# Predicting Drug–Target Interactions With Multi-Information Fusion

Lihong Peng, Bo Liao, Wen Zhu, Zejun Li, and Keqin Li

***Abstract*—Identifying potential associations between drugs and targets is a critical prerequisite for modern drug discovery and repurposing. However, predicting these associations is difficult because of the limitations of existing computational methods. Most models only consider chemical structures and protein sequences, and other models are oversimplified. Moreover, datasets used for analysis contain only true-positive interactions, and experimentally validated negative samples are unavailable. To overcome these limitations, we developed a semi-supervised based learning framework called NormMulInf through collaborative filtering theory by using labeled and unlabeled interaction information. The proposed method initially determines similarity measures, such as similarities among samples and local correlations among the labels of the samples, by integrating biological information. The similarity information is then integrated into a robust principal component analysis model, which is solved using augmented Lagrange multipliers. Experimental results on four classes of drug-target interaction networks suggest that the proposed approach can accurately classify and predict drug–target interactions. Part of the predicted interactions are reported in public databases. The proposed method can also predict possible targets for new drugs and can be used to determine whether atropine may interact with alpha1B- and beta1- adrenergic receptors. Furthermore, the developed technique identifies potential drugs for new targets and can be used to assess whether olanzapine and propiomazine may target 5HT2B. Finally, the proposed method can potentially address limitations on studies of multitarget drugs and multidrug targets.**

***Index Terms*—Drug similarity, drug–target interaction (DTI), local correlations among labels of samples, multi-information fusion, robust PCA, semi-supervised learning, similarities among samples, target similarity.**

Manuscript received August 25, 2015; revised October 31, 2015; accepted December 21, 2015. Date of publication December 30, 2015; date of current version March 3, 2017. This work was supported by the Program for New Century Excellent Talents in University under Grant NCET-10-0365, National Nature Science Foundation of China under Grant 60973082, Grant 11171369, Grant 61202462, Grant 61272395, Grant 61370171, Grant 61300128, and Grant 61572178, the National Nature Science Foundation of Hunan province under Grant 12JJ2041 and Grant 13JJ3091, and the Planned Science and Technology Project of Hunan Province under Grant 2012FJ2012, and the Project of Scientific Research Fund of Hunan Provincial Education Department under Grant 14B023.

L. Peng, B. Liao, W. Zhu, and Z. Li are with the Key Laboratory for Embedded and Network Computing of Hunan Province, the College of Information Science and Engineering, Hunan University, Changsha 410082, China (e-mail: plhhnu@163.com; dragonbw@163.com)

K. Li is with the Key Laboratory for Embedded and Network Computing of Hunan Province, the College of Information Science and Engineering, Hunan University, Changsha 410082, China, and also with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/JBHI.2015.2513200

## I. Introduction

### A. Motivation

IDENTIFYING potential interactions between drugs and targets is a critical prerequisite for modern drug discovery and repurposing [1], [2]. Systematic analysis of potential associations is used to detect multitarget drugs and multidrug targets [3], elucidate the underlying mechanism of action of existing drugs [4], distinguish genotype-based resistance or sensitivity of drugs [5], [6], prevent side effects of drugs [7], and design effective treatment scheme [5]. However, known drug-target interactions (DTIs) are limited [8]. PubChem [9] contains about 35 million compounds, approximately 7000 of which are link to target proteins [8]. This phenomenon impels the need for developing effective techniques to determine underlying associations between drugs and targets [10].

Current experimental methods of identifying new DTIs are expensive and time consuming [11], [12], and feature low success rates [13]. In this regard, computational approaches have been increasingly used as a complement for existing methods [12]. Drug and target data from different sources, such as DrugBank [14], KEGG [15], Metador [16], and ChEMBL [17] databases, can be used to analyze potential relationships between drugs and targets at the systematic level.

Conventional computational techniques include ligand-based [18], receptor-based [19], and text-mining methods [20]. Although these techniques are widely applied in biology, they present several limitations. Ligand-based methods rely on the number of known ligands [21]. Receptor-based methods cannot be used to infer DTIs when the 3D structures of the target proteins are unknown [19]. Text-mining methods, which are performed by searching related keywords, suffer from issues of compound/gene name redundancy in the literature [20]. Therefore, this study aims to develop integrative approaches combining machine learning and biological information to determine novel associations between drugs and targets [22], [23]. The proposed machine learning-based prediction methods are divided into two categories:

*Supervised Learning-Based Method:* Supervised learning methods are widely applied to discover potential drug-target relationships. Yamanishi *et al.* [24] used a two-step supervised learning approach to identify novel DTIs by integrating chemical and genomic information. Bleakley and Yamanishi [25] developed bipartite local models (BLM) to predict new DTIs. Although these approaches achieve high prediction accuracy, the unlabeled interactions in the training dataset are assumed as negative samples and cannot be identified [26]. The BLM

algorithm was improved by Yamanishi *et al.* [27], van Laarhoven *et al.* [28], Fakhraei *et al.* [29], and Mei *et al.* [12]. Cheng *et al.*[30] developed three supervised inference models based on drug similarities, target similarities, and DTI networks. Gönen [31] proposed a Bayesian matrix factorization algorithm to classify unlabeled DTIs. Wang and Zeng [11] proposed a restricted Boltzmann machine. Whereas Alaimo *et al.* [3] developed a bipartite network projection model to mine potential DTIs. Zu *et al.* [1] observed that previous studies ignored the competitive effects between drug chemical substructures or protein domains; as such, they developed a global optimization-based inference model to infer associations between chemical substructures and protein domains. This promising approach provides novel insights into predicting DTIs.

Supervised learning-based models exhibit satisfactory performance and is the representative method for predicting DTIs; however, these models exhibit the following limitations. 1) The majority of these methods measure drug and target similarities by using chemical structures and protein sequences only; the obtained information may not adequately reflect the characteristics that determine whether a drug acts on a target [2]. Moreover, these methods disregard significant information such as quantitative structure-affinity relationship [32] and dose dependence [33]. 2) Known DTIs are rare, and negative DTIs are difficult or even impossible to achieve because experimentally validated negative samples are not reported and unavailable [8], [21], [34]. 3) Model evaluations are usually performed by crossvalidation, which assumes that potential DTIs are randomly distributed in a known DTI network [33]. These evaluations may result in oversimplified formulation, overoptimistic performance, and selection bias of model parameters during prediction [33]. Furthermore, the rarity of an algorithm requires a time-based evaluation, except for those approaches proposed by Fakhraei *et al.* [29]. 4) The rarity of techniques is emphasized to predict interactions for new drugs without any known target information and for new targets without any known drug targeting information. Considering these limitations, Pahikkala *et al.* [33] concluded that problem model, nature of datasets, assessment procedures, and experimental setup may cause a significant discrepancy in prediction performance.

*Semi-supervised Learning Based Method:* Several semi-supervised based approaches have been recently applied to identify potential DTIs. Xia *et al.* [26] evaluated a manifold regularized Laplacian method and proposed Laplacian regularized least squares model (LapRLS) and LapRLS based on a network, which use labeled and unlabeled information; nevertheless, these methods only consider chemical structures and sequences to identify drug and target similarities, which may not adequately capture the characteristics that determine whether a drug acts on a target [2]. Chen *et al.* [35] assumed that similar drugs interact with similar targets, and thus, proposed a network-based random walk with restart on a heterogeneous network. This approach integrates drug similarity networks, protein similarity networks, and known DTI data into a heterogeneous network and implement the random walk on the network. However, when inferring possible target proteins for new drugs without any known target information, network-based drug and target similarity matrices are considered zero, thereby limiting their applications [21], [35]. Using the framework of random walk, Chen and Zhang [21] used a network-consistency-based prediction scheme, namely, NetCBP, to efficiently mine new DTIs by integrating labeled and unlabeled DTI data. This scheme highly relies on similarity measures [21]. Generally, improving prediction performance by using semi-supervised learning may exhibit less significant because of the rarity of positive samples, no experimentally validated negative samples [21], [34], and the imbalance of DTI data. Given this limitation, Xiao [36] balanced positive and negative samples through neighbor cleaning theory and synthetic minority oversampling.

### B. Study Contributions

In this study, a semi-supervised based inference method was developed and designated as NormMulInf. This method uses a small quantity of available labeled data and abundant unlabeled data and then integrates biological information related to drugs and targets into a convex optimization model to determine underlying DTIs. This approach is based on the assumption that similar drugs interact with similar targets [21], [34], [37]. This study has the following main contributions.

1) We propose a semi-supervised learning based DTI prediction approach to address difficulties in obtaining negative DTI samples in practical problems. We also discuss the rationale and analyze the validity of the proposed method.
2) Biological information, which constitute similarities between samples and the local correlations between labels of samples in the DTI network, is integrated into a unified framework to capture new DTIs.
3) The prediction method can be applied to new drugs without any known target information and new targets without any known drugs targeting information.

The remaining sections of this paper are organized as follows. Section II briefly presents a review of related works. Section III introduces the DTI prediction approach. Section IV describes the method used for comparative experiments. Section V presents the experimental results. Section VI indicates the conclusions of the study and provides directions for further research.

## II. Brief Review of Related Works

### A. DTI Prediction

Yu *et al.* [38] proposed a weak-label learning approach, namely, protein function prediction with weak-label learning (ProWL), through guilt-by-association rule by using correlations among features; this approach relies heavily on correlations among functions [39]. Wang *et al.* [40] assumed that biological processes are highly inter-related and proposed a network-based method, namely, function-function correlated multilabel learning approach (FCML); this approach cannot predict functions on completely unannotated proteins [38]. Based on Hilbert–Schmidt independence theory, Yu *et al.* [39] further developed a protein function prediction method by

using dependency maximization (ProDM) to replenish missing data. ProDM relies on relationships among functions [41]. These three methods are classical multilabel learning methods and can be applied to predict DTIs.

van Laarhoven *et al.* [28] introduced a Gaussian interaction profile kernel and used a regularized least squares classifier (RLS-Kron) to investigate DTIs by combing related features of the DTI network. However, this method cannot be applied to infer new interactions for drugs or targets without any known interactions [28]. Chen and Zhang [21] presented a semi-supervised based learning approach (NetCBP) based on random walk to rank DTI scores according to their correlations with the labeled data; this approach relies on similarity measures. Mei *et al.* [12] integrated an interaction-profile inferring (NII) method by using neighbor information through the existing BLM model (BLM-NII) to determine new DTIs. These three approaches are represent DTI prediction techniques; of which, BLM-NII is the current state-of-the-art approach for predicting DTIs.

### B. Multi-Information Fusion

Incorporating multiple available data sources related to drugs and targets can improve DTI prediction performance [22], [23], [42]. The challenge lies in mining and fusing these heterogeneous information [22], [23]. Wang *et al.* [22] integrated different types of information, such as chemical structures, pharmacological information, and therapeutic effects of drugs, as well as sequences of target proteins, and proposed kernel method based on an SVM predictor to determine novel DTIs. The functional annotation analysis showed that the DTIs predicted by this approach are worthy of further experimental validation. Perlman *et al.* [42] integrated multiple methods of measuring drug gene similarities into a similarity-based DTI inference framework by using a logistic regression model to develop a DTI prediction method named SITAR. Martínez-Jiménez and Marti-Renom [43] assumed that structurally similar binding sites are likely to bind similar ligands and developed a network-based inference method, namely, nAnnoLyze, by integrating biological knowledge into a bipartite network. The approach provides examples of DTI prediction at proteome scale and enables annotation and analysis of the associations on a large scale. Wang *et al.* [23] integrated DTIs, drug ATC codes, drug-disease interactions, and SVM-based algorithm into a unified framework to predict DTIs, infer associations between drug and its ATC codes, and identify drug-disease connections. This approach efficiently integrates various heterogeneous data sources and promotes related research in drug discovery. Fakhraei *et al.* [29] represented a DTI network through BLM augmented with drug target similarities information to predict unknown interactions by using probabilistic soft logic. These models yield improved prediction performance and are considered representative information fusion methods in predicting DTIs. Based on these methods, we propose a multi-information fusion approach.

### C. Robust Principal component analysis (PCA)

PCA is a prevalent tool for discovering and exploiting low-dimensional structures in high-dimensional data [44]. However, gross errors often occur in bioinformatics applications. The lack of robustness to gross corruption or outliers limits the performance and applicability of PCA; even a small portion of large errors can corrupt the estimation of low-rank structures for biological data [45]. Robust PCA, a modified PCA method, was developed to efficiently and accurately recover the low-rank matrix $\boldsymbol{A}$ from highly corrupted measurements.

$$\boldsymbol{D} = \boldsymbol{A} + \boldsymbol{E}. \tag{1}$$

The corrupted entries can be described as the additive error matrix $\boldsymbol{E}$, which are unknown and arbitrary in magnitude. Errors $\boldsymbol{E}$ are sparse and affect only a small portion of the entries of the observations $\boldsymbol{D}$ in robust PCA [45], [46] compared with that in classical setting in PCA, where low-rank matrix $\boldsymbol{A}$ is affected by small but dense noise. Robust PCA can be solved within polynomial-time via convex optimization by minimizing a nuclear norm for low-rank recovery and minimizing $\ell_1$-norm for error correction [47]:

$$\min_{\boldsymbol{A},\boldsymbol{E}} ||\boldsymbol{A}||_* + \lambda||\boldsymbol{E}||_1 \ \text{ subject to } \ \boldsymbol{D} = \boldsymbol{A} + \boldsymbol{E}. \tag{2}$$

Wright *et al.* [45] applied iterative thresholding to precisely recover the corrupted low-rank matrix; however, the technique converges extremely slowly [47]. As such, Lin *et al.* [48] proposed an accelerated proximal gradient method (APG), which can be applied to the primality and duality of the convex optimization model. The APG algorithm often leaves many small nonzero terms in the error matrix $\boldsymbol{E}$ and only obtains a close approximate solution [48]. In this regard, Lin *et al.* [47] used the augmented Lagrange multipliers (ALM) and proposed exact ALM and inexact ALM, which are two algorithms with high accuracy and converge Q-linearly to the optimal solution.

### D. Collaborative Filtering (CF)

As a widely used technique in building recommendation systems, CF can effectively solve problems of data sparsity and scalability and produce high-quality preferences for other users by using the preferred information of users [49]. Memory-based CF techniques [50]–[52] can be simply implemented and incrementally add new data. However, these methods exhibit reduced performance when data are sparse, limited scalability for large datasets, and inability to predict new interactions for new drugs and targets [49]. By contrast, model-based CF methods [53] can efficiently solve issues with regard to data sparsity and scalability, achieve improved prediction performance, and provide intuitive reasoning for prediction; nevertheless, these models are expensive [49], [53]. To address the limitations of these CF models and improve the prediction performance, researchers developed hybrid CF [54]. To optimize these methods, we integrated different types of information and measured drug and target similarities by vector cosine-based similarity [50], which is a representative similarity computation method in memory-based CF models. We then infer novel DTIs by using a robust PCA model based on CF [49], [53].

TABLE I
DATASET DESCRIPTIONS INVOLVING HUMAN ENZYMES (ENZ), ION CHANNELS (ION), GPCRS, AND NUCLEAR RECEPTORS (NUC) [24]

| Dataset | Enz | Ion | GPCRs | Nuc |
|---|---|---|---|---|
| drugs ($n$) | 445 | 210 | 223 | 54 |
| targets ($m$) | 664 | 204 | 95 | 26 |
| interactions | 2926 | 1476 | 635 | 90 |
| the ratio ($n/m$) | 0.67 | 1.03 | 2.35 | 2.08 |
| $N_{\mathrm{avetar}}$ | 6.58 | 7.03 | 2.85 | 1.67 |
| $N_{\mathrm{avedrug}}$ | 4.41 | 7.24 | 6.68 | 3.46 |

## III. MATERIALS AND METHODS

### A. Data Preparation

*1) Chemical Data:* Yamanishi *et al.* [24] achieved chemical structures of compounds from the DRUG and COMPOUND sections in the KEGG LIGAND database [15]. The chemical structure similarity among drugs was obtained with SIMCOMP [55], which denotes compounds as graphs and calculates the similarity score according to the number of the common substructures between two compounds. The chemical structure similarity between two compounds $d_i$ and $d_j$ can be calculated based on the Tanimoto coefficient as

$$\mathrm{Sim}_{\mathrm{StruDrug}}(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i \cup d_j|}. \tag{3}$$

The chemical structure similarity matrix of drug compounds is described as $\mathrm{Sim}_{\mathrm{StruDrug}}$.

*2) Genomic Data:* Yamanishi *et al.* [24] extracted sequence information of target proteins from the KEGG GENES database [15], and calculated sequence similarity of target proteins by using a normalized version of the Smith–Waterman score [56]. The sequence similarity can be calculated as

$$\mathrm{Sim}_{\mathrm{SeqTar}}(t_c, t_d) = \mathrm{SW}(t_c, t_d)/\sqrt{SW(t_c, t_c)SW(t_d, t_d)} \tag{4}$$

where $SW(t_c, t_d)$ denotes the canonical Smith–Waterman score between the target proteins $t_c$ and $t_d$. The sequence similarity matrix of the target proteins is denoted as $\mathrm{Sim}_{\mathrm{SeqTar}}$.

*3) DTI Data:* Yamanishi *et al.* [24] determined that 445, 210, 223, and 54 drugs interact with 664, 204, 95, and 26 proteins from human enzymes, ion channels, GPCRs, and nuclear receptors, respectively, with known interactions of 2926, 1476, 635, and 90, respectively. Table I presents the details and the number of drugs ($n$), number of targets ($m$), number of interactions, average number of targets interacting with each drug ($N_{\mathrm{avetar}}$), average number of drugs interacting with each target ($N_{\mathrm{avedrug}}$). We use four datasets as the "gold standard" to evaluate and compared the proposed method with previously reported methods [21], [24], [25], [27], [30], [31], [35].

### B. Problem Description

Given $n$ drugs and $m$ targets, suppose that the original DTI network $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n]$ represents $n$ drugs, where $b_{ij} = 1$ if the $i$th target interacts with the $j$th drug; otherwise, $b_{ij} = 0$. To recover the low-rank DTI matrix and identify new DTIs, we assume that the current DTI data are complete and mask part of interactions for each sample according to its masked DTI ratio (MDTIR). Given that MDTIR is 0.2, if a drug interacts with six targets and INT( 6*0.2 )=1, we can change one interaction from 1 to 0 and keep only five interactions for the drug. The masked DTI matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$, in which only part of interactions are kept, is obtained from the original DTI network $\boldsymbol{B}$. The interactions labeled 0 are unknown pairs that will be predicted. We represent matrices and vectors by boldface uppercase and boldface lowercase letters, respectively.

Robust PCA efficiently and precisely recovers the low-rank matrix $\boldsymbol{A}$ from highly corrupted measurements. DTI data are sparse, low-rank, and imbalanced. Only few labeled data (true-positive interactions) but abundant unlabeled data are available, and negative DTIs are difficult or even impossible to obtain because experimentally validated negative samples are not reported [8], [21], [34]. Furthermore, a certain degree of similarity exists among row (column) vectors in the DTI matrix. This similarity causes DTI matrix to become a low-rank matrix. Therefore, the characteristics of DTI data satisfy the condition of robust PCA. In this regard, we aim to recover the DTI matrix based on the robust PCA model.

We intend to identify novel DTIs based on the robust PCA model by using (5), which minimizes the discrepancy between the known DTI matrix $\boldsymbol{X}$ and the predicted associated matrix $\boldsymbol{Pre}$

$$\min_{\boldsymbol{Pre}, \boldsymbol{E}} \|\boldsymbol{Pre}\|_* + \lambda \|\boldsymbol{E}\|_1 \quad s.t. \quad \boldsymbol{X} = \boldsymbol{Pre} + \boldsymbol{E} \tag{5}$$

where $\|\boldsymbol{Pre}\|_*$ represents the nuclear norm of the predicted DTI matrix $\boldsymbol{Pre}$, $\|\boldsymbol{E}\|_1$ denotes the $\ell_1$-norm of the discrepancy matrix $\boldsymbol{E}$, the weight parameter $\lambda$ represents the weight sparse error term in the cost function, and $0 \leq \lambda \leq 1$. The optimization model can be solved using the Exact ALM method from a previous study [47] and expressed as

$$\boldsymbol{Pre} = \mathrm{RPCA}(\boldsymbol{X}_{\mathrm{Laplacian}}, \lambda). \tag{6}$$

### C. Methods for DTI Prediction

Nigam [57] reported that integrating unlabeled data into machine learning can effectively reduce errors of classifiers and obtain improved classification performance when using sparse labeled data. Therefore, we propose a semi-supervised learning framework by using labeled and unlabeled interaction information. Previous studies [12], [22], [23], [42] indicated that integrating multiple types of data can improve the prediction performance compared with techniques using unlabeled data. Therefore, we incorporate multiple types of biological information into a semi-supervised learning framework.

Ding *et al.* [8] performed systematic analysis and comparison to comprehensively review state-of-the-art similarity-based machine learning methods for predicting DTIs. The majority of the methods disregard the similarities between samples and the local correlations between the labels of samples in the DTI network. Information regarding a label may contribute to learning another related label, particularly when the training samples of

some labels are inadequate [58]. In contrast to similarity-based machine learning methods [8], the proposed technique measures drug and target similarities based on various biological information, particularly similarities among samples and local correlations among labels of samples. We integrate different information fusion methods and robust PCA solved by the augmented Lagrange approach [47] into a unified framework. Finally, we conduct extensive experiments to evaluate the performance of the proposed method compared with that of six state-of-the-art techniques in the "gold standard" datasets from human enzymes, ion channels, GPCRs, and nuclear receptors. The results demonstrate that the proposed approach exhibits superior performance. In addition, we observed that several strongly predicted DTIs are reported by public databases.

*1) NormDrug for DTI Prediction:* In this section, we consider drugs as samples and each target as a label. The proposed method assumes that drugs shared by many targets may be similar in the DTI network [21], [38], [39]. The prediction model based on drugs is presented by integrating biological information related to **drug**s (NormDrug) into robust PCA method, which minimizes the combination of nuclear **norm** for low-rank recovery and $\ell_1$-norm for error correction. The method is categorized into three parts: the first part masks part of interactions for each sample according to MDTIR; the second part computes the Laplacian matrix [59] by combining the chemical structure similarities between samples (drugs) and the local correlations between the labels of samples in the DTI network; and the third part achieves the predicted DTI matrix.

In contrast to similarity measures in a previous study [8], drug similarity is measured in the present study by considering each drug as a vector of the frequency of interaction with the targets; we then calculate the cosine value of the angle formed by two drug vectors [49], [50].

Suppose that $\mathrm{Sim}_{\mathrm{NetDrug}}$ denotes the drug similarity matrix according to the local correlations between the labels of samples in the DTI network, we calculate drug similarity by (7) through a vector cosine-based similarity method [49], [50]

$$\mathrm{Sim}_{\mathrm{NetDrug}}(i,j) = \frac{\boldsymbol{x}_i \boldsymbol{x}_j^T}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|}. \tag{7}$$

We can conclude that the value of $\mathrm{Sim}_{\mathrm{NetDrug}}(i,j)$ is higher than that of $\mathrm{Sim}_{\mathrm{NetDrug}}(i,k)$ if the $i$th and $j$th drugs are simultaneously associated with abundant targets; however, the $i$th and $k$th drugs act only on few targets or no targets, as shown in (7). We obtain the likelihood that a drug interacts with a target, considering that this drug interacts with another target by normalizing $\mathrm{Sim}_{\mathrm{NetDrug}}(i,j)$

$$\mathrm{Sim}_{\mathrm{NetDrugNorm}}(i,j) = \frac{\mathrm{Sim}_{\mathrm{NetDrug}}(i,j)}{\sum_{k=1}^{n} \mathrm{Sim}_{\mathrm{NetDrug}}(i,k)}. \tag{8}$$

By combining the similarity in the chemical structure of drugs and the local associations between the labels of drugs in the DTI network, we obtain the final drug similarity matrix by

$$\mathrm{Sim}_{\mathrm{Drug}} = \mathrm{Sim}_{\mathrm{NetDrugNorm}} + \alpha \mathrm{Sim}_{\mathrm{StruDrug}} \tag{9}$$

where the weighted parameter $\alpha$ balances the importance between the similarities in the chemical structures of drugs and the local associations of their labels

$$\alpha = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Sim}_{\mathrm{NetDrugNorm}}(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Sim}_{\mathrm{StruDrug}}(i,j)}. \tag{10}$$

We define the Laplacian matrix $\boldsymbol{L}_{\mathrm{Drug}}$ with (11) by using the final drug similarity matrix

$$\boldsymbol{L}_{\mathrm{Drug}} = \boldsymbol{I}_{\mathrm{Drug}} - \boldsymbol{D}_{\mathrm{Drug}}^{-\frac{1}{2}} \mathrm{Sim}_{\mathrm{Drug}} \boldsymbol{D}_{\mathrm{Drug}}^{-\frac{1}{2}} \tag{11}$$

where $\boldsymbol{I}_{\mathrm{Drug}}$ is an $n \times n$ identity matrix, $\boldsymbol{D}_{\mathrm{Drug}}$ is a diagonal matrix which entries

$$D_{\mathrm{Drug}}(i,i) = \sum_{j=1}^{n} \mathrm{Sim}_{\mathrm{Drug}}(i,j). \tag{12}$$

Suppose that (13) represents the association matrix by label propagation [60] after masking parts of the interactions for each sample

$$\boldsymbol{X}_{\mathrm{DrugLap}} = \boldsymbol{X}\boldsymbol{L}_{\mathrm{Drug}}. \tag{13}$$

We view DTI prediction as a special case of the model by (5) to identify potential interactions by using limited number of known interactions through robust PCA with

$$\boldsymbol{Pre}_{\mathrm{Drug}} = \mathrm{RPCA}(\boldsymbol{X}_{\mathrm{DrugLap}}, \lambda). \tag{14}$$

The model can be solved using the Exact ALM method from a previous study [47]. We summarize DTI prediction approaches based on drug information and develop Algorithm 1 to determine novel DTIs from the original DTI network $\boldsymbol{B}$.

**Algorithm 1:** NormDrug for DTI prediction

**Input:** $\mathrm{Sim}_{\mathrm{StruDrug}}$, $\boldsymbol{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n\} \in \Re^{m \times n}$, $\lambda$;
**Output:** $\boldsymbol{Pre}_{\mathrm{Drug}}$;
Obtain the masked DTI matrix $\boldsymbol{X}$;
Compute $\mathrm{Sim}_{\mathrm{Drug}}$ using (9);
Compute $\boldsymbol{L}_{\mathrm{Drug}}$ using (11);
Compute $\boldsymbol{X}_{\mathrm{DrugLap}}$ using (13);
Obtain $\boldsymbol{Pre}_{\mathrm{Drug}}$ through robust PCA model with (14) solved by using the Exact ALM method [47];
Sort DTIs in $\boldsymbol{Pre}_{\mathrm{Drug}}$ in descending order;
Return obtained DTI ranking list;

*2) NormTarget for DTI Prediction:* Similar to that in NormDrug, we consider targets as samples and each drug as a label. We predict novel DTIs by using biological information related to **Target**s (NormTarget) through robust PCA, which minimizes the combination of nuclear **norm** for low-rank recovery and $\ell_1$-norm for error correction. The method is categorized into three parts: The first and the third parts are similar to those in NormDrug. We compute the Laplacian matrix based on the target similarity by combining the similarities between the samples (targets) and the local correlations between the labels(drugs) of the samples in the second part.

Suppose that $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ represents the masked DTI matrix. $\mathrm{Sim}_{\mathrm{NetTar}}$ denotes the similar matrix between targets according to the local correlations between the labels of

samples in the DTI network. We calculate the matrix by (15) based on the vector cosine-based similarity measure method

$$\mathrm{Sim}_{\mathrm{NetTar}}(i,j) = \frac{\boldsymbol{X}_{i.}\boldsymbol{X}_{j.}^{T}}{\|\boldsymbol{X}_{i.}\|\,\|\boldsymbol{X}_{j.}\|} \tag{15}$$

where $\boldsymbol{X}_{i.}$ represents the $i$th row of $\boldsymbol{X}$. We then normalize $\mathrm{Sim}_{\mathrm{NetTar}}(i,j)$ with (16) as follows:

$$\mathrm{Sim}_{\mathrm{NetTarNorm}}(i,j) = \frac{\mathrm{Sim}_{\mathrm{NetTar}}(i,j)}{\sum_{k=1}^{m}\mathrm{Sim}_{\mathrm{NetTar}}(i,k)}. \tag{16}$$

By combining the sequence similarities of target proteins and the local correlations of labels between samples in the DTI network, we obtain the final target similarity matrix by

$$\mathrm{Sim}_{\mathrm{Tar}} = \mathrm{Sim}_{\mathrm{NetTarNorm}} + \beta\mathrm{Sim}_{\mathrm{SeqTar}} \tag{17}$$

where weighted parameter

$$\beta = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Sim}_{\mathrm{NetTarNorm}}(i,j)}{\sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Sim}_{\mathrm{SeqTar}}(i,j)}. \tag{18}$$

We determine the association matrix by label propagation [60] after masking parts of the interactions for each sample

$$\boldsymbol{X}_{\mathrm{TarLap}} = \boldsymbol{L}_{\mathrm{Tar}}\boldsymbol{X} \tag{19}$$

where the Laplacian matrix

$$\boldsymbol{L}_{\mathrm{Tar}} = \boldsymbol{I}_{\mathrm{Tar}} - \boldsymbol{D}_{\mathrm{Tar}}^{-\frac{1}{2}}\mathrm{Sim}_{\mathrm{Tar}}\boldsymbol{D}_{\mathrm{Tar}}^{-\frac{1}{2}} \tag{20}$$

and the calculations of $\boldsymbol{I}_{\mathrm{Tar}}$ and $\boldsymbol{D}_{\mathrm{Tar}}$ are similar to those in NormDrug.

*3) NormMulInf for DTI Prediction:* In the preceding two sections, NormDrug considers drugs as samples and targets as labels, whereas NormTarget uses targets as samples and drugs as labels. In this section, we consider all factors and propose NormMulInf based on NormDrug and NormTarget as follows:

$$\boldsymbol{Pre} = \boldsymbol{Pre}_{\mathrm{Drug}} + \gamma\boldsymbol{Pre}_{\mathrm{Tar}} \tag{21}$$

where $\boldsymbol{Pre}_{\mathrm{Tar}}$ denotes the DTI score matrix by NormTarget, $\gamma$ represents the balance between the score matrix $\boldsymbol{Pre}_{\mathrm{Drug}}$ by NormDrug and that of $\boldsymbol{Pre}_{\mathrm{Tar}}$ by NormTarget

$$\gamma = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Pre}_{\mathrm{Drug}}(i,j)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Pre}_{\mathrm{Tar}}(i,j)}. \tag{22}$$

## IV. Experiments

In this study, we conduct extensive experiments to compare the performance of the proposed method with those of the six state-of-the-art methods for determining possible DTIs. We confirm the predicted DTIs via retrieving public databases which are not applied in the learning stage. We conduct two cases, which predict targets of new drugs and drugs targeting new proteins, respectively, to elucidate the prediction performance of the proposed method on new drugs and targets.

### A. Experimental Setup and Evaluation Metrics

We compare the performance of NormMulInf with those of the six state-of-the-art methods, namely, FCML [40], ProWL [38], ProDM [39], RLS-Kron [28], NetCBP [21], and BLM-NII [12]. The parameters of these methods are set as proposed by the corresponding authors in their codes or in the papers. For NormDrug, NormTarget, and NormMulInf, we search the optimal $\lambda$ values within the range of [0.1, 1] with an interval of 0.05 and then set $\lambda$ as 0.6. The performances of these three methods does not obviously change when we vary $\lambda$ around the fixed value. We mask part of interactions for each sample according to MDTIR in the experiments, except for predicting targets of new drugs and drugs targeting new proteins.

DTI prediction can easily result in overfitting problem, and the prediction results are not accurate when the samples size is relatively small. Based on the method proposed by Yu *et al.* [38], we consider all samples within the dataset as training and testing data to decrease bias caused by small samples in the experiments.

Various evaluation metrics have been proposed to evaluate DTI prediction approaches; of which, AUC and AUPR are extensively used. AUC is the average area under the receiver operating characteristic curve and can be calculated using true positives as a function of false positives; this parameter is also a quality measure [61]. High AUC values result in improved performance. AUPR is the area under the precision-recall curve and calculated by the plot of the ratio of true interactions among all predicted DTIs for each given recall rate. AUPR is a quantitative measure that determines how well, on average, the predicted scores of true interactions are separated from the predicted scores of true noninteractions. Higher AUPR value results in improved performance. For DTI prediction, known interactions are relatively rare. As such, AUPR is a more effective quality assessment tool than AUC because the former adopts several measures to reduce the influence of predicted false DTI data among highest ranked scores [62]. In particular, the AUPR score is a more reasonable evaluation metric than the AUC score in certain instances [63]. We used these two metrics to evaluate the performance of the proposed method.

TABLE II
Prediction Performance Comparison on Enzyme Dataset

| Metric | MDTIR | FCML | NetCBP | ProWL | ProDM | RLS-Kron | BLM-NII | NormMulInf |
|---|---|---|---|---|---|---|---|---|
| AUC | 0.2 | .8827 | .8102 | .8739 | .9293 | .9589 | **.9643** | .9583 |
| | 0.4 | .8563 | .7694 | .8475 | .8912 | .9246 | **.9295** | .9251 |
| | 0.6 | .8126 | .7214 | .8093 | .8523 | .8687 | **.8859** | **.8862** |
| | 0.8 | .7459 | .6607 | .7438 | .7815 | .8030 | .8284 | **.8316** |
| AUPR | 0.2 | .8676 | .7342 | .8627 | .9063 | .8975 | .9217 | **.9324** |
| | 0.4 | .8164 | .6901 | .8252 | .8715 | .8649 | .8939 | **.9058** |
| | 0.6 | .7581 | .6454 | .7740 | .8273 | .8161 | .8506 | **.8635** |
| | 0.8 | .6952 | .5726 | .7218 | .7628 | .7512 | .8023 | **.8149** |

### B. Performance on Predicting Interactions Data

In this section, we performed experiments to evaluate and compare the performance of NormMulInf with FCML [40], NetCBP [21], ProWL [38], ProDM [39], RLS-Kron [28], and BLM-NII [12]. We varied the MDTIR from 0.2 to 0.8 for each sample, with an interval of 0.2. We performed the experiments 20 times and calculated the average performance. Tables II–V

TABLE III
PREDICTION PERFORMANCE COMPARISON ON ION CHANNEL DATASET

| Metric | MDTIR | FCML | NetCBP | ProWL | ProDM | RLS-Kron | BLM-NII | NormMulInf |
|---|---|---|---|---|---|---|---|---|
| AUC | 0.2 | .7508 | .7936 | .8828 | .9402 | .9097 | **.9683** | .9389 |
| | 0.4 | .7116 | .7418 | .8401 | .9087 | .8674 | **.9254** | .9112 |
| | 0.6 | .6673 | .6925 | .8014 | .8535 | .8256 | **.8819** | .8721 |
| | 0.8 | .5837 | .6053 | .7495 | .7818 | .7569 | **.8241** | **.8234** |
| AUPR | 0.2 | .7190 | .7501 | .8451 | .8833 | .8662 | **.9248** | .9125 |
| | 0.4 | .6826 | .7237 | .8094 | .8618 | .8450 | **.8917** | .8869 |
| | 0.6 | .6432 | .6754 | .7645 | .8182 | .8031 | **.8491** | **.8487** |
| | 0.8 | .5647 | .5780 | .6979 | .7296 | .7154 | .7839 | **.7862** |

TABLE IV
PREDICTION PERFORMANCE COMPARISON ON GPCRS DATASET

| Metric | MDTIR | FCML | NetCBP | ProWL | ProDM | RLS-Kron | BLM-NII | NormMulInf |
|---|---|---|---|---|---|---|---|---|
| AUC | 0.2 | .7852 | .8083 | . 8496 | .9247 | .8980 | **.9624** | .9481 |
| | 0.4 | .7474 | .7604 | .8113 | .8906 | .8568 | **.9287** | .9215 |
| | 0.6 | .7035 | .7156 | .7517 | .8419 | .8073 | .8812 | **.8824** |
| | 0.8 | .6296 | .6445 | .6764 | .7721 | .7397 | .8194 | **.8253** |
| AUPR | 0.2 | .7025 | .7551 | .7439 | **.8784** | .7752 | .8586 | **.8789** |
| | 0.4 | .6643 | .7130 | .7037 | .8340 | .7396 | .8235 | **.8467** |
| | 0.6 | .6130 | .6649 | .6445 | .7718 | .6881 | .7802 | **.8071** |
| | 0.8 | .5336 | .5962 | .5853 | .7052 | .6119 | .7164 | **.7458** |

TABLE V
PREDICTION PERFORMANCE COMPARISON ON NUCLEAR RECEPTOR DATASET

| Metric | MDTIR | FCML | NetCBP | ProWL | ProDM | RLS-Kron | BLM-NII | NormMulInf |
|---|---|---|---|---|---|---|---|---|
| AUC | 0.2 | .7689 | .8313 | .8616 | .9439 | .8725 | **.9529** | .9412 |
| | 0.4 | .7230 | .7992 | .8263 | **.9122** | .8367 | **.9134** | **.9125** |
| | 0.6 | .6695 | .7494 | .7782 | .8563 | .7829 | .8663 | **.8698** |
| | 0.8 | .5616 | .6514 | .6835 | .7685 | .7042 | .7962 | **.8051** |
| AUPR | 0.2 | .7175 | .7681 | .7958 | **.8583** | .6612 | .8532 | **.8569** |
| | 0.4 | .6602 | .7174 | .7469 | .8175 | .6201 | .8114 | **.8193** |
| | 0.6 | .6034 | .6616 | .6917 | .7725 | .5738 | .7638 | **.7745** |
| | 0.8 | .5326 | .5842 | .6335 | .6859 | .5123 | .7005 | **.7136** |

summarize the performance of all methods in terms of AUC and AUPR. The highest and comparable performances are presented in **boldface**. As shown in Tables II–V, NormMulInf generates promising performance under the majority of conditions or remains the same in the few remaining conditions.

As a state-of-the-art approach in predicting DTIs, NormMulInf performs more efficiently than the other methods and exhibits a significant advantage. The results explain that NormMulInf can efficiently mine underlying DTIs when known DTI data decrease. For example, AUPR values are used in the enzyme dataset. The AUPR values in NormMulInf increase by 6.95%, 21.26%, 7.48%, 2.80%, 3.74%, and 1.15% compared with those in FCML, NetCBP, ProWL, ProDM, RLS-Kron, and BLM-NII when MDTIR is 0.2; the values also increase by 9.9%, 23.81%, 8.90%, 3.79%, 4.52%, and 1.31%, respectively, when MDTIR is 0.4. The values also increase by 12.21%, 25.26%, 10.34%, 4.20%, 5.49%, and 1.49%, respectively, when MDTIR is 0.6 and further increase by 14.69%, 29.73%, 11.42%, 6.39%, 7.82% and 1.55%, respectively, when MDTIR is 0.8.

The efficiencies of these methods decrease gradually when the MDTIR increases from 0.2 to 0.8. However, the robust of NormMulInf performs more efficiently than the other comparative approaches when masked DTI increases. For example, AUPR values are used in the enzyme dataset. When the MDTIR increases from 0.2 to 0.8, the AUPR scores of FCML decreases by 6.27%, 7.69%, and 9.05%. NetCBP is reduced by 6.39%, 6.93%, and 12.71%. ProWL decreases at ratios of 4.54%, 6.59%, and 7.26%. ProDM decreases from 4.0% to 5.34% and then 8.46%. RLS-Kron declines by 3.77%, 5.98%, and 8.64%. BLM-NII declines by 3.11%, 5.1%, and 6.02%. The decreased ratios in NormMulInf are considerably lower than those of the other six methods, which are 2.94%, 4.90%, and 5.97%.

NormMulInf remains more efficient than BLM-NII, which is the current state-of-the-art DTI prediction approach, but is found to be inferior in the ion channel dataset. NormMulInf is distinctly superior to BLM-NII in GPCR and nuclear receptor datasets. Meanwhile, BLM-NII outperforms the other five competitors over the two evaluation metrics. ProDM significantly outperforms ProWL, which agrees with the conclusion in a previous study [38] and confirms the advantage of considering dependences between drugs and targets.

The performance of NormMulInf is improved at different levels among the different datasets. For instance, NormMulInf generally obtains higher significant improvement in the enzyme dataset and less distinct improvement in the nuclear receptor dataset than ProDM. In contrast to NetCBP, NormMulInf obtains a more remarkable improvement in the ion channel dataset and a less prominent improvement in the nuclear receptor dataset. These differences in the rate of improvement can be attributed to variation in data structures in the four datasets. Based on the comprehensive evaluation of the experimental results, NormMulInf performs the optimal performance, followed by BLM-NII, ProDM, RLS-Kron, ProWL, NetCBP, and FCML.

In the enzyme dataset, we predict that drug D00437 interacts with target hsa:1559; this pair obtains the highest score. D00437 is annotated as nifedipine(*JP16/USP/INN*), which acts mainly on vascular smooth muscle cells and is used for treatment of hypertension and chronic stable angina [14]. Hsa:1559 is annotated as cytochrome P450, family 2, subfamily C, polypeptide 9. Cytochrome P450, which consists of heme-thiolate monooxygenases, oxidizes various structurally unrelated compounds and contributes to the wide pharmacokinetics variability of drug metabolism [64]. This interaction was also predicted by Gönen [31], Xia [26], and Laarhoven [28], which is ranked 1, 3 and 5, respectively, and validated in the DrugBank, Metador, and ChEMBL databases. D00437 interacts with hsa:1555, hsa:1558, hsa:1562, hsa:1565, hsa:1571, hsa:1572, and hsa:1573 in the "gold standard" datasets. The target proteins are all annotated as cytochrome P450, family 2. Their functions are very similar to hsa:1559. Therefore, we conclude that D00437 may interact with hsa:1559.

In the ion channel dataset, we determine that the DTI pair with the highest score is D00538-hsa:6331. D00538 is annotated as zonisamide (*JAN/USAN/INN*), which is the approved adjunctive therapy in adults with partial onset seizures [14]. Hsa:6331 is annotated as sodium channel, voltage-gated, type V, alpha subunit. The protein mediates the voltage-dependent permeability of the sodium ions of the excitable membranes

[64]. This interaction was also predicted by van Laarhoven and Marchiori [65] and Gönen [31], which are both ranked 2, and reported in the ChEMBL and DrugBank databases. D00538 interacts with hsa:6323, hsa:6328, hsa:6329, and hsa:6336 in the "gold standard" datasets. The target proteins are all annotated as sodium channel protein and their functions are very similar to hsa:6331. Therefore, we conclude that D00538 may interact with hsa:6331.

In the GPCR dataset, we predict that the pair with the highest interaction score is D00283 and hsa:1814. D00283 is annotated as clozapine (*JAN/USP/INN*), which is an atypical antipsychotic agent that binds to several types of central nervous system receptors and exhibits a unique pharmacological profile. Hsa:1814 is annotated as dopamine receptor D3 [14], [17], whose activity is mediated by G proteins, and inhibits adenylyl cyclase. The protein promotes cell proliferation [64]. This interaction was also predicted by Gönen [31], Xia [26], Laarhoven [28], and Laarhoven [65], which is ranked 3, 5, 1, and 1, respectively, and can be retrieved from the ChEMBL, Metador, and DrugBank databases. D00283 interacts with hsa:1812, hsa:1813, and hsa:1815 in the "gold standard" datasets. The target proteins are all annotated as dopamine receptor and their functions are very similar to hsa:1814. Therefore, we conclude that D00283 may interact with hsa:1814.

In the nuclear receptor dataset, we predict that the interaction of D00348 with hsa:5915 obtains the highest score. D00348 is annotated as isotretinoin (*USP*), which is a compound used to treat severe acne and prevent certain skin cancers types [14]. The target protein hsa:5915 is annotated as the retinoic acid receptor. In the absence or presence of a hormone ligand, the protein acts mainly as gene expression activator because of weak binding to corepressors. Combined with RARG, it is required for skeletal growth, matrix homeostasis and growth plate function [64]. This interaction was also predicted by Xia *et al.*[26], van Laarhoven and Marchiori[65], and Gönen [31], which is ranked 1, 3, and 2, respectively, and reported in the ChEMBL and KEGG databases. Very similar to the function of hsa:5915, hsa:5914 is also annotated as the retinoic acid receptor. D00348 interacts with hsa:5914 in the "gold standard" datasets. Therefore, we conclude that D00348 may interact with hsa:5915.

### C. Other Performance Evaluations

In this section, we further analyze the performance of the proposed approach.

*1) Performance Comparison Considering Local Correlations Among Labels of Samples or Not:* In this section, we compare the method considering **local** correlations among the labels of samples in the DTI network (NormLocal) with the method that does **not** consider **local** correlations (NormNoLocal). NormNoLocal measures drug and target similarities by using the chemical structures of drugs and the sequences of target proteins. By contrast, NormLocal measures drug and target similarities by combining the chemical structures of drugs, the sequences of target proteins, and the local correlations among the labels of samples in the DTI network. We present the comparative results in the four datasets in terms of AUC and AUPR

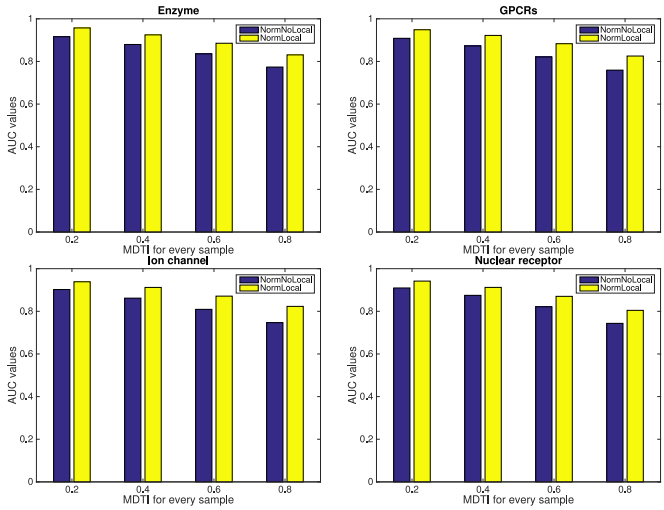



Fig. 1. Performance comparison of prediction considering local correlation of labels between samples or not in terms of AUC on four datasets.

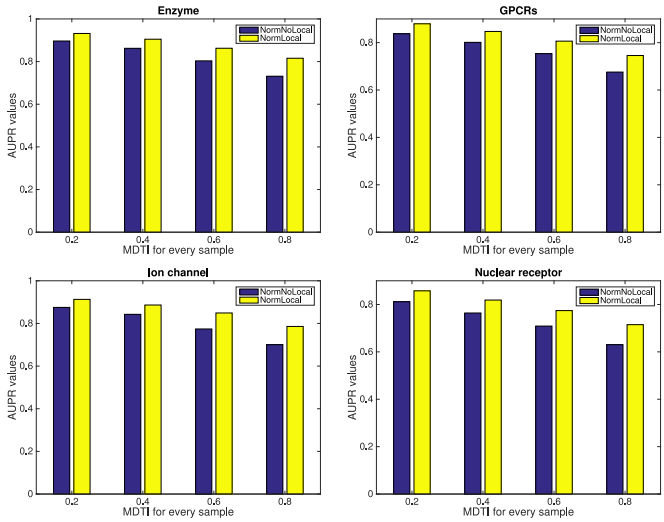



Fig. 2. Performance comparison of prediction considering local correlation of labels between samples or not in terms of AUPR on four datasets.

scores (see Figs. 1 and 2). The results confirm the feasibility of integrating local correlation information of the labels between the samples. As the number of masked interactions increases, the reliability of prediction efficiency decreases, and replenishment of the missing data becomes difficult.

*2) Performance Comparison Incorporating Various Information:* We investigate the performances of the proposed approaches, namely, NormDrug, NormTarget, and NormMulInf. Figs. 3–4 indicate that the performance of the three approaches gradually declines with decreasing MDTIR for each sample. NormMulInf is superior to NormDrug and NormTarget probably because it incorporates more information compared with the latter two. The experimental results confirm that the known biological information can improve prediction efficiency. Furthermore, NormTarget outperforms NormDrug in the ion channel, GPCRs, and nuclear receptor dataset, in which the average number of drugs for each target is higher than the average number of targets for each drug.

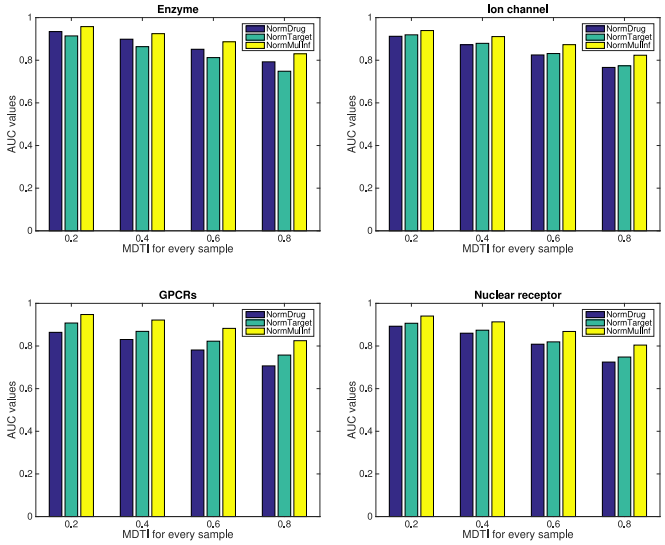



Fig. 3. Performance comparison of prediction in terms of AUC on four datasets based on multi-information fusion.

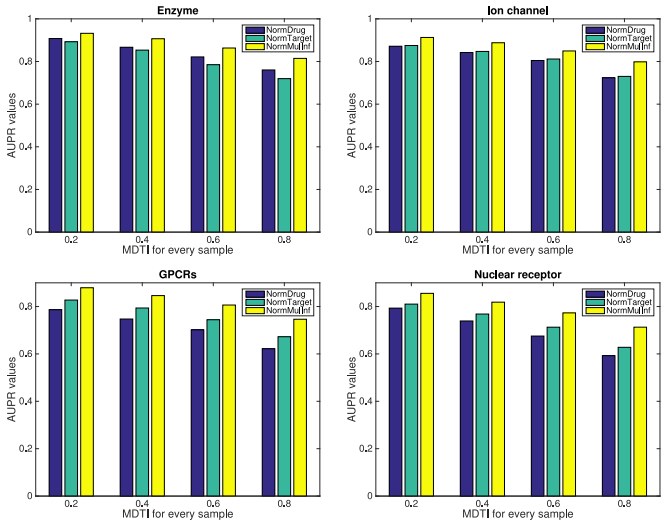



Fig. 4. Performance comparison of prediction in terms of AUPR on four datasets based on multi-information fusion.

*3) Case Predicting Targets of New Drugs:* To investigate the prediction performance of NormMulInf for new drugs, we conducted a case study on atropine, an antimuscarinic agent that binds and inhibits muscarinic acetylcholine receptors, thereby producing various anticholinergic effects. Adequate doses of atropine can eliminate various types of reflex vagal cardiac slowing or asystole [14]. Therefore, mining the potential targets of this drug is important.

Masking performed in this part differs from that in NormMulInf. We consider atropine as a new drug and keep all DTIs, except that the labels associated with the drug are set as 0 in the original DTI network. Thus, we do not know its targets and intend to identify them. The 95 potential targets from human GPCRs are scored according to NormMulInf. The five biochemical experimentally validated targets, namely, hsa:1128 (cholinergic receptor, muscarinic 1), hsa:1129 (cholinergic receptor, muscarinic 2), hsa:1131 (cholinergic receptor, muscarinic 3), hsa:1132 (cholinergic receptor, muscarinic 4), and hsa:1133 (cholinergic receptor, muscarinic 5), are ranked 1, 3, 15, 23, and 26, respectively. This observation indicates that two of the five targets are included in the top 4% of the 95 potential targets. All known targets are also included in the top 28% of the targets. Meanwhile, we predict that atropine interacts with hsa:147 (Alpha-1B adrenergic receptor) and hsa:153 (Beta-1 adrenergic receptor), which are ranked 2 and 4, respectively.

*4) Case Predicting Drugs Targeting New Proteins:* We also evaluated the prediction performance of NormMulInf for new targets. A case study about the target hsa:3357 (5-hydroxytryptamine receptor 2B, 5HT2B) was conducted. 5HT2B functions as a receptor for various ergot alkaloid derivatives and psychoactive substances and affects neural activity. 5HT2B regulates behavior, including impulsive behavior, and is involved in the adaptation of pulmonary arteries to chronic hypoxia. 5HT2B is also required for normal proliferation of embryonic cardiac myocytes and normal heart development to ensure normal osteoblast function and proliferation, as well as for maintaining normal bone density [64]. Therefore, identifying potential drugs targeting 5HT2B exhibits great significance.

We consider 5HT2B as a new target protein and keep all DTIs, except that the labels associated with the target are set as 0 in the original DTI network. Thus, we do not know its targeting drugs and intend to identify them. All 223 potential targeting drugs from human GPCRs are scored according to NormMulInf. The six biochemical experimentally validated targeting drugs, namely, D00283 (Clozapine (*JAN/USP/INN*)), D00451 (Sumatriptan (*JAN/USP/INN*)), D00513 (Pindolol (*JP16/USP/INN*)), D00726 (Metoclopramide (*JP16/INN*)), D01164 (Aripiprazole (*JAN/USAN/INN*)), D01973 (Eletriptan hydrobromide (*JAN/USAN*)), are ranked 1, 6, 3, 4, 15, and 19, respectively. This result indicates that four of the six targeting drugs are included in the top 3% of the 223 potential drugs. All known targeting drugs are also included in the top 9% of the drugs. We also predict that 5HT2B is targeted by drug olanzapine (*JAN/USAN/INN*) and propiomazine (*USAN/INN*), which are ranked 2 and 5, respectively.

## V. Discussion

In this section, we discuss the experimental results described in the preceding section.

In the "gold standard" datasets, the DTI data are sparse, low-rank, and imbalanced. The number of known interactions are lower than that of unknown ones. Therefore, various computational methods can be used to determine potential DTIs. We compare the performance of the proposed approach with those of other comparative methods on four benchmark datasets, which include human enzymes, ion channels, GPCRs, and nuclear receptors. The originality of the proposed approach remains, that is, making full use of unlabeled data, integrating various biological information, and applying robust PCA method, which minimizes the combination of nuclear norm and $\ell_1$-norm, to DTI prediction. The experimental results reveal the merits of the model. High increases in AUC and AUPR indicate that the DTIs predicted using the proposed approach are likely to be more accurate than those predicted by other methods.

NormMulInf can achieve superior results regardless of the AUC or AUPR results. This observation may be attributed to the following features of the algorithm. 1) The algorithm incorporates various biological information, particularly similarities among the samples and the local correlations among the labels of samples in the DTI network. 2) The method makes full use of unlabeled data in the DTI network. 3) Robust PCA solved by ALMs exhibits good convergence and can converge to the optimal solution [47]. 4) To decrease bias caused by small sample, the algorithm considers all samples in the dataset as training and testing data.

The proposed approach is also beneficial in the design and interpretation of pharmacological experiments, particularly in identifying novel DTIs and addressing problems related to determining multitarget drugs and multidrug targets. The technique can be further used to investigate other biological associations similar to DTI, such as microRNA-disease, gene-disease, and drug-complex associations.

## VI. Conclusion and Further Research

In this study, we developed a novel approach for DTI prediction, which integrates robust PCA with various biological information into a unified framework. We conducted a comparative evaluation of the proposed approach using four benchmark datasets. The experimental results suggest that the proposed approach can achieve superior classification results and can competitively predict DTIs. Further analysis showed that the DTIs predicted by the proposed method are worthy of further experimental validation.

Using large amount of biological information related to drugs and targets can improve the efficiency of the technique. Integrating various biological information can help identify new DTIs; however, in this study, we do not fully use this additional biological information. Therefore, with additional information related to drug and target validated by biochemical experiments, we will integrate a large amount of information in subsequent investigations, for example, drug–drug interactions, protein–protein interactions, and side effects of drugs. Furthermore, we will extend similarity measures as a regression to make model be more general.

There are a small quantity of available labeled data validated by biomedical experiments and abundant unlabeled data. We make a correct point about the unlabeled interactions, which are not truly negative DTIs and should be identified with an unsupervised model. However, Negative DTI data are not reported and are unavailable. When using AUPR and AUC for evaluation, part of unlabeled interactions are being assumed negative samples, which may affect the accuracy of the method. Therefore, another way to improve the performance is by building a negative dataset; investigation of this technique is currently underway.

## Acknowledgment

The corresponding author of this paper is Bo Liao (dragonbw@163.com).

## References

[1] S. Zu, T. Chen, and S. Li, "Global optimization-based inference of chemogenomic features from drug–target interactions," *Bioinformatics*, vol. 31, pp. 2323–2529, 2015.
[2] J.-Y. Shi, S.-M. Yiu, Y. Li, H. C. Leung, and F. Y. Chin, "Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering," *Methods*, vol. 83, pp. 98–104, 2015.
[3] S. Alaimo, V. Bonnici, D. Cancemi, A. Ferro, R. Giugno, and A. Pulvirenti, "Dt-web: A web-based application for drug–target interaction and drug combination prediction through domain-tuned network-based inference," *BMC Syst. Biol.*, vol. 9, no. Suppl 3, p. S4, 2015.
[4] G. Chevereau and T. Bollenbach, "Systematic discovery of drug interaction mechanisms," *Molecular Syst. Biol.*, vol. 11, no. 4, pp. 807–815, 2015.
[5] M. A. Heiskanen and T. Aittokallio, "Predicting drug–target interactions through integrative analysis of chemogenetic assays in yeast," *Molecular BioSyst.*, vol. 9, no. 4, pp. 768–779, 2013.
[6] R. A. Copeland, "Drug-target interactions: Stay tuned," *Nature Chem. Biol.*, vol. 11, pp. 451–452, 2015.
[7] Á. R. Perez-Lopez, K. Z. Szalay, D. Türei, D. Módos, K. Lenti, T. Korcsmáros, and P. Csermely, "Targets of drugs are generally, and targets of drugs having side effects are specifically good spreaders of human interactome perturbations," *Sci. Rep.*, vol. 5, 2015.
[8] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-based machine learning methods for predicting drug–target interactions: A brief review," *Briefings Bioinformat.*, vol. 15, no. 5, pp. 734–747, 2014.
[9] R. C. NCBI, "Database resources of the national center for biotechnology information." *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D7–D17, 2014.
[10] A. M. Wassermann, E. Lounkine, J. W. Davies, M. Glick, and L. M. Camargo, "The opportunities of mining historical and collective data in drug discovery," *Drug Discovery Today*, vol. 20, no. 4, pp. 422–434, 2015.
[11] Y. Wang and J. Zeng, "Predicting drug-target interactions using restricted Boltzmann machines," *Bioinformatics*, vol. 29, no. 13, pp. i126–i134, 2013.
[12] 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, 2013.
[13] P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review," *Pharmacol. Therapeutics*, vol. 138, no. 3, pp. 333–408, 2013.
[14] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "Drugbank 4.0: Shedding new light on drug metabolism," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1091–D1097, 2014.
[15] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "Kegg for linking genomes to life and the environment," *Nucleic Acids Res.*, vol. 36, no. suppl 1, pp. D480–D484, 2008.
[16] N. Bhardwaj, M. Källberg, W. Cho, H. Lu, Y. Pan, J. Wang, and M. Li, "Metador: Online resource and prediction server for membrane targeting peripheral proteins," *Algorithmic Artificial Intell. Methods Protein Bioinformat.*, pp. 481–494, 2013.
[17] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. P. Overington, "The chembl bioactivity database: An update," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1083–D1090, 2014.
[18] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature Biotechnol.*, vol. 25, no. 2, pp. 197–206, 2007.
[19] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, and E. S. Huang, "Structure-based maximal affinity model predicts small-molecule druggability," *Nature Biotechnol.*, vol. 25, no. 1, pp. 71–75, 2007.
[20] S. Zhu, Y. Okuno, G. Tsujimoto, and H. Mamitsuka, "A probabilistic model for mining implicit 'chemical compound–gene' relations from literature," *Bioinformatics*, vol. 21, no. suppl 2, pp. ii245–ii251, 2005.
[21] H. Chen and Z. Zhang, "A semi-supervised method for drug-target interaction prediction with consistency in networks," *PloS One*, vol. 8, no. 5, p. e62975, 2013.

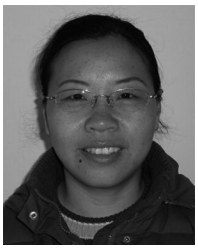
[22] Y.-C. Wang, C.-H. Zhang, N.-Y. Deng, and Y. Wang, "Kernel-based data fusion improves the drug–protein interaction prediction," *Comput. Biol. Chemistry*, vol. 35, no. 6, pp. 353–362, 2011.
[23] Y. C. Wang, N. Deng, S. Chen, and Y. Wang, "Computational study of drugs by integrating omics data with kernel methods," *Molecular Informat.*, vol. 32, nos. 11/12, pp. 930–941, 2013.
[24] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.
[25] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug–target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.
[26] Z. Xia, L.-Y. Wu, X. Zhou, and S. T. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," *BMC Syst. Biol.*, vol. 4, no. Suppl 2, p. S6, 2010.
[27] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, 2010.
[28] 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.
[29] S. Fakhraei, B. Huang, L. Raschid, and L. Getoor, "Network-based drug-target interaction prediction with probabilistic soft logic," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 5, pp. 775–787, Sep./Oct. 2014.
[30] 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 Comput. Biol.*, vol. 8, no. 5, p. e1002503, 2012.
[31] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.
[32] S. Funar-Timofei and L. Kurunczi, "Reply to quantitative structure–affinity relationship study of azo dyes for cellulose fibers by multiple linear regression and artificial neural network," *Dyes Pigments*, vol. 113, pp. 325–326, 2015.
[33] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, and T. Aittokallio, "Toward more realistic drug–target interaction predictions," *Briefings Bioinformat.*, vol. 16, pp. 325–337, 2014.
[34] H. Liu, J. Sun, J. Guan, J. Zheng, and S. Zhou, "Improving compound–protein interaction prediction by building up highly credible negative samples," *Bioinformatics*, vol. 31, no. 12, pp. i221–i229, 2015.
[35] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," *Molecular BioSyst.*, vol. 8, no. 7, pp. 1970–1978, 2012.
[36] X. Xiao, J.-L. Min, W.-Z. Lin, Z. Liu, X. Cheng, and K.-C. Chou, "idrug-target: Predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach," *J. Biomolecular Struct. Dyn.*, vol. 33, pp. 2221–2233, 2015.
[37] C. Wang, J. Liu, F. Luo, Z. Deng, and Q.-N. Hu, "Predicting target-ligand interactions using protein ligand-binding site and ligand substructures," *BMC Syst. Biol.*, vol. 9, no. Suppl 1, p. S2, 2015.
[38] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction with incomplete annotations," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 3, pp. 579–591, May/Jun. 2014.
[39] G. Yu, C. Domeniconi, H. Rangwala, and G. Zhang, "Protein function prediction using dependence maximization," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2013, pp. 574–589.
[40] H. Wang, H. Huang, and C. Ding, "Function–function correlated multi-label protein function prediction over interaction networks," *J. Comput. Biol.*, vol. 20, no. 4, pp. 322–343, 2013.
[41] G. Yu, H. Zhu, C. Domeniconi, and J. Liu, "Predicting protein function via downward random walks on a gene ontology," *BMC Bioinformat.*, vol. 16, no. 1, pp. 271–283, 2015.
[42] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, "Combining drug and gene similarity measures for drug-target elucidation," *J. Comput. Biol.*, vol. 18, no. 2, pp. 133–145, 2011.
[43] F. Martínez-Jiménez and M. A. Marti-Renom, "Ligand-target prediction by structural network biology using nannolyze," *PLoS Comput. Biol.*, vol. 11, no. 3, p. e1004157, 2015.
[44] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Wiley, 2002.
[45] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.
[46] T. Bouwmans and E. H. Zahzah, "Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance," *Comput. Vis. Image Understanding*, vol. 122, pp. 22–34, 2014.
[47] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *PLoS One*, vol. 9, 2010.
[48] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," in *Proc. Comput. Adv. Multi-Sensor Adaptive Process.*, 2009, vol. 61.
[49] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, p. 4, 2009.
[50] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1986.
[51] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 285–295.
[52] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel, "Probabilistic memory-based collaborative filtering," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 1, pp. 56–69, Jan. 2004.
[53] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proc. 14th Conf. Uncertainty Artificial Intell.*, 1998, pp. 43–52.
[54] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles, "Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach," in *Proc. 16th Conf. Uncertainty Artif. Intell.*, 2000, pp. 473–480.
[55] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways," *J. Amer. Chem. Soc.*, vol. 125, no. 39, pp. 11853–11865, 2003.
[56] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Molecular Biol.*, vol. 147, no. 1, pp. 195–197, 1981.
[57] K. P. Nigam, "Using unlabeled data to improve text classification," Ph.D. dissertation, School Comput. Sci, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2001.
[58] S.-J. Huang, Z.-H. Zhou, and Z. Zhou, "Multi-label learning by exploiting label correlations locally," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 949–955.
[59] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[60] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," School Comput. Sci, Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CALD-02-107, 2002.
[61] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
[62] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.
[63] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.
[64] U. Consortium, "Reorganizing the protein space at the universal protein resource (uniprot)," *Nucleic Acids Res.*, vol. 40, pp. D71–D75, 2011.
[65] 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.

**Lihong Peng** was born in Hunan, China. She is currently working toward the Ph.D. degree in the College of Information Science and Engineering, Hunan University, Changsha, China.

Her research interests include machine learning, data mining, and bioinformatics.

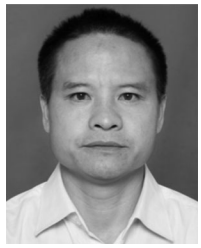
**Bo Liao** received the Ph.D. degree in computational mathematics from the Dalian University of Technology, Dalian, China, in 2004.

He is currently working in Hunan University as a Professor. He worked in the Graduate University of Chinese Academy of Sciences as a Postdoctorate from 2004 to 2006. His current research interests include bioinformatics, data mining, and machine learning.



**Wen Zhu** received the M.Sc. degree in computer science and technology from the Hunan University, China, in 2010.

He is currently working in Hunan University as a Lecturer. Her current research interest includes bioinformatics, data mining, and machine learning.



**Zejun Li** is currently working toward the Ph.D. degree in the College of Information Science and Engineering, Hunan University, Changsha, China.

His research interests include machine learning and bioinformatics.

**Keqin Li** is a SUNY Distinguished Professor of computer science. His current research interests include parallel computing and high-performance computing, distributed computing, energy-efficient computing and communication, heterogeneous computing systems, cloud computing, big data computing, CPU-GPU hybrid and cooperative computing, multicore computing, storage and file systems, wireless communication networks, sensor networks, peer-to-peer file sharing systems, mobile computing, service computing, Internet of things and cyber-physical systems. He has published more than 390 journal articles, book chapters, and refereed conference papers, and has received several best paper awards. He is currently or has served on the editorial boards of IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON CLOUD COMPUTING, *Journal of Parallel and Distributed Computing*.