

## RESEARCH ARTICLE

# Drug-target interaction prediction using Multi Graph Regularized Nuclear Norm Minimization

Aanchal Mongia<sup>1</sup> , Angshul Majumdar<sup>2\*</sup>

**1** Dept. of Computer Science and Engineering, IIIT-Delhi, Delhi, India, **2** Dept. of Electronics and Communications Engineering, IIIT-Delhi, Delhi, India

\* [aanchalm@iiitd.ac.in](mailto:aanchalm@iiitd.ac.in)



## OPEN ACCESS

**Citation:** Mongia A, Majumdar A (2020) Drug-target interaction prediction using Multi Graph Regularized Nuclear Norm Minimization. PLoS ONE 15(1): e0226484. <https://doi.org/10.1371/journal.pone.0226484>

**Editor:** Ivan Olier, Liverpool John Moores University, UNITED KINGDOM

**Received:** January 24, 2019

**Accepted:** November 27, 2019

**Published:** January 16, 2020

**Copyright:** © 2020 Mongia, Majumdar. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The data were simulated from public databases KEGG BRITE [43], BRENDA [44] SuperTarget [7] and DrugBank [4] and are publically available at the given link: <http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>. Additionally, the software is publicly available at <https://github.com/aanchalMongia/MGRNNMforDTI>.

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

The identification of potential interactions between drugs and target proteins is crucial in pharmaceutical sciences. The experimental validation of interactions in genomic drug discovery is laborious and expensive; hence, there is a need for efficient and accurate in-silico techniques which can predict potential drug-target interactions to narrow down the search space for experimental verification. In this work, we propose a new framework, namely, Multi-Graph Regularized Nuclear Norm Minimization, which predicts the interactions between drugs and target proteins from three inputs: known drug-target interaction network, similarities over drugs and those over targets. The proposed method focuses on finding a low-rank interaction matrix that is structured by the proximities of drugs and targets encoded by graphs. Previous works on Drug Target Interaction (DTI) prediction have shown that incorporating drug and target similarities helps in learning the data manifold better by preserving the local geometries of the original data. But, there is no clear consensus on which kind and what combination of similarities would best assist the prediction task. Hence, we propose to use various multiple drug-drug similarities and target-target similarities as multiple graph Laplacian (over drugs/targets) regularization terms to capture the proximities exhaustively. Extensive cross-validation experiments on four benchmark datasets using standard evaluation metrics (AUPR and AUC) show that the proposed algorithm improves the predictive performance and outperforms recent state-of-the-art computational methods by a large margin. Software is publicly available at <https://github.com/aanchalMongia/MGRNNMforDTI>.

## Introduction

The field of drug discovery in Pharmaceutical Sciences is plagued with the problem of high attrition rate. The task is to find effective interactions between chemical compounds (drugs) and amino-acid sequences/ proteins (targets). This is traditionally done through wet-lab experiments which are costly and laborious.

An effective and appropriate alternative to avoid costly failures is to computationally predict the interaction probability. A lot of algorithms have been proposed for DTI (Drug-target interaction) prediction in recent years [1, 2], which use small number of experimentally validated interactions in existing databases such as ChEMBL [3], DrugBank [4], KEGG DRUG [5], STITCH [6] and SuperTarget [7]. Identification of drug-target pairs leads to improvements in different research areas such as drug discovery, drug repositioning, polypharmacology, drug resistance and side-effect prediction [8].

For example, Drug repositioning [9, 10] (reuse of existing drugs for new indications) can grant polypharmacology (multi-target effect) to a drug. One of the many successfully repositioned drugs is Gleevec (imatinib mesylate). Earlier, it was known to interact only with the Bcr-Abl fusion gene which is indicative of leukemia. However, later discoveries showing that it also interacts with PDGF and KIT, repositioned it for the treatment of gastrointestinal stromal tumors [11, 12].

There are three major classes of computational methods for predicting DTI: Ligand-based, Docking based, and Chemogenomic approaches. Ligand-based approaches leverage the similarity between target proteins' ligands to predict interactions [13]. The idea is that molecules with similar structure/property would bind similar proteins [14]. But, the reliability of results might get compromised due to limited information about known ligands per protein. Docking-based approaches use the 3D structure of both drugs and proteins to predict the interaction likelihood [15–17]. This, although is well-accepted, but is very time-consuming and hence, cannot be used for protein families for which the 3D structure is either difficult to predict or is unavailable [18] like the G-protein coupled receptors (GPCRs).

Chemogenomic approaches overcome the challenges of traditional methods and thus, have recently gained much attention. The approaches under this category can work on the huge amount of biological data, publicly available in existing online databases and can process meta-data (chemical structures and genomic sequences) for both the drug and target, respectively. These approaches can further sub-classified based on the representation of the input data: Feature-based techniques and Similarity-based techniques. Feature-based techniques take their inputs in the form of features and class labels (binary values here) and leverage machine learning for classifying if an input instance corresponds to a positive interaction or a negative one. Examples of typical feature based methods include Decision Trees (DT), Random Forests (RF) and Support Vector Machines (SVM) to build classification models based on the labeled feature vectors [19]. In the training set, positive instances are the experimentally known interactions while the negative ones are either non-interactions or unknown interactions. The other category of chemogenomic techniques, Similarity-based methods, use two similarity matrices corresponding to drug and target similarity, respectively, along with the drug-target interaction matrix.

Let us discuss the similarity between the said DTI problem and the problem of collaborative filtering (CF). CF is a standard problem in information retrieval. It is used in recommendations systems (e.g. in Netflix movie recommendations and Amazon product recommendations). There is a database of user's and their ratings on items (movies, products, etc.). Obviously, not all the ratings are available; users typically rate only a small subset of items. The objective is to estimate the ratings of all the users on all the items. If that can be done accurately, recommendation accuracy increases. The similarity between DTI and CF should be straightforward now; the drugs play the role of users and the targets play the role of items. The interactions are similar to the ratings. Over the years, many approaches originally developed for CF have been leveraged to solve the DTI problems.

In both CF [20] and DTI [21–23], the initial techniques were based on simple neighborhood-based models. In order to predict the interaction of a (active) drug on a target, the first

step is to find out similar (neighbor) drugs by computing some kind of a similarity score. Once the neighborhood is obtained, the interaction value from the drugs in the neighborhood are weighted (by the normalized similarity score) to interpolate the interaction of the active drug on the target. This is similar to KNN (K-nearest neighbor) based approaches in DTI prediction literature [24]. Another approach was based on bipartite local models. In such models, a local model is built for every drug and target. For example in [25] an SVM was trained for each to predict the interaction of each drug on all targets and each target on all drugs. Finally, the decision from the two was fused. This is just an example, there are other techniques falling under this generic approach like [26, 27]. The above-mentioned methods can be categorized as classification based approaches where the chemical/biological information is used to generate features for drugs and targets individually and these two types of features are then concatenated and the corresponding interaction is assumed to the class corresponding to this feature. Any standard classifier is generally used for the final classification. In such class of techniques, the emphasis is on different feature selection mechanisms [28, 29]. Both semi-supervised [30, 31] and supervised [32, 33] classification based prediction approaches have been leveraged in Drug-Target interaction prediction.

The second category is based on network diffusion models. One technique for DTI prediction based on such models is based on a random walk on the network with a predefined transition matrix [34]. Another work falling under this category, predicts interactions by finding a simple path (without loops) between nodes of the network.

The third approach is based on matrix factorization. These techniques were originally developed for collaborative filtering [35]. It is assumed that drugs and targets are characterized by latent factors. The probability of interaction is high when the latent factors match; i.e. when the inner product has a high value. Therefore, it is logical to express the interaction matrix as a (an inner) product of drug and target latent factors. This allows matrix factorization (and its variants) to be applied [36, 37].

In a very recent review paper [2] it was empirically shown that matrix factorization based techniques yields by far the best results. The fundamental assumption behind matrix factorization to work is that there are very few (latent) factors that are responsible for drug target interactions. This is the reason, one can factor the DTI matrix into a tall (drug) latent factor matrix and a fat (target) latent factor matrix. Mathematically speaking, the assumption is that the DTI matrix is of low-rank. Matrix factorization is being used to model low-rank matrices for the past two decades since the publication of Lee and Seung's seminal paper [38]. However, matrix factorization is a bi-linear non-convex problem; there are no convergence guarantees. In order to ameliorate this problem, mathematicians proposed an alternative approach based on nuclear norm minimization [39–41]. The nuclear norm is the closest convex surrogate to the rank minimization (known to be NP-hard) problem and there are provable mathematical guarantees on its equivalence to rank minimization.

The standard versions of both the matrix factorization and nuclear norm minimization techniques are unable to incorporate similarity information of the drugs and the targets. In recent studies [37, 42], it was shown that the best results are obtained when these technique incorporate graph regularization penalties into them. But, these works regularize the objective function by taking into account, just the standard chemical structure similarity for drugs ( $S_d$ ) and the genomic sequence similarity for targets ( $S_t$ ). No study in literature gives a clear picture of which kind of similarities would be the best for DTI prediction. We, therefore, incorporate different other kinds of similarities and a combination of them as a multi-graph Laplacian regularization with Nuclear Norm Minimization for DTI prediction. The algorithm uses four new similarity measures over the drugs and targets, apart from the standard similarities to construct the graph Laplacians. The four newly incorporated similarities are computed from the

interaction matrix and take into account the Cosine similarity, Correlation, Hamming distance and Jaccard similarity between the drugs and targets. To the best of our knowledge, this is the first work on multiple graph laplacian regularized nuclear norm minimization for DTI prediction.

## Materials and methods

### Dataset description

We use the four benchmark datasets introduced in [21] having four different classes of proteins: enzymes (Es), ion channels (ICs), G protein- coupled receptors (GPCRs) and nuclear receptors (NRs). The data was simulated from public databases KEGG BRITE [43], BRENDA [44] SuperTarget [7] and DrugBank [4] and is publically available at the given link: <http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>.

The data from each of these databases is formatted as an adjacency matrix, called interaction matrix between drugs and targets, encoding the interaction between  $n$  drugs and  $m$  targets as 1 if the drug  $d_i$  and target  $t_j$  are known to interact and 0, otherwise.

Along with the interaction matrix, drug similarity matrix  $S_d$  and a target similarity matrix  $S_t$  are also provided. In  $S_d$ , each entry represents the pairwise similarity between the drugs and is measured using SIMCOMP [45]. It represents the chemical structure similarity between drugs; measured using the number of shared substructures within the chemical structures of two drugs. In  $S_t$ , the similarity score between two proteins is the genomic sequence similarity. It is based on the amino acid sequences of the target protein and is computed using normalized Smith–Waterman [46].

The similarity matrices  $S_d$  and  $S_t$  constitute the most standard similarities that have been used in the DTI prediction task hitherto. We use these similarities along with the following four more similarities computationally derived from the drug-target interaction matrix to form the graph laplacian terms:

- Cosine similarity: measures the cosine of the angle between two drug/target vectors projected in a multi-dimensional space. Its value ranges from -1 (exactly opposite) to 1 (exactly the same). Given two  $n$ -dimensional drug (or target) vectors (A and B), the cosine similarity is calculated as follows:

$$S^{\text{cos}} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Here,  $A_i$  and  $B_i$  denote the components of vector A and B.

- Correlation: computes the Pearson’s linear correlation coefficient indicating the extent to which two variables are linearly related. It has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. For a pair (say A and B) of drugs/targets with sample size  $n$ , it is given by:

$$S^{\text{cor}} = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^n (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^n (B_i - \bar{B})^2}}$$

where

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n A_i \text{ and } \bar{B} = \frac{1}{n} \sum_{i=1}^n B_i$$

- Hamming similarity: has been computed using Hamming distance. For any two  $n$ -dimensional drugs/targets ( $A$  and  $B$ ), the hamming distance is the percentage of interaction positions that differ. We calculate Hamming distance based similarity by simply subtracting hamming distance from 1, giving us its complementary (the percentage of common interaction positions for a pair of drugs/targets). It can be calculated as follows:

$$S^{ham} = 1 - \frac{\#(A_i \neq B_i)}{n}$$

- Jaccard similarity: is defined as the percentage of common non-zero interaction positions for the two given sample sets of drugs/target vectors.

$$S^{jac} = \frac{\#[(A_i = B_i) \cap ((A_i \neq 0) \cup (B_i \neq 0))]}{\#[(A_i \neq 0) \cup (B_i \neq 0)]}$$

Table 1 summarizes the statistics of all four datasets.

### Nuclear norm minimization

Let us assume that  $X$  is the adjacency matrix where each entry denotes interaction between a drug and target (1 if they interact, 0 otherwise). Unfortunately, we only observe this matrix partially because all interactions are not known. If  $M$  denotes the partially observed adjacency matrix, the mathematical relation between  $X$  and  $M$  is expressed as:

$$M = A(X) \tag{1}$$

In the above equation,  $A$  denoted the sub-sampling operator, element-wise multiplied to  $X$ . It is nothing but a binary matrix or a mask that has 0's where the interaction  $X$  has not been observed or is unknown and 1's where they have been. The task is to recover  $X$ , given the observations  $M$ , and the sub-sampling mask  $A$ . It is known that  $X$  is of low-rank. Ideally,  $X$  should be recovered by (2).

$$\min_X \text{rank}(X) \text{ such that } M = A(X) \tag{2}$$

Unfortunately, rank minimization is an NP-hard problem with doubly exponential complexity, therefore solving it directly is not feasible.

Traditionally, a low-rank matrix has been modeled as a product of a thin and a fat matrix and recovered using Matrix Factorization techniques [38]. But, Matrix Factorization is a bi-

Table 1. Drugs, targets and interactions in each dataset used for validation.

Datasets	Nuclear Receptor (NR)	G-Protein Coupled Receptor (GPCR)	Ion channel (IC)	Enzyme (E)
# Interactions	90	635	1476	2926
# Drugs	54	223	210	445
# Targets	26	95	204	664

<https://doi.org/10.1371/journal.pone.0226484.t001>

linear non-convex problem, therefore there is no guarantee for global convergence. In the past decade, mathematicians showed that the rank minimization problem can be relaxed by its convex surrogate (nuclear norm minimization) with provable guarantees [39, 40] This turns out to be a convex problem that can be solved by Semi-Definite Programming. More efficient solvers have also been proposed. Problem (2) is expressed as (3)

$$\min_X \|X\|_* \text{ such that } M = A(X) \tag{3}$$

Here the nuclear norm ( $\|\cdot\|_*$ ) is defined as the sum of singular values of data matrix  $X$ . It is the  $l_1$  norm (sum of absolute values) of the singular values of  $X$  and is the tightest convex relaxation of the rank of the matrix, and hence its ideal replacement.

Here, (3) is a constrained formulation for the noiseless scenario, usually its relaxed version, (4) is solved.

$$\min_X \|M - A(X)\|_F^2 + \lambda \|X\|_* \tag{4}$$

One of the efficient solvers for Nuclear Norm minimization is the Singular value shrinkage (SVS) algorithm [47].

**Algorithm 1** Singular value shrinkage

- 1: **procedure** MATRIX-SVS ( $M, A, \lambda$ )
- 2:   **Initialize:**  $X = rand, a$
- 3:   **For loop,** iterate ( $k$ )
- 4:      $B_k = X_{k-1} + \frac{1}{a} A^T (M - A \circ X_{k-1})$
- 5:     Compute SVD of  $B$ :  $B_k = USV^T$
- 6:     Soft threshold the singular values:  $\Sigma = soft(S, \lambda/2)$
- 7:      $X_k = U\Sigma V^T$
- 8:      $X_k \leftarrow X_k^+$
- 9: **End loop 1**

### Multi Graph Regularized Nuclear Norm Minimization

Nuclear Norm based Low-rank Matrix Completion is not our contribution, it has been around since the past decade. The problem with standard Nuclear norm minimization (NNM) is that it cannot accommodate associated information such as Similarity matrices for Drugs and Targets. But, it has been seen in recent studies that accommodating the similarity information is crucial for improving the DTI prediction results. The current works have incorporated the standard similarity measures for drugs and targets in matrix factorization [37] and Matrix completion [42] frameworks. It is imperative that NNM should be capable of taking into account more types and combinations of similarities. To achieve this, we have augmented four other types of similarities between drugs/targets and presented Multi-Graph regularized Nuclear Norm Minimization (MGRNNM).

Graph regularization assumes that data points which are in the neighborhood of each other in the original space should also be close to each other in the learned manifold (**Local Invariance assumption**). So, Graph regularization would allow/enable the algorithm to learn manifolds for the drug and target spaces in which the data is assumed to lie. The multi graph regularized version of Nuclear norm minimization, aims to prevent over fitting and greatly enhance the generalizing capabilities. It is incorporated into the formulation/objective

function as Laplacian weights corresponding to drugs and targets:

$$\min_X \|M - A(X)\|_F^2 + \lambda \|X\|_* + \mu_1 \text{Tr}(X^T \sum_{i=1}^{nos} L_d^i X) + \mu_2 \text{Tr}(X \sum_{i=1}^{nos} L_t^i X^T) \tag{5}$$

where  $\lambda \geq 0, \mu_1 \geq 0$  and  $\mu_2 \geq 0$  are parameters balancing the reconstruction error of NNM in the first two terms and graph regularization in the last two terms,  $\text{Tr}(\cdot)$  denotes the trace of the matrix,  $nos$  stands for number of similarity matrices ( $nos = 5$  in our case).

If, say we consider a single similarity matrix for drugs ( $S_d$ ) and that for targets ( $S_t$ ), then  $L_d = D_d - S_d$  and  $L_t = D_t - S_t$  are the graph Laplacians [48] for  $S_d$  (drug similarity matrix) and  $S_t$  (target similarity matrix), respectively, and  $D_d^i = \sum_j S_d^{ij}$  and  $D_t^i = \sum_j S_t^{ij}$  are degree matrices.

Problem (5) is solved using a variable splitting approach [49]. The augmented Lagrangian is expressed as (6). We introduce two new proxy variables  $Z$  and  $Y$  such that  $Z^T = X$  and  $Y = X$ .

$$\begin{aligned} \min_{X,Y,Z} \|M - A(X)\|_F^2 + \lambda \|X\|_* + \mu_1 \text{Tr}(Z \sum_{i=1}^{nos} L_d^i Z^T) + \mu_2 \text{Tr}(Y \sum_{i=1}^{nos} L_t^i Y^T) + \\ v_1 \|Z^T - X\|_F^2 + v_2 \|Y - X\|_F^2 \end{aligned} \tag{6}$$

The variables are updated using ADMM [50, 51]. This leads to the following subproblems (7), (8) and (9)

$$X \leftarrow \min_X \|M - A(X)\|_F^2 + v_1 \|Z^T - X\|_F^2 + v_2 \|Y - X\|_F^2 + \lambda \|X\|_* \tag{7}$$

$$Y \leftarrow \min_Y \mu_2 \text{Tr}(Y \sum_{i=1}^{nos} L_t^i Y^T) + v_2 \|Y - X\|_F^2 \tag{8}$$

$$Z \leftarrow \min_Z \mu_1 \text{Tr}(Z \sum_{i=1}^{nos} L_d^i Z^T) + v_1 \|Z^T - X\|_F^2 \tag{9}$$

Problem (7) can be expressed as a standard NNM problem (by column stacking the variables).

$$\left\| \begin{pmatrix} M \\ \sqrt{v_1} Z^T \\ \sqrt{v_2} Y \end{pmatrix} - \begin{pmatrix} A \\ \sqrt{v_1} I \\ \sqrt{v_2} I \end{pmatrix} X \right\|_F^2 + \lambda \|X\|_* \tag{10}$$

To solve for  $Y$  and  $Z$ , we differentiate (8) and (9) wrt  $Y$  and  $Z$ , respectively.

$$Y = \arg \min_Y (F_1) \text{ where } F_1 = \mu_2 \text{Tr}(Y \sum_{i=1}^{nos} L_t^i Y^T) + v_2 \|Y - X\|_F^2 \tag{11}$$

$$Z = \arg \min_Z (F_2) \text{ where } F_2 = \mu_1 \text{Tr}(Z \sum_{i=1}^{nos} L_d^i Z^T) + v_1 \|Z^T - X\|_F^2 \tag{12}$$

$$\frac{\partial F_1}{\partial Y} = \mu_2 (Y (\sum_{i=1}^{nos} L_t^i)^T + Y \sum_{i=1}^{nos} L_t^i) + 2v_2 (Y - X) \tag{13}$$

$$\frac{\partial F_1}{\partial Y} = \mu_2 Y \left[ \sum_{i=1}^{nos} (L_i^T + L_i) \right] + 2v_2(Y - X) \tag{14}$$

Since  $L_i$  is a symmetric matrix,  $L_i^T = L_i$ . So,

$$\frac{\partial F_1}{\partial Y} = 2\mu_2 Y \sum_{i=1}^{nos} L_i + 2v_2(Y - X)$$

Equating the derivative to zero, we get:

$$v_2 Y + \mu_2 Y \sum_{i=1}^{nos} L_i = v_2 X \tag{15}$$

The matrix equation of this form ( $AT+TB = C$ ) cannot be solved directly for variable T and is called Sylvester equation. Such an equation has a unique solution when the eigenvalues of A and -B are distinct.

A similar Sylvester equation and update step for Z can be obtained by differentiating  $F_2$  and equating to 0.

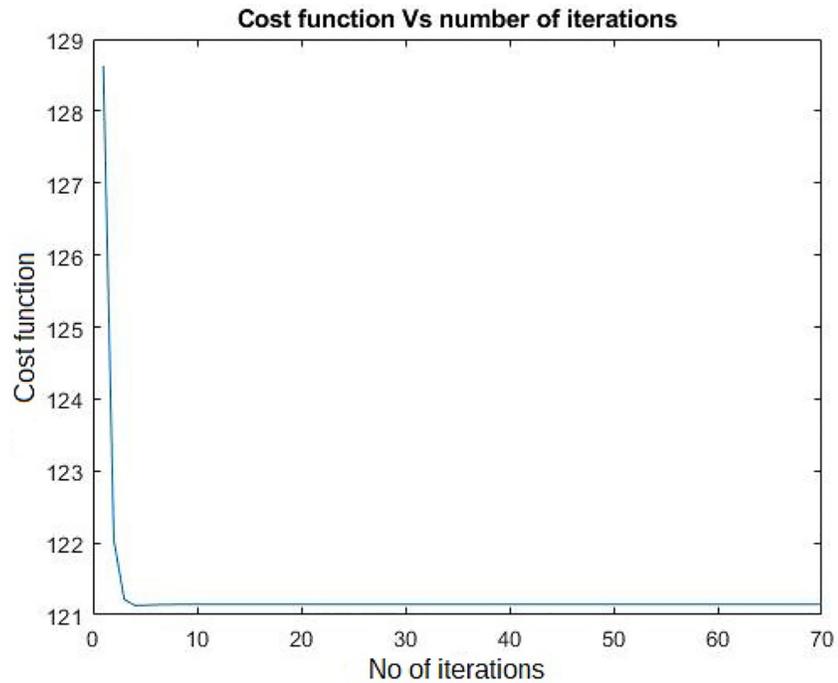
$$v_1 Z + \mu_1 Z \sum_{i=1}^{nos} L_d^i = v_1 X^T \tag{16}$$

It can be shown that computing the sum of the Graph Laplacians is equivalent to computing the Laplacian from the sum of various similarity matrices involved. For instance, consider the sum of drug Graph Laplacians:

$$\begin{aligned} & \sum_{i=1}^{nos} L_d^i \\ &= \sum_{i=1}^{nos} (D_d^i - S_d^i) \\ &= \sum_{i=1}^{nos} D_d^i - \sum_{i=1}^{nos} S_d^i \\ &= \sum_{i=1}^{nos} \text{diag} \left( \sum_j S_d^j \right) - \sum_{i=1}^{nos} S_d^i \\ &= \text{diag} \left( \sum_j \left( \sum_{i=1}^{nos} S_d^i \right)^j \right) - \sum_{i=1}^{nos} S_d^i \end{aligned}$$

Let  $\sum_{i=1}^{nos} S_d^i = S_d^{COM}$  where  $S_d^{COM}$  stands for combined similarity for drugs. Essentially,

$$S_d^{COM} = S_d + S_d^{cos} + S_d^{cor} + S_d^{ham} + S_d^{jac} \tag{17}$$



**Fig 1. Converge plot of the MGRNNM algorithm for NR dataset with cross validation setting CVS1 (drug-target pair prediction).**

<https://doi.org/10.1371/journal.pone.0226484.g001>

Then,

$$\sum_{i=1}^{nos} L_d^i = \text{diag}(\sum_j S_d^{COM}) - S_d^{COM} = D_d^{COM} - S_d^{COM} = L_d^{COM} \tag{18}$$

Here,  $D_d^{COM}$  and  $L_d^{COM}$  denote combined degree matrix and combined Laplacian matrix (sum of graph laplacians) for drugs. Of note, the individual Laplacians or the similarities can be weighted unequally to give more or less emphasis on a specific type of similarity. The pseudo-code for MGRNNM has been given in Algorithm 2.

The standard NNM is a convex problem and the introduced graph regularization penalties are also convex, so entire formulation (5), being a sum of convex functions, is convex. Therefore it is bound to converge to a global minima. We chose the number of iterations such that the algorithm halts when the objective function does not change with iterations. A sample convergence plot for one of the datasets for drug-target pair prediction has been shown in Fig 1.

**Algorithm 2** Multi Graph regularized Nuclear Norm Minimization

- 1: **procedure** MGRNNM ( $M, A, S_d^{COM}, S_t^{COM}$ )
- 2: **Sparsify:**  $S_d^{COM}, S_t^{COM}$
- 3: **Initialize:**  $\lambda, \mu_1, \mu_2, v_1, v_2, L_d^{COM}, L_t^{COM}, Y = M, Z = M^T$
- 4:  $AA \leftarrow \begin{pmatrix} A \\ \sqrt{v_1}I \\ \sqrt{v_2}I \end{pmatrix}$
- 5: **For loop,** iterate (k)
- 6:  $YY_k \leftarrow \begin{pmatrix} M \\ \sqrt{v_1}Z^T \\ \sqrt{v_2}Y \end{pmatrix}$

```

7:       $X_k \leftarrow MATRIX - SVS(Y Y_k^T, AA, \lambda)$ 
8:       $Y_k \leftarrow \text{solve - sylvester}(v_1 I, \mu_1 L_d^{COM}, v_1 X_k)$ 
9:       $Z_k \leftarrow \text{solve - sylvester}(v_2 I, \mu_2 L_t^{COM}, v_2 X_k)$ 
10:   End loop 1

```

## Time complexity of MGRNNM

The algorithm is iterative, so we only discuss the time complexity per iteration. In each iteration, we solve the two Sylvester equations and one NNM. NNM itself is an iterative algorithm that requires solving Singular value decomposition in each iteration, the order of complexity of which is  $O(n^3)$ . The complexity of solving Sylvester equation is  $O(n^b \cdot \log(n))$  [52] where  $b$  is between 2 and 3.

## Results and discussion

### Experimental setup

We validated our proposed method by comparing it with recent and well-performing prediction methods proposed in the literature. Out of the 7 approaches with which we compare,

- Five are specifically designed for DTI task (WGRMF: Weighted Graph Regularized Matrix Factorization, CMF: Collaborative Matrix Factorization, RLS\_WNN: Regularized Least square Nearest neighbor profile, NRLMF: Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction and TMF: Triple matrix factorization) [23, 37, 53–55]. Of note, the code available for TMF does not reproduce the results stated in the corresponding paper; the results obtained after running their code has been reported in this work.
- One being traditional matrix completion (MC: matrix completion) [47] and
- Last one being a naive solution to our problem, available as an unpublished work (MCG: matrix completion on graphs). Of note, the Space complexity of MCG is  $O(n^4)$  while that of MGRNNM is  $O(n^2)$ . [56])

All baselines designed for DTI problem are recent and are already compared against older methods.

We have performed 5 runs of 10-fold cross-validation (CV) for each of the algorithms under three cross-validation setting (CVS) [2]:

- CVS1/Pair prediction: random drug–target pairs are left out for testing set to be used in prediction. It is the conventional setting for validation and evaluation.
- CVS2/Drug prediction: the complete drug profiles are left out for the testing set. It tests the algorithm’s ability to predict interactions for novel drugs i.e. drugs for which no interaction information is available.
- CVS3/Target prediction: the complete target profiles are left out for the testing set. It tests the algorithm’s ability to predict interactions for novel targets.

In 10-fold CV, the given data was divided into 10 folds and out of those 10 folds, one was left out for testing whereas the remaining 9 folds were used as the training set. As the evaluation metrics, we have used AUC (Area under ROC curve) and AUPR (Area under the precision-recall curve). In biological drug discovery, AUPR is a practically more important metric since it penalizes high ranked false positive interactions much more than

AUC. This is because those pairs would be biologically validated later in the drug discovery process.

### Preprocessing

Each of the drug and target similarity matrices were summed up to compute the combined similarity matrices  $S_d^{COM}$  and  $S_t^{COM}$  (Eq (17)). The combined similarity matrices were further sparsified by using p-nearest neighbor graph which is obtained by taking into account only the similarity values which correspond to the nearest neighbors for each drug/target. The usage of such a pre-processing, as shown by [37], helps learn a data manifold on or near to which the data is assumed to lie which, in turn, is expected to preserve the local geometries of the original data and hence give more accurate results.

$$\forall i, j$$

$$N_{ij} = \begin{cases} 1, & j \in N_p(i) \& i \in N_p(j) \\ 0, & j \notin N_p(i) \& i \notin N_p(j) \\ 0.5, & \text{else} \end{cases}$$

where  $N_p(i)$  is the set of  $p$  nearest neighbors to drug  $d_i$ . Similarity matrix sparsification is done by element-wise multiplying it with  $N_{ij}$ . In the next step, the combined graph laplacian terms are computed. Also, instead of graph laplacians ( $L_d^{COM}$  and  $L_t^{COM}$ ), we have used normalized graph laplacians ( $(D_d^{COM})^{-1/2} L_d^{COM} (D_d^{COM})^{-1/2}$  and  $(D_t^{COM})^{-1/2} L_t^{COM} (D_t^{COM})^{-1/2}$ ) instead as normalized graph Laplacians are known to perform better in many cases [57].

### Parameter settings

For setting the parameters of our algorithm, we performed cross-validation on the training set on the parameters  $p, \lambda, \mu_1, \mu_2, \nu_1, \nu_2$  to find the best parameter combination for each dataset, under each cross-validation setting. As mentioned earlier, the individual laplacians or the similarities can be weighted unequally to give more or less emphasis on a specific type of similarity, we weigh the Cosine, Correlation and Jaccard similarities heavily (4 times) relative to Hamming similarity. This was done because hamming similarity showed the least improvement in prediction accuracy as compared to the other three similarities when taken into account along with standard similarities. For the other methods, we set the parameters to their optimal (which were found to be already optimal) in [2].

### Interaction prediction

Tables 2 and 3 show the AUPR results and the AUC results for the CVS1 cross validation setting respectively. The following tables (Tables 4, 5, 6 and 7) report the CVS2 and CVS3 validation setting results. The second column in each table shows the results of our algorithm when only the standard similarity matrices ( $S_d$ : chemical structure similarity for drugs,  $S_t$ : Genomic sequence similarity for target proteins) were used for prediction.

The results clearly show that the interaction prediction in MGRNNM, not only shows great improvement on incorporation of new similarity types but also outperforms all the state-of-the-art prediction methods in terms of AUC and AUPR evaluation metrics in almost all test cases except in CVS3 (where the difference between AUPR obtained by the best-performing method: TMF and MGRNNM is not much).

Table 2. AUPR results for interaction prediction under validation setting CVS1.

AUPR	MGRNNM	standard	MC	MCG	WGRMF	RLS_WNN	CMF	NRLMF	TMF
E	<b>0.9660</b> ( <b>0.0006</b> )	0.9014 (0.0018)	0.7882 (0.0022)	0.7621 (0.0025)	0.8768 (0.0020)	0.8093 (0.0045)	0.8837 (0.0026)	0.8749 (0.0017)	0.7886 (0.0020)
IC	<b>0.9585</b> ( <b>0.0013</b> )	0.9298 (0.0026)	0.8868 (0.0028)	0.8346 (0.0025)	0.9225 (0.0022)	0.8459 (0.0106)	0.9373 (0.0019)	0.8674 (0.0056)	0.8654 (0.0041)
GPCR	<b>0.8515</b> ( <b>0.0033</b> )	0.7483 (0.0039)	0.6481 (0.0116)	0.5956 (0.0102)	0.7370 (0.0024)	0.6933 (0.0226)	0.7543 (0.0017)	0.7115 (0.0144)	0.6600 (0.0059)
NR	<b>0.8791</b> ( <b>0.0019</b> )	0.6408 (0.0234)	0.3950 (0.0298)	0.4558 (0.0202)	0.6016 (0.0378)	0.7072 (0.0290)	0.6383 (0.0149)	0.7390 (0.312)	0.4248 (0.0163)
Average	<b>0.9138</b>	0.8051	0.6795	0.6620	0.7845	0.7639	0.8034	0.7982	0.6847

<https://doi.org/10.1371/journal.pone.0226484.t002>

Table 3. AUC results for interaction prediction under validation setting CVS1.

AUC	MGRNNM	standard	MC	MCG	WGRMF	RLS_WNN	CMF	NRLMF	TMF
E	<b>0.9955</b> ( <b>0.0003</b> )	0.9798 (0.0004)	0.8753 (0.0023)	0.9596 (0.0015)	0.9647 (0.0013)	0.9635 (0.0014)	0.9705 (0.0013)	0.9761 (0.0017)	0.8943 (0.0030)
IC	<b>0.9947</b> ( <b>0.0004</b> )	0.9829 (0.0012)	0.9415 (0.0015)	0.9539 (0.0010)	0.9747 (0.0022)	0.9786 (0.0026)	0.9832 (0.0008)	0.9838 (0.0009)	0.9433 (0.0017)
GPCR	<b>0.9785</b> ( <b>0.0020</b> )	0.9531 (0.0028)	0.8110 (0.0055)	0.8977 (0.0047)	0.9432 (0.0010)	0.9458 (0.0044)	0.9493 (0.0031)	0.9620 (0.0023)	0.8373 (0.0038)
NR	<b>0.9660</b> ( <b>0.0056</b> )	0.9083 (0.0058)	0.5882 (0.0253)	0.8315 (0.0165)	0.8892 (0.0153)	0.9329 (0.0114)	0.8679 (0.0124)	0.9479 (0.0045)	0.5496 (0.0296)
Average	<b>0.9837</b>	0.9560	0.8040	0.9107	0.9429	0.9552	0.9427	0.9674	0.8061

<https://doi.org/10.1371/journal.pone.0226484.t003>

Table 4. AUPR results for interaction prediction under validation setting CVS2.

AUPR	MGRNNM	standard	MC	MCG	WGRMF	RLS_WNN	CMF	NRLMF	TMF
E	<b>0.8603</b> ( <b>0.0095</b> )	0.4089 (0.0104)	0.0114 (0.0005)	0.0457 (0.0008)	0.4019 (0.0128)	0.2409 (0.0272)	0.3848 (0.0094)	0.3582 (0.0101)	0.3748 (0.0113)
IC	<b>0.9026</b> ( <b>0.0197</b> )	0.3650 (0.0178)	0.0473 (0.0035)	0.0925 (0.0013)	0.3666 (0.0169)	0.3090 (0.0200)	0.3538 (0.0137)	0.3414 (0.0148)	0.3371 (0.0112)
GPCR	<b>0.8538</b> ( <b>0.0112</b> )	0.4175 (0.0076)	0.0404 (0.0017)	0.1091 (0.0044)	0.4247 (0.0113)	0.3463 (0.0106)	0.4059 (0.0104)	0.3671 (0.0067)	0.3866 (0.0078)
NR	<b>0.8773</b> ( <b>0.0125</b> )	0.5620 (0.0262)	0.1120 (0.0206)	0.2404 (0.0337)	0.5695 (0.0136)	0.5373 (0.0216)	0.5203 (0.0250)	0.5296 (0.0348)	0.4912 (0.0230)
Average	0.8735	0.4384	0.0528	0.1219	0.4407	0.3584	0.4162	0.3990	0.3974

<https://doi.org/10.1371/journal.pone.0226484.t004>

Table 5. AUC results for interaction prediction under validation setting CVS2.

AUC	MGRNNM	standard	MC	MCG	WGRMF	RLS_WNN	CMF	NRLMF	TMF
E	<b>0.9460</b> ( <b>0.0033</b> )	0.8260 (0.0108)	0.5060 (0.0090)	0.7413 (0.0118)	0.7982 (0.0144)	0.7755 (0.0093)	0.7952 (0.0110)	0.8151 (0.0062)	0.8204 (0.0111)
IC	<b>0.9714</b> ( <b>0.0095</b> )	0.7913 (0.0090)	0.5512 (0.0034)	0.7196 (0.0071)	0.7902 (0.0149)	0.7669 (0.0140)	0.7576 (0.0125)	0.7881 (0.0140)	0.8030 (0.0184)
GPCR	<b>0.9567</b> ( <b>0.0084</b> )	0.8805 (0.0024)	0.5855 (0.0039)	0.7745 (0.0027)	0.8800 (0.0025)	0.8524 (0.0072)	0.8067 (0.0067)	0.8841 (0.0054)	0.8452 (0.0044)
NR	<b>0.9533</b> ( <b>0.0127</b> )	0.8452 (0.0215)	0.5294 (0.0200)	0.6992 (0.0244)	0.8615 (0.0244)	0.8390 (0.0261)	0.8124 (0.0228)	0.8804 (0.0179)	0.8435 (0.0225)
Average	0.9568	0.8357	0.5430	0.7337	0.8325	0.8085	0.7930	0.8419	0.8280

<https://doi.org/10.1371/journal.pone.0226484.t005>

Table 6. AUPR results for interaction prediction under validation setting CVS3.

AUPR	MGRNNM	standard	MC	MCG	WGRMF	RLS_WNN	CMF	NRLMF	TMF
E	<b>0.9041</b> ( <b>0.0125</b> )	0.8087 (0.0156)	0.0124 (0.0005)	0.0691 (0.0009)	0.8070 (0.0185)	0.5465 (0.0144)	0.7808 (0.0131)	0.8112 (0.0166)	0.8000 (0.0167)
IC	<b>0.9029</b> ( <b>0.0024</b> )	0.8079 (0.0096)	0.0421 (0.0043)	0.2256 (0.0038)	0.8128 (0.0069)	0.7437 (0.0088)	0.7786 (0.0108)	0.7753 (0.0072)	0.7893 (0.0073)
GPCR	<b>0.7228</b> ( <b>0.0323</b> )	0.5963 (0.0336)	0.0549 (0.0105)	0.1061 (0.0027)	0.6093 (0.0314)	0.5397 (0.0193)	0.5989 (0.0323)	0.5515 (0.0234)	0.6001 (0.0243)
NR	<b>0.5418</b> ( <b>0.0309</b> )	0.4356 (0.0177)	0.0850 (0.0227)	0.2669 (0.0288)	0.4643 (0.0183)	0.4907 (0.0326)	0.4774 (0.0173)	0.5207 (0.0247)	0.4709 (0.0256)
Average	0.7679	0.6621	0.0486	0.1669	0.6734	0.5801	0.6589	0.6646	0.6650

<https://doi.org/10.1371/journal.pone.0226484.t006>

Table 7. AUC results for interaction prediction under validation setting CVS3.

AUC	MGRNNM	standard	MC	MCG	WGRMF	RLS_WNN	CMF	NRLMF	TMF
E	<b>0.9683</b> ( <b>0.0043</b> )	0.9246 (0.0091)	0.5234 (0.0057)	0.8065 (0.0012)	0.9338 (0.0071)	0.9067 (0.0105)	0.9272 (0.0050)	0.9465 (0.0052)	0.9436 (0.0072)
IC	0.9541 (0.0019)	0.9346 (0.0041)	0.4724 (0.0065)	0.7871 (0.0069)	0.9460 (0.0034)	0.9286 (0.0046)	0.9368 (0.0032)	<b>0.9587</b> ( <b>0.0027</b> )	0.9476 (0.0042)
GPCR	0.8975 (0.0093)	0.8798 (0.0134)	0.5683 (0.0310)	0.6289 (0.0151)	0.8892 (0.0110)	0.8694 (0.0146)	0.8966 (0.0073)	<b>0.9205</b> ( <b>0.0052</b> )	0.8735 (0.0160)
NR	0.7502 (0.0285)	0.7263 (0.0211)	0.3767 (0.0204)	0.6522 (0.0297)	0.7967 (0.0132)	0.8124 (0.0202)	0.8373 (0.0083)	<b>0.8613</b> ( <b>0.0097</b> )	0.8407 (0.0202)
Average	0.8909	0.8618	0.4575	0.7486	0.8922	0.8826	0.9004	0.9217	0.9013

<https://doi.org/10.1371/journal.pone.0226484.t007>

The degradation in performance in CVS3 setting can be attributed to the comparatively unstable results obtained for mainly NR dataset. This can be due to its excessively small size, as also concluded by [58].

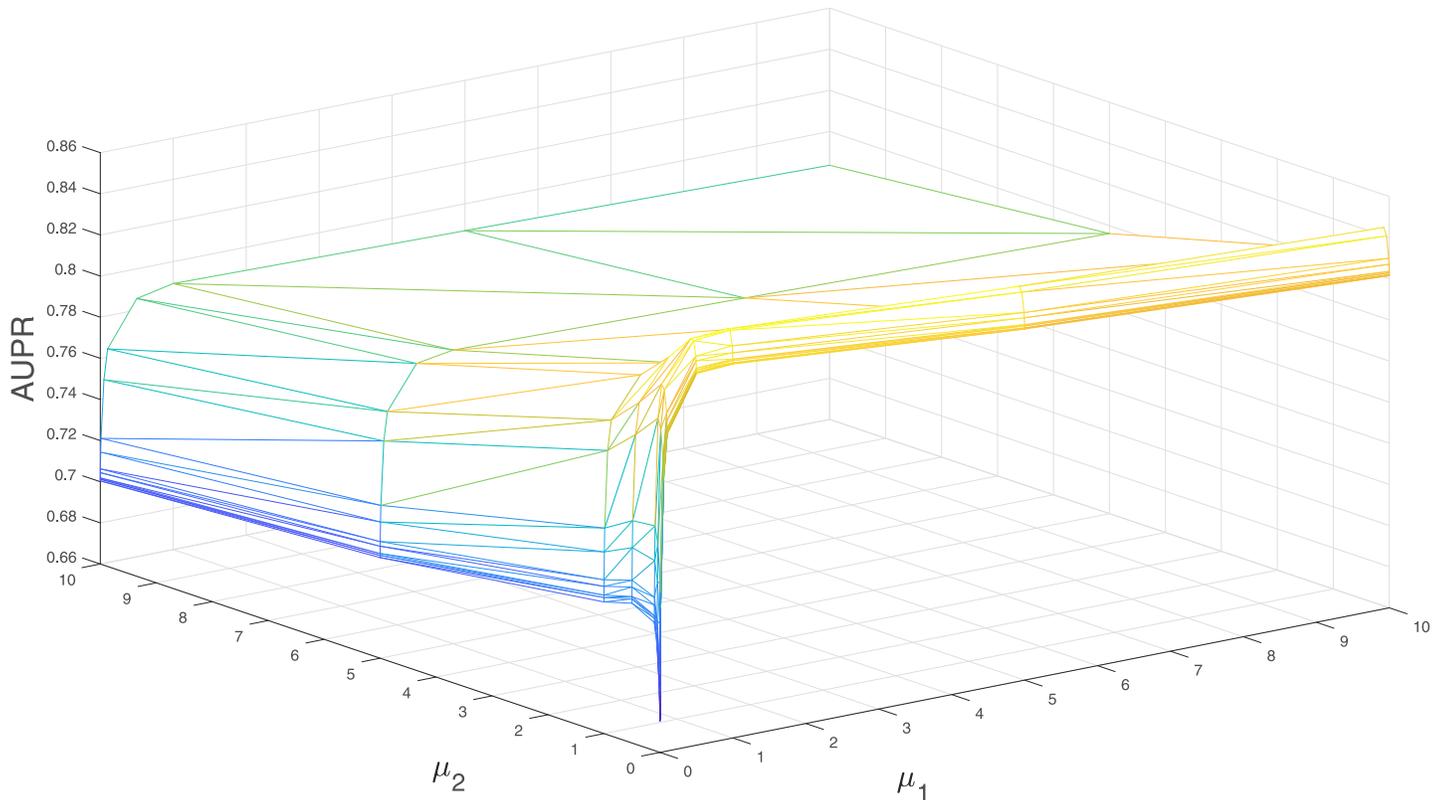
It is also observed that inference of MGRNNM under CVS1 is always better than in CVS2/ CVS3 because novel drugs/target proteins have no interaction available and hence CVS1 validation setting provides more information in the training data.

If we compare the performance of MGRNNM under CVS2 and CVS3 in all the datasets, an important factor which influence the results in these two cross-validation settings is the “drug to target ration” (say DTR). DTR for NR, GPCR, IC and E datasets are 54:26, 223:95, 105:102 and 445:664 respectively. Since, more information is the prior condition to achieve better inferences, performance under CVS2 should be better than in CVS3 for NR and GPCR datasets, performance under CVS2 and CVS3 should be similar for IC dataset and performance under CVS3 should be better than in CVS2. The results from MGRNNM perfectly follow this trend.

We also analyze the performance of MGRNNM with the two regularization parameters  $\mu_1$  and  $\mu_2$ , which govern the incorporation of the two graph laplacian terms in our algorithm. As an example, Fig 2 shows how these two parameters affect the prediction in case of GPCR dataset under cross-validation setting CVS1. When  $\mu_1$  and  $\mu_2$  are close to zero, the value of AUPR is 0.67; whereas when  $\mu_1$  and  $\mu_2$  gradually increase, the value of AUPR improves (more steep increase with  $\mu_1$  than with  $\mu_2$ ), achieving the best value (0.85) at  $\mu_1 = 0.5$  and  $\mu_2 = 0.1$  validating the effectiveness of multiple graph Laplacian components.

### Validation of multiple similarities

To precisely analyze the consequence of multiple similarities incorporation, we observed the mean AUPR for several cases:



**Fig 2. Three-dimensional mesh depicting the variation of AUPR with the parameters  $\mu_1$  and  $\mu_2$  for drug-target interaction prediction using MGRNNM.**

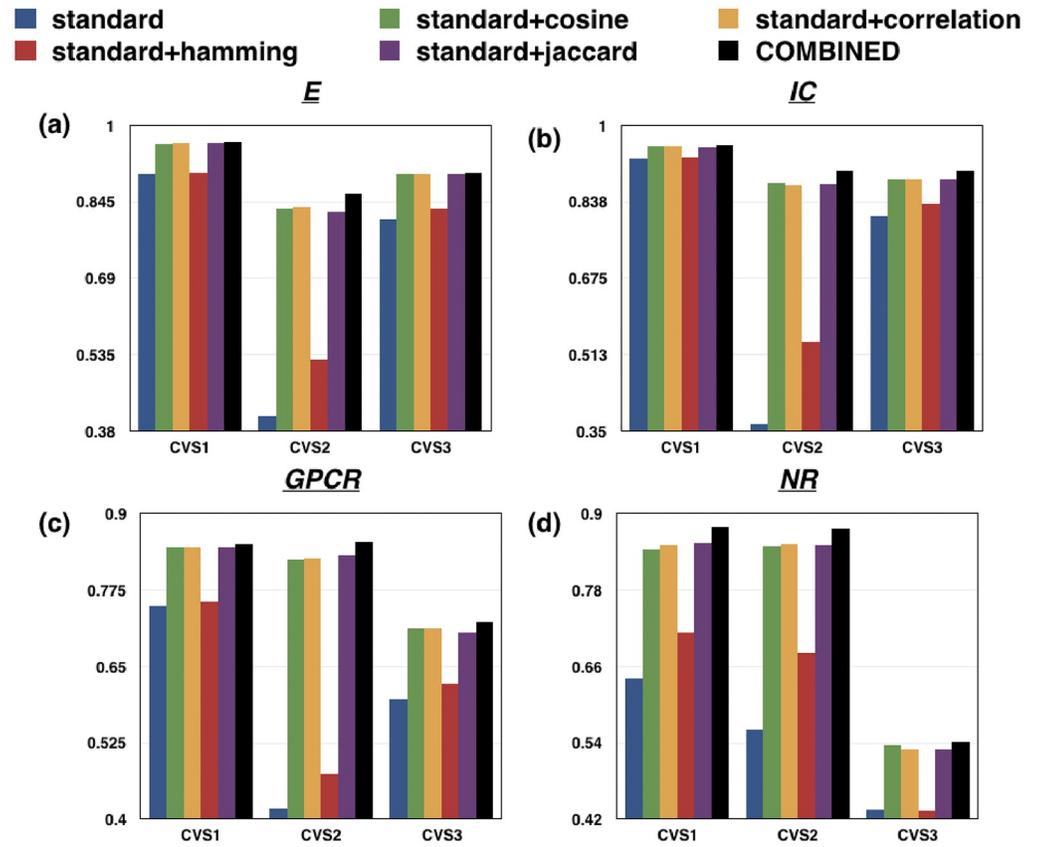
<https://doi.org/10.1371/journal.pone.0226484.g002>

- standard: When only the standard similarity matrices ( $S_d$ : chemical structure similarity for drugs,  $S_t$ : Genomic sequence similarity for target proteins) were used for prediction.
- standard+Cosine: When Cosine similarity between each pair of drugs/targets ( $S_d^{cos}$ ,  $S_t^{cos}$ ) was taken into account along with standard similarities.
- standard+Correlation: When Pearson's linear Correlation between each pair of drugs/targets ( $S_d^{cor}$ ,  $S_t^{cor}$ ) was taken into account along with standard similarities.
- standard+Hamming: When Hamming similarity between each pair of drugs/targets ( $S_d^{ham}$ ,  $S_t^{ham}$ ) was taken into account along with standard similarities.
- standard+Jaccard: When Jaccard similarity between each pair of drugs/targets ( $S_d^{jac}$ ,  $S_t^{jac}$ ) was taken into account along with standard similarities.
- COMBINED: When all five similarity types between each pair of drugs/targets ( $S_d^{COM}$ ,  $S_t^{COM}$ ) were taken into account.

The analysis was carried out for every dataset under all the three cross-validation settings. Fig 3 clearly depicts that incorporating all the similarities for drugs and targets for prediction task yields the best results.

## Conclusion

Drug-target interaction prediction is a crucial task in genomic drug discovery. Many computational techniques have been proposed in the literature. In this work, we presented a novel



**Fig 3.** Bar plots depicting that incorporating all the similarities for drugs and targets for prediction task yields best results for every dataset (a) E (b) IC (c) GPCR and (d) NR under the three cross-validation settings in comparison to the cases where each type of similarity was considered separately. Here, “standard” represents the case when only the chemical structure similarity for drugs and genomic sequence similarity for targets were taken into account and “COMBINED” refers to the use case where all the similarity matrices (standard similarity, Cosine similarity, Correlation, Hamming similarity and Jaccard similarity) were considered.

<https://doi.org/10.1371/journal.pone.0226484.g003>

chemogenomic approach for predicting the drug-target interactions, MGRNNM (Multi-Graph regularized Nuclear Norm Minimization). It is a graph regularized version of the traditional Nuclear Norm Minimization algorithm which incorporates multiple Graph Laplacians over the drugs and targets into the framework for an improved interaction prediction. The algorithm is generic and can be used for prediction in protein-protein interaction [59], RNA-RNA interaction [60], etc.

The evaluation was performed using three different cross-validation settings, namely CVS1 (random drug-target pairs left out), CVS2 (entire drug profile left out) and CVS3 (entire target profile left out) to compare our method with 5 other state-of-the-art methods (three specifically designed for DTI prediction). In almost all of the test cases, our algorithm shows the best performance, outperforming the baselines. This work can be extended by accounting for more types of drug and target similarities which could be either chemically/biologically driven or obtained from the metadata itself to improve the prediction accuracy even further.

## Supporting information

**S1 File.** Supplementary file showing the experimental results on the analysis of the improvements achieved by the MGRNNM algorithm and statistical significance of the

**improvement of MGRNNM over the other methods.**  
(PDF)

## Author Contributions

**Conceptualization:** Angshul Majumdar.

**Data curation:** Aanchal Mongia.

**Formal analysis:** Aanchal Mongia.

**Methodology:** Angshul Majumdar.

**Software:** Aanchal Mongia, Angshul Majumdar.

**Supervision:** Angshul Majumdar.

**Writing – original draft:** Aanchal Mongia.

**Writing – review & editing:** Aanchal Mongia, Angshul Majumdar.

## References

1. Dai YF, Zhao XM. A survey on the computational approaches to identify drug targets in the postgenomic era. *BioMed research international*. 2015; 2015. <https://doi.org/10.1155/2015/239654>
2. Ezzat A, Wu M, Li XL, Kwok CK. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics*. 2018.
3. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*. 2011; 40(D1):D1100–D1107. <https://doi.org/10.1093/nar/gkr777> PMID: 21948594
4. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*. 2007; 36(suppl\_1):D901–D906. <https://doi.org/10.1093/nar/gkm958> PMID: 18048412
5. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*. 2011; 40(D1):D109–D114. <https://doi.org/10.1093/nar/gkr988> PMID: 22080510
6. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, Von Mering C, Jensen LJ, et al. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic acids research*. 2013; 42(D1):D401–D407. <https://doi.org/10.1093/nar/gkt1207> PMID: 24293645
7. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, et al. SuperTarget and Matador: resources for exploring drug–target relationships. *Nucleic acids research*. 2007; 36(suppl\_1):D919–D922. <https://doi.org/10.1093/nar/gkm862> PMID: 17942422
8. Masoudi-Nejad A, Mousavian Z, Bozorgmehr JH. Drug–target and disease networks: polypharmacology in the post-genomic era. *In silico pharmacology*. 2013; 1(1):17. PMID: 25505661
9. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*. 2004; 3(8):673. <https://doi.org/10.1038/nrd1468> PMID: 15286734
10. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, et al. Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS computational biology*. 2012; 8(5):e1002503. <https://doi.org/10.1371/journal.pcbi.1002503> PMID: 22589709
11. McLean SR, Gana-Weisz M, Hartzoulakis B, Frow R, Whelan J, Selwood D, et al. Imatinib binding and cKIT inhibition is abrogated by the cKIT kinase domain I missense mutation Val654Ala. *Molecular cancer therapeutics*. 2005; 4(12):2008–2015. <https://doi.org/10.1158/1535-7163.MCT-05-0070> PMID: 16373716
12. Frantz S. *Drug discovery: playing dirty*; 2005.
13. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nature biotechnology*. 2007; 25(2):197. <https://doi.org/10.1038/nbt1284> PMID: 17287757
14. Johnson A, Wiley-Interscience MM. *Concepts and Applications of Molecular Similarity*. Edited; 1991.

15. Xie L, Evangelidis T, Xie L, Bourne PE. Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS computational biology*. 2011; 7(4):e1002037. <https://doi.org/10.1371/journal.pcbi.1002037> PMID: 21552547
16. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, et al. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic acids research*. 2006; 34(suppl\_2):W219–W224. <https://doi.org/10.1093/nar/gkl114> PMID: 16844997
17. Pujadas G, Vaque M, Ardevol A, Blade C, Salvado M, Blay M, et al. Protein-ligand docking: A review of recent advances and future perspectives. *Current Pharmaceutical Analysis*. 2008; 4(1):1–19. <https://doi.org/10.2174/157341208783497597>
18. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, et al. Structure-based maximal affinity model predicts small-molecule druggability. *Nature biotechnology*. 2007; 25(1):71. <https://doi.org/10.1038/nbt1273> PMID: 17211405
19. Nagamine N, Shirakawa T, Minato Y, Torii K, Kobayashi H, Imoto M, et al. Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. *PLoS computational biology*. 2009; 5(6):e1000397. <https://doi.org/10.1371/journal.pcbi.1000397> PMID: 19503826
20. Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. In: *Proceedings of the 10th international conference on World Wide Web*. ACM; 2001. p. 285–295.
21. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*. 2008; 24(13):i232–i240. <https://doi.org/10.1093/bioinformatics/btn162> PMID: 18586719
22. Shi JY, Yiu SM. SRP: A concise non-parametric similarity-rank-based model for predicting drug-target interactions. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2015. p. 1636–1641.
23. van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PloS one*. 2013; 8(6):e66952. <https://doi.org/10.1371/journal.pone.0066952> PMID: 23840562
24. Shi JY, Li JX, Lu HM, Zhang Y. Predicting Drug-Target Interactions Between New Drugs and New Targets via Pairwise K-nearest Neighbor and Automatic Similarity Selection. In: He X, Gao X, Zhang Y, Zhou ZH, Liu ZY, Fu B, et al., editors. *Intelligence Science and Big Data Engineering. Big Data and Machine Learning Techniques*. Cham: Springer International Publishing; 2015. p. 477–486.
25. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*. 2009; 25(18):2397–2403. <https://doi.org/10.1093/bioinformatics/btp433> PMID: 19605421
26. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*. 2011; 27(21):3036–3043. <https://doi.org/10.1093/bioinformatics/btr500> PMID: 21893517
27. Mei JP, Kwok CK, Yang P, Li XL, Zheng J. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*. 2012; 29(2):238–245. <https://doi.org/10.1093/bioinformatics/bts670> PMID: 23162055
28. He Z, Zhang J, Shi XH, Hu LL, Kong X, Cai YD, et al. Predicting drug-target interaction networks based on functional groups and biological features. *PloS one*. 2010; 5(3):e9603. <https://doi.org/10.1371/journal.pone.0009603> PMID: 20300175
29. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, et al. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PloS one*. 2012; 7(5):e37608. <https://doi.org/10.1371/journal.pone.0037608> PMID: 22666371
30. Chen H, Zhang Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PloS one*. 2013; 8(5):e62975. <https://doi.org/10.1371/journal.pone.0062975> PMID: 23667553
31. Shi JY, Li JX, Lu HM. Predicting existing targets for new drugs base on strategies for missing interactions. *BMC bioinformatics*. 2016; 17(8):282. <https://doi.org/10.1186/s12859-016-1118-2> PMID: 27585458
32. Shi JY, Liu Z, Yu H, Li YJ. Predicting drug-target interactions via within-score and between-score. *BioMed research international*. 2015; 2015. <https://doi.org/10.1155/2015/350983>
33. Shi JY, Yiu SM, Li Y, Leung HC, Chin FY. Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods*. 2015; 83:98–104. <https://doi.org/10.1016/j.ymeth.2015.04.036> PMID: 25957673
34. Chen X, Liu MX, Yan GY. Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*. 2012; 8(7):1970–1978. <https://doi.org/10.1039/c2mb00002d> PMID: 22538619

35. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. *Computer*. 2009;(8):30–37. <https://doi.org/10.1109/MC.2009.263>
36. Cobanoglu MC, Liu C, Hu F, Oltvai ZN, Bahar I. Predicting drug–target interactions using probabilistic matrix factorization. *Journal of chemical information and modeling*. 2013; 53(12):3399–3409. <https://doi.org/10.1021/ci400219z> PMID: 24289468
37. Ezzat A, Zhao P, Wu M, Li XL, Kwok CK. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*. 2017; 14(3):646–656. <https://doi.org/10.1109/TCBB.2016.2530062>
38. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999; 401(6755):788. <https://doi.org/10.1038/44565> PMID: 10548103
39. Candès EJ, Tao T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*. 2010; 56(5):2053–2080. <https://doi.org/10.1109/TIT.2010.2044061>
40. Candès EJ, Recht B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*. 2009; 9(6):717. <https://doi.org/10.1007/s10208-009-9045-5>
41. Recht B. A simpler approach to matrix completion. *Journal of Machine Learning Research*. 2011; 12 (Dec):3413–3430.
42. Wang M, Tang C, Chen J. Drug-Target Interaction Prediction via Dual Laplacian Graph Regularized Matrix Completion. *BioMed Research International*. 2018; 2018. <https://doi.org/10.1155/2018/1425608>
43. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*. 2006; 34(suppl\_1):D354–D357. <https://doi.org/10.1093/nar/gkj102> PMID: 16381885
44. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, et al. BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research*. 2004; 32(suppl\_1):D431–D433. <https://doi.org/10.1093/nar/gkh081> PMID: 14681450
45. Hattori M, Tanaka N, Kanehisa M, Goto S. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic acids research*. 2010; 38(suppl\_2):W652–W656. <https://doi.org/10.1093/nar/gkq367> PMID: 20460463
46. AMoZ J. Identification of Common Molecular Subsequences.
47. Majumdar A, Ward RK. Some empirical advances in matrix completion. *Signal Processing*. 2011; 91 (5):1334–1338. <https://doi.org/10.1016/j.sigpro.2010.12.005>
48. Chung FR. Spectral graph theory (CBMS regional conference series in mathematics, No. 92). 1996.
49. Combettes PL, Pesquet JC. Proximal splitting methods in signal processing. In: *Fixed-point algorithms for inverse problems in science and engineering*. Springer; 2011. p. 185–212.
50. Nishihara R, Lessard L, Recht B, Packard A, Jordan MI. A general analysis of the convergence of ADMM. *arXiv preprint arXiv:150202009*. 2015.
51. Boyd S. Alternating direction method of multipliers. In: *Talk at NIPS workshop on optimization and machine learning*; 2011.
52. Kirrinnis P. Fast algorithms for the Sylvester equation  $AX - XBT = C$ . *Theoretical Computer Science*. 2001; 259(1-2):623–638. [https://doi.org/10.1016/S0304-3975\(00\)00322-4](https://doi.org/10.1016/S0304-3975(00)00322-4)
53. Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2013. p. 1025–1033.
54. Liu Y, Wu M, Miao C, Zhao P, Li XL. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS computational biology*. 2016; 12(2):e1004760. <https://doi.org/10.1371/journal.pcbi.1004760> PMID: 26872142
55. Shi JY, Zhang AQ, Zhang SW, Mao KT, Yiu SM. A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization. *BMC systems biology*. 2018; 12(9):136. <https://doi.org/10.1186/s12918-018-0663-x> PMID: 30598094
56. Kalofolias V, Bresson X, Bronstein M, Vandergheynst P. Matrix completion on graphs. *arXiv preprint arXiv:14081717*. 2014.
57. Belkin M, Niyogi P, Sindhvani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*. 2006; 7(Nov):2399–2434.
58. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Sz wajda A, Tang J, et al. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*. 2014; 16(2):325–337. <https://doi.org/10.1093/bib/bbu010> PMID: 24723570

59. Albert I, Albert R. Conserved network motifs allow protein–protein interaction prediction. *Bioinformatics*. 2004; 20(18):3346–3352. <https://doi.org/10.1093/bioinformatics/bth402> PMID: 15247093
60. Alkan C, Karakoc E, Nadeau JH, Sahinalp SC, Zhang K. RNA–RNA interaction prediction and anti-sense RNA target search. *Journal of Computational Biology*. 2006; 13(2):267–282. <https://doi.org/10.1089/cmb.2006.13.267> PMID: 16597239