Drug-Protein-Disease Association Prediction and Drug Repositioning Based on Tensor Decomposition

Ran Wang*, Shuai Li, Man Hon Wong, Kwong Sak Leung
Department of Computer Science and Engineering
The Chinese University of Hong Kong, Hong Kong
{rwang, shuaili, mhwong, ksleung}@cse.cuhk.edu.hk

***Abstract*—The old paradigm "one gene, one drug, one disease" of drug discovery is challenged in many cases, where many drugs act on multiple targets and diseases rather than only one. Drug repositioning, which aims to discover new indications of known drugs, is a useful and economical strategy for drug discovery. It is also important to identify the functional clustering of target proteins, drugs and diseases, and to understand the pathological reasons for their interactions among these clusters and individuals. In this study, we propose a novel computational method to predict potential associations among drugs, proteins and diseases based on tensor decomposition. First, we collect pairwise associations between drugs, proteins and diseases, and integrate them into a three-dimensional tensor, representing the drug-protein-disease triplet associations. Then, we carry out tensor decomposition on the association tensor together with some additional information, and get three factor matrices of drugs, proteins and diseases respectively. Finally, we reconstruct the association tensor by the factor matrices to derive new predictions of triplet associations. We compare our method with some baseline methods and find our method outperforming the others. We validate our top ranked predictions by literature search and computational docking. In addition, we cluster the drugs, proteins and diseases using the factor matrices, which reflect the functional patterns of the drugs, proteins and diseases. Comparing our clustering to existing classifications/clusters, we find some agreement between them and that the factor matrices indeed reflect the functional patterns.**

***Index Terms*—drug repositioning, drug discovery, tensor decomposition, drug-protein-disease association**

I. Introduction

Drug discovery is the process through which potential new medicines are discovered. It is one of the primary objectives in pharmaceutical science. The conventional experimental method aims to design exquisitely selective ligands against a single target, under the paradigm "one gene, one drug, one disease". However, its process is very time-consuming and expensive, and is plagued by the high attrition rate. Furthermore, the paradigm may fail in many cases: many drugs act on multiple targets and diseases rather than only one, which is intended in the drug design stage. The targets that are not originally intended are called "off-target". Such off-target interactions, though may cause adverse drug reactions (ADRs), can provide opportunities to seek new use of existing drugs in drug discovery.

In general, the computational methods of drug repositioning can be categorized into three main groups: docking simulation approaches, ligand-based approaches and machine learning approaches. Most of the current studies focus on only one aspect of drug repositioning, such as drug-target interaction prediction [1], [2] or drug indication (i.e. drug-disease association) prediction [3], which is a limitation of only investigating part of the associations among drugs, targets and diseases. However, although some drugs act by binding to specific proteins, most of the FDA-approved drugs were developed without knowing the molecular mechanisms responsible for their indicated diseases [4]. The study of drug-protein-disease triplet association is able to uncover the underlying molecular mechanisms. We model the triplet associations as an association tensor with its three dimensions representing drug, protein and disease, respectively, and discover new associations by tensor decomposition. Tensor decomposition has been applied to several problems in drug study, such as IC50 prediction [5], drug toxicogenomics [6], identifying drug target genes in gene expression profiles [7], etc. However, our knowledge of associations among drugs, proteins and diseases is very limited, resulting in a very sparse association tensor. To deal with this problem, we adopt additional information to support the decomposition of our association tensor, since it can provide more information about the triplets.

In this study, we propose a novel computational drug repositioning method based on tensor decomposition, in order to not only discover new drug-protein-disease triplet associations, but also reveal the underlying functional patterns. First, we collect drugs, proteins, diseases and corresponding drug-protein interactions, drug indications and protein-disease associations from public databases, and integrate the pairwise associations/interactions into a three-dimensional association tensor. Second, we carry out tensor decomposition on the integrated association tensor together with additional information, e.g. similarity of drugs and proteins, drug-drug interactions (DDI) and protein-protein interactions (PPI). The association tensor is decomposed into three lower rank factor matrices, representing the latent factors of triplet associations and the loadings of drugs, proteins and diseases on the latent factors. Last, the factor matrices are used to reconstruct the original association tensor and provide new triplet associations. We compare our method with several baselines and validate some of our new predictions. In addition, the latent factor loadings are used to cluster the drugs, proteins and diseases, revealing the functional patterns of them.

* Corresponding author.

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

II. RELATED WORK

Docking simulation and ligand-based approaches are two conventional computational methods in drug repositioning. The docking simulation approaches [8] use the structural information of targets to predict potential drug-target interactions, which is time-consuming and faces the problem of lacking structural information. The ligand-based approaches [9] compare a candidate ligand with the known ligands of a target protein to discover potential drug-target interactions. However, for targets with only a small number of interacted ligands, ligand-based methods do not work well. In recent years, it shows that machine learning approaches are more effective and efficient than these two traditional approaches.

The drug-target interaction prediction problem can be modeled as a binary classification problem, regarding the drug-target pairs as samples and the chemical structure of the drugs and the amino acid sequence of targets as features. Different kernels have been used to build the classifiers, such as similarity of drugs and proteins [10] and Gaussian interaction profile (GIP) kernels [11]. To deal with the problem of sparsity and cold-start, semi-supervised learning [12] and neighbor information [1], [13] are adopted in some studies.

Deep learning technique shows its strong predictive power in many fields, such as image processing [14], natural language processing [15] and speech recognition [16]. It is also applied to this field in recent years. Various deep learning models have been used, such as multi-layer perceptron [17], [18], deep belief network [19], stacked auto-encoder [20], [21]. However, there are still some limitations when applying deep learning techniques in drug repositioning. (1) It is hard to select negative samples in the training stage, since there are rarely experimentally verified negative samples. (2) Most studies use empirical features to represent proteins and drugs, which does not make full use of the automatic learning power of deep learning. In both traditional machine learning and deep learning approaches, it is common that the biological meaning of the models is hard to interpret.

The drug-target interaction prediction problem can be also modeled as a recommendation problem to select potential protein targets for a given drug. Matrix decomposition, a commonly used approach in recommendation systems, has been applied to drug repositioning in its different forms, such as Bayesian matrix decomposition [22], probabilistic matrix decomposition [2], logistic matrix decomposition [23] and collective matrix decomposition [24], [25]. Similar to matrix, associations/interactions among drugs, proteins and diseases can be represented in networks. Network-based inference methods have been used to build such networks and infer new associations/interactions. Some additional information, such as structural similarity of drugs and proteins [4], [26] and side-effect information [27], are integrated into the networks to predict association/interactions more accurately. Zheng et al.[24] combined multiple similarities into one method and demonstrated that different similarities had different performance on different datasets.

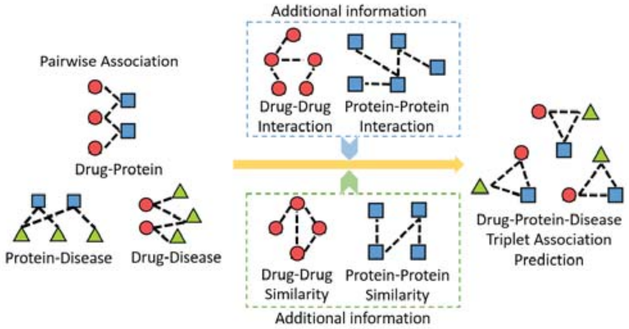

Fig. 1. Problem definition. Red circles, blue rectangles and green triangles represent drugs, proteins and diseases, respectively.

III. MATERIALS AND METHODS

A. Problem Definition

In this paper, we aim to predict new drug-protein-disease triplet associations given known pairwise relations, i.e. drug-protein interactions, drug-disease associations and protein-disease associations. To improve the performance, different kinds of additional information can be used, including similarity of drugs and proteins, as well as DDIs and PPIs. Thus, the input data is the known pairwise associations/interactions between drugs, proteins and diseases (Fig. 1 left), together with some additional information about them (Fig. 1 middle), and the output is new discoveries of drug-protein-disease triplet association (Fig. 1 right).

B. Data Preparation

We use the dataset provided by Luo et al. [28]. In their dataset, there are 708 drugs extracted from DrugBank database [29], $1,512$ proteins from human protein reference database (HPRD) [30] and $5,603$ diseases from comparative toxicogenomics database (CTD) [31]. They collect corresponding drug-target interactions and drug-drug interactions from DrugBank, protein-protein interactions from HPRD, drug-disease and protein-disease associations from CTD. We remove drugs and proteins with no drug-protein interactions and diseases with less than 100 drug-disease associations or 300 protein-disease associations, due to much more associated proteins than drugs of each disease. There are 549 drugs, 424 proteins and 340 diseases left. We extract $1,923$ corresponding drug-protein interactions, $73,075$ drug-disease associations, $129,563$ protein-disease associations, $6,078$ drug-drug interactions and $1,029$ protein-protein interactions.

The most commonly used additional information in drug-target interaction prediction is similarity of drugs and proteins. In this paper, the similarity of drugs is the Tanimoto coefficient of the product-graphs of their chemical structures [28]. Similarity of proteins is calculated using Smith-Waterman score based on their amino acid sequences.

C. Tensor Integration

To model the triplet associations of drugs, proteins and diseases, we use a third-order tensor χ, which we call it association tensor. The three dimensions of the tensor represent drug, protein and disease respectively. An entry χ_{ijk} in the association tensor χ means the triplet association of drug i, protein j and disease k. If $\chi_{ijk} = 1$, the triplet association of drug i, protein j and disease k is observed; otherwise, the triplet association does not exist or is still unobserved. Since most of the related databases only provide pairwise associations/interactions between drugs, proteins and diseases, we construct the association tensor by integrating the pairwise associations/interactions. Intuitively, $\chi_{ijk} = 1$ if and only if all of the three conditions are satisfied: (1) drug i interacts with protein j, (2) protein j is associated with disease k, and (3) drug i is associated with disease k. Otherwise, $\chi_{ijk} = 0$. The tensor constructed using this strategy is represented as χ^{tri}.

However, our knowledge of associations among drugs, proteins and diseases is very limited. The above strategy of tensor integration results in a very sparse association tensor ($\sim 0.33\%$). To generate a denser tensor and improve the prediction performance, we adopt an assumption used in some studies [32] that if drug A interacts with protein B and protein B is associated with disease C, then we can infer that drug A is associated with disease C. Thus, we propose another tensor integration strategy that $\chi_{ijk} = 1$ if (1) drug i interacts with protein j, and (2) protein j is associated with disease k; otherwise, $\chi_{ijk} = 0$. Using the second strategy, we derive another association tensor χ^{bi}. In total, there are $598,804$ entries ($\sim 7.6\%$) with observed or inferred triplet associations in χ^{bi}.

D. Drug-Protein-Disease Association Prediction

We model the drug-protein-disease triplet association prediction problem as tensor decomposition with additional information:

$$\begin{aligned}\min_{C,P,D,d_i} \; & \omega_{main}\|\chi - [\![C,P,D]\!]\|_F^2 \\ & + \Sigma_{i=1}^{M}\omega_i\|S_i - A_i d_i B_i^T\|_F^2 \\ & + \omega_{reg}(\|C\|_F^2 + \|P\|_F^2 + \|D\|_F^2 + \Sigma_{i=1}^{M}\|d_i\|_F^2)\end{aligned} \tag{1}$$

where C, P and D are factor matrices of drugs, proteins and diseases, respectively. M is the number of additional information matrices used. S_i is one of the additional information matrix which is decomposed to A_i, d_i and B_i. In our case, both of A_i and B_i are factor matrices of χ. d_i is used to capture scaling differences between difference data source. The first part of the formula aims to approximate the original association tensor using the factor matrices. The second part tries to approximate the additional information. The third part is regularization part, which makes the lower rank matrices as small as possible. ω_{main}, ω_i and ω_{reg} are weights for different parts, indicating their importance in optimization.

The factor matrices C, P and D share the latent factor dimension with length R (the number of latent factors). The decomposition of the triplet tensor and the additional information matrices share the same factor matrices in the corresponding dimension. By collectively factorizing the association tensor with additional information matrices, the factor matrices of drugs, proteins and diseases are derived from not only the triplet associations, but also the additional information, fusing different information source together.

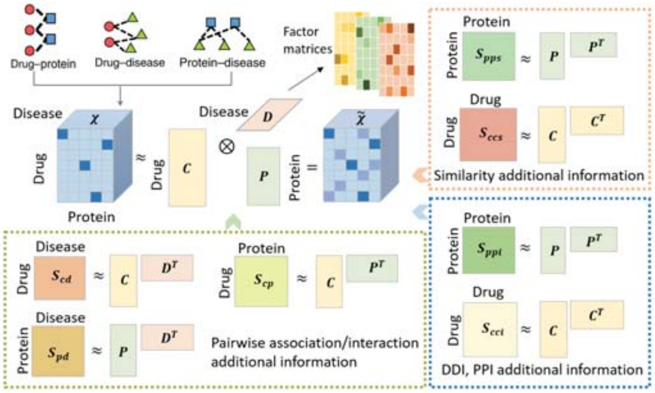

Fig. 2. Workflow of our method. First we integrate the association tensor χ using pairwise associations/interactions. Then we decompose the tensor together with different additional information and get new predictions in the reconstructed tensor $\widetilde{\chi}$. Finally, we analyze the factor matrices C, P and D.

In this paper, we use different additional information, including similarities of drugs and proteins, pairwise associations/interactions, as well as DDIs and PPIs (Fig. 2 bottom and right). The reason for using pairwise associations/interactions is that some information is lost in tensor integration stage, and we hope to use pairwise associations/interactions as additional information to compensate for that. The detailed model of tensor decomposition using different additional information are listed as follows:

1) Drug-drug and protein-protein similarity

$$\begin{aligned}\min_{C,P,D,d_{ccs},d_{pps}} \; & \omega_{main}\|\chi - [\![C,P,D]\!]\|_F^2 \\ & + \omega_1\|S_{ccs} - C d_{ccs} C^T\|_F^2 \\ & + \omega_2\|S_{pps} - P d_{pps} P^T\|_F^2 \\ & + \omega_{reg}(\|C\|_F^2 + \|P\|_F^2 + \|D\|_F^2 \\ & \qquad + \|d_{ccs}\|_F^2 + \|d_{pps}\|_F^2)\end{aligned} \tag{2}$$

2) Pairwise associations/interactions

$$\begin{aligned}\min_{C,P,D,d_{cp},d_{cd},d_{pd}} \; & \omega_{main}\|\chi - [\![C,P,D]\!]\|_F^2 \\ & + \omega_1\|S_{cp} - C d_{cp} P^T\|_F^2 \\ & + \omega_2\|S_{cd} - C d_{cd} D^T\|_F^2 \\ & + \omega_3\|S_{pd} - P d_{pd} D^T\|_F^2 \\ & + \omega_{reg}(\|C\|_F^2 + \|P\|_F^2 + \|D\|_F^2 \\ & \qquad + \|d_{cd}\|_F^2 + \|d_{cp}\|_F^2 + \|d_{pd}\|_F^2)\end{aligned} \tag{3}$$

3) Drug-drug and protein-protein interactions

$$\begin{aligned}\min_{C,P,D,d_{cci},d_{ppi}} \; & \omega_{main}\|\chi - [\![C,P,D]\!]\|_F^2 \\ & + \omega_1\|S_{cci} - C d_{cci} C^T\|_F^2 \\ & + \omega_2\|S_{ppi} - P d_{ppi} P^T\|_F^2 \\ & + \omega_{reg}(\|C\|_F^2 + \|P\|_F^2 + \|D\|_F^2 \\ & \qquad + \|d_{cci}\|_F^2 + \|d_{ppi}\|_F^2)\end{aligned} \tag{4}$$

where S_{ccs} and S_{pps} are similarities of drugs and proteins, respectively. S_{cp}, S_{cd} and S_{pd} are drug-protein interactions, drug-disease associations and protein-disease associations, respectively. S_{cci} and S_{ppi} are DDIs and PPIs, respectively. After decomposing the association tensor, we derive a new tensor $\widetilde{\chi}$ reconstructed from the three latent matrices, i.e. C, P and D, and get new predictions from $\widetilde{\chi}$ (Fig.2 middle). We use the package Tensorlab [33] for tensor decomposition.

IV. Results

A. Performance of Our Method

We use 10-fold cross-validation to test the performance of triplet association prediction. We use area under the receiver operating characteristic curve (AUC) and area under precision-recall curve (AUPR) to evaluate the performance. Pairwise associations/interactions contain information related to test data. Thus, when using them as additional information, we remove all pairwise associations/interactions of each triplet associations in the test set from S_{cp}, S_{cd} and S_{pd}. We set the value of w_{main} to be 1, w_1, w_2 and w_3 to be $5e-3$, and w_{reg} to be $1e-6$, which are default settings in Tensorlab.

We first investigate the relation between performance and the number of latent factors (the value of R). In Fig.3, we show the AUC and AUPR as functions of R when decomposing the association tensor χ^{bi} using similarity as additional information. We find that the AUC and AUPR increase with the value of R and then converge when R approaches 250. The performance drops after that because it overfits to the observed data, instead of extracting the functional patterns. Similar trend occurs when using other additional information, as well as in the decomposition of tensor χ^{tri}.

Then we compare the performance of decomposing tensors integrated from different strategies, i.e. χ^{tri} and χ^{bi} (Fig.4). We find that the AUC of decomposing χ^{tri} and χ^{bi} are comparable, while the AUPR of decomposing χ^{tri} is much lower than that of χ^{bi}. As we know, in such an imbalanced prediction problem with large proportion of negative samples, i.e. unobserved associated or not associated drug-protein-disease triplets, AUPR is considered to be more informative than AUC, because AUC may give an excessively optimistic picture of the algorithm with many false positives, and therefore low precision. The decomposition of χ^{tri} gives more false positive predictions, especially when using similarity as additional information. This perhaps results from the sparsity of χ^{tri} and the noise involved by similarity measure.

We further investigate the effects of using different additional information in χ^{tri} decomposition, including similarity, pairwise associations/interactions, DDIs and PPIs (Fig.5). We find that additional information contributes more to AUC than AUPR. The performance of using different or no additional information become comparable when the value of R is large enough. However, it requires much more computational resource using larger R. Using similarity as additional information performs the best, achieving about 2% higher AUC when the value of R is small. It supports that similar drugs are likely to interact with similar protein targets, which is widely accepted in the community and is used as a foundational assumption in drug-protein interaction prediction. Since we remove all pairwise associations/interactions related to test samples when using them as additional information, there is little additional information remained to achieve better performance than using no additional information. The DDIs and PPIs seem to involve some useful but indirect information, which is not that powerful as similarity in drug-protein-disease association prediction.

B. Comparison with Baseline Methods

1) Baseline Methods

Since most of the related studies investigate drug-protein interactions or directly drug-disease associations, very few of them study the triplet associations. Thus we compare our method to some pairwise association/interaction prediction methods, including SNScore [34] and another two baseline methods, which are random walk on heterogeneous network [35] and collective matrix decomposition.

a) SNScore: SNScore [34] algorithm calculates the probability of connections based on the number of in-between nodes connecting two nodes and the weights of the connections. In their reported performance, the AUC of drug-protein interaction, drug-disease association and protein-disease association prediction is 0.937, 0.868 and 0.871, respectively. Since only the trained platform is provided and their data covers all of our data, in evaluation, we randomly select 10% entries in χ^{bi}, project these triplet associations into pairwise associations/interactions, and retrieve the SNScore of them. The probability of a triplet association is calculated by (5), where P_{drug_pro} and P_{pro_dis} are the probability of corresponding drug-protein interaction and protein-disease association, respectively. AUC and AUPR are computed based on the inferred probability $P_{drug\_pro\_dis}$ of the triplet associations.

$$P_{drug\_pro\_dis} = P_{drug\_pro} \times P_{pro\_dis} \quad (5)$$

b) Random walk on heterogeneous network: Chen et al. [35] proposed a drug-target interaction prediction algorithm called Network-based Random Walk with Restart on Heterogeneous network (NRWRH). Similar algorithms are used in some other studies [36], [37], [38]. We generalize the method to network with three kind of nodes, i.e. drugs, proteins and diseases, and construct the heterogeneous network as Fig.6a. The diffusion states of the nodes are used as the probability of pairwise associations/interactions. We use grid search to find the optimal values of three important parameters in NRWRH, which are the number of maximum iteration, restart probability and transition probability. Finally, we set their values to be 400, 0.5 and 0.2, respectively. In evaluation, since the number of associations in χ^{bi} is much larger than that of the network, we set the number of test samples in NRWRH to be 10% of the pairwise associations/interactions. To generate test samples, we randomly select N_{test} (calculated by (6)) entries in χ^{bi}, project each triplet association into three pairwise associations/interactions, randomly remove one

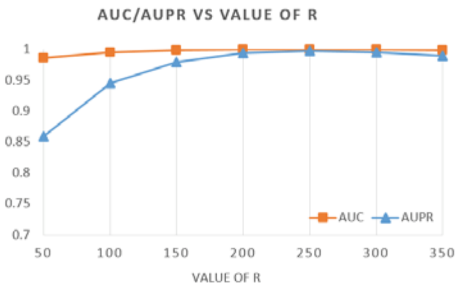

Fig. 3. Relation of AUC/AUPR and the value of R.

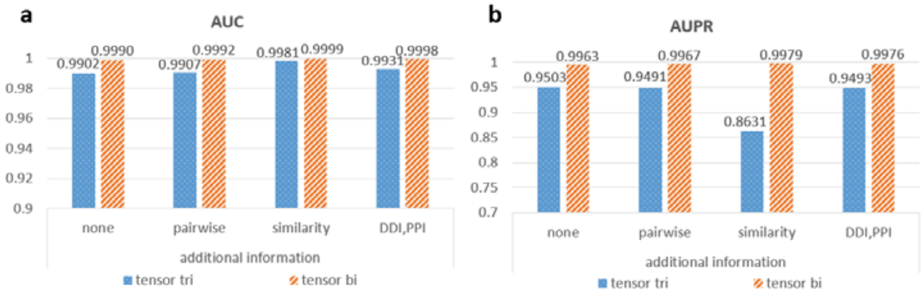

Fig. 4. Performance comparison of χ^{tri} and χ^{bi} with different additional information.

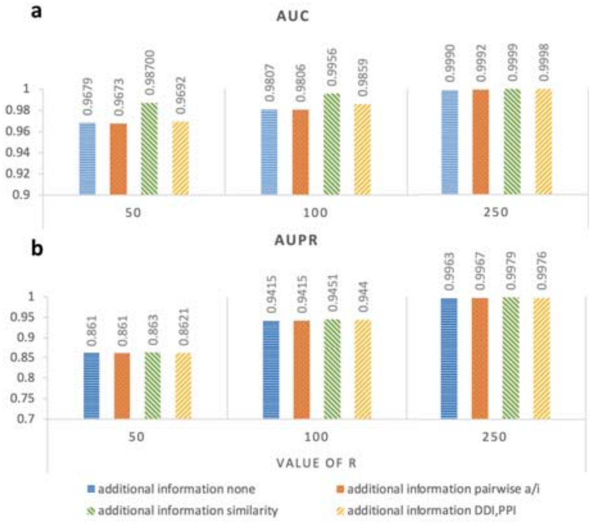

Fig. 5. Comparison of using different additional information.

TABLE I
PAIRWISE ASSOCIATIONS/INTERACTIONS IN TOP 50 PREDICTIONS

Pairwise associations	Known	Unknown
Drug-protein interactions	50	0
Drug-disease associations	14	36
Protein-disease associations	0	50

TABLE II
PAIRWISE ASSOCIATIONS/INTERACTIONS IN TOP 10,000 PREDICTIONS

Pairwise associations	Known	Unknown
Drug-protein interaction	$8,608$	$1,392$
Drug-disease association	$4,130$	$5,870$
Protein-disease association	$1,392$	$8,608$

of them, and calculate the probability of them by NRWRH. The triplet associations are also calculated by (5).

$$N_{test} = 0.1 \times (N_{drug} \times N_{protein} + N_{drug} \times N_{disease} + N_{protein} \times N_{disease}) \tag{6}$$

where N_{drug}, $N_{protein}$ and $N_{disease}$ are the number of drugs, proteins and diseases, respectively.

c) Collective matrix decomposition: Similar to our method but using lower order matrices, we collectively decompose five matrices in Fig.6b. The three pairwise association/interaction matrices are used as main matrices, while the two similarity matrices are used as additional information. We set the number of latent factors to 100, whose performance is better than that of 200. We evaluate this baseline method in the same way as NRWRH.

2) Performance Comparison

We compare our method with the three baseline methods described above and show the result in Fig.7, where RW and MD are random walk method and collective matrix decomposition method, respectively. In our method, we set $R = 250$ and use similarity of drugs and proteins as additional information, which means we use exactly the same data with RW and MD methods. Our method outperforms all the others in both AUC and AUPR. Our method gets about 4.07% higher AUC and 40.31% higher AUPR compared to the second best method RW. The performance of SNScore predicting for triplet associations is much worse than that of their reported pairwise interaction/association prediction. In addition, we find that although RW and MD get relatively high AUC, they get much lower AUPR than our method.

C. New Prediction Validation

To further validate the ability of discovering new associations of our method, we investigate 50 top-ranked new predictions, which are the entries in the reconstructed tensor $\widetilde{\chi}^{bi}$ with highest value while their values in the original integrated tensor χ^{bi} is zero (unobserved before). In this section, we use the result of tensor decomposition with $R = 250$, and similarity as additional information. Since it is hard to validate the new predictions in its triplet manner, we project them to pairwise associations/interactions, and try to validate the new predictions indirectly by validating these pairwise associations/interactions. We check the top 50 triplet predictions, and show the statistics of their projected pairwise associations/interactions in Table I. “Known” means that the projected pairwise associations/interactions exist in the input data, while “unknown” means they are unobserved in the input data (the corresponding values are “0”), which are new pairwise predictions. In the top 50 triplet predictions, all of the related drug-protein interactions are known, all of the related protein-disease associations are unknown, and most of the related drug-disease associations are unknown. It means that it is easier to find new triplet associations with related known drug-protein interactions.

We find support for these new pairwise predictions in the top 50 triplet predictions by literature search. Since our input

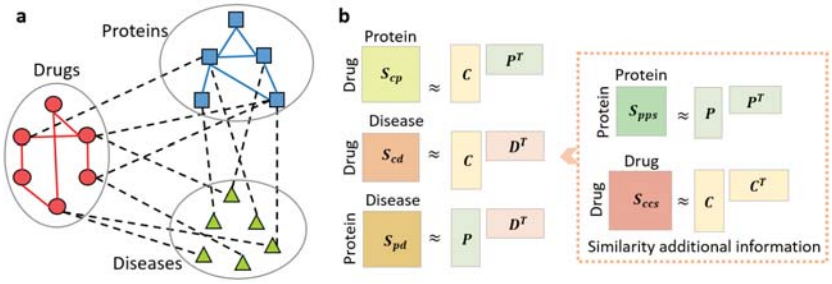

Fig. 6. Baseline methods. a. Heterogeneous network. b. Collective matrix decomposition.

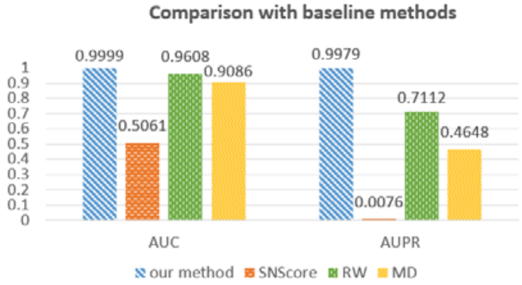

Fig. 7. Comparison with baseline methods.

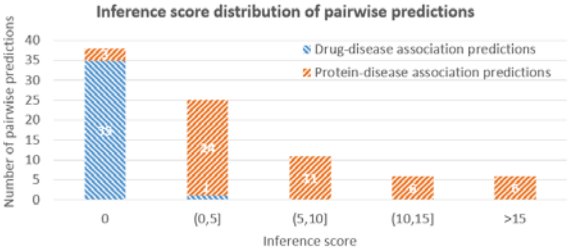

Fig. 8. Distribution of inference score of the pairwise association/interaction predictions in top 50 triplet predictions. Blue and orange bars represent the inference score distribution of new drug-disease associations and protein-disease associations, respectively.

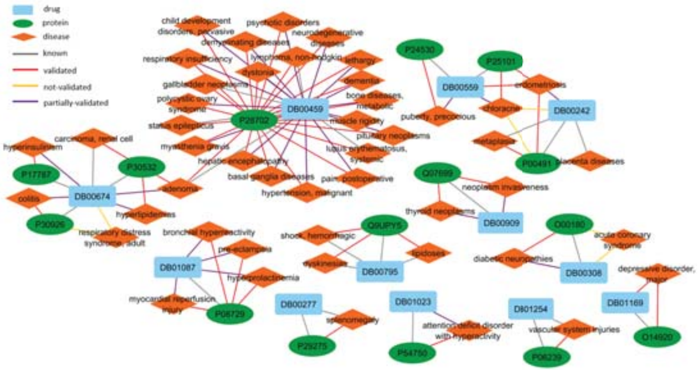

Fig. 9. Network visualization of top 50 predictions. Blue rectangles, green ellipses and orange diamonds represent drugs, proteins and diseases, respectively. Grey, red, purple and yellow lines represent known, validated, partial-validated and not validated associations/interactions, respectively.

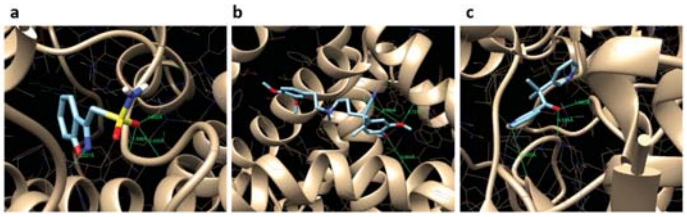

Fig. 10. The docked poses of three top ranked drug-protein interaction predictions without support. a. Zonisamide vs. AQP1. b. Verapamil vs. KCNK1. c. Metyrapone vs. FDX1.

data is collected from CTD update 2013, and it released a new version in 2017 [39], we first search for the new pairwise predictions in the updated version of CTD. In total, there are 36 new drug-disease associations and 50 new protein-disease associations. The distribution of their inference score in CTD 2017 is illustrated in Fig.8. "0" means we do not find the pairwise associations/interactions mentioned in CTD 2017. We find support for more than half of the pairwise new predictions, including one drug-disease association, and 47 protein-disease associations. Some of them get relatively high inference score in CTD 2017. Almost all of the predictions with literature support come from the protein-disease associations. We label the pairwise associations/interactions with support as "validated". We further set the drug-disease associations with corresponding known or validated drug-protein interactions and protein-disease associations to "partial-validated" according to our previous assumption. We visualize the pairwise predictions related to the top 50 predictions in a network shown as Fig.9.

From Fig.9, we find that there are many new predicted pairwise associations/interactions of protein "P28702", i.e. "Retinoic acid receptor beta", and drug "DB00459", i.e. "Acitretin". "Retinoic acid receptor beta" is a member of the thyroid-steroid hormone receptor superfamily of nuclear transcriptional regulators. It binds retinoic acid, the biologically active form of vitamin A, which mediates cellular signalling in embryonic morphogenesis, cell growth and differentiation. It regulates gene expression in various biological processes. Acitretin activates nuclear retinoic acid receptors, resulting in induction of cell differentiation, inhibition of cell proliferation, and inhibition of tissue infiltration by inflammatory cells. So far, we have not find any support for the three "not-validated" triplet associations (yellow lines in Fig.9). They may be new discoveries and will be further studied.

There is no triplet association predictions related to unknown drug-protein interactions in the top 50 predictions. To validate the ability of our method predicting for drug-protein interactions, we check top 10,000 predictions and find some new drug-protein interactions (Table II). We select the top 10 drug-protein interaction predictions and search in CTD and Pubmed, finding references for five of them. We perform computational docking for all of the top 10 predictions using Autodock Vina [40]. However, there are two proteins with no 3D structure provided, i.e."CHRM5" and "PTGER2". We take the 11th-ranked drug-protein pair "Metyrapone" and "FDX1" instead of "PTGER2" and "Bimatoprost", which has no literature support. The support references and docking scores are listed in Table III. The docked poses and predicted Hydrogen bonds of three pairs are illustrated in Fig.10.

TABLE III
SUPPORTING REFERENCES AND DOCKING SCORES OF THE TOP DRUG-PROTEIN INTERACTION PREDICTIONS

Drug	Protein(Gene)	Reference/Docking score
Miglitol	SI	[41] / −6.4 kcal/mol
Tropicamide	CHRM5	CTD/ –
temsirolimus	FKBP1A	[42] / −10 kcal/mol
Tacrolimus	MTOR	[43] / −7.4 kcal/mol
Dopamine	ALDH2	CTD / −6.6 kcal/mol
Amifostine	NT5C2	−5.6 kcal/mol
Vigabatrin	GABBR2	−5.1 kcal/mol
Verapamil	KCNK1	−7.9 kcal/mol
Zonisamide	AQP1	−7.1 kcal/mol
Metyrapone	FDX1	−8.7 kcal/mol

D. Latent Factor Analysis

After tensor decomposition, we get three factor matrices of drugs, proteins and diseases, respectively. These factor matrices reflect the functional patterns of them by condensing the original association tensor and additional information into lower rank matrices. Thus, by analyzing the factor matrices, we find functionally similar drugs, proteins and diseases, and cluster them into different functional groups. We show the result of drug clustering in Fig.11a. We cluster all of the drugs using the factor matrix of drugs, i.e. D, and compare our clustering to the chemical structure classes of the drugs in DrugBank. We find that some of the drugs from the same chemical structure class are clustered together, such as drugs in “Lipids and lipid-like molecules” class (Fig.11b, yellow) and “Organoheterocyclic compounds” class (Fig.11c, green), which reflects that these structurally similar drugs have similar functions. On the other hand, some drugs from different chemical structure classes are clustered together, meaning that they have similar functional patterns despite of having different chemical structures. For example, Brinzolamide and Methazolamide belong to class “Organoheterocyclic compounds”, and Furosemide belongs to class “Benzenoids”. These three drugs are clustered together (Fig.11d). We find that all of them share the same protein target “Carbonic anhydrase 2”. Adenosine, belonging to class “Adenosine or adenosine derivatives”, is reported to interact with protein target “Adenosine receptor A2a”, same as another two drugs, i.e. Theophylline and Pentoxifylline in class “Organoheterocyclic compounds”, which are closely clustered with Adenosine (Fig.11e).

V. CONCLUSION

In this paper, we have proposed a drug-protein-disease triplet association prediction method based on tensor decomposition, by integrating pairwise associations and decomposing it together with different additional information. Comparing our method to some baseline methods, we have found our method outperforming the others. Moreover, we have validated most of the top 50 new predictions by literature search. In addition, we have checked the top 10 newly predicted drug-protein interactions (which are not in the above top 50 predictions). We have found support for five of them by literature search and validated the others by computational docking. Last, we have analyzed the factor matrices derived from the tensor decomposition and used them to cluster the drugs. By comparing our clustering with existing chemical structure classes of the drugs and analyzing several cases, we have found that the derived latent factors indeed reflect the functional pattern of the drugs.

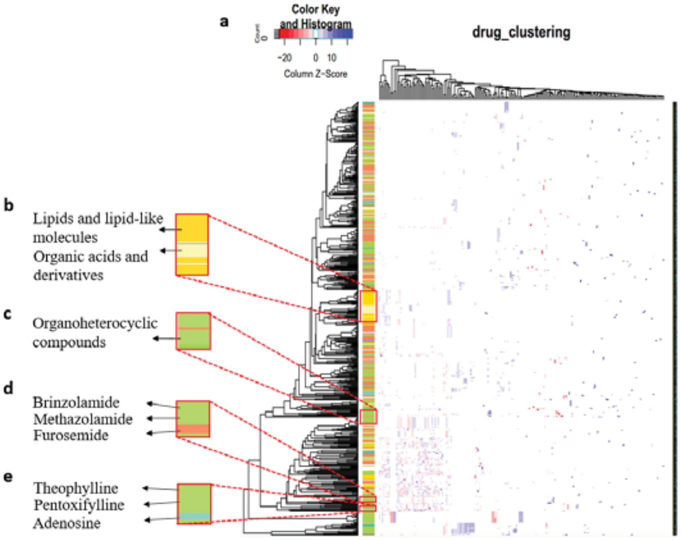

Fig. 11. Drug Clustering using the factor matrix of drugs. In a, each row is a drug, while each column is a latent factor. The hierarchical clustering of the drugs are shown on the left. The color bar shows the chemical structure class of the corresponding drug.

There are some future directions based on our study. First, since tensor decomposition requires huge computational resources, currently we can only investigate a relatively small number (about 500) of drugs, proteins and diseases. To perform large-scale association study, paralleled tensor decomposition technique would help. Second, other than the widely used similarity of drugs and proteins in drug repositioning, checking whether the use of disease similarity as additional information will help improve the performance is an interesting future work. Next, the associations and results need to be further investigated and explained biologically and pathologically, in order to discover some important associations in real life applications and drug study.

ACKNOWLEDGMENT

This work was supported by The Chinese University of Hong Kong Direct Grant [4055073].

REFERENCES

[1] T. van Laarhoven and E. Marchiori, “Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile,” *PloS one*, vol. 8, no. 6, p. e66952, 2013.

[2] M. C. Cobanoglu, C. Liu, F. Hu, Z. N. Oltvai, and I. Bahar, “Predicting drug–target interactions using probabilistic matrix factorization,” *Journal of chemical information and modeling*, vol. 53, no. 12, pp. 3399–3409, 2013.

[3] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, “Computational drug repositioning using low-rank matrix approximation and randomized algorithms,” *Bioinformatics*, vol. 34, no. 11, 2018.

[4] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS computational biology*, vol. 8, no. 5, p. e1002503, 2012.

[5] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau, "Macau: scalable bayesian multi-relational factorization with side information using mcmc," *arXiv preprint arXiv:1509.04610*, 2015.

[6] S. A. Khan, E. Leppäaho, and S. Kaski, "Bayesian multi-tensor factorization," *Machine Learning*, vol. 105, no. 2, pp. 233–253, 2016.

[7] Y.-H. Taguchi, "Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and drugmatrix datasets," *Scientific reports*, vol. 7, no. 1, p. 13733, 2017.

[8] Y. Y. Li, J. An, and S. J. Jones, "A computational approach to finding novel targets for existing drugs," *PLoS computational biology*, vol. 7, no. 9, p. e1002139, 2011.

[9] K. Wang, J. Sun, S. Zhou, C. Wan, S. Qin, C. Li, L. He, and L. Yang, "Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity," *PLoS computational biology*, vol. 9, no. 11, p. e1003315, 2013.

[10] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug–target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[11] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[12] Z. Xia, L.-Y. Wu, X. Zhou, and S. T. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," in *BMC systems biology*, vol. 4, no. 2. BioMed Central, 2010, p. S6.

[13] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Drug–target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2012.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513–520.

[16] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[17] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov, "Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data," *Molecular pharmaceutics*, vol. 13, no. 7, pp. 2524–2530, 2016.

[18] L. Xie, Z. Zhang, S. He, X. Bo, and X. Song, "Drugtarget interaction prediction with a deep-learning-based model," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 469–476.

[19] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu, "Deep-learning-based drug–target interaction prediction," *Journal of proteome research*, vol. 16, no. 4, pp. 1401–1409, 2017.

[20] M. Bahi and M. Batouche, "Drug-target interaction prediction in drug repositioning based on deep semi-supervised learning," in *Computational Intelligence and Its Applications: 6th IFIP TC 5 International Conference, CIIA 2018, Oran, Algeria, May 8-10, 2018, Proceedings 6*. Springer, 2018, pp. 302–313.

[21] L. Wang, Z.-H. You, X. Chen, S.-X. Xia, F. Liu, X. Yan, Y. Zhou, and K.-J. Song, "A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network," *Journal of Computational Biology*, vol. 25, no. 3, pp. 361–373, 2018.

[22] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[23] Y. Liu, M. Wu, C. Miao, P. Zhao, and X.-L. Li, "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," *PLoS computational biology*, vol. 12, no. 2, p. e1004760, 2016.

[24] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1025–1033.

[25] H. Lim, A. Poleksic, Y. Yao, H. Tong, D. He, L. Zhuang, P. Meng, and L. Xie, "Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing," *PLoS computational biology*, vol. 12, no. 10, p. e1005135, 2016.

[26] S. Alaimo, A. Pulvirenti, R. Giugno, and A. Ferro, "Drug–target interaction prediction through domain-tuned network-based inference," *Bioinformatics*, vol. 29, no. 16, pp. 2004–2008, 2013.

[27] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, 2008.

[28] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature communications*, vol. 8, no. 1, p. 573, 2017.

[29] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu *et al.*, "Drugbank 3.0: a comprehensive resource for omics research on drugs," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D1035–D1041, 2010.

[30] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal *et al.*, "Human protein reference database2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–D772, 2008.

[31] A. P. Davis, C. G. Murphy, R. Johnson, J. M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, M. C. Rosenstein, T. C. Wiegers *et al.*, "The comparative toxicogenomics database: update 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D1104–D1114, 2012.

[32] R. Sawada, H. Iwata, S. Mizutani, and Y. Yamanishi, "Target-based drug repositioning using large-scale chemical–protein interactome data," *Journal of chemical information and modeling*, vol. 55, no. 12, pp. 2717–2730, 2015.

[33] L. Sorber, M. V. Barel, and L. D. Lathauwer, "Structured data fusion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 586–600, 2015.

[34] H. S. Lee, T. Bae, J.-H. Lee, D. G. Kim, Y. S. Oh, Y. Jang, J.-T. Kim, J.-J. Lee, A. Innocenti, C. T. Supuran *et al.*, "Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug," *BMC systems biology*, vol. 6, no. 1, p. 80, 2012.

[35] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," *Molecular BioSystems*, vol. 8, no. 7, pp. 1970–1978, 2012.

[36] Y. Li and J. C. Patra, "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, 2010.

[37] A. Seal, Y.-Y. Ahn, and D. J. Wild, "Optimizing drug–target interaction prediction based on random walk on heterogeneous networks," *Journal of cheminformatics*, vol. 7, no. 1, p. 40, 2015.

[38] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 4, pp. 905–915, 2017.

[39] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: update 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D972–D978, 2016.

[40] O. Trott and A. J. Olson, "Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of computational chemistry*, vol. 31, no. 2, pp. 455–461, 2010.

[41] K. Tan, C. Tesar, R. Wilton, R. P. Jedrzejczak, and A. Joachimiak, "The interaction of anti-diabetic α-glucosidase inhibitors and gut bacteria α-glucosidase," *Protein Science*, 2018.

[42] K. Eberhart, O. Oral, and D. Gozuacik, "Induction of autophagic cell death by anticancer agents," in *Autophagy: Cancer, Other Pathologies, Inflammation, Immunity, Infection, and Aging*. Elsevier, 2014, pp. 179–202.

[43] F. Shihab, U. Christians, L. Smith, J. R. Wellen, and B. Kaplan, "Focus on mtor inhibitors and tacrolimus in renal transplantation: Pharmacokinetics, exposure–response relationships, and clinical outcomes," *Transplant immunology*, vol. 31, no. 1, pp. 22–32, 2014.