018c) downloaded the chemical-disease associations from the CTD. Then they mapped the chemicals into the DrugBank (Knox et al., 2011; Law et al., 2014; Wishart et al., 2018) database, a comprehensive knowledge base for drugs, to obtain approved drugs and some biological features for drugs, such as chemical substructures, targets, enzymes, pathways, and drug-drug interactions. The diseases were matched into the Medical Subject Headings (MeSH), a vocabulary thesaurus for biomedicine controlled by the National Library of Medicine, to collect the MeSH descriptors of diseases for the use of calculating disease semantic similarity. We use this dataset as our benchmark dataset to evaluate the performance of models.

As we described above, chemical-disease associations in CTD are annotated as therapeutic or marker/mechanism. Therapeutic associations mean that chemicals play a therapeutic role in diseases, while marker/mechanism associations mean that chemicals correlate with diseases. In this study, we can easily label these associations as therapeutic associations or non-therapeutic (marker/mechanism) associations. Extremely few associations are simultaneously annotated as two association types. Without loss of statistical properties of the data, we only label the extreme cases as therapeutic associations. Finally, the benchmark dataset contains 18,416 drug-disease associations involving 269 drugs and 598 diseases. Among these associations, 6,244 associations are therapeutic associations and 12,172 associations are non-therapeutic associations.

## Similarities for Drugs and Diseases

Let $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$ denote the set of drugs and $\mathcal{D} = \left\{d_1, d_2, \ldots, d_n\right\}$ represent the set of diseases. In this section, we introduce the drug-drug similarity and the disease-disease semantic similarity.

## Drug-Drug Similarity

A feature of a drug is a collection of entities or attributes related to the drug. Thus, we can use the Tanimoto score (Tanimoto, 1958) [also known as Jaccard index (Jaccard, 1908) for measurement of similarity between two sets] to calculate the drug-drug similarity. Let $\Gamma_i$ and $\Gamma_j$ denote features of two drugs, the Jaccard index is described as:

$$S_{Jaccard}\left(\mathit{\Gamma}_i, \mathit{\Gamma}_j\right) = \frac{\left|\mathit{\Gamma}_i \cap \mathit{\Gamma}_j\right|}{\left|\mathit{\Gamma}_i \cup \mathit{\Gamma}_j\right|} = \frac{\left|\mathit{\Gamma}_i \cap \mathit{\Gamma}_j\right|}{\left|\mathit{\Gamma}_i\right| + \left|\mathit{\Gamma}_j\right| - \left|\mathit{\Gamma}_i \cap \mathit{\Gamma}_j\right|} \tag{1}$$

where $|\cdot|$ is the number of elements in the set.

Let $\Gamma = \bigcup_i^m \Gamma_i$ represent the union set of features of $m$ drugs and $|\Gamma| = c$, and then the drug feature can be encoded as a $c$-dimensionality binary vector, e.g., the drug $r_i$ is encoded as $x_i \in \{0, 1\}^c$ where the $i$th element is set to 1 if the $i$th descriptor in $\Gamma$ belongs to the set $\Gamma_i$; otherwise, it is set to zero. Obviously, the Equation (1) can be rewritten as:

$$S_{drug}\left(r_i, r_j\right) = \frac{\left\langle x_i, x_j\right\rangle}{\left\langle x_i, e\right\rangle + \left\langle x_j, e\right\rangle - \left\langle x_i, x_j\right\rangle} \tag{2}$$

where $\langle\cdot,\cdot\rangle$ is the inner product of two vectors and $e$ is a vector with all elements equal to 1.

## Disease-Disease Similarity

As described in Wang et al. (2010) Gong et al. (2019), Zhang et al. (2019b), the hierarchical MeSH descriptors of diseases can be compiled as Directed Acyclic Graphs (DAGs), where vertexes represent the diseases and edges represent the relationships between different diseases. For a disease $d$, the DAG is denoted as $DAG_d = (V_d, E_d)$, where $V_d$ is the set of all ancestors of $d$ (including itself) and $E_d$ is the set of links from ancestor disease to their children. The semantic contribution of disease $t \in V_d$ to disease $d$ is defined as:

$$SC_d(t) = \begin{cases} 1 & if\ t = d \\ max\left\{\ \Delta \times SC_d\left(t'\right) \middle|\ t' \in C(t)\right\} & if\ t \neq d \end{cases} \tag{3}$$

where $C(t)$ is the set of children nodes of $t$, is the semantic contribution factor. Then the semantic value of disease $d$ is calculated by:

$$SV_d = \sum t \in V_d SC_d(t) \tag{4}$$

Finally, the semantic similarity between two diseases $d_i$ and $d_j$ is calculated by:

$$S_{disease}\left(d_i, d_j\right) = \frac{\sum_{t \in V_{d_i} \cap V_{d_j}} \left(SC_{d_i}(t) + SC_{d_j}(t)\right)}{SV_{d_i} + SV_{d_j}} \tag{5}$$

## Collective Matrix Factorization-Based Multi-Task Learning Method

### Multi-Task Learning

Multi-task learning is an inductive transfer learning approach that captures the connections amongst multiple related learning tasks as an inductive bias by a specific shared mechanism (Ando and Zhang, 2005), and exploits the task relatedness as prior knowledge to improve generalization capabilities (Caruana, 1997). During the learning process of multi-task learning, these related tasks are learned in parallel and complement each other, which is saying that what is learned for each task can help other tasks be learned better. In this work, we formulate predicting drug-disease therapeutic associations and non-therapeutic associations as two related tasks and put them in a multi-task setting for better predictive performance.

### Overview

The workflow of the collective matrix factorization-based multi-task learning method (CMFMTL) is demonstrated in **Figure 1**. The CMFMTL involves several critical steps to construct a prediction model for predicting two types of drug-disease associations. First, a drug-disease association network is constructed based on known associations and their types: therapeutic and non-therapeutic. Second, the drug-disease association network is divided into two subnetworks: one subnetwork involves links representing therapeutic associations and the other contains links representing non-therapeutic associations. Third, two binary matrices regarding the two subnetworks are simultaneously factorized into the product of three low-dimensional matrices which are served as latent

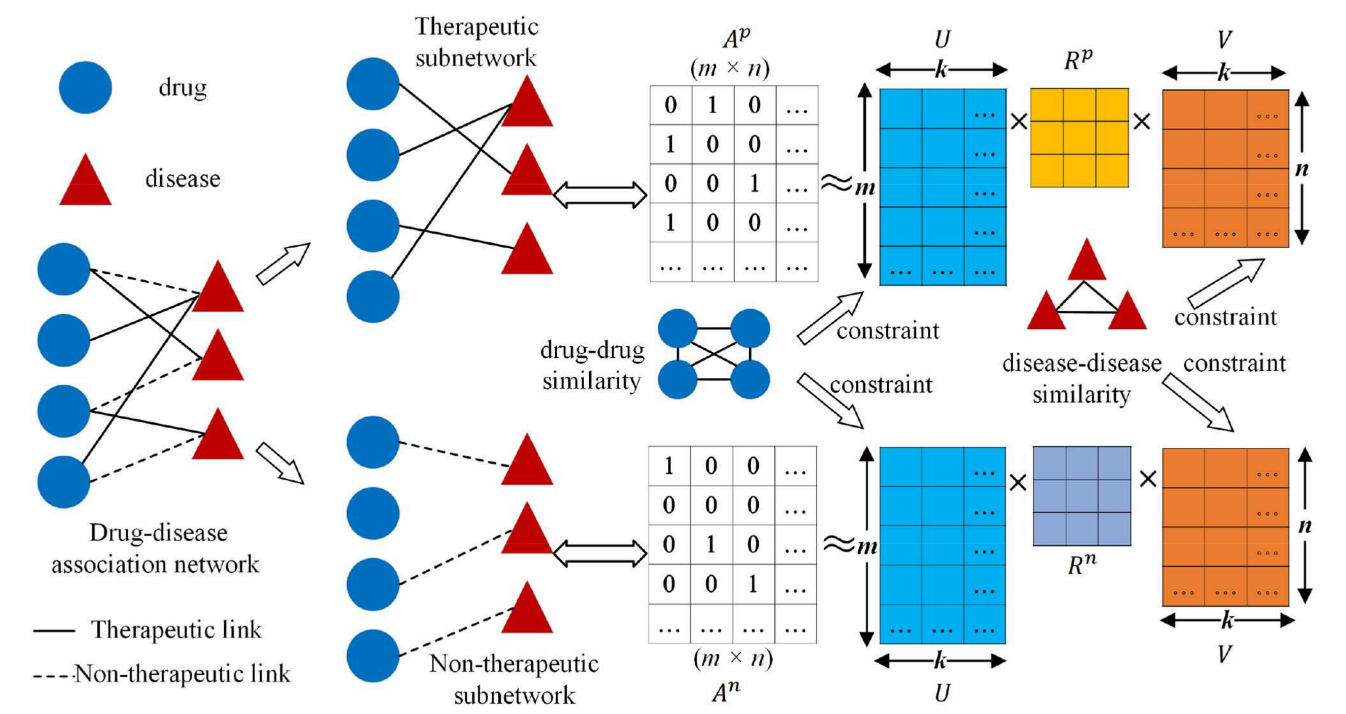

**FIGURE 1 |** Workflow of collective matrix factorization-based multi-task learning method (CMFMTL): $A^p$ is the corresponding binary matrix for the therapeutic subnetwork; $A^n$ is the corresponding binary matrix for non-therapeutic subnetwork; $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ are, respectively, the low-dimensional representations for drugs and diseases; $R^p$ and $R^n$ are coefficient matrices.

components for drugs and diseases, and coefficient matrices measuring the level of interaction between latent components. The latent representations of drugs and diseases are shared in the two factorization tasks for capturing the relatedness of these two tasks, and the different coefficient matrices maintain the specificity of two tasks. Finally, the graph Laplacian regularizations (Cai et al., 2011) based on the biological features of drugs and diseases are introduced to further enhance interpretability and generalization.

## Objective Function of CMFMTL

Given a set of drugs $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$ and a set of diseases $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, we can construct a relation network $\mathcal{G}$, which uses $\mathcal{R}$ and $\mathcal{D}$ as two disjoint sets of nodes. There are two types of links between nodes in $\mathcal{R}$ and nodes in $\mathcal{D}$. The link between drug $r_i$ and disease $d_j$ is labeled as therapeutic link if the drug $r_i$ has a therapeutic effect on the disease $d_j$; the edge is labeled as a non-therapeutic link if the drug $r_i$ has a non-therapeutic effect on the disease $d_j$. Then the drug-disease association network $\mathcal{G}$ can be divided into a therapeutic subnetwork $\mathcal{G}_p$ and a non-therapeutic subnetwork $\mathcal{G}_n$. $A^p \in \{0,1\}^{m\times n}$ is the corresponding binary matrix for $\mathcal{G}_p$, where $A^p_{ij} = 1$ if the drug $r_i$ has a therapeutic link to the disease $d_j$, otherwise $A^p_{ij} = 0$. Similarly, $A^n \in \{0,1\}^{m\times n}$ is the corresponding binary matrix for $\mathcal{G}_n$, where $A^n_{ij} = 1$ if the drug $r_i$ has a non-therapeutic link to the disease $d_j$, otherwise $A^n_{ij} = 0$. We employ the matrix tri-factorization technique to model $A^p$ and $A^n$, respectively, and map the drugs (diseases) into common latent representations shared in two tasks. Specifically, we approximate the association matrices $A^p$ and $A^n$ by minimizing the reconstruction errors:

$$\min_{U,V,R^p,R^n} \frac{1}{2}\left(\left\|A^p - UR^pV^T\right\|_F^2 + \left\|A^n - UR^nV^T\right\|_F^2\right) \tag{6}$$

where $\|\cdot\|_F^2$ is the Frobenius norm; $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ are the low-dimensional representations for drugs and diseases, respectively; $R^p$ and $R^n$ are coefficient matrices which model how the latent representations interact in the respective association type; $k < \min(m, n)$ is the dimensionality of the low-dimensional space.

Since Equation (6) maps drugs and diseases into a low-dimensional space, a natural idea occurs that the low-dimensional representations should preserve the underlying interconnection information of drugs and diseases. Studies on manifold learning (Belkin et al., 2006; Ma and Fu, 2012; Zhang et al., 2018a), spectral graph theory (Chung, 1997; Rana et al., 2015) and their applications (Zhang et al., 2016a, 2017a,b,c, 2018b; Ruan et al., 2017) have shown that the learning performance can be signally enhanced, if the local topological invariant properties are preserved. Cai et al. (2011) proposed Laplacian regularizations to achieve this goal. Here, we introduce the regularization terms based on biological features about drugs and diseases to incorporate similarity information in our model. We denote the drug-drug similarity matrix as $W^r \in \mathbb{R}^{m\times m}$ where the $(i,j)$th entry $w^r_{ij} = S_{drug}(r_i, r_j)$ and the disease-disease semantic similarity as $W^d \in \mathbb{R}^{n\times n}$ where the $\left(i,j\right)$th entry $w_{ij}^d = S_{disease}\left(d_i, d_j\right)$, which are previously calculated in Equations (2) and (5). Then, the graph Laplacian matrices are constructed as $L^U = D^r - W^r$ and $L^V = D^d - W^d$, where $D^r$ and $D^d$ are, respectively, diagonal matrices whose diagonal elements are corresponding row sums of $W^r$ and $W^d$. The graph Laplacian regularizations are formulated as:

$$\begin{aligned}\mathcal{R}_1 &= tr\left(U^T L^U U\right) = \frac{1}{2}\sum_{i,j=1}^{m}\left\|U\left(i,:\right) - U\left(j,:\right)\right\|_2^2 w_{ij}^r \\ \mathcal{R}_2 &= tr\left(V^T L^V V\right) = \frac{1}{2}\sum_{i,j=1}^{n}\left\|V\left(i,:\right) - V\left(j,:\right)\right\|_2^2 w_{ij}^d \qquad (7)\end{aligned}$$

where $tr\left(\cdot\right)$ denotes the trace of a square matrix; $U\left(i,:\right)$ and $V\left(i,:\right)$ are the $i$th row vector of $U$ and $V$, respectively; more details for the second equality can be referred to in Cai et al. (2011). Obviously, minimizing $\mathcal{R}_1$ (or $\mathcal{R}_2$) will lead to a result that the drug $r_i$ (the disease $d_i$) is closer to the drug $r_j$ (the disease $d_j$) in the low-dimensional space if the similarity between them $w_{ij}^r$ ($w_{ij}^d$) is higher. Additionally, we introduce the $L_2$ regularizations to reinforce the smoothness of $U$, $V$, $R^p$, and $R^n$. Therefore, we obtain the optimization objective of the CMFMTL by combining the $L_2$ regularizations, Equations (6) and (7):

$$\begin{aligned}\min_{U,V,R^p,R^n} \frac{1}{2}&\left(\left\|A^p - UR^pV^T\right\|_F^2 + \left\|A^n - UR^nV^T\right\|_F^2\right) \\ &+\frac{\alpha}{2}tr\left(U^T L^U U\right) + \frac{\beta}{2}tr\left(V^T L^V V\right) \\ &+\frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2 + \left\|R^p\right\|_F^2 + \left\|R^n\right\|_F^2\right) \qquad (8)\end{aligned}$$

where $\alpha$, $\beta$ and $\lambda$ are the regularization parameters.

## Optimization

To efficiently solve problem (8), we equivalently convert it into an equation constrained optimization problem:

$$\begin{aligned}\min_{U,V,R^p,R^n} \frac{1}{2}&\left(\left\|A^p - UR^pV^T\right\|_F^2 + \left\|A^n - UR^nV^T\right\|_F^2\right) \\ &+\frac{\alpha}{2}tr\left(W^T L^U W\right) + \frac{\beta}{2}tr\left(J^T L^V J\right) \\ &+\frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2 + \left\|R^p\right\|_F^2 + \left\|R^n\right\|_F^2\right) \\ &s.t.\ J = V,\ \ W = U \qquad (9)\end{aligned}$$

Then, the augmented Lagrangian function $\mathcal{L}$ of Equation (9) is introduced as follows:

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\left(\left\|A^p - UR^pV^T\right\|_F^2 + \left\|A^n - UR^nV^T\right\|_F^2\right) \\ &+ \frac{\alpha}{2}tr\left(W^T L^U W\right) + \frac{\beta}{2}tr\left(J^T L^V J\right) + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right. \\ &\left.+ \left\|R^p\right\|_F^2 + \left\|R^n\right\|_F^2\right) + tr\left(Z^T\left(W - U\right)\right) \\ &+ \frac{\rho_1}{2}\|W - U\|_F^2 + tr\left(Y^T\left(J - V\right)\right) + \frac{\rho_2}{2}\|J - V\|_F^2 \qquad (10)\end{aligned}$$

where $J$ and $W$ are the auxiliary variables; $\rho_1 > 0$, $\rho_2 > 0$ are called as the penalty parameters; $Z$ and $Y$ are the Lagrange multipliers. We resort to the alternating direction method of multipliers (ADMM) framework (Boyd et al., 2011) to devise an alternately updating rule for optimizing Equation (10).

Next, differentiating $\mathcal{L}$ with respect to $J$, $W$, $U$, and $V$, respectively, and setting the partial derivatives to zero, we have the following updating rule:

$$\begin{aligned}J &= \left(\beta L^V + \rho_2 I\right)^{-1}\left(\rho_2 V - Y\right) \\ W &= \left(\alpha L^U + \rho_1 I\right)^{-1}\left(\rho_1 U - Z\right) \\ U &= \left(A^pVR^{p^T} + A^nVR^{n^T} + Z + \rho_1 W\right)\left(R^pV^TVR^{p^T}\right. \\ &\qquad \left.+ R^nV^TVR^{n^T} + \lambda I + \rho_1 I\right)^{-1} \\ V &= \left(A^{p^T}UR^p + A^{n^T}UR^n + Y + \rho_2 J\right)\left(\left(UR^p\right)^T UR^p\right. \\ &\qquad \left.+ \left(UR^n\right)^T UR^n + \lambda I + \rho_2 I\right)^{-1} \qquad (11)\end{aligned}$$

where $I$ represents the identity matrix with an adaptive dimensionality in different equations. When fixing other variables, the objective function for $R^p$ is simplified as:

$$\min_{R^p}\frac{1}{2}\left\|A^p - UR^pV^T\right\|_F^2 + \frac{\lambda}{2}\left\|R^p\right\|_F^2 \qquad (12)$$

Equation (12) can be efficaciously solved by the algorithm proposed in Yu et al. (2014) which leverages the conjugate gradient method (CG) to improve the efficiency of the solver. Here, we omit the details about the algorithm, and denote the solution for the Equation (12) solved by the algorithm as $CG\left(R^p\right)$. The objective function with regard to $R^n$ shares the same optimization structure with the Equation (12), and thus we denote the solution as $CG\left(R^n\ \right)$.

Finally, the Lagrange multipliers and the penalty parameter are updated as follows:

$$\begin{aligned}Y &= Y + \rho_2\left(J - V\right) \\ Z &= Z + \rho_1\left(W - U\right) \\ \rho_1 &= \mu\rho_1 \\ \rho_2 &= \mu\rho_2 \qquad (13)\end{aligned}$$

We alternatively update all variables until convergence and the whole process are summarized in Algorithm 1. According to Yu et al. (2014), the main operation in each iteration of the conjugate gradient procedure is a multiplication of three matrices, which can be done in $O\left(\min(m,\ n)k^2 + mnk + k^3\right)$ time. We set the maximal iterative number in conjugate gradient procedure as $t$. In each iteration of ADMM, the main operations contain several matrix inverse calculations [in Equation (11)] that cost $O\left(n^3 + m^3 + k^3\right)$, several matrix multiplications [in Equation (11) and the initialization for conjugate gradient procedure] that cost $O\left(n^2k + m^2k + mk^2 + nk^2 + k^3 + mnk\right)$ and the conjugate gradient procedure that cost $O\left(\left(\min(m,\ n)k^2 + mnk + k^3\right)t\ \right)$.

**Algorithm 1:** The updated process of CMFMTL.

```
Input: known drug-disease therapeutic association matrix,
        $A^p \in \{0,1\}^{m\times n}$;
       known drug-disease non-therapeutic association matrix,
        $A^n \in \{0,1\}^{m\times n}$;
       drug similarity matrix, $w^r \in \mathbb{R}^{m\times m}$;
       disease similarity matrix, $w^d \in \mathbb{R}^{n\times n}$;
       dimensionality of the embedded space, $k < \min(m, n)$;
       regularization parameters, $\alpha > 0$, $\beta > 0$ and $\lambda > 0$
Output: the prediction matrices $A^{p*}$, $A^{n*}$
Initialize $V \in \mathbb{R}^{n\times k}$ and $U \in \mathbb{R}^{m\times k}$ in the interval [0, 1] randomly; $Y = 0$ and $Z = 0$;
    $\rho_1 = \rho_2 = 1$
Repeat
    Update $R^p$ and $R^n$ using
        $R^p = CG\left(R^p\right), R^n = CG\left(R^n\right)$
    Update $J$, $W$, $U$ and $V$ via the equation (11)
    Update $Y$, $Z$, $\rho_1$ and $\rho_2$ via the equation (13)
End until convergence
Output $A^{p*}$, $A^{n*}$ using
        $A^{p*} = UR^pV^T, A^{n*} = UR^nV^T$
```

# RESULTS AND DISCUSSION

## Evaluation Metrics

In our experiment, 5-fold cross validation (5-CV) experiments are conducted to systematically evaluate prediction models. Considering assessing models in two tasks, where predicting drug-disease therapeutic associations is called task 1 and the other is called task 2, we respectively split known therapeutic associations and non-therapeutic associations into five equal-sized parts at random. In each task, one of the five subsets is considered as the testing set in turn, and the remaining four subsets are combined as the training set. The metrices can be calculated in each fold, and the average of five evaluations is adopted.

Several evaluation metrics, such as sensitivity (SE, also known as recall), specificity (SP), accuracy (ACC), precision (PRE), and F-measure (F), are calculated. Since they depend on a threshold to classify predictions as positive or negative, we adopt the threshold which produces the max F-measure. Moreover, the area under the receiver-operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR) are adopted as the primary metrics.

## Parameter Setting

The collective matrix factorization-based multi-task learning method (CMFMTL) has four key parameters: the dimensionality of the common latent space $k$, and the regularization coefficients $\alpha$, $\beta$, and $\lambda$. These parameters may have great impact on the performances of the CMFMTL, so analysis of parameters is necessary. For simplicity, we set $\alpha = \beta$, $\lambda \in \{2, 4, 6, 8, 10\}$ and $k \in \{5, 10, 15, 20, 25, 30, 35, 40\}$. Note that we have several kinds of drug features as mentioned in section Dataset. We use drug substructures to calculate drug-drug similarity for better performance. For the calculation of disease-disease similarity, we set semantic contribution factor = 0.5 (Zhang et al., 2019b). For the growth factor $\mu$ of the penalty parameters $\rho_1$ and $\rho_2$ in Equation (13), we set $\mu = 1.1$. By grid-search, we obtain the best results with an AUPR of 0.2122 in task 1 when $\alpha = \beta = 8$, $\lambda = 4$ and $k = 30$; and with an AUPR of 0.1838 in task 2 when $\alpha = \beta = 10$, $\lambda = 6$ and $k = 35$. **Figures 2A,C** show the influence of regularization coefficients on the performance of the CMFMTL in task 1 and task 2, respectively. **Figures 2B,D** correspond to the impact of dimensionality in the two tasks. From some observations, the $L_2$ regularization coefficient $\lambda$ may control the trade-off between the two tasks, e.g., greater $\lambda$ produces better performance in task 2 than task 1. When the dimensionality $k$ is too low, models perform poorly. The cause may be that vital data information fails to be fully embedded in the latent representations.

## Comparison With State-of-the-Art Association Prediction Methods

As we discussed above, CMFMTL is a multi-task learning method that simultaneously predicts therapeutic and non-therapeutic associations between drugs and diseases. Existing methods only predict a certain type of drug-disease associations, such as drug indications and side effects. For this reason, we conduct each of several association prediction methods, respectively, on two tasks, and then compare the performance of them with our proposed CMFMTL model.

Here, we consider three state-of-the-art association prediction methods: TL-HGBI, LRSSL, and DRRS, which are the classic or latest works of predicting drug-disease associations. TL-HGBI (Wang et al., 2014) bridged drugs to targets and linked them to diseases to depict a three-layer heterogeneous network. Then, a similarity-based information diffusion method was used to estimate the probabilities of unknown drug-disease associations. LRSSL (Liang et al., 2017) modeled the prediction of drug indications as a joint optimization problem by combining Laplacian regularization with a sparse learning framework, and then an iteratively updating algorithm was implemented to obtain a locally optimal solution. DRRS (Luo et al., 2018) stated drug repositioning as a recommendation problem and utilized a matrix completion algorithm on a block matrix which was concatenated by a drug-disease association matrix, a drug-drug similarity matrix, and a disease-disease similarity matrix. In addition, we use a reduced version of our model (CMFMTL-R) as a baseline method with only one matrix tri-factorization term in Equation (8). CMFMTL-R is a single-task version of CMFMTL, which acquires the result in each task by separately factorizing each corresponding data matrix, e.g., factorizing $A^p$ without decomposing $A^n$ in task 1. We also retain the graph regularizations and $L_2$ regularizations, and use the same algorithm and parameter setting in CMFMTL-R as in CMFMTL for fair comparison.

All methods are evaluated by 5-CV, and results are shown in **Tables 1**, **2**. Clearly, CMFMTL produces better results than TL-HGBI, LRSSL, and DRRS in the two tasks. It is observed that TL-HGBI and LRSSL perform poorly on our dataset. The

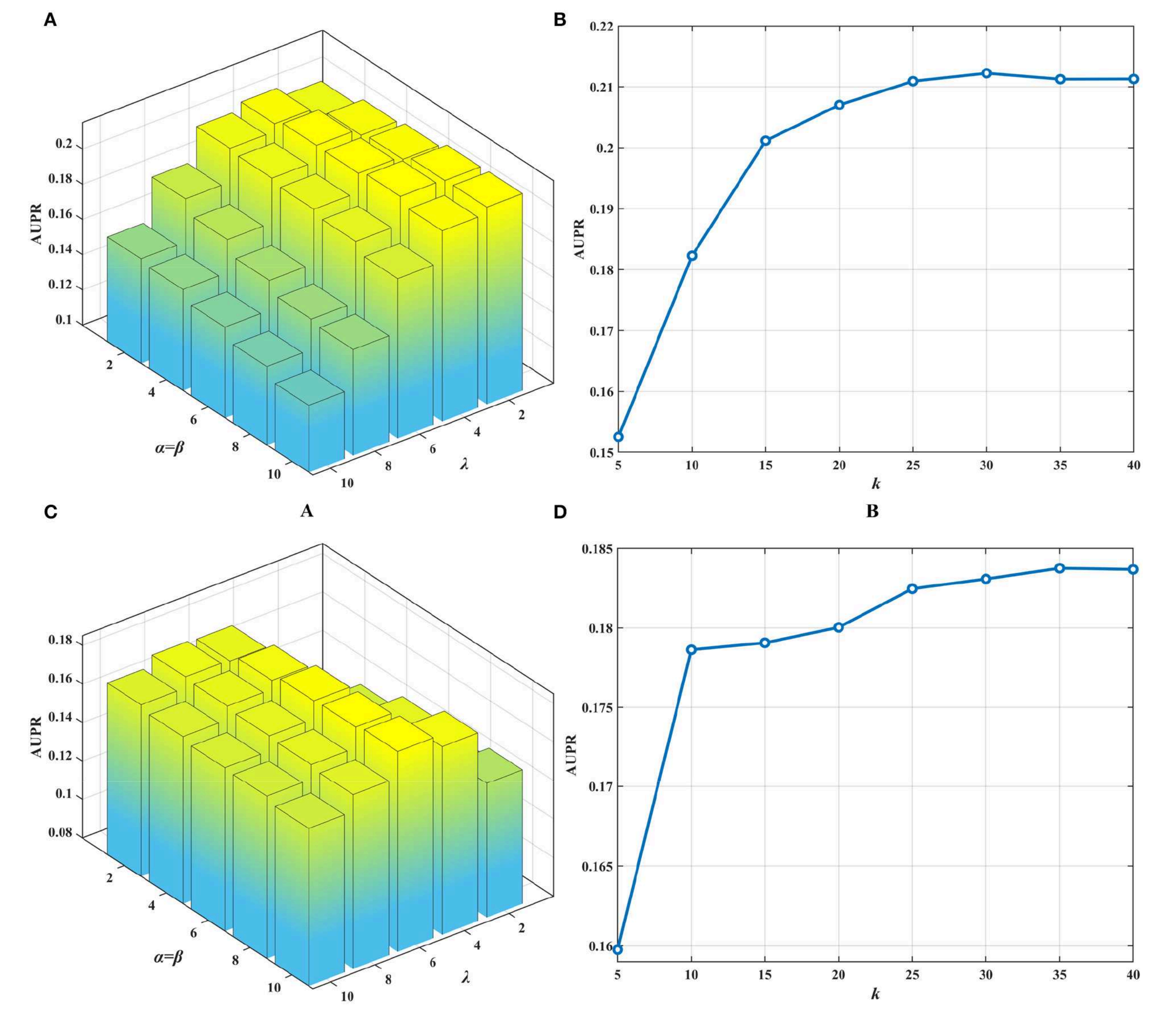

**FIGURE 2 |** Influence of parameters on the performance of CMFMTL involving two tasks: **(A)** shows the influence of $\alpha, \beta, \lambda$ on the AUPR score in task 1. **(B)** indicates the effect of $k$ on the AUPR score in task 1. **(C)** illustrates the impact of $\alpha, \beta, \lambda$ on the AUPR score in task 2. **(D)** demonstrates the perturbation of $k$ on the AUPR score in task 2.

**Table 1 |** Performances of Prediction Models in Task 1.

| Methods | AUPR | AUC | SE | SP | PRE | ACC | F |
|---|---|---|---|---|---|---|---|
| CMFMTL | 0.2122 | 0.8898 | 0.2888 | 0.9926 | 0.2544 | 0.9866 | 0.2690 |
| CMFMTL-R | 0.1217 | 0.8543 | 0.2135 | 0.9905 | 0.1644 | 0.9839 | 0.1849 |
| TL-HGBI | 0.0444 | 0.7444 | 0.1265 | 0.9827 | 0.0624 | 0.9753 | 0.0808 |
| LRSSL | 0.0420 | 0.7341 | 0.1489 | 0.9745 | 0.0490 | 0.9674 | 0.0731 |
| DRRS | 0.1735 | 0.8893 | 0.2756 | 0.9917 | 0.2292 | 0.9856 | 0.2468 |

**Table 2 |** Performances of Prediction Models in Task 2.

| Methods | AUPR | AUC | SE | SP | PRE | ACC | F |
|---|---|---|---|---|---|---|---|
| CMFMTL | 0.1838 | 0.8661 | 0.3091 | 0.9798 | 0.2091 | 0.9686 | 0.2473 |
| CMFMTL-R | 0.1465 | 0.8449 | 0.2623 | 0.9798 | 0.1812 | 0.9679 | 0.2139 |
| TL-HGBI | 0.0635 | 0.7469 | 0.1839 | 0.9653 | 0.0840 | 0.9523 | 0.1140 |
| LRSSL | 0.0606 | 0.7393 | 0.1812 | 0.9644 | 0.0801 | 0.9514 | 0.1106 |
| DRRS | 0.1150 | 0.8570 | 0.3105 | 0.9690 | 0.1454 | 0.9580 | 0.1979 |

most possible reason is that these models are unstable and the performances of them highly rely on their datasets. DRRS is a matrix completion method, which is thought to be able to obtain better results on sparse data. Thereby, DRRS performs better on fewer therapeutic associations than on denser non-therapeutic associations. In contrast, CMFMTL-R performs more steadily

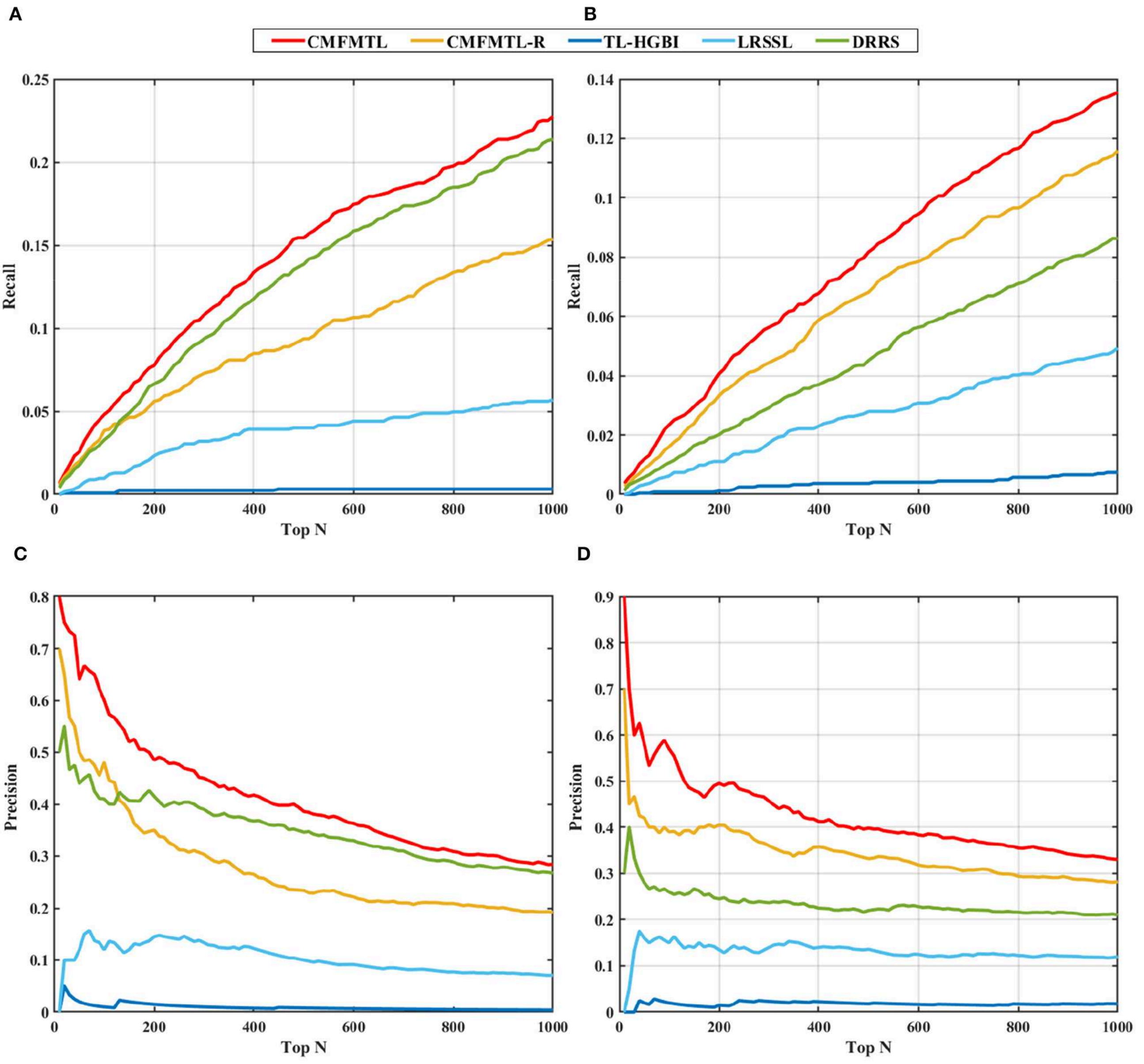

**FIGURE 3 |** Top-N ranked recall and precision of all methods in two tasks: **(A)** shows the top-N ranked recall in task 1. **(B)** displays the top-N ranked recall in task 2. **(C)** demonstrates the top-N ranked precision in task 1. **(D)** illustrates the top-N ranked precision in task 2.

in two tasks. Compared with other methods, CMFMTL successfully makes use of all useful association information by collaboratively learning from two tasks. Such advantages make CMFMTL generally outperform other single-task learning methods.

In practical application, one may be concerned about how many true associations can be recovered by the predictive models from highly ranked predictions. We evaluate the capabilities of all models for top-N predictions. Recall that we randomly select 20% of known therapeutic associations and 20% known non-therapeutic associations, and remove them in 1-fold of 5-CV. We can then investigate the recall scores and precision scores of all models in top predictions ranging from top 10 to top 1,000 (in a step size of 10), and the results are shown in **Figure 3**. Overall, in both tasks, the proposed CMFMTL method performs best among all methods in terms of both precision and recall at each value of N. Especially, there are more than 50% associations precisely predicted by the CMFMTL within top-100 predictions in both tasks. We ascribe the poor performance of the TL-HGBI to the weak predictive power of the network-based method which heavily relies on the

**Table 3 |** Top 10 Drug-Disease Associations Predicted by CMFMTL.

| Drug name | Disease name | Type | Evidence |
|---|---|---|---|
| Chloroquine | Bradycardia | −1 | Don Michael and Aiwazzadeh, 1970 |
| Chlorpromazine | Coma | −1 | N.A. |
| Risperidone | Anxiety disorders | 1 | Ravindran et al., 2007 |
| Clozapine | Headache | −1 | https://en.wikipedia.org/wiki/Clozapine |
| Methotrexate | Neoplasms | 1 | https://en.wikipedia.org/wiki/Methotrexate |
| Valproic Acid | Fatigue | −1 | N.A. |
| Amitriptyline | Confusion | −1 | https://en.wikipedia.org/wiki/Amitriptyline |
| Ibuprofen | Drug hypersensitivity | −1 | Nanau and Neuman, 2010 |
| Tamoxifen | Diarrhea | −1 | N.A. |
| Vincristine | Neoplasms | 1 | https://en.wikipedia.org/wiki/Vincristine |

*N.A. means that the predicted association cannot be confirmed. Type 1 denotes the therapeutic associations and type −1 refers to non-therapeutic associations.*

network structure. All the results indicate that CMFMTL absorbs complementary information from two tasks for better performance.

## Case Study

In this section, we use case studies to demonstrate the practical usefulness of CMFMTL in predicting therapeutic and non-therapeutic associations. CMFMTL makes predictions by collective learning, and also shares predictive signals across two tasks. Hence, the prediction scores that the CMFMTL simultaneously generates for two tasks are able to measure the probabilities that drugs associate diseases in a certain association type. We use all drug-disease associations in our dataset to train the CMFMTL model and then rank the prediction scores of all unknown entries which remain unrecorded in the dataset. Then, we focus on the top predicted (drug, disease, association type) triples. We list top 10 ranked predictions in **Table 3** and then check up on these associations according to the literature, publications and credible websites. As shown in **Table 3**, we find evidence to confirm seven associations as well as the corresponding association type. For example, Risperidone, a safe and effective atypical antipsychotic medication, has been frequently used off-label by clinicians to treat Anxiety Disorders (Ravindran et al., 2007). Drug Hypersensitivity is an allergy to a drug and is a form of adverse drug reaction, and the study (Nanau and Neuman, 2010) presented an Ibuprofen-induced clinical manifestation of Hypersensitivity syndrome.

# CONCLUSION

In this work, to simultaneously predict two types of drug-disease association, we present a novel model named collective matrix factorization-based multi-task learning (CMFMTL). Different from existing methods that focus on the existence of drug-disease associations, CMFMTL aims to predict the drug-disease associations and their corresponding association type. Since drug-disease associations are annotated into two categories, predicting each type of association can be served as one individual task. The underlying relatedness across the tasks is a vital piece of prior knowledge that can greatly improve learning abilities. CMFMTL captures the relations between two tasks and successfully utilizes all useful information to achieve high-accuracy and robust performances. The experimental results show that CMFMTL outperforms other state-of-the-art association prediction methods. Case studies demonstrate CMFMTL can find out novel associations and accurately infer the association type.

Nevertheless, CMFMTL still has limitations. CMFMTL predicts the probabilities of therapeutic associations and non-therapeutic associations for all non-interaction drug-disease pairs. However, we notice that some drug-disease associations are included in the top prediction of therapeutic associations as well as the top prediction of non-therapeutic associations. It means that these associations are predicted by CMFMTL to be both therapeutic and non-therapeutic, which is conflicting. The possible reason is that these drugs and diseases are very popular and have a great number of associations. Then, the model learns the data bias. In future work, we will optimize the proposed model to avoid this conflict. Note that similarity integration methods are usually able to achieve high-accuracy performance in similar bioinformatics issues (Zhang et al., 2018d, 2019a,c). We should also consider redesigning our model to integrate several resources of drug feature information.

# DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The Comparative Toxicogenomics Database (CTD) https://github.com/LoseHair/CMFMTL, DrugBank https://www.drugbank.ca/, Medical Subject Headings (MeSH) http://www.bioinfotech.cn/SCMFDD/.

# AUTHOR CONTRIBUTIONS

FH and SL designed the project and wrote the manuscript. YQ and QL performed the experiments and analyzed the results. SL and FN supervised and conceived the study. All authors read and approved the manuscript.

# FUNDING

This work was supported by National Key Research and Development Program of China (2018YFC1604000), the Fundamental Research Funds for the Central Universities (2662019QD011).

## REFERENCES

Adams, C. P., and Brantner, V. V. (2006). Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff.* 25, 420–428. doi: 10.1377/hlthaff.25.2.420

Ando, R. K., and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* 6, 1817–1853. Available online at: http://www.jmlr.org/papers/v6/ando05a.html

Auclair, D., Miller, D., Yatsula, V., Pickett, W., Carter, C., Chang, Y., et al. (2007). Antitumor activity of sorafenib in FLT3-driven leukemic cells. *Leukemia* 21, 439–445. doi: 10.1038/sj.leu.2404508

Belkin, M., Partha, N., Sindhwani, V. (2006). Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* 7, 2399–2434. Available online at: http://www.jmlr.org/papers/v7/belkin06a.html

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Cai, D., He, X., Han, J., and Huang, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1548–1560. doi: 10.1109/TPAMI.2010.231

Caruana, R. (1997). Multitask learning. *Mach. Learn.* 28, 41–75. doi: 10.1023/A:1007379606734

Cheng, F., Li, W., Wang, X., Zhou, Y., Wu, Z., Shen, J., et al. (2013). Adverse drug events: database construction and in silico prediction. *J. Chem. Inf. Model.* 53, 744–752. doi: 10.1021/ci4000079

Chung, F. R. K. (1997). *Spectral Graph Theory*. Providence, R.I.: Published for the Conference Board of the mathematical sciences by the American Mathematical Society.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., Mcmorran, R., et al. (2017). The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.* 45, D972–D978. doi: 10.1093/nar/gkw838

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., et al. (2013). The comparative toxicogenomics database: update 2013. *Nucleic Acids Res.* 41, D1104–D1114. doi: 10.1093/nar/gks994

Dimasi, J. A. (2001). New drug development in the United States from 1963 to 1999. *Clin. Pharmacol. Ther.* 69, 286–296. doi: 10.1067/mcp.2001.115132

Don Michael, T. A., and Aiwazzadeh, S. (1970). The effects of acute chloroquine poisoning with special reference to the heart. *Am. Heart J.* 79, 831–842. doi: 10.1016/0002-8703(70)90371-6

Gong, Y., Niu, Y., Zhang, W., and Li, X. (2019). A network embedding-based multiple information integration method for the MiRNA-disease association prediction. *BMC Bioinform.* 20:468. doi: 10.1186/s12859-019-3063-3

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7:496. doi: 10.1038/msb.2011.26

Huang, Y. F., Yeh, H. Y., and Soo, V. W. (2013). Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC Med. Genomics* 6 (Suppl. 3):S4. doi: 10.1186/1755-8794-6-S3-S4

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44, 223–270.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 39, D1035–D1041. doi: 10.1093/nar/gkq1126

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097. doi: 10.1093/nar/gkt1068

Liang, X., Zhang, P., Li, J., Fu, Y., Qu, L., Chen, Y., et al. (2019). Learning important features from multi-view data to predict drug side effects. *J. Cheminform.* 11:79. doi: 10.1186/s13321-019-0402-3

Liang, X., Zhang, P., Yan, L., Fu, Y., Peng, F., Qu, L., et al. (2017). LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning. *Bioinformatics* 33, 1187–1196. doi: 10.1093/bioinformatics/btw770

Luo, H., Li, M., Wang, S., Liu, Q., Li, Y., and Wang, J. (2018). Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 34, 1904–1912. doi: 10.1093/bioinformatics/bty013

Ma, Y., and Fu, Y. (2012). *Manifold Learning Theory and Applications*. Boca Raton, FL: CRC; Taylor & Francis distributor. doi: 10.1201/b11431

Moghadam, H., Rahgozar, M., and Gharaghani, S. (2016). Scoring multiple features to predict drug disease associations using information fusion and aggregation. *SAR QSAR Environ. Res.* 27, 609–628. doi: 10.1080/1062936X.2016.1209241

Nanau, R. M., and Neuman, M. G. (2010). Ibuprofen-induced hypersensitivity syndrome. *Transl. Res.* 155, 275–293. doi: 10.1016/j.trsl.2010.01.005

Oh, M., Ahn, J., and Yoon, Y. (2014). A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. *PLoS ONE* 9:e111668. doi: 10.1371/journal.pone.0111668

Pauwels, E., Stoven, V., and Yamanishi, Y. (2011). Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinform.* 12:169. doi: 10.1186/1471-2105-12-169

Rana, B., Juneja, A., Saxena, M., Gudwani, S., Kumaran, S. S., Behari, M., et al. (2015). Graph-theory-based spectral feature selection for computer aided diagnosis of Parkinson's disease using T1-weighted MRI. *Int. J. Imag. Syst. Technol.* 25, 245–255. doi: 10.1002/ima.22141

Ravindran, A. V., Bradbury, C., Mckay, M., and Da Silva, T. L. (2007). Novel uses for risperidone: focus on depressive, anxiety and behavioral disorders. *Expert Opin. Pharmacother.* 8, 1693–1710. doi: 10.1517/14656566.8.11.1693

Ruan, C. Y., Wang, Y., Zhang, Y. C., Ma, J. G., Chen, H. J., Aickelin, U., et al. (2017). "THCluster:herb supplements categorization for precision traditional chinese medicine," in *2017 IEEE International Conference on Bioinformatics And Biomedicine* (Kansas City, MO: BIBM), 417–424. doi: 10.1109/BIBM.2017.8217685

Tanimoto, T. T. (1958). "Elementary mathematical theory of classification and prediction," in *IBM Internal Report*. doi: 10.2208/jscej1949.1958.54_35

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, W., Yang, S., Zhang, X., and Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 30, 2923–2930. doi: 10.1093/bioinformatics/btu403

Wang, Y., Chen, S., Deng, N., and Wang, Y. (2013). Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS ONE* 8:e78518. doi: 10.1371/journal.pone.0078518

Wilson, J. F. (2006). Alterations in processes and priorities needed for new drug development. *Ann. Intern. Med.* 145, 793–796. doi: 10.7326/0003-4819-145-10-200611210-00024

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Xuan, P., Cao, Y., Zhang, T., Wang, X., Pan, S., and Shen, T. (2019). Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics* 35, 4108–4119. doi: 10.1093/bioinformatics/btz182

Yamada, M., Chiba, T., Sasabe, J., Terashita, K., Aiso, S., and Matsuoka, M. (2008). Nasal colivelin treatment ameliorates memory impairment related to Alzheimer's disease. *Neuropsychopharmacology* 33, 2020–2032. doi: 10.1038/sj.npp.1301591

Yang, L., and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PLoS ONE* 6:e28025. doi: 10.1371/journal.pone.0028025

Yu, H.-F., Jain, P., Kar, P., and Dhillon, I. S. (2014). "Large-scale multi-label learning with missing labels," in *Proceedings of the 31st International Conference on International Conference on Machine Learning*. (Beijing: JMLR.org).

Zhang, W., Chen, Y., and Li, D. (2017a). Drug-target interaction prediction through label propagation with linear neighborhood information. *Molecules* 22:2056. doi: 10.3390/molecules22122056

Zhang, W., Chen, Y., Tu, S., Liu, F., and Qu, Q. (2016a). "Drug side effect prediction through linear neighborhoods and multiple data source integration," in *2016 IEEE International Conference on Bioinformatics and Biomedicine* (Shenzhen: BIBM), 427–434. doi: 10.1109/BIBM.2016.7822555

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). SFLLN: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug-drug interactions. *Inf. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2931546. [Epub ahead of print].

Zhang, W., Liu, F., Luo, L., and Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinform.* 16:365. doi: 10.1186/s12859-015-0774-y

Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., and Li, X. (2018a). Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing* 287, 154–162. doi: 10.1016/j.neucom.2018.01.085

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018b). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhang, W., Tang, G., Zhou, S., and Niu, Y. (2019c). LncRNA-miRNA interaction prediction through sequence-derived linear neighborhood propagation method with information combination. *BMC Genomics* 20, 1–12. doi: 10.1186/s12864-019-6284-y

Zhang, W., Yue, X., Chen, Y., Lin, W., Li, B., Liu, F., et al. (2017b). "Predicting drug-disease associations based on the known association bipartite network," in *2017 IEEE International Conference on Bioinformatics and Biomedicine* (BIBM), 503–509. doi: 10.1109/BIBM.2017.8217698

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018c). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinform.* 19:233. doi: 10.1186/s12859-018-2220-4

Zhang, W., Yue, X., Liu, F., Chen, Y. L., Tu, S. K., and Zhang, X. N. (2017c). A unified frame of predicting side effects of drugs by using linear neighborhood similarity. *BMC Syst. Biol.* 11 (Suppl. 6):101. doi: 10.1186/s12918-017-0477-2

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018d). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol.* 14:e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., and Xiao, W. (2016b). Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* 173, 979–987. doi: 10.1016/j.neucom.2015.08.054

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.