Computational prediction of Drug-Disease association based on Graph-regularized one bit Matrix completion

Aanchal Mongia, Emilie Chouzenoux, Angshul Majumdar

Computational prediction of Drug-Disease association based on Graph-regularized one bit Matrix completion

Aanchal Mongia, Emilie Chouzenoux and Angshul Majumdar*

Abstract—Investigation of existing drugs is an effective alternative to the discovery of new drugs for treating diseases. This task of drug re-positioning can be assisted by various kinds of computational methods to predict the best indication for a drug given the open-source biological datasets. Owing to the fact that similar drugs tend to have common pathways and disease indications, the association matrix is assumed to be of low-rank structure. Hence, the problem of drug-disease association prediction can be modeled as a low-rank matrix completion problem.
In this work, we propose a novel matrix completion framework that makes use of the side-information associated with drugs/diseases for the prediction of drug-disease indications modeled as neighborhood graph: Graph regularized 1-bit matrix completion (GR1BMC). The algorithm is specially designed for binary data and uses parallel proximal algorithm to solve the aforesaid minimization problem taking into account all the constraints including the neighborhood graph incorporation and restricting predicted scores within the specified range. The results have been validated on two standard databases by evaluating the AUC across the 10-fold cross-validation splits. The usage of the method is also evaluated through a case study where top 5 indications are predicted for novel drugs, which then are verified with the CTD database.

Index Terms—Drug-Disease association, Graph regularization, Matrix Completion

1 Introduction

Inspite of the large financial investment in pharmaceutical industry, the number of drugs approved over the past few decades is limited [1]. This can be attributed to the time (10-15 years) and effort it takes to test a therapeutic compound and declare it as a market-ready drug. The problem calls for an alternative to drug discovery: "drug-repositioning" or "drug-repurposing". This essentially means that an existing, already approved drug is identified to seek its new indications. The benefits include shorter drug-development timelines, established safety and savings on money for launching the drug. Also, the strategy of drug-repurposing offers an opportunity to overcome the threats associated with antimicrobial resistance (AMR) [2]. Some examples of re-positioned drugs include chlorocyclizine, an anti-allergic drug re-purposed as an antiviral [3], sertraline, an antidepressant drug as an antifungal [4] and disulfiram, an anti-alcoholic drug repurposed as an antibacterial [5], [6].

There have been some successfully re-positioned drugs through manual and rational investigations but this is not an efficient and scalable way given the huge space of drug interactions. Therefore, computational approaches have been used over the past years to systematically predict the indications, pruning down the massive search space for researchers and saving huge amounts of effort, time and cost. This explains the immense importance of predicting new associations between drugs and diseases using statistical and machine learning-based methods .

Initial attempts to predict novel indications were based on gene expression profiles [7], [8], [9]. [7] proposed a database having ranked drug response gene expression which was queried with a gene signature specific to a disease. The drug response profiles which either correlate or anti-correlate were identified. This approach lacks validation on a large scale dataset and may not be precise enough owing to different conditions under which expression profiles are generated.

Other sets of approaches captured the notion of similarity [10] where it was assumed that alternative for one of the two diseases which are treated by the same drug, may also be used as a potential treatment for the other disease.

Later, network-based models were proposed. [11], [12] proposed PREDICT, a method which computationally predicts drug-disease associations using integrated drug and disease information. Various kinds of drug and disease similarities are calculated to find the feature vectors for the candidate associations which are further used to train a classification model using logistic regression. [13] created a 3-layer heterogeneous network, corresponding to drug, disease and targets. Edge weights between the nodes of same type (i.e. intra-connections) correspond to similarity between them while those between different types of nodes are associated with the relationship or association between

- *Aanchal Mongia is with Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, New Delhi 110020, India. E-mail: see aanchalm@iiitd.ac.in*
- *Emilie Chouzenoux is with CVN, Inria Saclay, Univ. Paris Saclay, 91190 Gif-sur-Yvette, France. Email: emilie.chouzenoux@centralesupelec.fr*
- **Angshul Majumdar (corresponding author) is with the Department of Electronics Engineering, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, New Delhi 110020, India. Email: angshul@iiitd.ac.in*

the nodes i.e. drug-target or drug-disease relationship. The missing edges of this network are inferred using guilt-by-association principle. In a similar fashion, [14] integrated information from drugs, diseases and targets and proposed a network-based prioritization method for predicting new drug indications and novel disease treatments. Another work, [15] integrates molecular structure, molecular activity, and phenotype data and constructs a kernel function to correlate drugs with diseases, and finally train an SVM (Support vector machine) classifier for the prediction of drug-disease interaction. [16] identifies the drug/disease modules by clustering the drug network and disease network and then connecting drug-disease module pairs. Very recently, a new network-based approach was proposed by Yang at al [17] where the authors employ heterogeneous network embedding for the characterization of drug-disease association and trains an SVM for predicting novel associations.

There have also been several machine learning and deep learning techniques used for association prediction apart from the ones (clustering and classification methods) used in few of the works mentioned above [18]. Very recently, [19] trained a dual convolutional neural network on two association layers simultaneously, one encoding the drug-disease characteristics while another one, the associated neighborhood information. [20] applied graph convolution operation with attention mechanism to the drug-disease heterogenous network to learn the embeddings of drugs and diseases. [21] extracted feature descriptors from drug and disease Gaussian interaction profile based and other similarities using autoencoder and trained a random forest classifier to predict drug-disease associations. [22] trained a neural network on the aggregated neighborhood information with the drugs and diseases association and similarity matrices; they minimize the loss between initial and recovered matrices while training the neural network on the heterogeneous data.

Drug-disease association prediction can also be modelled intuitively as a collaborative filtering problem. The objective of this class of approaches is to recover a complete matrix from its sampled entries by exploiting its low-rank structure. The low-rank assumption stems from the idea that similar drugs affect biological systems in a similar way and have common indications [23].

The underlying techniques which aim to solve collaborative filtering problem via matrix completion are majorly based on matrix factorization or nuclear norm minimization. Matrix factorization has been employed in the community over the past few years. It assumes that the number of latent (or hidden) features which may determine the association between a drug and a disease (such as substructures, targets, enzymes, pathways, MeSH information, etc) is very few and highly correlated. [24] used probabilistic matrix factorization on causal networks connecting drug–target–pathway–gene–disease to classify drug-disease associations. [25] integrates genomic space into the matrix factorization framework to exploit the molecular biological information using gene interaction network and then predicts novel indications. [12] projects the association information to two low-rank latent spaces, while taking into account the topological information of drug and disease data points by using the similarity information of drugs and diseases in the objective function of matrix factorization. Very recently, [26] deploys multi-similarities bilinear matrix factorization to predict indications (diseases) for existing and novel drugs. Matrix factorization is a bilinear non-convex problem, which makes it challenging to solve, as spurious local minima usually occur. This problem can be overcome by an alternate approach for matrix completion: Nuclear norm minimization [27]. Minimizing the nuclear norm (sum of singular values of a matrix) is the closest convex surrogate to minimizing the rank (number of singular values of a matrix) of that matrix, which is known to be a NP-hard problem. There are relatively few works modelling the prediction task using nuclear norm minimization. [28] and [29] deploy nuclear norm minimization on a heterogeneous network matrix obtained by integrating drug similarity, disease similarity, association matrix and its transpose; the latter work additionally handles the noise originating from similarities which violate the low-rankness and restrict the predicted values to be in range [0,1]. But, the low-rank property of the heterogeneous matrix is unexplained in both the works, although it is a crucial assumption behind nuclear norm minimization. This heterogeneous matrix comprises of associations between drugs and diseases as well as drug-drug and disease-disease similarities. The authors clearly explain validity of the low-rank assumption in association matrix but not for the heterogeneous matrix.

In this work, we formulate drug disease association prediction as a one-bit matrix completion problem. Furthermore, we introduce graph regularization to exploit the similarities between drugs and diseases. The objective function is minimized using parallel proximal algorithm (PPXA) [30]. PPXA is an iterative proximal splitting algorithm that paralelly solves for each of the non-necessarily smooth terms in the objective function, while benefiting from sounded convergence guarantees. The novelty of our approach lies in

- Modelling the drug-disease association prediction as graph-regularized matrix completion problem.
- Restricting the association scores in range [0,1] for obtaining meaningful biological scores.
- Solving the optimization problem using PPXA which has guaranteed convergence properties [31].

A schematic overview of GR1BMC is shown in Figure 1.

2 Material and Methods

2.1 Dataset

We have used two gold standard databases to validate our approach. The first one, called *F dataset*, proposed by [11] has 313 diseases, 593 drugs and 1933 drug-disease associations from various sources. The second dataset, called *Cdataset* is a larger one with 663 drugs, 409 diseases and 2532 associations [32]. In the remaining of the paper, we denote n_1 the number of drugs and n_2, the number of diseases.

For both datasets, the drug information is obtained from DrugBank [33], an exhaustive database containing comprehensive information about drugs and targets. The disease information was assembled from human phenotypes listed in public database, OMIM (Online Mendelian Inheritance in

Man) database [34], which has information on human genes and diseases.

The similarity information among drugs, calculated as Tanimato score [35], is extracted using Chemical Development Kit (CDK) [36] based on the chemical structures of drugs in SMILES (Simplified Molecular-Input Line-Entry System) [37] format, obtained from DrugBank. MimMiner [38] provides the similarities among diseases using the medical descriptors of diseases from OMIM database by measuring the number of MeSH (medical subject headings vocabulary) terms. Both kinds of similarites are encoded as a value in the range $[0, 1]$.
The information on number of drugs, diseases and the known associations among them has been summarized in Table 1, for both datasets.

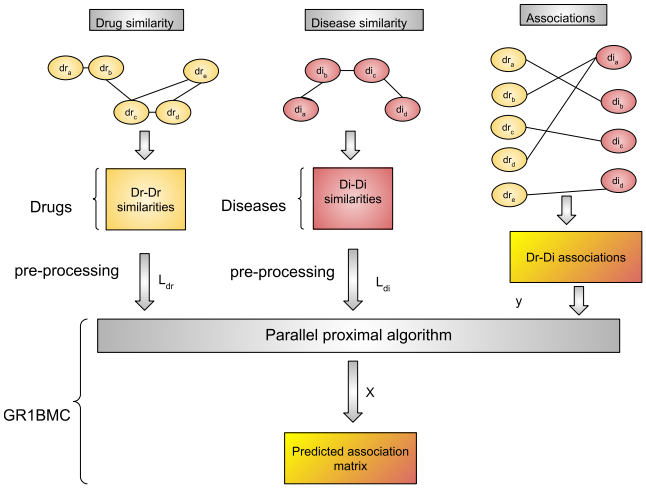

Fig. 1: A schematic overview of GR1BMC for predicting drug-disease assocations

2.2 Preprocessing

As an input, we are given two similarity matrices, namely $S_{dr} \in [0,1]^{n_1 \times n_1}$ and $S_{di} \in [0,1]^{n_2 \times n_2}$. We describe hereafter the pre-processing of these quantities, so as to obtain the Laplacian matrices $L_{dr} \in \mathbb{R}^{n_1 \times n_1}$ and $S_{di} \in \mathbb{R}^{n_2 \times n_2}$, associated to the proposed graph regularization strategy.

2.2.1 Similarity compression:

First, in order to ensure that the local geometries of the association data are preserved during the learning process, both similarity matrices are compressed by keeping only $p-$nearest neighbor of each drug/disease profile in the drug/disease similarity matrix. This is done by multiplying each similarity matrix element-wise (i.e., S_{dr} and S_{di}) with so-called "neighborhood matrices" representing the p-nearest neighbor graphs of drugs (i.e., N_{dr}) and diseases (i.e., N_{di}). The row i and column j of these matrices are given by:

$$N^{ij} = \begin{cases} 1, & \text{if} \quad j \in N_p(i) \quad \text{and} \quad i \in N_p(j), \\ 0, & \text{if} \quad j \notin N_p(i) \quad \text{and} \quad i \notin N_p(j), \\ 0.5, & \text{elsewhere} \end{cases}$$

where $N_p(i)$ is the set of p nearest neighbors to the i-th element (drug or disease). We have set $p = 5$ here. This leads to the compressed similarity matrices:

$$\widehat{S}_{di} = N_{di} \odot S_{di}, \tag{1}$$
$$\widehat{S}_{dr} = N_{dr} \odot S_{dr}, \tag{2}$$

with $\odot$ the element-wise product.

2.2.2 Normalized graph Laplacians:

We use normalized versions of the graph Laplacians [39] associated to $\widehat{S}_{di}$ and $\widehat{S}_{dr}$, given by:

$$L_{di} = (D_{di})^{-1/2}(D_{di} - \widehat{S}_{di})(D_{di})^{-1/2}, \tag{3}$$
$$L_{dr} = (D_{dr})^{-1/2}(D_{dr} - \widehat{S}_{dr})(D_{dr})^{-1/2}, \tag{4}$$
$$\tag{5}$$

with $D_{di} \in \mathbb{R}^{n_1 \times n_1}$ a diagonal matrix with i-th entry equals to $\sum_{j=1}^{n_1}(\widehat{S}_{di})^{ij}$, and $D_{dr} \in \mathbb{R}^{n_2 \times n_2}$ a diagonal matrix with i-th entry equals to $\sum_{j=1}^{n_2}(\widehat{S}_{dr})^{ij}$.

2.3 Proposed Algorithm

2.3.1 Problem formulation:

Our aim is to learn the drug-disease association matrix $X \in [0,1]^{n_1 \times n_2}$, from m known associations, and prior similarity knowledge encoded in two Laplacian matrices L_{dr} and L_{di}. The available associations are stacked into a vector $y \in [0,1]^m$. We furthermore introduce the binary-valued linear operator $R \in \{0,1\}^{m \times n_1 n_2}$ such that the product $R\text{vec}(X)$, with $\text{vec}(X) \in \mathbb{R}^{n_1 n_2}$ stacking the columns of X, contains m elements expected to be closed to the observed ones in y. In a nutshell, our aim is to estimate X such that $y \approx R\,\text{vec}(X)$, and X satisfies some prior knowledge. Namely, we seek for X as a low-rank matrix with elements in the range $[0, 1]$, and with rows (resp. columns) correlated through L_{di} (resp. L_{dr}). Since rank evaluation function leads to NP-hard minimization problems, we make use of its closest convex surrogate i.e. nuclear norm, $\|\cdot\|_*$, defined as the sum of the absolute singular values of a matrix. Note that restricting the association scores in range $[0, 1]$ aims at obtaining meaningful biological scores. To incorporate the disease and drug similarities into our framework, we finally introduce Laplacian graph regularization terms [40], [41]. This leads to the following optimization problem:

$$\underset{X \in \mathbb{R}^{n_1 \times n_2}}{\text{minimize}} \left(\frac{1}{2}\|y - R\,\text{vec}(X)\|^2 + \lambda \|X\|_* + \mu_1 \text{Tr}(X^\top L_{di} X) + \mu_2 \text{Tr}(X L_{dr} X^\top) \right) \quad \text{such that} \quad X \in [0,1]^{n_1 \times n_2}. \tag{6}$$

Here, $\text{Tr}(\cdot)$ denotes the trace operator. The above formulation is a convex, but non-smooth optimization problem that can be solved efficiently using the parallel proximal

TABLE 1: A summary of the number of associations, drugs and diseases in each dataset used.

Datasets	# Associations	# Drugs (n_1)	# Diseases (n_2)
# Fdataset	1933	593	313
# Cdataset	2532	663	409

algorithm (PPXA) [31], [42] which can be seen as a parallel version of ADMM [43]. PPXA benefits from sounded convergence properties [31] and leads to great practical performance, for instance in [44] in the context of biochemistry).

In PPXA algorithm, we solve (6), by introducing $\theta = 5$ proxy variables, associated to each of the five terms in (6) [42]. For each iteration $k \in \mathbb{N}$, we compute the proximity operators, associated to each of these variables:

$$\widehat{X}_1^{(k)} = \underset{X \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left(\frac{\theta}{2} ||y - R\,\mathrm{vec}(X)||^2 + \frac{1}{2} ||X_1^{(k-1)} - X||_F^2 \right), \tag{7}$$

$$\widehat{X}_2^{(k)} = \underset{X \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left(\lambda\theta ||X||_* + \frac{1}{2} ||X_2^{(k-1)} - X||_F^2 \right), \tag{8}$$

$$\widehat{X}_3^{(k)} = \mathrm{Proj}_{[0,1]^{n_1 \times n_2}} \left(X_3^{(k-1)} \right), \tag{9}$$

$$\widehat{X}_4^{(k)} = \underset{X \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left(\theta\mu_1 \mathrm{Tr}(X^\top L_{di} X) + \frac{1}{2} ||X_4^{(k-1)} - X||_F^2 \right), \tag{10}$$

$$\widehat{X}_5^{(k)} = \underset{X \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left(\theta\mu_2 \mathrm{Tr}(X L_{dr} X^\top) + \frac{1}{2} ||X_5^{(k-1)} - X||_F^2 \right). \tag{11}$$

Note that we perform a projection step onto the proxy variable X_3 to ensure that the predicted scores lie in range $[0, 1]$. This projection is equivalent to the proximity operator of the indicator function for this constraint. The next iterate $X^{(k)}$ is finally obtained by averaging over the five proximal values, as follows:

$$\widehat{X}^{(k)} = \frac{1}{\theta} (\widehat{X}_1^{(k)} + \widehat{X}_2^{(k)} + \widehat{X}_3^{(k)} + \widehat{X}_4^{(k)} + \widehat{X}_5^{(k)}) \tag{12}$$

with $\theta = 5$. Furthermore, each of the proxy variables is updated via the following update rule:

$$X_i^{(k)} = X_i^{(k-1)} + 2\widehat{X}^{(k)} - \widehat{X}^{(k-1)} - \widehat{X}_i^{(k)}, i = 1, \ldots, 5. \tag{13}$$

2.3.2 Proximity steps:

Hereafter, we provide the solution for the proximity subproblems (7) to (11):

• We first need to solve for $\widehat{X}_1^{(k)}$. This amounts to solve:

$$\begin{aligned} &\theta(-R^\top)(y - R\widehat{x}_1^{(k)}) + (\widehat{x}_1^{(k)} - x_1^{(k-1)}) = 0 \\ \Leftrightarrow\ &\theta R^\top R \widehat{x}_1^{(k)} - \theta R^\top y + \widehat{x}_1^{(k)} - x_1^{(k-1)} = 0 \\ \Leftrightarrow\ &(\theta R^\top R + I)\widehat{x}_1^{(k)} = x_1^{(k-1)} + \theta R^\top y \\ \Leftrightarrow\ &\widehat{x}_1^{(k)} = (\theta R^\top R + I)^{-1}(x_1^{(k-1)} + \theta R^\top y) \end{aligned} \tag{14}$$

where I is the identity matrix, $\widehat{x}_1^{(k)} = \mathrm{vec}(\widehat{X}_1^{(k)})$ and $x_1^{(k-1)} = \mathrm{vec}(X_1^{(k-1)})$. Then, $\widehat{X}_1^{(k)} = \mathrm{mat}(\widehat{x}_1^{(k)})$, with mat the reciprocal operation of vec.

• The computation of $\widehat{X}_2^{(k)}$ requires to solve the proximity operator of a spectral function (here, the nuclear norm). This problem has been studied, for instance, in [45], [46]. The result can be obtained by soft-thresholding, element-wise, the singular values of $X_2^{(k-1)}$ and multiplying the thresholded singular value matrix by the left and right singular vector matrices of $X_2^{(k-1)}$ i.e.

$$X_2^{(k-1)} = U S^{(k-1)} V^\top \tag{15}$$

$$\widehat{S}^{(k-1)} = \mathrm{soft}(S^{(k-1)}, \lambda\theta/2) \tag{16}$$

$$\widehat{X}_2^{(k)} = U \widehat{S}^{(k-1)} V^\top \tag{17}$$

with

$$\begin{aligned} &\mathrm{soft}(S^{(k-1)}, \lambda\theta/2) = \\ &\qquad \mathrm{sign}(S^{(k-1)}) \max(0, |S^{(k-1)}| - \lambda\theta/2), \end{aligned} \tag{18}$$

where $S^{(k-1)}$ is the singular value diagonal matrix of $X_2^{(k-1)}$, and the sign and $|\cdot|$ operations must be understood element-wise. Moreover, U (resp. V) are the left (resp. right) singular matrices, associated to the SVD decomposition of $X_2^{(k-1)}$.

• The update of $\widehat{X}_3^{(k)}$ is performed element-wise, by capping the entries of $X_3^{(k-1)}$ onto the range $[0, 1]$:

$$\widehat{X}_3^{(k)} = \min(\max(X_3^{(k-1)}, 0), 1). \tag{19}$$

• The resolution of (10) amounts to solving:

$$\begin{aligned} &\theta\mu_1(L_{di}\widehat{X}_4^{(k)} + L_{di}^\top \widehat{X}_4^{(k)}) + (\widehat{X}_4^{(k)} - X_4^{(k-1)}) = 0 \\ \Leftrightarrow\ &2\theta\mu_1 L_{di}\widehat{X}_4^{(k)} + \widehat{X}_4^{(k)} = X_4^{(k-1)} \\ \Leftrightarrow\ &\widehat{X}_4^{(k)} = (2\theta\mu_1 L_{di} + I)^{-1} X_4^{(k-1)} \end{aligned} \tag{20}$$

• Similarly, update step for $\widehat{X}_5^{(k)}$ can be obtained as follows:

$$\widehat{X}_5^{(k)} = X_5^{(k-1)}(2\theta\mu_2 L_{dr} + I)^{-1}. \tag{21}$$

2.3.3 GR1BMC algorithm:

The complete algorithm is given in Algorithm 1 [1]. The convergence of the sequence $(\widehat{X}^{(k)})_{k \in \mathbb{N}}$ to a solution to (6) is ensured, according to [31]. We display in Figures 2 and 3, example of convergence plots (i.e. evolution of objective function along iterations) for Fdataset and Cdataset, respectively.

It should be noted that it would be possible to include other disease and drug features in our framework. For instance, one can modify the trace terms in the proposed formulation so that the Laplacian matrix (L_{di} or L_{dr}) is replaced by the summation of Laplacians derived from individual drug/disease graph similarities/features as was done in [47]. Moreover, the proposed technique can be used in cases when drugs/diseases are not observed (by simply modifying the operator R), provided that their similarities to the drugs/diseases already in the dataset are available and can be incorporated into the graph regularization Laplacian terms.

2.4 Parameter settings

The matrices $X_1^{(0)}, X_2^{(0)}, X_3^{(0)}, X_4^{(0)}$ and $X_5^{(0)}$ are initialized randomly, through a uniform law in $[0, 1]^{n_1 \times n_2}$. The GR1BMC algorithm is run for a fixed number of iterations K ($K = 20$ here) that appears sufficient to reach practical

1. The code of GR1BMC is available at https://github.com/aanchalMongia/GROBMC-PPXA-DDA

Algorithm 1 GR1BMC (y, R, S_{di}, S_{dr})

Set parameters: p, μ_1, μ_2, λ.
Initialize: $X_1^{(0)}, X_2^{(0)}, X_3^{(0)}, X_4^{(0)}, X_5^{(0)}$.
Preprocessing:
Compute N_{di}, N_{dr}, $\widehat{S}_{di} = N_{di}^{ij} \odot S_{di}, \widehat{S}_{dr} = N_{dr}^{ij} \odot S_{dr}$, $(D_{di})^{ii} = \sum_{j=1}^{n_1} (\widehat{S}_{di})^{ij}$, $(D_{dr})^{ii} = \sum_{j=1}^{n_2} (\widehat{S}_{dr})^{ij}$, $(\forall i)$.
Define $L_{di} = (D_{di})^{-1/2}(D_{di} - \widehat{S}_{di})(D_{di})^{-1/2}$ and $L_{dr} = (D_{dr})^{-1/2}(D_{dr} - \widehat{S}_{dr})(D_{dr})^{-1/2}$.
For $k = 1, \ldots, K$
    $\widehat{X}_1^{(k)} = \text{mat}\left((5R^\top R + I)^{-1}(\text{vec}(X_1^{(k-1)}) + 5R^\top y)\right)$
    $X_2^{(k-1)} = US^{(k-1)}V^\top$
    $\widehat{S}^{(k-1)} = \text{sign}(S^{(k-1)}) \max(0, |S^k| - 5\lambda/2)$
    $\widehat{X}_2^{(k)} = U\widehat{S}^{(k-1)}V^\top$
    $\widehat{X}_3^{(k)} = \min(\max(X_3^{(k-1)}, 0), 1)$
    $\widehat{X}_4^{(k)} = (10\mu_1 L_{di} + I)^{-1} X_4^{(k-1)}$
    $\widehat{X}_5^{(k)} = X_5^{(k-1)}(10\mu_2 L_{dr} + I)^{-1}$
    $\widehat{X}^{(k)} = \frac{1}{5}(\widehat{X}_1^{(k)} + \widehat{X}_2^{(k)} + \widehat{X}_3^{(k)} + \widehat{X}_4^{(k)} + \widehat{X}_5^{(k)})$
    $X_i^{(k)} = X_i^{(k-1)} + 2\widehat{X}^{(k)} - \widehat{X}^{(k-1)} - \widehat{X}_i^{(k)}, i = 1, \ldots, 5$
End
Return: $\widehat{X}^{(K)}$

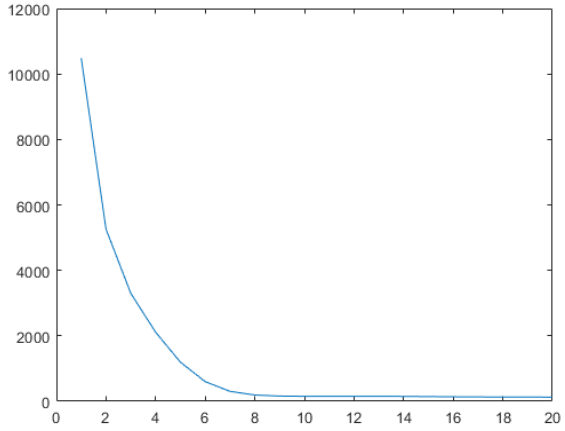

Fig. 2: Convergence plot for GR1BMC on Fdataset

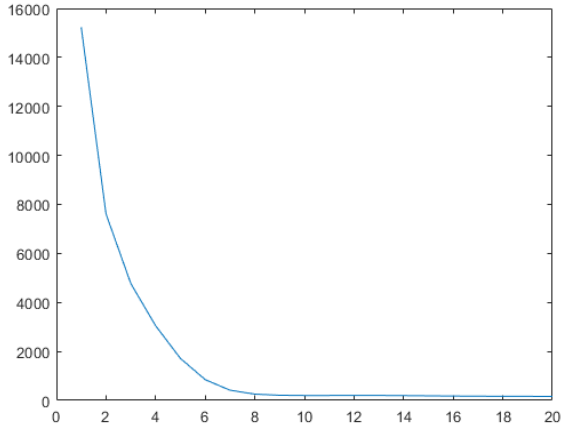

Fig. 3: Convergence plot for GR1BMC on Cdataset

stabilization of the objective function, as can be seen in Figs. 2 and 3. The running time is in the order of seconds; PPXA takes approximately 4 and 6 seconds on Fdataset and Cdataset respectively on a single core machine with a clock speed of 2.8 GHz, 64 GB RAM (Intel(R) Xeon(R) CPU E5-1603 v3 processor). We must determine suitable values for the hyperparameters λ, μ_1 and μ_2, in order to weight the importance of nuclear norm term and the trace terms in our objective function for each of the two datasets. The values of μ_1 and μ_2 determine the weights given to each of the drug and disease laplacians, hence exhibiting the importance of neighborhood information of drugs and targets in our framework for a dataset. The optimal values of these parameters are found by performing cross validation on the training set and taking the value of parameters from the set $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. The values of λ, μ_1 and μ_2 are robust across the datasets and are found to be $0.1, 0.05, 0.1$, respectively, for both datasets.

3 Results and Discussion

3.1 Evaluation criteria

To experimentally evaluate the prediction performance of GR1BMC, we use κ-fold cross validation strategy ($\kappa = 10$). κ-fold cross validation is an evaluation method where we divide the known associations into κ equal subsets (called folds). Out of all the subsets, one of them is treated as a testing set, while the remaining ones constitute the training set. The associations in training set are given as input to the algorithm which then returns the fully imputed association matrix.

After performing the matrix completion task, the predictions on testing set and other candidate associations for all drugs are ranked in descending order of scores. TPR (True Positive Rate)/Recall, FPR (False Positive Rate) and PPV (Positive predicted value)/Precision are calculated for every rank threshold. These values at every threshold are used to plot an ROC (Receiver Operating Characteristic) curve with FPR on x-axis and TPR on y-axis. In a similar way, a Precision-Recall curve is obtained by plotting Recall/TPR on x-axis and Precision on y-axis. The area under both these curves called Area under the ROC curve (AUC) and Area under the precision-recall curve (AUPR) are used to assess the performance of the methods used to predict drug-disease associations.

The above procedure is repeated κ-times and the average of AUC/AUPR across all the κ folds is reported. Figures 4 and 5 show the ROC curves obtained on all the 10 folds of cross validation experiment after running GR1BMC on both the datasets. The average AUC and AUPR across all the folds has been highlighted in black in the figures and shown in Tables 2 and 3. As can be observed from the table, GR1BMC performs better than the benchmarks techniques on both the datasets, especially in terms of AUPR. It should be noted that AUPR is a relatively more important metric in this problem since it heavily punishes highly ranked non-associations (false positives), which is crucial in this application as false positive indications would lead to wastage of resources if the proposed indications were tested in clinical experiments.

We assess the efficacy of the proposed technique in predicting disease association of drugs with no known disease interactions in the database (novel drugs) by finding the precision and recall at top k diseases (Pre@k and Rec@k, k=3, 5 and 7) for drugs while implementing Leave-one-out-cross validation (LOOCV) by hiding (i.e., leaving out) the association profile of every drug in table 4. This shows the performance of GR1BMC algorithm for predicting diseases for novel drugs is reasonably good. Notably, Recall @7 for both the dataset is in range 0.4-0.45 showing that on an average, 40-45% of associated diseases appear in top 7 predicted diseases.

To demonstrate the usefulness of the proposed graph-based regularization terms, we remove either of the two disease and drug graph regularization terms (by setting μ_1 or μ_2 to zero), and compare these models with the one with both graph regularization terms (Table 5). We observe that the addition of the graph regularization term corresponding to drugs is degrading the results. This may be because different drugs having widely different molecular structures can be used to achieve the same goal; they would operate via different pathways. For example, both Clonazepam [48] and Melatonin [49] are used for treating sleep disorder. However, they have very different structures as shown in Supplementary Table 1. The graph Laplacian is not able to account for the overall effect from the structural similarity. We believe this may be one reason for poor performance while trying to account for similarity arising from drug structure. On the other hand, studies have shown that drugs having structural similarity routinely have very different effects [50]. The graph Laplacian for drugs is trying to enforce similar action for structurally similar drugs. This may be the reason why the corresponding penalty term is degrading the results.

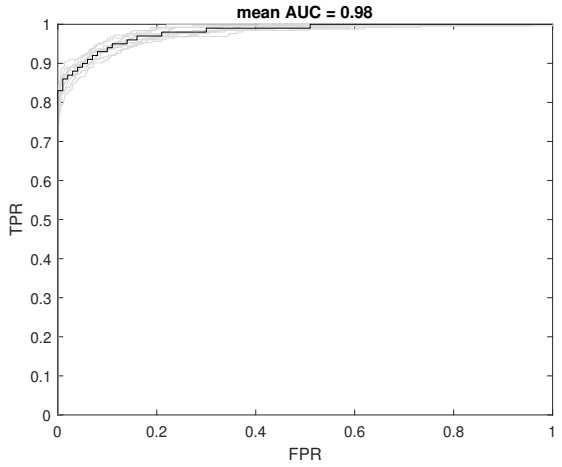

Fig. 4: ROC curves obtained for all the 10 folds after applying GR1BMC on Fdataset

3.2 Comparison with benchmark techniques

To evaluate the performance of GR1BMC, we compare the results of cross-validation experiments with those of the latest methods proposed for drug-disease association prediction: Bounded nuclear norm regularization (BNNR)

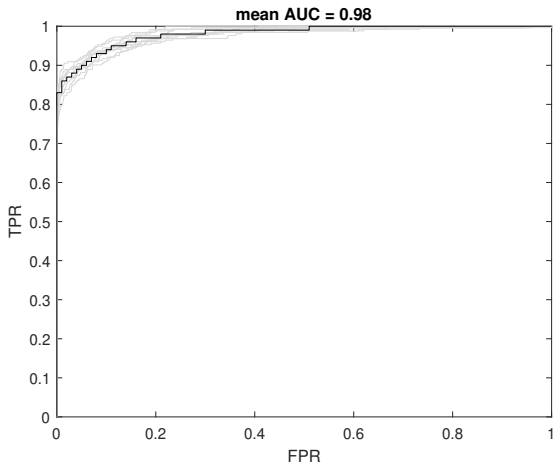

Fig. 5: ROC curves obtained for all the 10 folds after applying GR1BMC on Cdataset

[29], Heterogeneous Network for drug-Disease association prediction (HNRD) [22] and drug repositioning recommendation system (DRRS) [28]. BNNR and DRRS are the closest in terms of formulation used to model the problem. Both the methods deploy nuclear norm minimization on a heterogeneous network matrix obtained by integrating drug similarity, disease similarity, association matrix and its transpose; BNNR additionally handles the noise originating from similarities which violate the low-rankness and restrict the predicted values to be in range [0,1]. However, the choice for imposing the low-rank property of the heterogeneous matrix remains unexplained in both aforementioned works, although it is a crucial assumption when resorting to nuclear norm minimization.

The results of 10-fold cross-validation have been shown in tables 2 and 3. It can be seen that our proposed approach shows competetive performance in terms of area under the ROC curve and is better than the benchmark techniques in temrms of precision and recall also.

3.3 Case study to predict novel associations

To asses the practical usage of the proposed algorithm, we perform a case study where we chose 5 candidate drugs to look for their novel indications (dummy drug

TABLE 2: Average AUC across 10-fold cross-validation for various techniques while predicting drug-disease associations.

Datasets	GR1BMC	BNNR	HNRD	DRRS
Fdataset	**0.9773**	0.9330	0.9420	0.9300
Cdataset	**0.9807**	0.9480	0.9500	0.9470

TABLE 3: Average AUPR across 10-fold cross-validation for various techniques while predicting drug-disease associations.

Datasets	GR1BMC	BNNR	HNRD	DRRS
Fdataset	**0.7247**	0.4410	0.5720	0.3780
Cdataset	**0.7537**	0.4710	0.6700	0.4020

TABLE 4: Precision@k and Recall@k for association prediction for k=3, 5 and 7 with LOOCV for novel drugs

Metric	Datasets	k=3	k=5	k=7
Precision	Fdataset	0.2901	0.2246	0.1795
	Cdataset	0.3097	0.2531	0.2038
Recall	Fdataset	0.3307	0.4060	0.4458
	Cdataset	0.2897	0.3762	0.4153

TABLE 5: Comparison of prediction results (average AUPR) after removing disease or drug Laplacian term versus results obtained with both terms.

Datasets	GR1BMC($\mu_1 = 0$)	GR1BMC($\mu_2 = 0$)	GR1BMC
Fdataset	0.7102	0.9405	0.7247
Cdataset	0.7288	0.9415	0.7537

re-positioning) after predicting the associations using our proposed approach.

We train our model on the known associations on Fdataset. After the matrix completion is done, we rank the remaining candidate diseases for each drug in descending order of predicted association scores.

These rankings or predictions of novel indications for drugs is verified by validating the top-5 indications for any 5 drugs with the public database comparative toxicogenomics database (CTD) [51]. We show the validation on the following 5 drugs: Levodopa, Doxorubicin, Amantadine, Flecainide and Metformin.

The indications predicted by GR1BMC and the evidence from CTD is shown in table 6. It can be seen that at least 3 indications are confirmed with the CTD database for 4 out of 5 drugs and a total of 17 out of 25 predicted associations have evidence in CTD database. Also, the indications which are not verified could be potential candidates for drug-repositioning and could be explored by medical researchers. Let us note that the training data in Fdataset and CTD database have been collected and used independently. The presence of predicted associations using Cdataset in the CTD database shows their overlap but nowhere in the training process, the CTD dataset was used.

4 Conclusion

The huge amount of time and efforts taken for the development drugs calls for the need for efficient and reliable computational methods to assist drug re-positioning. In this paper, we present a novel approach to predict drug-disease indications based on parallel proximal algorithm, which benefits from guaranteed convergence and great numerical performance. Cross validation and experiments on gold standard dataset demonstrate the superiority of the proposed approach over the benchmark techniques. The practical usage is also validated by a case study where novel indications for existing drugs are found and majority are validated with the CTD database. The proposed method is generic and can be applied to other association/interaction prediction problems such as protein-protein interaction prediction, human microbe-disease association (MDA) prediction, gene-disease association prediction, etc.

Acknowledgements

This manuscript has been submitted to the preprint server-bioRxiv. Aanchal Mongia and Angshul Majumdar were partially supported by the Infosys Center for AI at IIIT-Delhi, India.

References

[1] W. P. Walters, J. Green, J. R. Weiss, and M. A. Murcko, "What do medicinal chemists actually make? a 50-year retrospective," *Journal of medicinal chemistry*, vol. 54, no. 19, pp. 6405–6416, 2011.

[2] G. Kaul, M. Shukla, A. Dasgupta, and S. Chopra, "Update on drug-repurposing: is it useful for tackling antimicrobial resistance?" 2019.

[3] S. He, B. Lin, V. Chu, Z. Hu, X. Hu, J. Xiao, A. Q. Wang, C. J. Schweitzer, Q. Li, M. Imamura *et al.*, "Repurposing of the antihistamine chlorcyclizine and related compounds for treatment of hepatitis c virus infection," *Science translational medicine*, vol. 7, no. 282, pp. 282ra49–282ra49, 2015.

[4] H. Villanueva-Lozano, R. d. J. Treviño-Rangel, G. M. González, P. A. Hernández-Rodríguez, A. Camacho-Ortiz, L. Castillo-Reyna, S. G. Galindo-Alvarado, and M. F. Martínez-Reséndez, "Clinical evaluation of the antifungal effect of sertraline in the treatment of cryptococcal meningitis in hiv patients: a single mexican center experience," *Infection*, vol. 46, no. 1, pp. 25–30, 2018.

[5] S. Das, T. Garg, S. Chopra, and A. Dasgupta, "Repurposing disulfiram to target infections caused by non-tuberculous mycobacteria," *Journal of Antimicrobial Chemotherapy*, vol. 74, no. 5, pp. 1317–1322, 2019.

[6] R. Thakare, M. Shukla, G. Kaul, A. Dasgupta, and S. Chopra, "Repurposing disulfiram for treatment of staphylococcus aureus infections," *International journal of antimicrobial agents*, vol. 53, no. 6, pp. 709–715, 2019.

[7] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[8] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PloS one*, vol. 4, no. 8, 2009.

[9] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, "Pgcn: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks," *bioRxiv*, p. 532226, 2019.

[10] A. P. Chiang and A. J. Butte, "Systematic evaluation of drug–disease relationships to identify leads for novel drug uses," *Clinical Pharmacology & Therapeutics*, vol. 86, no. 5, pp. 507–510, 2009.

[11] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "Predict: a method for inferring novel drug indications with application to personalized medicine," *Molecular systems biology*, vol. 7, no. 1, 2011.

[12] W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, and F. Liu, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–12, 2018.

[13] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[14] V. Martinez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "Drugnet: Network-based drug–disease prioritization by integrating heterogeneous data," *Artificial intelligence in medicine*, vol. 63, no. 1, pp. 41–49, 2015.

[15] Y. Wang, S. Chen, N. Deng, and Y. Wang, "Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data," *PloS one*, vol. 8, no. 11, 2013.

[16] L. Yu, X. Ma, L. Zhang, J. Zhang, and L. Gao, "Prediction of new drug indications based on clinical data and network modularity," *Scientific reports*, vol. 6, p. 32530, 2016.

[17] K. Yang, X. Zhao, D. Waxman, and X.-M. Zhao, "Predicting drug-disease associations with heterogeneous network embedding," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 12, p. 123109, 2019.

[18] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, "Machine learning for drug-target interaction prediction," *Molecules*, vol. 23, no. 9, p. 2208, 2018.

TABLE 6: Top 5 predicted diseases for Levodopa, Doxorubicin, Amantadine, Flecainide and Metformin with their evidence in CTD database

DRUG INFORMATION		DISEASE INFORMATION		
Drug name	DrugBank ID	Disease name	OMIM ID	Confirmation
Levodopa	DB01235	(PARKINSON DISEASE, LATE-ONSET; PD)	168600	CTD confirmed
		(DEMENTIA/PARKINSONISM WITH NON-ALZHEIMER AMYLOID PLAQUES)	D125320	CTD confirmed
		(DYSTONIA 9; DYT9	D601042	CTD confirmed
		(DEMENTIA, LEWY BODY; DLB)	D127750	
		(RENAL FAILURE, PROGRESSIVE, WITH HYPERTENSION; RFH1)	D161900	
Doxorubicin	DB00997	(COLORECTAL CANCER; CRC)	D114500	CTD confirmed
		(DOHLE BODIES AND LEUKEMIA)	D223350	
		(RETICULUM CELL SARCOMA)	D267730	CTD confirmed
		(RENAL CELL CARCINOMA, NONPAPILLARY; RCC)	D144700	CTD confirmed
		(LEUKEMIA, CHRONIC LYMPHOCYTIC, SUSCEPTIBILITY TO, 2)	D109543	CTD confirmed
Amantadine	DB00915	(PARKINSON DISEASE, LATE-ONSET; PD)	D168600	CTD confirmed
		(DEMENTIA/PARKINSONISM WITH NON-ALZHEIMER AMYLOID PLAQUES)	D125320	CTD confirmed
		(ALZHEIMER DISEASE, FAMILIAL EARLY-ONSET, WITH COEXISTING AMYLOID AND PRION PATHOLOGY)	D605055	CTD confirmed
		(DEMENTIA, LEWY BODY; DLB)	D127750	CTD confirmed
		(ALZHEIMER DISEASE; AD)	D104300	CTD confirmed
Flecainide	DB01195	(ATRIAL FIBRILLATION, FAMILIAL, 1; ATFB1)	D608583	CTD confirmed
		(HYPERTENSION, DIASTOLIC, RESISTANCE TO)	D608622	CTD confirmed
		(RENAL FAILURE, PROGRESSIVE, WITH HYPERTENSION; RFH1)	D161900	
		(INSENSITIVITY TO PAIN WITH HYPERPLASTIC MYELINOPATHY)	D147530	
		(STROKE, ISCHEMIC)	D601367	
Metformin	DB00331	(DIABETES MELLITUS, INSULIN-DEPENDENT, 2)	D125852	CTD confirmed
		(COLORECTAL CANCER; CRC)	D114500	CTD confirmed
		(HYPERLIPOPROTEINEMIA, TYPE V)	D144650	CTD confirmed
		(ENDOMETRIOSIS, SUSCEPTIBILITY TO, 1)	D131200	
		(UTERINE ANOMALIES)	D192000	

[19] P. Xuan, H. Cui, T. Shen, N. Sheng, and T. Zhang, "Heterodualnet: A dual convolutional neural network with heterogeneous layers for drug-disease association prediction via chou's five-step rule," *Frontiers in pharmacology*, vol. 10, 2019.

[20] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug–disease associations through layer attention graph convolutional network," *Briefings in Bioinformatics*, 2020.

[21] H.-J. Jiang, Y.-A. Huang, and Z.-H. You, "Predicting drug-disease associations via using gaussian interaction profile and kernel-based autoencoder," *BioMed research international*, vol. 2019, 2019.

[22] Y. Wang, G. Deng, N. Zeng, X. Song, and Y. Zhuang, "Drug-disease association prediction based on neighborhood information aggregation in neural networks," *IEEE Access*, vol. 7, pp. 50 581–50 587, 2019.

[23] E. Jadamba and M. Shin, "A systematic framework for drug repositioning from integrated omics and drug phenotype profiles using pathway-drug network," *BioMed research international*, vol. 2016, 2016.

[24] J. Yang, Z. Li, X. Fan, and Y. Cheng, "Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization," *Journal of chemical information and modeling*, vol. 54, no. 9, pp. 2562–2569, 2014.

[25] W. Dai, X. Liu, Y. Gao, L. Chen, J. Song, D. Chen, K. Gao, Y. Jiang, Y. Yang, J. Chen *et al.*, "Matrix factorization-based prediction of novel drug indications by integrating genomic space," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[26] M. Yang, G. Wu, Q. Zhao, Y. Li, and J. Wang, "Computational drug repositioning based on multi-similarities bilinear matrix factorization," *Briefings in Bioinformatics*, 2020.

[27] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[28] 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, pp. 1904–1912, 2018.

[29] M. Yang, H. Luo, Y. Li, and J. Wang, "Drug repositioning based on bounded nuclear norm regularization," *Bioinformatics*, vol. 35, no. 14, pp. i455–i463, 2019.

[30] N. Pustelnik, C. Chaux, and J.-C. Pesquet, "Parallel proximal algorithm for image restoration using hybrid regularization," *IEEE transactions on Image Processing*, vol. 20, no. 9, pp. 2450–2462, 2011.

[31] P. Combettes and J. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Problems*, vol. 24, no. 27, 2008.

[32] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu, and Y. Pan, "Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, 2016.

[33] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D668–D672, 2006.

[34] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 30, no. 1, pp. 52–55, 2002.

[35] T. T. Tanimoto, "An elementary mathematical theory of classification and prediction. 1958," *International Business Machines Corporation*, 1958.

[36] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics," *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 493–500, 2003.

[37] H. Öztürk, E. Ozkirimli, and A. Özgür, "A comparative study of smiles-based compound similarity functions for drug-target interaction prediction," *BMC bioinformatics*, vol. 17, no. 1, p. 128, 2016.

[38] M. A. Van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, "A text-mining analysis of the human phenome," *European journal of human genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[39] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of machine learning research*, vol. 7, no. Nov, pp. 2399–2434, 2006.

[40] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[41] A. Ezzat, P. Zhao, M. Wu, X. Li, and C. Kwoh, "Drug-target interaction prediction with graph regularized matrix factorization," *IEEE/ACM Transactions on Computational Biolology and Bioinformatics*, vol. 14, no. 3, pp. 646–656, 2017.

[42] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212.

[43] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[44] A. Cherni, E. Chouzenoux, and M.-A. Delsuc, "Palma, an improved algorithm for dosy signal processing," *Analyst*, vol. 142, no. 5, pp. 772–779, 2017.

[45] A. Benfenati, E. Chouzenoux, and J.-C. Pesquet, "Proximal approaches for matrix optimization problems: Application to robust precision matrix estimation," *Signal Processing*, no. 169, p. 107417, Apr. 2020.

[46] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2011.
[47] A. Mongia and A. Majumdar, "Matrix completion on multiple graphs: Application in collaborative filtering," *Signal Processing*, vol. 165, pp. 144–148, 2019.
[48] A. Kales, R. L. Manfredi, A. N. Vgontzas, C. F. Baldassano, K. Kostakos, and J. D. Kales, "Clonazepam: sleep laboratory study of efficacy and withdrawal." *Journal of clinical psychopharmacology*, 1991.
[49] N. Zisapel, "The use of melatonin for the treatment of insomnia," *Neurosignals*, vol. 8, no. 1-2, pp. 84–89, 1999.
[50] Y. C. Martin, J. L. Kofron, and L. M. Traphagen, "Do structurally similar molecules have similar biological activity?" *Journal of medicinal chemistry*, vol. 45, no. 19, pp. 4350–4358, 2002.
[51] 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, 2013.

Aanchal Mongia is currently a post doctoral researcher at University of Pennsylvania, USA. She defended in 2020 her Ph.D. thesis at IIIT-Delhi, India. Her research interests include bioinformatics and machine learning.

Emilie Chouzenoux (IEEE Senior Member) received the engineering degree from Ecole Centrale, Nantes, France, in 2007, and the Ph.D. degree in signal processing from the Institut de Recherche en Communications et Cybernétique (IRCCyN, UMR CNRS 6597), Nantes, in 2010. Between 2011 and 2019, she was a Maître de conferences at the University of Paris-Est Marne-la-Vallée, Champs-sur-Marne, France (LIGM, UMR CNRS 8049). Since September 2019, she has been a Researcher at Inria Saclay, within the project team OPIS, in Centre pour la Vision Numérique, CentraleSupélec. She is an Associate Editor of IEEE Transactions in Signal Processing and of SIAM Journal on Mathematics of Data Science. Since January 2020, she has been the PI of the ERC Starting Grant MAJORIS. Her research interests are in large scale optimization algorithms for inverse problems and machine learning problems of image processing.

Anghsul Majumdar is currently an associate professor at IIIT-Delhi, India. He holds joint appointments in ECE and CSE departments. He has been associated with the institute since 2012, when he joined as an assistant professor right after finishing his PhD. Prior to that, Anghsul did his Master's and PhD from UBC in 2009 and 2012 respectively. He is a senior member of IEEE. Currently he serves as an associated editor for IEEE TCSVT and IEEE OJSP. He also serves as a member-at-large for IEEE SPS Education Board.