Modeling Relational Drug-Target-Disease Interactions via Tensor Factorization with Multiple Web Sources

Huiyuan Chen
Electrical Engineering and Computer Science
Case Western Reserve University
hxc501@case.edu

Jing Li
Electrical Engineering and Computer Science
Case Western Reserve University
jingli@cwru.edu

ABSTRACT

Modeling the behaviors of drug-target-disease interactions is crucial in the early stage of drug discovery and holds great promise for precision medicine and personalized treatments. The growing availability of new types of data on the internet brings great opportunity of learning a more comprehensive relationship among drugs, targets, and diseases. However, existing methods often consider drug-target interactions or drug-disease interactions separately, which ignores the dependencies among these three entities. Also, many of them cannot directly incorporate rich heterogeneous information from diverse sources.

In this work, we investigate the utility of tensor factorization to model the relationships of drug-target-disease, specifically leveraging different types of online data. Our motivation is two-fold. First, in human metabolic systems, many drugs interact with protein targets in cells to modulate target activities, which in turn alter biological pathways to promote healthy functions and to treat diseases. Instead of binary relationships of <drug, disease> or <drug, target>, a tighter triple relationships <drug, target, disease> should be exploited to better understand drug mechanism of actions (MoAs). Second, medical data could be collected from different sources (i.e., drug's chemical structure, target's sequence, or expression measurements). Therefore, effectively exploiting the complementarity among multiple sources is of great importance. Our method elegantly explores a <drug, target, disease> tensor together with complementarity among different data sources, thus improves prediction accuracy. We achieve this goal by formulating the problem into a coupled tensor-matrix factorization problem and directly optimize it on the nonlinear manifold. Experimental results on real-world datasets show that the proposed model outperforms several competitive methods. Our model opens up opportunities to use large Web data to predict drugs' MoAs in pharmacological studies.

CCS CONCEPTS

• **Computing methodologies** → **Factorization methods**; • **Applied computing** → **Health informatics**.

KEYWORDS

Tensor factorization; Multi-view learning; Manifold optimization; Grassmann manifold; Drug discovery; Disease analysis;

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313476

ACM Reference Format:
Huiyuan Chen and Jing Li. 2019. Modeling Relational Drug-Target-Disease Interactions via Tensor Factorization with Multiple Web Sources. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308558.3313476

1 INTRODUCTION

Targeted therapies and personalized treatments are the most promising approaches to treat complex human diseases such as cancer. Clear understanding of drugs' mechanism of actions (MoAs) is a critical step in drug discovery [21]. Traditional high-throughput screening methods are desirable to identify a new drug against a chosen target (most time a druggable protein) for a special disease. However, the process has been costly and lengthy. A conservative estimate is that it takes $ 2.87 billion and more than 10 years to bring a new drug into market [26]. On the other hand, statistical and machine learning models provide an alternative way to accelerate the process of understanding drugs' MoAs by mining bioinformatics, cheminformatics, and Web data sources [3, 4, 7, 16, 25, 32, 37]. Learning tasks have been proposed from different angles, such as exploring drug behaviors [15, 38, 46], assessing target activities [19, 34, 37], and understanding disease models [33, 44, 47]. Two of the most prominent formulations among these recent advancements are the drug-target [16] prediction and drug-disease prediction, also known as drug repositioning [25]. In these two tasks, researchers have attempted to collect a variety of omics data from scientific literature and online Web sources, and discover new interactions between drugs and targets/diseases through different statistical models [9, 24, 35, 46]. For example, a network-based inference method was proposed to infer new targets for known drugs by leveraging the drug-target bipartite network [13]. The method Decagon modeled drug polypharmacy side-effect via graph convolutional networks [46]. Another approach, PREDICT, integrated multiple drug-drug and disease-disease similarities to infer potential drug-disease interactions [20]. Approaches based on multiple kernel learning were developed to predict drug-target [30] or drug-disease [9] relationships, by integrating additional heterogeneous information sources.

However, despite the obvious connections, most existing work treat drug-disease and drug-target predictions as two independent tasks, which is viewed as a major shortcoming. Indeed, the therapeutic effect of drugs on a disease is through their abilities to bind and to modulate biological targets that involve the disease pathways, which in turn promote healthy function of the metabolic system and cure the disease [21]. Therefore, instead of a binary relationship such as <drug, disease> or <drug, target>, a strong

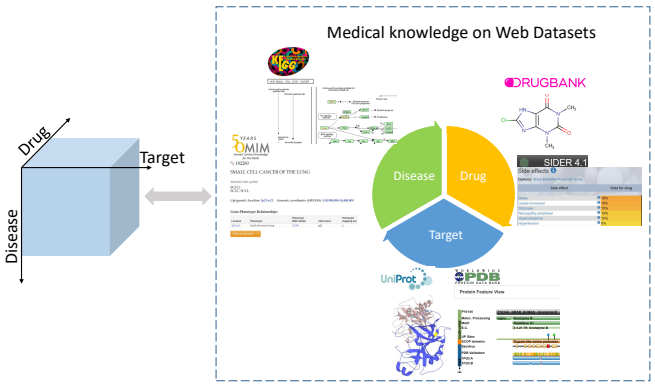

Figure 1: Illustration of DTD. Our model jointly explores the $drug \times target \times disease$ tensor along with rich existing medical knowledge on the Web.

triple relationship <drug, target, disease> should be considered to better model their interplay. In addition, the growing availability of new types of data on the web (Figure 1) brings new opportunities to learn a more comprehensive relationship among drugs, targets, and diseases [3, 4, 11, 30, 32, 45]. Incorporating such heterogeneous information can significantly improve our understanding of the underlying biological processes. For example, joint analysis of drugs' chemical structures, drugs' side-effects, and protein-protein networks can improve success rates of finding novel drug-target interactions [20, 30, 45].

In this work, we propose a tensor completion method, termed **DTD**, to model **D**rug-**T**arget-**D**isease interactions, with the help of existing data from the Web. DTD explicitly learns a third-order $drug \times target \times disease$ tensor using Tucker Decomposition [23]. In addition, to alleviate the data sparsity issue, DTD incorporates multiple auxiliary information sources of drugs, targets, and diseases. For example, in Figure 1, in addition to $drug \times target \times disease$ tensor, there are rich data on the Web describing drugs (e.g., chemical structures in DrugBank and side-effects in SIDER database), targets (e.g., protein sequence in Uniprot), and diseases (phenotype description in OMIM or KEGG). The tensor and multiple feature matrices are coupled in the "drug", "target", and "disease" mode, respectively. These feature matrices from Web sources are very useful to learn extra static information about each mode of the tensor. Fusing those datasets together can lead to better interpretations of the complex biological processes. To achieve this goal, DTD further explores the correlations among the latent matrices from the tensor and these multiple data sources via a coupled tensor-matrix factorization. This ensures that knowledge in the tensor aligns more closely with existing medical knowledge of each of the entities.

Another critical challenge is how to solve the tensor completion problem effectively. Because of the nonlinear and non-convex orthogonality constraints in the tensor Tucker model, the solutions of the popular HOSVD and HOOI Euclidean solvers are not unique, which makes it hard to couple with auxiliary information [23]. Motivated by the fast growing Riemannian optimization [1, 22, 43], we cast the coupled tensor-matrix factorization problem as a nonlinear program with the factor matrices constrained to the Grassmann manifold. Rather than a special non-uniqueness solution in Euclidean solvers, DTD obtains an equivalence class of matrices on Grassmann manifold, which leads to more meaningful subspace representations of factor matrices and can be well coupled with auxiliary information. Moreover, empirical studies have shown that nonlinear Riemannian solvers are significantly faster comparing to the Euclidean solvers [22, 43]. Our contributions are summarized as follows.

- We investigate the relationships of drug-target-disease by utilizing a tensor completion method. Integration of all these biological entities naturally leads to more meaningful interpretations of the results.
- We further explore the correlations among latent matrices from $drug \times target \times disease$ tensor and multiple data sources via a coupled tensor-matrix factorization, which ensures that the sparse tensor is supplemented using existing knowledge of the three entities.
- We directly optimize the tensor completion problem by adopting an optimization technique on Grassmann manifold, which obtains more reliable subspace representations of the tensor and can be well coupled with spectral embeddings of auxiliary information.
- We conduct extensive experiments on real-life datasets. Experimental results demonstrate DTD outperforms several state-of-the-art comparative models. It shows that DTD can effectively utilize rich information on the Web to ease the issue of sparsity.

The rest of the paper is organized as follows. Sec. 2 gives the background and task. Sec. 3 and Sec. 4 introduce DTD method in details. Sec. 5 evaluates the experimental performance. Sec. 6 provides a short review of related work. Sec. 7 concludes the paper.

2 BACKGROUND AND TASK DESCRIPTION

2.1 Tensor Algebra

In this study, we follow the notations introduced by Kolda and Bader [23]. We denote matrices by boldface uppercase letters (e.g., $\mathbf{A}$). *Tensors* are multidimensional arrays that extend the concept of matrices, and they are represented by boldface caligraphic letters (e.g., $\boldsymbol{\mathcal{X}}$). The *order* of a tensor is the number of its dimensions, also known as ways or modes. A *fiber* is a vector extracted from a tensor by fixing every index but one. A *slice* is a matrix extracted from a tensor by fixing all but two indices. Note that an N-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ reduces to a vector when $N = 1$, and a matrix when $N = 2$. The $(i_1, \ldots, i_N)$-th element of $\boldsymbol{\mathcal{X}}$ is denoted as $\boldsymbol{\mathcal{X}}_{i_1,\ldots,i_N}$. *Matricization*, also known as unfolding or flattening, is the process of reordering the elements of a tensor into a matrix. The mode-n matricization of an N-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is represented as $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times I_1 \ldots I_{n-1} \times I_{n+1} \ldots I_N}$ and is arranging the mode-n fibers of the tensor as columns of the long matrix. The *n-mode matrix product* of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with a matrix $\mathbf{U} \in \mathbb{R}^{J \times I_n}$ is denoted as $\boldsymbol{\mathcal{X}} \times_n \mathbf{U}$ with size of $I_1 \cdots \times I_{n-1} \times J \times I_{n+1} \cdots \times I_N$. We also give the definition of Tucker decomposition and coupled tensors for third-order tensors.

Definition 1. ***(Tucker Decomposition)***. *The Tucker decomposition is a form of higher-order PCA. It decomposes a tensor into a core tensor multiplied by a matrix along each mode. For a three-way tensor* $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$*, its Tucker decomposition is*

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$

where $\mathbf{A} \in \mathbb{R}^{I \times R_1}$, $\mathbf{B} \in \mathbb{R}^{J \times R_2}$ *and* $\mathbf{C} \in \mathbb{R}^{K \times R_3}$ *are the factor matrices (which are usually orthogonal) and can be regarded as the principal components in each mode. The tensor* $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ *is the core tensor and captures interactions between factor matrices.*

Definition 2. ***(Coupled Tensors)***. *If a tensor shares one or more modes with other matrices or other tensors, they are called coupled tensors or coupled tensor-matrices. For example, in a recommender system, a triple relationship user* $\times$ *movie* $\times$ *review tensor and a user* $\times$ *movie rating matrix are coupled because they share user and movie modes.*

2.2 Task Description

Our aim is to infer potential drug-target-disease interactions for rational drug repositioning. We formulate the task as a tensor completion problem. To be specific, the input can be organized as a three-way *drug* $\times$ *target* $\times$ *disease* tensor $\mathcal{X}$ of size $n_1 \times n_2 \times n_3$, in which n_1, n_2, and n_3 denote the number of drugs, targets, and diseases, respectively. An entry $\mathcal{X}_{ijk} = 1$ if drug i binds to target j and treats disease k. Otherwise, the entries are set to 0. In practice, we can only observe part of the tensor $\mathcal{X}$ and this partially observed tensor is denoted as $\mathcal{T}$. Our goal is to predict the potential concurrences of <drug, target, disease>, which can be achieved by completing tensor $\mathcal{X}$ given the incomplete tensor $\mathcal{T}$.

In real-world applications, tensor $\mathcal{T}$ is often sparse with a large number of unknown entries. Recovering tensor $\mathcal{X}$ is challenging when relying only on the observed tensor $\mathcal{T}$. Fortunately, for *drug* $\times$ *target* $\times$ *disease*, there exist many additional data sources on the Web to describe drugs, targets, and diseases. For example, for most drugs, their chemical structures and side-effects can be obtained from online databases, which represent different views of drugs. Datasets regarding targets and diseases are also available online. In most cases, one can further summarize such auxiliary data as drug-drug, target-target, and disease-disease similarity/kernel matrices, which can be incorporated in the learning process. Formally, let $\mathcal{S}_A = \{\mathbf{S}_A^{(1)}, \ldots, \mathbf{S}_A^{(n_a)}\}$ denote the similarity matrices constructed from n_a views of drugs. $\mathcal{S}_B = \{\mathbf{S}_B^{(1)}, \ldots, \mathbf{S}_B^{(n_b)}\}$ and $\mathcal{S}_C = \{\mathbf{S}_C^{(1)}, \ldots, \mathbf{S}_C^{(n_c)}\}$ are defined similarly from n_b views of targets and n_c views of diseases, respectively. All of them are assumed to be symmetric and non-negative. The details of their constructions will be presented in Sec. 5. The proposed DTD framework jointly explores the main tensor $\mathcal{X}$ together with multi-view auxiliary information $\mathcal{S}_A$, $\mathcal{S}_B$ and $\mathcal{S}_C$ to predict more meaningful interactions of drug-target-disease.

3 THE DTD MODEL

In this section, we propose an effective coupled tensor-matrix factorization formulation named DTD and show how different data sources can be included in a principled way. Notations used throughout the paper are summerized in Table 1.

Table 1: Main Notation

Symbol	Description
$\mathcal{X}$, $\mathcal{T}$	Full recovery tensor and observed tensor
$\mathcal{G}$	Core tensor in Tucker model
$\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$	Drug, target, disease factor matrix
$\mathbf{L}_A^{(i)}$, $\mathbf{A}^{(i)}$	Laplacian matrix and its spectral embedding from i-th view of drug auxiliary information
$\mathbf{L}_B^{(j)}$, $\mathbf{B}^{(j)}$	Laplacian matrix and its spectral embedding from j-th view of target auxiliary information
$\mathbf{L}_C^{(k)}$, $\mathbf{C}^{(k)}$	Laplacian matrix and its spectral embedding from k-th view of disease auxiliary information

3.1 Recover the Main Tensor

To complete the tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ described in Sec. 2.2, we adopt the Tucker Decomposition (Definition 1), which can be represented by the following optimization problem:

$$\min_{\mathcal{X}, \mathcal{G}, \mathbf{A}, \mathbf{B}, \mathbf{C}} \|\mathcal{X} - \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}\|_F^2 \quad s.t. \quad \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{T}) \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm. Recall that $\mathcal{G}$ is the core tensor; $\mathbf{A} \in \mathbb{R}^{n_1 \times r_1}$, $\mathbf{B} \in \mathbb{R}^{n_2 \times r_2}$, and $\mathbf{C} \in \mathbb{R}^{n_3 \times r_3}$ are the factor matrices with respect to the drug, target, and disease mode. Ω is the set that contains the indices of observed entities and $\mathcal{P}_\Omega(\cdot)$ keeps the entries in Ω and zeros out others [22, 41]. The equality constraint ensures that the corresponding elements of the recovering tensor $\mathcal{X}$ should match the observed elements in tensor $\mathcal{T}$.

In addition, the factor matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are orthogonal matrices, i.e., $\mathbf{A}^T\mathbf{A} = \mathbf{B}^T\mathbf{B} = \mathbf{C}^T\mathbf{C} = \mathbf{I}$ and $\mathbf{I}$ is the identity matrix. The orthogonal constraints ensure that matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are well defined on the so-called *Stiefel manifold* [1], which contains a set of $n \times m$ orthonormal matrices and is defined as:

$$St(m, n) = \{\mathbf{U} \in \mathbb{R}^{n \times m} | \mathbf{U}^T\mathbf{U} = \mathbf{I}_m\}$$

We will show the advantage of Stiefel manifold when coupling the tensor with auxiliary information.

3.2 Coupled with Auxiliary Information

To incorporate multi-view auxiliary information, we adopt the idea of spectral clustering, due to its flexibility and ease of implementation [28, 31]. We first give a brief introduction of spectral clustering. Suppose $\mathbf{S} \in \mathbb{R}^{N \times N}$ is the similarity/affinity matrix for N objects, where $\mathbf{S}_{ij}$ measures the similarity between object i and object j. One can then compute the normalized Laplacian matrix: $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$, in which $\mathbf{D}$ is the diagonal matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{N} \mathbf{S}_{ij}$. The spectral clustering is to solve the following optimization:

$$\min_{\mathbf{U} \in \mathbb{R}^{N \times k}} \quad \langle \mathbf{U}\mathbf{U}^T, \mathbf{L} \rangle \quad s.t. \quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the matrix inner product. $\mathbf{U}$ can be regarded as the low-dimensional spectral embedding of N objects. For clustering task, the k-means algorithm can be then applied to $\mathbf{U}$ to get the clustering indicators. Furthermore, the orthogonal constraint indicates that the spectral embedding $\mathbf{U}$ is also well defined on the Stiefel manifold. Therefore, it is reasonable to jointly consider the factor matrices in Eq. (1) together with the spectral embedding in Eq. (2) when performing the coupled tensor-matrix factorization in the sense that both of embeddings are on the Stiefel manifold.

For the drug mode of tensor $\mathcal{X}$, considering drug factor matrix $\mathbf{A} \in St(r_1, n_1)$ and its multi-view auxiliary information $\mathcal{S}_A = \{\mathbf{S}_A^{(1)}, \ldots, \mathbf{S}_A^{(n_a)}\}$, we extend the single-view spectral embedding to the follow multi-view co-training optimization function as [10, 28]:

$$\min_{\mathbf{A}, \mathbf{A}^{(i)} \in St(r_1, n_1)} \sum_{i=1}^{n_a} (\langle \mathbf{A}^{(i)} \mathbf{A}^{(i)^T}, \mathbf{L}_A^{(i)} \rangle + \|\mathbf{A}^{(i)} \mathbf{A}^{(i)^T} - \mathbf{A}\mathbf{A}^T\|_F^2) \quad (3)$$

where $\mathbf{L}_A^{(i)}$ is the normalized Laplacian matrix of $\mathbf{S}_A^{(i)}$ and $\mathbf{A}^{(i)}$ is the corresponding spectral embedding. The key idea behind Eq. (3) is that all the spectral embeddings $\mathbf{A}^{(i)}$ $(0 \le i \le n_a)$ from auxiliary information should be close to the drug factor matrix $\mathbf{A}$ since they all represent the same drugs. We achieve this by minimizing the disagreements $d(\mathbf{A}^{(i)}, \mathbf{A}) = \|\mathbf{A}^{(i)} \mathbf{A}^{(i)^T} - \mathbf{A}\mathbf{A}^T\|_F^2$. The reason for choosing $d(\mathbf{A}^{(i)}, \mathbf{A})$ is two-fold: (i) the reconstruction of similarity/kernel matrix $\mathbf{A}^{(i)} \mathbf{A}^{(i)^T}$ from spectral embedding is expected to be consistent with similarity $\mathbf{A}\mathbf{A}^T$ from tensor factor matrix, which is our assumption. (ii) we later show that the joint optimization is further defined on the Grassmann manifold, the quotient space of Stiefel manifold [1]. $d(\mathbf{A}^{(i)}, \mathbf{A})$ exactly measures the *geodesic distance* between two Grassmannian points $\mathbf{A}^{(i)}$ and $\mathbf{A}$ [14].

The same analysis can be applied to the target and disease factor matrices $\mathbf{B}$ and $\mathbf{C}$ with auxiliary information $\mathcal{S}_B$ and $\mathcal{S}_C$. We leave the details in the unified model in Eq. (4) later.

3.3 Overall Model

We propose the coupled tensor-matrix factorization model by combining the loss functions in Eq. (1) and Eq. (3). In addition to two functions, to better represent drugs, we further add l_1-norm on $\mathbf{A}\mathbf{A}^T$ [28]. We argue that this type of regularization does make sense. It has been shown that the new affinity/similarity matrix $\mathbf{A}^{(i)} \mathbf{A}^{(i)^T}$ in Eq. (3) implies the true membership of data clusters, and it is naturally *sparse* in an ideal case [28, 31]. Because the geodesic distance measurement $d(\mathbf{A}^{(i)}, \mathbf{A})$ is included in the overall objective function for minimization, the matrix $\mathbf{A}\mathbf{A}^T$ is also expected to be sparse. Similar arguments can be made for the target's and disease's factor matrices. Putting everything together, our method DTD tries to minimize the following onjective function:

$$\begin{aligned}
\min f = {} & \|\mathcal{X} - \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}\|_F^2 + \rho \cdot (\|\mathbf{A}\mathbf{A}^T\|_1 + \|\mathbf{B}\mathbf{B}^T\|_1 + \|\mathbf{C}\mathbf{C}^T\|_1) \\
& + \alpha \cdot \sum_{i=1}^{n_a} (\langle \mathbf{A}^{(i)} \mathbf{A}^{(i)^T}, \mathbf{L}_A^{(i)} \rangle + \|\mathbf{A}^{(i)} \mathbf{A}^{(i)^T} - \mathbf{A}\mathbf{A}^T\|_F^2) \\
& + \beta \cdot \sum_{j=1}^{n_b} (\langle \mathbf{B}^{(j)} \mathbf{B}^{(j)^T}, \mathbf{L}_B^{(j)} \rangle + \|\mathbf{B}^{(j)} \mathbf{B}^{(j)^T} - \mathbf{B}\mathbf{B}^T\|_F^2) \\
& + \gamma \cdot \sum_{k=1}^{n_c} (\langle \mathbf{C}^{(k)} \mathbf{C}^{(k)^T}, \mathbf{L}_C^{(k)} \rangle + \|\mathbf{C}^{(k)} \mathbf{C}^{(k)^T} - \mathbf{C}\mathbf{C}^T\|_F^2) \\
s.t. \quad & \mathbf{A}, \mathbf{A}^{(i)} \in St(r_1, n_1);\ \mathbf{B}, \mathbf{B}^{(j)} \in St(r_2, n_2);\ \mathbf{C}, \mathbf{C}^{(k)} \in St(r_3, n_3) \\
& \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{T});\ \mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}
\end{aligned} \quad (4)$$

where parameter ρ controls the sparsity of factor matrices. And α, β and γ represent the impact of auxiliary information on each mode of tensor, i.e., how important such knowledge is to improve the performance. At first glance, the objective function f is complicated with nonlinear and non-convex orthogonality constraints. Next, we provide an effective algorithm to solve the problem.

4 OPTIMIZATION ALGORITHM

In this section, we develop an alternating minimization algorithm to optimize the objective function in Eq. (4). To be specific, the objective function f is successively minimized with respect to one variable while fixing others until convergence. To deal with orthogonality constraints, we directly optimize it on a Grassmann manifold, which is an emerging topic in nonlinear programming that leverages the smooth geometry of the search space with guaranteed convergence [1, 22, 39, 43].

Updating core tensor $\mathcal{G}$: The objective with respect to $\mathcal{G}$ is:

$$\min f(\mathcal{G}) = \|\mathcal{X} - \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}\|_F^2$$

The core tensor $\mathcal{G}$ is obtained as the closed form solution

$$\mathcal{G} = \mathcal{X} \times_1 \mathbf{A}^T \times_2 \mathbf{B}^T \times_3 \mathbf{C}^T \quad (5)$$

Updating factor matrices A, B and C: For matrix $\mathbf{A}$, the objective $f(\mathbf{A})$ can be regarded as an unconstrained manifold optimization problem on the Stiefel manifold:

$$\begin{aligned}
\min_{\mathbf{A} \in St(r_1, n_1)} f(\mathbf{A}) = {} & \|\mathcal{X} - \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}\|_F^2 + \rho \|\mathbf{A}\mathbf{A}^T\|_1 \\
& + \alpha \sum_{i=1}^{n_a} \|\mathbf{A}^{(i)} \mathbf{A}^{(i)^T} - \mathbf{A}\mathbf{A}^T\|_F^2
\end{aligned}$$

The above optimization can be further converted into the following [23]:

$$\min_{\mathbf{A} \in St(r_1, n_1)} f(\mathbf{A}) = -\|\mathbf{A}^T \mathbf{W}\|_F^2 + \alpha \sum_{i=1}^{n_a} \|\mathbf{A}^{(i)} \mathbf{A}^{(i)^T} - \mathbf{A}\mathbf{A}^T\|_F^2 + \rho \|\mathbf{A}\mathbf{A}^T\|_1 \quad (6)$$

where $\mathbf{W} = \mathbf{X}_{(1)}(\mathbf{C} \otimes \mathbf{B})$, in which $\mathbf{X}_{(1)}$ is the mode-1 matricization of tensor $\mathcal{X}$ and $\otimes$ is the Kronecker product. However, simply optimizing problem (6) on Stiefel manifold may result in identifiability issue [1]. We next analyze problem (6) more deeply.

Consider the r_1-order group $\mathcal{S}(r_1) = \{\mathbf{Q} \in \mathbb{R}^{r_1 \times r_1} | \mathbf{Q}^T \mathbf{Q} = \mathbf{I}\}$ that contains all the $r_1 \times r_1$ orthogonal matrices. With $\mathcal{S}(r_1)$, we can define an equivalent relation $\sim$ on the Stiefel manifold $St(r_1, n_1)$ in the sense that $\mathbf{A} \sim \mathbf{A}'$ if there exists a $\mathbf{Q} \in \mathcal{S}(r_1)$ such that $\mathbf{A} = \mathbf{A}'\mathbf{Q}$. The quotient space of Stiefel manifold $St(r_1, n_1)$ under this equivalence relation is exactly the *Grassmann manifold* $Gr(r_1, n_1)$ [1], which consists of all the r_1-dimensional subspaces in n_1-dimensional Euclidean space $\mathbb{R}^{n_1}$ $(0 \le r_1 \le n_1)$. Moreover, it is interesting to observe that for any $\mathbf{Q} \in \mathcal{S}(r_1)$, we have the following invariance property:

$$f(\mathbf{A}) = f(\mathbf{A}\mathbf{Q})$$

Above equivalence indicates that the function $f(\mathbf{A})$ is independent from the choice of basis spanned by $\mathbf{A}$ and it is thus well defined on the Grassmann manifolds. Instead of optimizing $f(\mathbf{A})$ on Stiefel manifold, a better strategy is thus to regard the problem (6) as an unconstrained Grassmann manifold optimization problem:

$$\min_{\mathbf{A} \in Gr(r_1, n_1)} f(\mathbf{A}) = \underbrace{-\|\mathbf{A}^T \mathbf{W}\|_F^2 + \alpha \sum_{i=1}^{n_a} \|\mathbf{A}^{(i)} \mathbf{A}^{(i)^T} - \mathbf{A}\mathbf{A}^T\|_F^2}_{\mathcal{J}_1(\mathbf{A})} + \rho \|\mathbf{A}\mathbf{A}^T\|_1 \quad (7)$$

Problem (7) can then be efficiently solved by using standard gradient descent algoirthms on the Grassmann manifold such as

Riemannian conjugate gradient descent algorithm or trust region algorithm [1]. For Grassmann manifold, its Riemannian gradient is the projection of Euclidean gradient into relevant tangent space of the manifold. We next compute the Euclidean gradient of function $f(\mathbf{A})$.

For the first two terms in the problem (7), we have:

$$\nabla \mathcal{J}_1(\mathbf{A}) = -2\mathbf{W}\mathbf{W}^T\mathbf{A} + \alpha \sum_{i=1}^{n_a}(4\mathbf{A}\mathbf{A}^T\mathbf{A} - 4\mathbf{A}^{(i)}{\mathbf{A}^{(i)}}^T\mathbf{A}) \tag{8}$$

The third term $\|\mathbf{A}\mathbf{A}^T\|_1$ in objective function (7) is not differentiable when the elements of $\mathbf{A}\mathbf{A}^T$ are zeros, we consider the sub-differential. According to the chain rule:

$$\mathrm{vec}\left(\frac{\partial \|\mathbf{A}\mathbf{A}^T\|_1}{\partial(\mathbf{A})}\right)^T = \mathrm{vec}(\mathrm{sgn}(\mathbf{A}\mathbf{A}^T))^T \frac{\partial \mathbf{A}\mathbf{A}^T}{\partial \mathbf{A}} \tag{9}$$

where $\mathrm{sgn}(\cdot)$ denotes sign function and $\mathrm{vec}(\cdot)$ is vectorize operator that stacks all columns of a matrix into a long vector. Also, from the partial equation: $\partial(\mathbf{A}\mathbf{A}^T) = (\partial\mathbf{A})\mathbf{A}^T + \mathbf{A}\partial(\mathbf{A}^T)$, we can get:

$$\begin{aligned}\partial\,\mathrm{vec}(\mathbf{A}\mathbf{A}^T) =& (\mathbf{A}\otimes\mathbf{I}_{n_1})\partial\,\mathrm{vec}(\mathbf{A}) + (\mathbf{I}_{n_1}\otimes\mathbf{A})\partial\,\mathrm{vec}(\mathbf{A}^T)\\ =& \left((\mathbf{A}\otimes\mathbf{I}_{n_1}) + \mathbf{K}_{(n_1^2,n_1^2)}(\mathbf{A}\otimes\mathbf{I}_{n_1})\right)\partial\,\mathrm{vec}(\mathbf{A})\\ =& (\mathbf{I}_{n_1^2} + \mathbf{K}_{(n_1^2,n_1^2)})(\mathbf{A}\otimes\mathbf{I}_{n_1})\partial\,\mathrm{vec}(\mathbf{A})\end{aligned}$$

From above equation, the derivative part in Eq. (9) is

$$\frac{\partial \mathbf{A}\mathbf{A}^T}{\partial \mathbf{A}} = (\mathbf{I}_{n_1^2} + \mathbf{K}_{(n_1^2,n_1^2)})(\mathbf{A}\otimes\mathbf{I}_{n_1})$$

where $\mathbf{I}_{n_1^2}$ is the identity matrix with size $n_1^2 \times n_1^2$ and $\mathbf{K}_{(n_1^2,n_1^2)}$ is the commutation matrix. Although both matrix $\mathbf{I}_{n_1^2}$ and $\mathbf{K}_{(n_1^2,n_1^2)}$ are large, they are very sparse with a large number of zeros. In practice, they can be encoded in a sparse matrix form, which can greatly save the computational space.

Define the column vector $\mathbf{d}$ as

$$\mathbf{d} = (\frac{\partial \mathbf{A}\mathbf{A}^T}{\partial \mathbf{A}})^T \mathrm{vec}(\mathrm{sgn}(\mathbf{A}\mathbf{A}^T)) \tag{10}$$

Combining Eq. (8) and (10), the Euclidean gradient of the objective function $f(\mathbf{A})$ is

$$\nabla f(\mathbf{A}) = \nabla \mathcal{J}_1(\mathbf{A}) + \rho \cdot \mathrm{ivec}(\mathbf{d}) \tag{11}$$

where $\mathrm{ivec}(\cdot)$ is the inverse vectorize operator, i.e., $\mathrm{ivec}(\mathrm{vec}(\mathbf{X})) = \mathbf{X}$. With the Euclidean gradient $\nabla f(\mathbf{A})$, the Riemannian gradient with respect to $\mathbf{A} \in Gr(r_1, n_1)$ can be computed as

$$\mathrm{grad}\, f(\mathbf{A}) = (\mathbf{I} - \mathbf{A}\mathbf{A}^T)\cdot\nabla f(\mathbf{A}) \tag{12}$$

With the Riemannian gradient, we can use the popular nonlinear ManOpt solver to solve problem (7) effectively [6].

For updating factor matrices $\mathbf{B}$ and $\mathbf{C}$, their objective function $f(\mathbf{B})$ and $f(\mathbf{C})$ share the similar optimization structure as $f(\mathbf{A})$. Therefore, they can be solved in the same way with corresponding Riemannian gradients. The details are omitted here.

Updating spectral embedding $\mathbf{A}^{(i)}$,$\mathbf{B}^{(j)}$ and $\mathbf{C}^{(k)}$: For matrix $\mathbf{A}^{(i)}$, the objective $f(\mathbf{A}^{(i)})$ is

$$\min_{\mathbf{A}^{(i)}\in St(r_1,n_1)} f(\mathbf{A}^{(i)}) = \langle \mathbf{A}^{(i)}{\mathbf{A}^{(i)}}^T, \mathbf{L}_A^{(i)}\rangle + \|\mathbf{A}^{(i)}{\mathbf{A}^{(i)}}^T - \mathbf{A}\mathbf{A}^T\|_F^2 \tag{13}$$

The objective function $f(\mathbf{A}^{(i)})$ is also invariant to $\mathbf{Q} \in \mathcal{S}(r_1)$, i.e., $f(\mathbf{A}^{(i)}) = f(\mathbf{A}^{(i)}\mathbf{Q})$. Therefore, problem (13) can be also regarded as an unconstrained Grassmann manifold optimization problem. The Euclidean gradient of the objective function $f(\mathbf{A}^{(i)})$ is

$$\nabla f(\mathbf{A}^{(i)}) = 2\mathbf{L}_A^{(i)}\mathbf{A}^{(i)} + 4\mathbf{A}^{(i)}{\mathbf{A}^{(i)}}^T\mathbf{A}^{(i)} - 4\mathbf{A}\mathbf{A}^T\mathbf{A}^{(i)} \tag{14}$$

The Riemannian gradient grad $f(\mathbf{A}^{(i)})$ can be obtained in the same way as in Eq. (12). The Riemannian trust-regions algorithm is then used to compute the optimal solution since the objective function $f(\mathbf{A}^{(i)})$ in Eq. (13) is smooth. The initial $n_1 \times r_1$ spectral embedding $\mathbf{A}^{(i)}$ can be approximated by the r_1 eigenvectors of $\mathbf{L}_A^{(i)}$ corresponding to the first r_1 smallest eigenvalues. The objective function $f(\mathbf{B}^{(j)})$ and $f(\mathbf{C}^{(k)})$ can be solved in a similar fashion.

Updating tensor $\mathcal{X}$: The optimization problem with respect to $\mathcal{X}$ is formulated as follows:

$$\min_{\mathcal{X}} \|\mathcal{X} - \mathcal{G}\times_1\mathbf{A}\times_2\mathbf{B}\times_3\mathbf{C}\|_F^2 \quad s.t. \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{T})$$

The optimal solution is given by:

$$\mathcal{X} = \mathcal{P}_\Omega(\mathcal{T}) + \mathcal{P}_{\Omega^c}(\mathcal{G}\times_1\mathbf{A}\times_2\mathbf{B}\times_3\mathbf{C}) \tag{15}$$

where Ω^c is the complement of Ω, i.e., the set of indexes of the unobserved elements.

According to above analysis, we summarize the algorithm for solving the tensor completion problem (4) in Algorithm 1.

Remark: Comparison with Euclidean solvers that usually deal with the orthogonality constraints by solving eigenvalue decomposition problem [23], our solver seeks to find concise subspace on the Grassmann manifold for all factor matrices. As pointed out in spectral clustering [31], when the leading eigenvalues are almost equal, the best spectral embedding is better determined by the subspace rather than a particular eigenvector as most of Euclidean solvers do. The same analysis can be analogously applied to solve Tucker decomposition in Euclidean space, i.e., using eigenvalue decomposition. The Riemannian solver is thus more interpretable in a straightforward way.

5 EXPERIMENTS

5.1 Datasets

We collect data from a variety of public sources. We first download the drug-disease associations from the Comparative Toxicogenomics Database [1] (CTD). The original dataset contains 1,048,547 pairs of drug-disease associations. Here, we only focus on those drugs with DrugBank [2] identifier and diseases with OMIM [3] identifier for conveniently integrating with auxiliary information on other public datasets. As discussed before, the drug's targets, usually proteins that drugs can bind, through which drugs interact with biological pathways and possibly change their behaviors and functions, are crucial in drug discovery. The drug's targets can be collected from the DrugBank database to obtain drug-target interactions. To get dense data, we only include those drugs that interact with at least two targets. We then merge the drug-disease and drug-target interactions into data schema $< drug, target, disease >$. The

[1] http://ctdbase.org/downloads/
[2] https://www.drugbank.ca/
[3] https://www.omim.org/

Algorithm 1: DTD

Input: $\mathcal{T}$, Ω, $\{\mathbf{S}_A^{(i)}\}_{i=1}^{n_a}$, $\{\mathbf{S}_B^{(j)}\}_{j=1}^{n_b}$, $\{\mathbf{S}_C^{(k)}\}_{k=1}^{n_c}$ and tol parameters α, β, γ ρ and rank (r_1, r_2, r_3).

Compute the all Laplacian matrix $\mathbf{L}_A^{(i)}$, $\mathbf{L}_B^{(j)}$ and $\mathbf{L}_C^{(k)}$
Initialize $\mathcal{X}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathcal{G}$ randomly, set $\mathcal{X}_\Omega = \mathcal{T}_\Omega$
Initialize $\mathbf{A}_0^{(i)}$ with r_1 eigenvectors of $\mathbf{L}_A^{(i)}$ corresponding to the first r_1 smallest eigenvalues. Same process to initialize $\mathbf{B}_0^{(j)}$ and $\mathbf{C}_0^{(k)}$
repeat
  Update $\mathcal{G}_{t+1}$ by Eq. (5)
  Compute Riemannian gradient grad $f(\mathbf{A}_t)$ by Eq. (12)
  Update $\mathbf{A}_{t+1}$ by using Conjugate-gradient solver
  Update $\mathbf{B}_{t+1}$ and $\mathbf{C}_{t+1}$ the same way as updating $\mathbf{A}_{t+1}$
  for $i \leftarrow 1$ *to* n_a **do**
    Update $\mathbf{A}_{t+1}^{(i)}$ by using trust-regions solver
  end
  for $j \leftarrow 1$ *to* n_b **do**
    Update $\mathbf{B}_{t+1}^{(j)}$ by using trust-regions solver
  end
  for $k \leftarrow 1$ *to* n_c **do**
    Update $\mathbf{C}_{t+1}^{(k)}$ by using trust-regions solver
  end
  Update $\mathcal{X}_{t+1}$ by Eq. (15)
until *Objective:* $\|f_{t+1} - f_t\|_F \leq tol$
return $\mathcal{X}$

Table 2: Dataset statistics

#drug	#target	#disease	#interaction	sparsity rate
450	708	1,267	188,479	0.047%

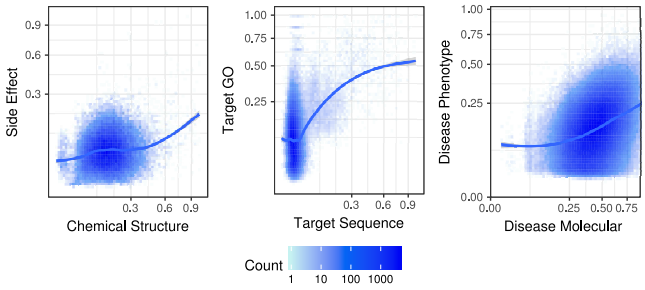

Figure 2: The pairwise scatter plots among different views of drugs, targets and diseases.

dataset contains 188,479 drug-target-disease interactions, involving 450 drugs, 708 targets and 1,267 diseases. Data statistics are shown in Table 2.

The data are then encoded into a $drug \times target \times disease$ tensor $\mathcal{X}$. An entry $\mathcal{X}_{ijk} = 1$ in the tensor indicates that drug i binds to target j and it can treat disease k; $\mathcal{X}_{ijk} = 0$ otherwise. The tensor is very sparse (with sparsity rate 0.047%). We further collect auxiliary information to align each mode of tensor with existing medical knowledge.

Auxiliary information: Following several previous studies [7, 9, 20, 30, 45], we collect auxiliary information from different datasets with respect to drugs, targets, and diseases. For drugs, we define two drug-drug similarities based on drugs' chemical structures and drugs' side effects [7]. For targets, two target-target similarities are considered based on target amino acid sequences and Gene Ontology (GO) terms, both of which can directly be obtain from Uniprot [4] database [20, 45]. For diseases, two disease-disease similarities are computed according to disease molecular profiles from biomedical literatures [8] and disease phenotypes on OMIM datasets [9].

Each view of data contains its own features and emphasizes different aspects. Different views may share some consistency and complementary properties. We begin our investigation by examining the strength of the associations among those different views of drugs, targets, and diseases. The scatter plots of the pairwise similarity scores are shown in Figure 2. We further assess the strength of their associations by calculating the correlation coefficient between two views of similarity scores, respectively. All correlations are positive with the correlation coefficient values of 0.148 (drug chemical structures vs. drug side effects), 0.372 (target sequences vs. target GO terms), and 0.374 (disease molecular profiles vs. disease phenotypes). Despite small values in some of the coefficients, all of them are significant because of the large sample sizes (e.g., 802, 011 data instances in the disease scatter plot). Therefore, it's reasonable to integrate the multi-view auxiliary information in the DTD framework and to improve the overall performance for modeling drug-target-disease interactions, as shown later.

5.2 Comparison Methods

As mentioned earlier, many existing methods can only handle the binary relationships of drug-target or drug-disease. We mainly compare with existing tensor completion approaches due to their abilities to represent the triple relationships of drug-target-disease. To show the effectiveness of the proposed DTD model, we compare with several existing models as described below.

- **Tucker decomposition** (TD): It only decomposes the tensor without any auxiliary information for each mode.
- **CMTF** [2]: CMTF constructs common latent factors shared by a tensor and single-view of auxiliary information by using coupled matrix-tensor CP-decomposition.
- **ConCMTF** [5]: A novel constrained tensor model with non-negativity, sparsity and orthogonality constraints. Similar to CMTF, it can incorporate single-view of auxiliary information but with Tucker tensor decomposition as base.
- **FaLRTC** [27]: FaLRTC can estimate the low rank structure by imposing the trace norm on its unfolding matrices and the algorithm is based upon alternating direction method of multipliers (ADMM).
- **TFAI** [29]: it recovers the tensor by incorporating within-mode auxiliary information and adopt alternating least squares algorithm to solve its problem.
- **Rubik** [41]: A novel knowledge-guided tensor factorization and completion framework to fit electronic health record data with non-negativity and sparsity constraints. It is also solved by ADMM algorithm.

[4] https://www.uniprot.org/

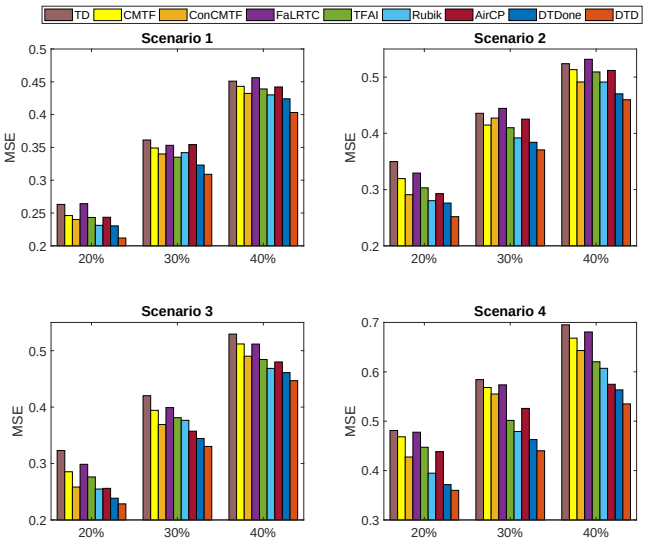

Figure 3: Comparison of recovery results over four scenarios with 20%, 30% and 40% test dataset.

- **AirCP** [18]: Another tensor model that integrates single-view auxiliary information using Laplacian regularization and it is optimized by ADMM algorithm.
- **DTDone**: A degraded version of our DTD model with single-view auxiliary information for each mode of tensor.

Note that the approaches TD and FaLRTC cannot integrate any auxiliary information when completing the tensor. The rest of comparison methods can only incorporate single-view auxiliary information for each mode. One has to either only include single-view for each mode, or a straightforward concatenation (e.g., average) of all similarity matrices into one. We tries both ways but the performance do not show much differences. The reason is that concatenation of all views may not be physically meaningful because each view has its owns specific statistical property. For these single-view models, we thus chose the drug chemical structure, target sequence, and disease phenotypes, all of which have long been considered valuable knowledge in drug discovery [20, 30, 45].

For all Tucker-based tensor models, we set the rank of core tensor $\mathcal{G}$, $r_i = 0.05n_i$ $(1 \le i \le 3)$; For CP-based tensor models, we set CP-rank $r = 0.05 \min(n_1, n_2, n_3)$. The regularization parameters of all comparison methods are tuned using the grid-based search algorithm for optimal performance. All methods use the same stopping threshold for a fair comparison, *i.e.*, the variation of two consecutive objective value is less than 10^{-4}. For the proposed DTD and DTDone model, we manually set the parameters $\alpha = \beta = \gamma = \rho = 0.01$. The impact of the regularization parameters on the performance of DTD will be discussed later.

5.3 Experimental Performance

To investigate the performance of different methods, we randomly selected a subset $\Omega' \in \Omega$ to be used as the unobserved elements (that is the test dataset), and evaluated the *mean squared errors* (MSE) between true values $\mathcal{T}_{\Omega'}$ and predicted values $\mathcal{X}_{\Omega'}$ [29]. We mainly consider the following random selection scenarios [18]: (i) Scenario 1 (random selection): we randomly select a fraction of elements across drug-target-disease tensor and assume that these are unobservable. (ii) Scenario 2 (random selection in drug mode): we randomly select a fraction of drugs, and assume that those drugs are completely unobservable across the whole tensor. (iii) Scenario 3 (random selection in target mode): similarly, we randomly select a fraction of targets. (iv) Scenario 4 (random selection in disease mode): A fraction of diseases are randomly selected. For each method, the experiment is repeated ten times independently and the average result is reported.

Figure 3 shows the performance of all the methods with different sampling ratios 20%, 30% and 40% (i.e., the fractions of test dataset). DTD consistently performs better than all the comparison methods within a wide range of sampling ratios. In addition, we have the following observations. First, the methods that integrate single-view or multi-view auxiliary information achieve better performance than TD and FaLRTC, indicating the importance of those auxiliary information. With the guidance of extra knowledge, these tensor models keep their performance even with a very sparse input tensor. Second, ConCMTF (Tucker decomposition) perform slightly better than CMTF (CP decomposition) in many situations. The main difference between them is that ConCMTF seeks to find as many non-overlapping structures as possible. Such non-overlapping latent structures are more concise and become specially favorable when jointly decomposing the tensor and matrices. The proposed DTD and DTDone also preserve such non-overlapping structures by imposing orthogonality constraint on factor matrices, which leads to better representations of data. Third, the proposed DTDone further achieves better performance than ConCMTF with an average improvement of 5.76%. As discussed before, the solutions of DTDone, both factor matrices and spectral embeddings, are exactly on the Grassmann manifold and can be well coupled together with each other to obtain better performance. And such advantages make DTDone generally outperform other single-view tensor completion methods such as TEAI, Rubik and AirCP. Finally, the proposed DTD method integrating all available auxiliary information achieves better performance than those only integrating single-view auxiliary information, which illustrates that DTD successfully makes use of all useful information sources to perform effective recovery for the drug-target-disease tensor. In summary, DTD can effectively predict potential interactions of drug, target and disease by leveraging multiple auxiliary data sources and has great potential to accelerate the drug development.

Top-k Prediction: In clinics, given a disease, it is critical to know which drugs can treat this disease as well as the targets involved in the disease pathway. These pairs of <drug, target> related to a special disease are very important in personalized treatments. We further evaluate the performance of top-k prediction for Scenario 4. Recall that we randomly select a subset of diseases and remove all of their interactions in the Scenario 4. For each disease, we can predict a top-k list of <drug, target> candidates.

We adopt Precision@k and Recall@k as our evaluation metrics for different methods. Both of them have been widely used to evaluate the quality of top-k predictions [18, 36]. In our experiments,

Table 3: Precision@k and Recall@k for different methods.

Methods	Prec@5	Rec@5	Prec@10	Rec@10
CMTF	0.287	0.347	0.264	0.283
ConCMTF	0.340	0.371	0.270	0.305
TEAI	0.334	0.359	0.254	0.284
Rubik	0.291	0.362	0.269	0.296
AirCP	0.289	0.352	0.261	0.298
DTDone	0.342	0.378	0.273	0.316
DTD	**0.353**	**0.394**	**0.292**	**0.331**

Table 4: Top 10 novel triple relationship of (drug→target→disease) by DTD model.

Carbamazepine→ Nuclear receptor subfamily 1 group I member 2 → Osteoporosis
Testosterone→ Estrogen receptor→Myocardial infarction
Nefazodone→D(2) dopamine receptor→Schizophrenia
Raloxifene→Estrogen receptor→Obesity
Fenofibrate→Matrix metalloproteinase→Psoriatic arthritis
Acetazolamide→Estrogen receptor→Amyotrophic lateral sclerosis
Raloxifene→Androgen receptor→Breast cancer
Amoxapine→Potassium voltage-gated channel subfamily H member 2 →Hepatocellular carcinoma
Nefazodone→Histamine H1 receptor→Schizophrenia
Promethazine→Muscarinic acetylcholine receptor M3→Obesity

we evaluate those tensor models with auxiliary information and test for k at 5 and 10 for both precision and recall.

Table 3 shows the top-k prediction performance of all the methods with 20% fractions of test dataset. Overall, the proposed DTD model performs the best among all methods in terms of both precision and recall with different values of k. The trend is very similar to the results based on the MSE metric. For example, DTD performs better than CMTF with an average improvement of 16.8% in precision and 15.3% in recall. We attribute the poor performance of the CMTF approach to the sparseness of tensor and its weak abilities of coupling tensor with auxiliary matrices (e.g., overlap structure). Moreover, DTD has a better performance than DTDone, indicating the superiority of a tensor model integrating rich multi-view auxiliary information. All these results illustrate that the proposed method can successfully predict top-k <drug, target> pairs by exploiting the compatible and complementary information from multi-view data sources.

Given these encouraging results, we use DTD model to predict novel triple relationships of drug-target-disease by leveraging all the observed data. The top-10 novel triple relationships are listed in Table 4. The results can then be evaluated by domain experts to see whether such interactions are clinically meaningful. Also, such top-k candidates can be further validated *in vitro* and *in vivo*. In summary, our proposed model efficiently predict potential interaction candidates with high accuracy, providing a systematic approach to narrow down the search space for further wet-lab investigations.

5.4 Parameter Studies

We further analyze the effect of key hyperparameters, the rank of core tensor (r_1, r_2, r_3) and the regularized parameters α, β, γ, and ρ.

5.4.1 The impact of tensor rank: In Tucker-based tensor decomposition, the hyperparameter (r_1, r_2, r_3) controls the rank of the core tensor $\mathcal{G}$, which also decides the number of latent features for each mode, i.e., **A**, **B** and **C**. In contrast to Tucker model, the CP decomposition only have one hyperparameter r, the CP-rank. In the experiments, we set the rank $r_i = \delta \cdot n_i$ $(1 \le i \le 3)$ for Tucker-based model and $r = \delta \cdot \min(n_1, n_2, n_3)$ for CP-based model. The δ is the ratio varying within $[0.05, 0.1, 0.15, 0.2, 0.25]$. We then evaluate all methods for scenario 4 with 20% as the test dataset. Figure 4 shows that our model benefits slightly from larger numbers of latent dimensions in terms of MSE and Prec@5 metric. But generally speaking, both DTD and DTDone are fairly robust with respect to δ.

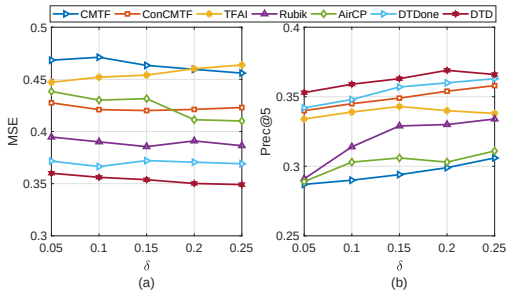

Figure 4: Effect of the tensor rank in scenario 4.

5.4.2 The impact of regularized parameters. Finally, we explore the impact of the regularization parameters on the quality of tensor completion. Recall that α, β and γ control the contributions of auxiliary information of drugs, targets, and disease, respectively. ρ controls the sparsity of factor matrices. In order to better understand the effect of these parameters, we vary their values in the range $[0.001, 0.01, 0.1, 1, 10]$, and then evaluate DTDone and DTD for scenario 4 with 20% as testing data. When studying one variable (e.g., α), we fix all the rest variables to be 0.01. As shown in Figure 5, DTDone and DTD are stable over a wide range of α, β and γ in terms of MSE. Specifically, a relatively low MSE can be achieved when those regularized parameters are around 0.01.

6 RELATED WORK

We review related research on computational predictions of drugs, targets, and diseases, as well as on tensor completion for high-dimensional structured data.

Modeling drug-target-disease. Treating human diseases caused by complex biological processes involves activities of many biological entities such as drugs and targets. Methods in computational pharmacology aim to find associations among those entities, and understand drug's MoAs [21]. Two excellent surveys [25], [16] provide a very detailed overview of different computational methods

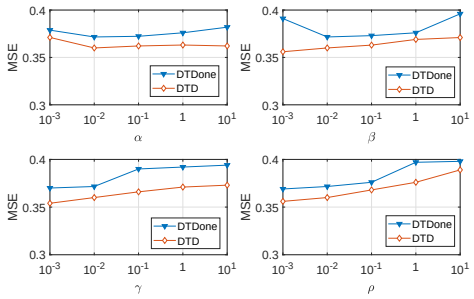

Figure 5: Effect of regularized parameters:α, β, γ and ρ .

for predicting drug-disease and drug-target relationships, respectively. Among existing approaches, one very popular method is network-based inference models, e.g., a bipartite network consisting of one layer for drugs and one layer for diseases (targets). Different machine learning algorithms, such as random-walk [12], matrix factorization [17], and support vector machine [42], have been applied to predict novel interactions of drug-disease (or target). In addition to network topology, existing medical knowledge, such as a variety of omics data, on the Web can be incorporated to better understand complex human metabolic systems [9, 30, 45, 46]. For example, a multiple kernel learning method has been developed to predict drug-target relationships by integrating multiple similarities of drugs and targets [30]. Decagon models drug polypharmacy side effects via graph convolutional networks with additional drug-drug and protein-protein interactions [46].

However, current studies generally considered drug-disease and drug-target predictions as two independent tasks and the relationships of drug-target-disease is typically ignored. There exist several studies incorporating target information in the task of drug-disease predictions [13, 40, 46]. However, their extensions are unsatisfactory because their goals are still on modeling binary relationships, not on the triple drug-target-disease patterns we aim to learn.

Tensor completion. Tensors are very powerful in real-world applications because of their ability to represent multi-aspects or high-dimensional data [23]. In many applications, we are often interested in analyzing tensors together with matrices from additional information, such as computational phenotyping [41], recommender system [18], and time-evolving topic discovery [5]. This coupled tensor-matrix factorization method has been developed over the years. For example, Acar et. al. proposed a gradient-based algorithm for coupled matrix-tensor factorization [2]. Narita et. al. incorporated valuable auxiliary information to further improve the quality of tensor recovery [29]. Wang et. al. proposed a knowledge guided tensor factorization model for health data analysis [41]. Ge et. al. proposed a spatiotemporal dynamics recovery framework, which could capture the latent relationships among locations, memes, and times by coupled factorization [18]. However, none of these methods are guaranteed to non-overlapping structures in both tensor and matrices and can only integrate single-view of additional information for *each mode* of tensor. Some models [5, 23, 29] try to impose orthogonality constraint on factor matrices. However, their solutions are usually not unique, which is hard to couple with auxiliary information. Recently, concrete tensor/matrix representations on the manifold is a fast growing research topic [1, 22, 43]. Kasai et. al. recently developed a novel Riemannian manifold preconditioning approach for tensor completion [43], but they did not take the auxiliary information into account. The proposed approach is innovative and properly addresses the challenges in coupled tensor-matrix factorization with multi-view auxiliary information.

7 CONCLUSION

Modeling drug-target-disease interactions is important for understanding drugs' MoAs in drug development. The problem has conventionally been addressed separately by considering associations of drug-target or drug-disease, and most existing methods cannot leverage intrinsic interactions among these three biological entities. Here we present a novel approach DTD, which explicitly explores a three-way drug-target-disease tensor via coupled tensor-matrix factorization. Completing such tensor is challenging because of its large sparsity. DTD elegantly integrates multiple Web data to align existing knowledge with the tensor. With the guidance of auxiliary information, DTD infers the concurrence of drug-target-disease more accurately than do baselines. Another distinguishing aspect of DTD is that it directly optimizes on the Grassmann manifold, which is more effective than Euclidean solvers. Experimental results on a real-world dataset demonstrate the effectiveness and efficiency of the proposed method.

The proposed DTD can easily incorporate additional domain knowledge and can be extended to detect high-order tensor (e.g., four-way *drug-target-gene-disease* tensor) with relatively little effort. Future work should aim to extend this model to simulate human metabolic systems by considering more biological entities (e.g., gene) as well as other domain knowledge (e.g., gene expression data). Deep understanding of those entities opens up opportunities to use rich Web data to assist follow-up analysis via formal pharmacological studies.

Acknowledgements

This work has been supported in part by NIH grants R03HG008632, NSF CCF1815139, and a Faculty Investment Fund from CWRU.

REFERENCES

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. 2009. *Optimization algorithms on matrix manifolds.* Princeton University Press.
[2] Evrim Acar, Tamara G Kolda, and Daniel M Dunlavy. 2011. All-at-once optimization for coupled matrix and tensor factorizations. *arXiv preprint arXiv:1105.3422* (2011).
[3] Benjamin M Althouse, Samuel V Scarpino, Lauren Ancel Meyers, John W Ayers, Marisa Bargsten, Joan Baumbach, John S Brownstein, Lauren Castro, Hannah Clapham, Derek AT Cummings, et al. 2015. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science* 4, 1 (2015), 17.
[4] Matheus Araujo, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto. 2017. Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *WebSci.*
[5] Sanaz Bahargam and E Papalexakis. 2018. Constrained Coupled Matrix-Tensor Factorization and its Application in Pattern and Topic Detection. In *ASONAM.*
[6] Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. 2014. Manopt, a Matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research* 15, 1 (2014), 1455–1459.
[7] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. 2008. Drug target identification using side-effect similarity. *Science* 321, 5886 (2008), 263–266.

[8] Horacio Caniza, Alfonso E Romero, and Alberto Paccanaro. 2015. A network medicine approach to quantify distance between hereditary disease modules on the interactome. *Scientific reports* 5 (2015), 17658.

[9] Huiyuan Chen and Jing Li. 2017. A flexible and robust multi-source learning algorithm for drug repositioning. In *BCB*. ACM.

[10] Huiyuan Chen and Jing Li. 2017. Learning Multiple Similarities of Users and Items in Recommender Systems. In *ICDM*.

[11] Huiyuan Chen and Jing Li. 2018. DrugCom: Synergistic Discovery of Drug Combinations Using Tensor Decomposition. In *ICDM*.

[12] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. 2012. Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems* 8, 7 (2012), 1970–1978.

[13] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology* 8, 5 (2012), e1002503.

[14] Alan Edelman, Tomás A Arias, and Steven T Smith. 1998. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications* 20, 2 (1998), 303–353.

[15] Tewodros Eguale, David L Buckeridge, Aman Verma, Nancy E Winslade, Andrea Benedetti, James A Hanley, and Robyn Tamblyn. 2016. Association of off-label drug use and adverse drug events in an adult population. *JAMA internal medicine* 176, 1 (2016), 55–63.

[16] Ali Ezzat, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. 2018. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics* (2018).

[17] Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. 2017. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 14, 3 (2017), 646–656.

[18] Hancheng Ge, James Caverlee, Nan Zhang, and Anna Squicciarini. 2016. Uncovering the spatio-temporal dynamics of memes in the presence of incomplete information. In *CIKM*.

[19] Graciela Gonzalez, Juan C Uribe, Luis Tari, Colleen Brophy, and Chitta Baral. 2007. Mining gene-disease relationships from biomedical literature: weighting protein–protein interactions and connectivity measures. In *Biocomputing 2007*. World Scientific, 28–39.

[20] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* 7, 1 (2011), 496.

[21] Alexander S Hauser, Misty M Attwood, Mathias Rask-Andersen, Helgi B Schiöth, and David E Gloriam. 2017. Trends in GPCR drug discovery: new agents, targets and indications. *Nature Reviews Drug Discovery* 16, 12 (2017), 829.

[22] Hiroyuki Kasai and Bamdev Mishra. 2016. Low-rank tensor completion: a Riemannian manifold preconditioning approach. In *ICML*.

[23] Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM review* 51, 3 (2009), 455–500.

[24] Stephanie N Lewis, Elaine Nsoesie, Charles Weeks, Dan Qiao, and Liqing Zhang. 2011. Prediction of disease and phenotype associations from genome-wide association studies. *PLoS One* 6, 11 (2011), e27175.

[25] Jiao Li, Si Zheng, Bin Chen, Atul J Butte, S Joshua Swamidass, and Zhiyong Lu. 2015. A survey of current trends in computational drug repositioning. *Briefings in bioinformatics* 17, 1 (2015), 2–12.

[26] Craig W Lindsley. 2014. New statistics on the cost of new drug development and the trouble with CNS drugs.

[27] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. 2013. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2013), 208–220.

[28] Canyi Lu, Shuicheng Yan, and Zhouchen Lin. 2016. Convex sparse spectral clustering: Single-view to multi-view. *IEEE Transactions on Image Processing* 25, 6 (2016), 2833–2843.

[29] Atsuhiro Narita, Kohei Hayashi, Ryota Tomioka, and Hisashi Kashima. 2012. Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery* 25 (2012).

[30] André CA Nascimento, Ricardo BC Prudêncio, and Ivan G Costa. 2016. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics* 17, 1 (2016), 46.

[31] Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *NIPS*.

[32] Michael J Paul and Mark Dredze. 2013. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 168–178.

[33] Nicola Perra, Duygu Balcan, Bruno Gonçalves, and Alessandro Vespignani. 2011. Towards a characterization of behavior-disease models. *PloS one* 6, 8 (2011), e23084.

[34] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. 2013. A large-scale evaluation of computational protein function prediction. *Nature methods* 10, 3 (2013), 221.

[35] Bisakha Ray, Elodie Ghedin, and Rumi Chunara. 2016. Network inference from multimodal data: a review of approaches from infectious disease transmission. *Journal of biomedical informatics* 64 (2016), 44–54.

[36] Steffen Rendle, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. 2009. Learning optimal ranking with tensor factorization for tag recommendation. In *KDD*.

[37] Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. 2017. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery* 16, 1 (2017), 19.

[38] Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics* 53 (2015), 196–207.

[39] Qiong Wang, Junbin Gao, and Hong Li. 2017. Grassmannian manifold optimization assisted sparse spectral clustering. In *CVPR*.

[40] Wenhui Wang, Sen Yang, Xiang Zhang, and Jing Li. 2014. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 30, 20 (2014), 2923–2930.

[41] Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel Kho, You Chen, Bradley A Malin, and Jimeng Sun. 2015. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *KDD*.

[42] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, 13 (2008), i232–i240.

[43] Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. 2016. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *NIPS*.

[44] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. 2017. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *WWW*.

[45] Xiaodong Zheng, Hao Ding, Hiroshi Mamitsuka, and Shanfeng Zhu. 2013. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *KDD*.

[46] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* (2018).

[47] Bin Zou, Vasileios Lampos, and Ingemar Cox. 2018. Multi-Task Learning Improves Disease Models from Web Search. In *WWW*.