Non-negative Matrix Tri-Factorization for Data Integration and Network-based Drug Repositioning

Gaia Ceddia
Dipartimento di Elettronica,
Informazione e Bioingegneria
Politecnico di Milano
Piazza L. Da Vinci, 32,
Milan, Italy
gaia.ceddia@polimi.it

Pietro Pinoli
Dipartimento di Elettronica,
Informazione e Bioingegneria
Politecnico di Milano
Piazza L. Da Vinci, 32,
Milan, Italy
pietro.pinoli@polimi.it

Stefano Ceri
Dipartimento di Elettronica,
Informazione e Bioingegneria
Politecnico di Milano
Piazza L. Da Vinci, 32,
Milan, Italy
stefano.ceri@polimi.it

Marco Masseroli
Dipartimento di Elettronica,
Informazione e Bioingegneria
Politecnico di Milano
Piazza L. Da Vinci, 32,
Milan, Italy
marco.masseroli@polimi.it

***Abstract*—Drug discovery is a high cost and high risk process, thus finding new uses for approved drugs, i.e. drug repositioning, via computational methods has become increasingly interesting. In this study, we present a new network-based approach for predicting potential new indications for existing drugs through their connections with other biological entities. For this aim, we first built a large network integrating drugs, proteins, biological pathways and drugs' categories as nodes of the network, and connections between such nodes as links of the network. Our method leverages the Non-Negative Matrix Tri-Factorization reconstruction of adjacency matrices in order to predict novel category-drug links, i.e. a new category (or use) associated with a drug, taking the entire network information into account. We tested our method on a set of 1,120 drugs labeled with ten categories; when we hide to the method the 10% of the drug-category associations, it was able to infer those missing values with a recall of 60% and a precision of 70%. Precision and recall remain higher than a Random Classifier in case of larger percentage of hidden links, demonstrating the robustness of the method. Also, we were able to predict novel drug-label associations not yet reported in the repository. Finally, we favorably compared our method with a state of the art method for drug repositioning; the NMTF method achieved an average precision score of 0.68 vs. the 0.55 score of the state of the art method.**

***Index Terms*—drug repositioning, protein-protein networks, pathways, interaction networks, data integration, link prediction**

I. Introduction

One of the most relevant steps in drug development consists in inferring potential indications for novel molecules and in the repositioning of approved drugs [1]. Especially drug repositioning has the benefit of starting from well-characterized molecules, hence reducing the risks in clinical phases and the cost of trials [1]. Drug repositioning's success and application derives from polypharmacology, i.e., the importance of the multi-target approach vs. the single target one in drug discovery [2]. Namely, drugs specifically designed for targeting one molecules may have other effects on other targets.

Computational tools that scan databases of approved drugs and predict novel indications drastically reduce the cost and the time of drug development, driving the need for more new approaches. The majority of computational methods for drug repositioning are based on drug similarity, which assumes that similar drugs are indicated for similar diseases [3]. For instance, in [4] the authors used side effects to identify drug-targets connections that further lead them to drug repositioning. Conversely, in [5] another approach based on integrated information from protein interaction and literature mining is reported to infer drug connections in particular diseases. More recently, [3] and [6] presented a 'Guilt by Association' approach to predict novel drug uses based on the drugs-disease relationships and drugs-protein relationships respectively.

Differently from similarity-based methods, we use a network-based approach integrating both drug-protein and protein-pathway connections, as well as previous knowledge about drugs, i.e., drugs' categories, to predict category-drug links. In this study, we propose a novel method based on Non-negative Matrix Tri-Factorization (NMTF) for data integration and the inference of indications about both new and approved molecules. The NMTF has proven its effectiveness in several fields for the integration and clustering of heterogeneous datasets. For examples, the NMTF has demonstrated a great potential in addressing various biological problems, such as disease association prediction [17] and protein-protein interaction prediction [18]. In a recent study, NMTF was applied to find patient-specific treatment for a particular cancer, analyzing a tripartite complex network [8]. For its wide applicability and good results reported in the literature, we used the NMTF graph regularized method in our application and we innovatively propose its applicability as a multilabel classifier, where categories (or labels) are part of the network as nodes. In other words, we predict drugs connection to a certain category and by doing that we classify those drugs according to the category.

The rest of the paper is organized as follows. Section II provides a detailed description of the NMTF method, section III describes used datasets, drug similarity comparative method and evaluation metrics, section IV reports the performance

This work was supported by the ERC Advanced Grant 693174 "Data-Driven Genomic Computing (GeCo)".

978-1-7281-1462-0/19/$31.00 ©2019 IEEE

curves and results for drug repositioning, finally section V contains conclusions and future work.

II. Methods

A. Non-negative Matrix Tri-Factorization

For the sake of simplicity, we describe the Non-negative Matrix Tri-Factorization (NMTF) [7] in the context of decomposition of incidence matrices. Consider two datasets D_1 and D_2 connected by some kind of relation, an incidence matrix is a two-dimensional matrix $R \in \mathbb{R}_+^{|D_1|\times|D_2|}$ such that each entry $R_{i,j}$ is positive if the i-th element of D_1 is connected with the j-th element of D_2, zero otherwise. Thus, the NMTF is able to work on any bipartite graph, independently of the type of elements of the two connected datasets. This property contributed in the establishment of NMTF as a co-clustering technique, with an important role in presence of heterogeneous datasets [8] (e.g., the result of a data integration framework).

We introduce NMTF starting from the Non-negative Matrix Factorization (NMF). In general, NMF decomposes R into two non-negative matrices:

$$R \approx G_1 G_2^T$$

where $G_1 \in \mathbb{R}_+^{|D_1|\times k}$ and $G_2 \in \mathbb{R}_+^{|D_2|\times k}$ and $k < min(|D_1|, |D_2|)$.

We can further constrain G_1 and G_2 to be orthogonal and to minimize the Frobenius norm:

$$\underset{G_1\geq 0, G_2 \geq 0}{min} \parallel R - G_1 G_2^T \parallel^2, s.t. \quad G_1^T G_1 = I, G_2^T G_2 = I$$

This has been proven to correspond to the simultaneous K-means clustering of the rows and the columns of R [10], with G_1 being the cluster indicator matrix for clustering rows and G_2 the cluster indicator matrix for clustering columns.

However, the double orthogonality constraint showed to be too restrictive for the low-rank approximation, thus Ding and colleagues [7] proposed to factorize R in three components:

$$R \approx G_1 S_{12} G_2^T,$$

where $G_1 \in \mathbb{R}_+^{|D_1|\times k_1}$, $G_2 \in \mathbb{R}_+^{|D_2|\times k_2}$, $S_{12} \in \mathbb{R}_+^{k_1\times k_2}$, both G_1 and G_2 are orthogonal, and $k_1, k_2 < min(|D_1|, |D_2|)$.

Finding optimal G_1, S_{12} and G_2 matrices such that their product is equal to R is recognized to be a NP-hard problem [13]. Thus, an approximate solution is computed by minimizing the Frobenius norm between the input relation matrix and the product of low-dimensional matrix factors [7]:

$$\underset{G_1\geq 0, G_2\geq 0, S\geq 0}{min} J = \underset{G_1\geq 0, G_2\geq 0, S\geq 0}{min} \parallel R_{ij} - G_1 S_{12} G_2^T \parallel^2, \\ s.t. \quad G_1^T G_1 = I, G_2^T G_2 = I$$

The minimisation of the objective function J for the computation of G_1, G_2 and S_{12} is performed by means iterative update rules [14]. Starting from a random initialization of the three matrices, we progressively compute the solution by iterating the following rules:

$$G_{1_{(i,j)}} \leftarrow G_{1_{(i,j)}} \sqrt{\frac{(R G_2 S_{12}^T)_{i,j}}{(G_1 G_1^T R G_2 S_{12}^T)_{i,j}}}$$

$$G_{2_{(i,j)}} \leftarrow G_{2_{(i,j)}} \sqrt{\frac{(R^T G_1 S_{12})_{i,j}}{(G_2 G_2^T R^T G_1 S_{12})_{i,j}}}$$

$$S_{12_{(i,j)}} \leftarrow S_{12_{(i,j)}} \sqrt{\frac{(G_1^T R G_2)_{i,j}}{(G_1^T G_1 S_{12} G_2^T G_2)_{i,j}}}$$

At each iteration the above rules decrease the value of J and are applied until the convergence criterion $\frac{|J_{n+1} - J_n|}{|J_n|} < 10^{-5}$ is reached [8].

B. Extension of NMTF to multiple datasets

Now we consider the case in which we have three datasets D_1, D_2 and D_3 and two association matrices $R_{12} \in \mathbb{R}_+^{|D_1|\times|D_2|}$, which connects elements of D_1 to elements of D_2, and $R_{23} \in \mathbb{R}_+^{|D_2|\times|D_3|}$, which connects elements of D_2 to elements of D_3.

We can extend the NMTF to this context and compute a set of positive matrices G_1, G_2, G_3, S_{12} and S_{23} such all of the G_i are orthogonal and they minimize the objective function:

$$\underset{G_1\geq 0, G_2\geq 0, S\geq 0}{min} \parallel R_{12} - G_1 S_{12} G_2^T \parallel^2 + \\ + \parallel R_{23} - G_2 S_{23} G_3^T \parallel^2 \tag{1}$$

Operatively, we can compute the matrices using the exact same update rules, except for G_2, that has to consider both approximation errors in 1. In this case we have to rewrite the update rule as follows:

$$G_{2_{(i,j)}} \leftarrow G_{2_{(i,j)}} \sqrt{\frac{(R_{12}^T G_1 S_{12} + R_{23} G_3 S_{23}^T)_{i,j}}{(G_2 G_2^T R_{12}^T G_1 S_{12} + G_2 G_2^T R_{23} G_3 S_{23}^T)_{i,j}}}$$

Following the same intuition, it is possible to extend the NMTF to any number N of datasets, by minimizing a coherent cost function:

$$J = \sum_{i=1}^{N-1} \parallel R_{i,i+1} - G_i S_{i,i+1} G_{i+1}^T \parallel^2$$

C. Integrating a priori information

Moreover, the NMTF can accept prior information embedded in the objective function to guide the co-clustering, leading to a semi-supervised method. The relation matrix R_{ij} describes inter-type relationships among heterogeneous datasets; yet, relationships within the same dataset (intra-type) can also occur. Such intra-type connections can be represented as Laplacian matrices L and embedded in the objective function as constraints [12]. By including the constraint matrices L_i and L_j in the objective function, we force connected objects of the same type to belong to the same cluster [15], [16]. These additional terms (named *graph regularization terms* [15], [16]) can be used in the objective function as follows:

$$\min_{G_i \geq 0, G_j \geq 0} \sum_{i,j}^{N} \parallel R_{ij} - G_i S_{ij} G_j^T \parallel^2 + tr(G_i^T L_i G_i) + \\ + tr(G_j^T L_j G_j), s.t. G_i^T G_i = I, G_j^T G_j = I \qquad (2)$$

where $tr(\cdot)$ denotes the trace of the matrix.

D. NMTF reconstruction for link prediction

In addition to co-clustering, NMTF is also used for matrix completion [12]. Namely, after obtaining the three matrix factors, G_i, S_{ij} and G_j^T, the reconstructed data matrix (obtained from the product of the three matrix factors) is more complete than the initial data matrix, R_{ij}, featuring new links, not present in the data, and emerged from the latent structure captured by the matrix factors [12]. Therefore, NMTF has the unique property of modelling heterogeneous network data and predicting unobserved links.

III. Data and Implementation

We applied NMTF to a quadripartite network (Figure 1), integrating three different relation matrices and two intra-type relationship matrices. The network connects drug categories to drugs, drugs to proteins and proteins to pathway. We then use the computed G_1, S_{12} and G_2 to reconstruct the first relation matrix R_{12} (i.e., the one associating drugs to categories) in order to predict missing links and, therefore, novel potential indications (drug repositioning). Since all of the three matrices are positive, also the reconstructed matrix $\bar{R}_{12}$ has positive values; we then set a threshold $\delta \in [0, 1]$ and we consider the *j-th* drugs to be associated to the *i-th* category if $\bar{R}_{12(i,j)} > \delta$. In order to show the efficiency of our method, we performed a 10-fold cross validation, where, in each fold, a certain amount of links between category and drugs was randomly deleted. We measure how well the method performed in inferring the hidden links in terms of precision and recall. Finally we compared NMTF to the drug similarity method for drug repositioning that was proposed in [3].

A. Datasets

We considered five datasets including category-drug, drug-target, protein-pathway, protein-protein and pathway-pathway connections, respectively extracted form different databases. We used them to construct five different networks, which we then merged in a single quadripartite network for its evaluation with the NMTF method. Specifically:

- The category-drug network (CDN) is obtained by using information from the DrugBank database [19]. We selected the drugs whose DrugBank's category is annotated as at least one of the following: Anti-Infective Agents, Nervous System Agents, Anti-Bacterial Agents, Immunosuppressive Agents, Hormones, Analgesics, Anti-Inflammatory Agents, Antibodies, Membrane Transport Modulators, or Respiratory System Agents. Therefore, CDN represents drugs' annotation according to DrugBank and $R_{12}^{n_1 \times n_2}$ is the adjacency matrix of CDN, where n_1 = 10 are the selected categories and n_2 = 1,120 are the drugs. If the drug d is labeled as the category c the link in R_{12} is 1, or zero otherwise. The training set for the NMTF classifier is built deleting 10 percent of the category-drug links (dashed lines in Figure 1).
- The drug-target network (DTN) is obtained searching for protein targets of the n_2 = 1,120 drugs. This resulted in a DTN of 5,191 interactions between n_2 = 1,120 drugs and n_3 = 1,012 target proteins. We represented these interactions through a binary relationship matrix $R_{23}^{n_2 \times n_3}$, which encodes drug-target interactions such as $R_{23}[d][t] = 1$ if the drug d has a relation with the protein target t or zero otherwise.
- The protein-pathway network (PPN) is built by considering 17,037 connections between the n_3 = 1,012 proteins and n_4 = 1,563 pathways, selected from the Reactome database [20]. Relationships between targets and pathways are represented as a zero-one matrix $R_{34}^{n_3 \times n_4}$ (with $R_{34}[t][p] = 1$ if the target t and pathway p are related, or zero otherwise).
- The target-target interactions (TTIs, or protein-protein interactions - PPIs) for the selected proteins are extracted from the UniProt database [21]. This resulted in 901 target-target interactions among the n_3 = 1,012 proteins. PPIs were defined by a Laplacian matrix $L_3^{n_3 \times n_3}$, computed as: $L_3 = D_3 - A_3$, where A_3 is the adjacency matrix of the PPIs and D_3 is the diagonal degree matrix of the PPIs (i.e., a diagonal matrix, whose elements on the diagonal are the row sums of A_3).
- The relationships within pathways are obtained from the Reactome database [20]. Reactome pathways are structured hierarchically, from the most general pathway to the most specific one and according to the biological events within the cell. The pathway-pathway hierarchy defines the relationships between a parent pathway and a child pathway. We built the pathway hierarchy network among the extracted n_4 = 1,563 pathways and defined pathway-pathway interactions (PaHs) by a Laplacian matrix $L_4^{n_4 \times n_4}$, computed as: $L_4 = D_4 - A_4$, where A_4 is the adjacency matrix of the PaHs and D_4 is the diagonal degree matrix of the PaHs.

The merged network has four different types of nodes, specifically: n_1 category, n_2 drug, n_3 protein and n_4 pathway nodes. Category-drug, drug-protein, protein-pathway, protein-protein and pathway-pathway relations are encoded in the R_{12}, R_{23}, R_{34}, L_3 and L_4 matrices, respectively (Figure 1).

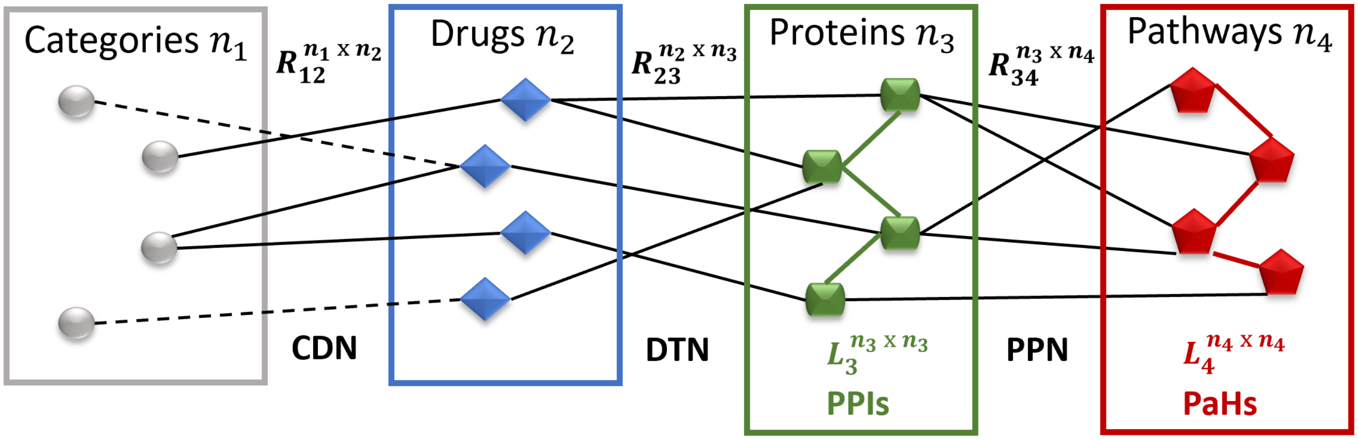

Fig. 1. Schematic illustration of the datasets used and the constructed network. Four kinds of nodes are shown: categories (grey circles), drugs (blu diamonds), target proteins (green squares) and pathways (red pentagons). Category-drug links (CDN) are represented by the R_{12} relation matrix (solid lines describe the training set for the NMTF, whereas dashed lines are the test set), drug-target connections (DTN) are encoded in the relationship matrix R_{23}, whereas protein-pathway connections (PPN) are encoded in the relationship matrix R_{34}. Links between proteins (PPIs) and between pathways (PaHs) are encoded in the L_3 and L_4 matrices, respectively.

B. NMTF objective function and parameter choice

The objective function for the NMTF method applied to the quadripartite network in Figure 1 is defined as:

$$\begin{aligned} min(&\parallel R_{12} - G_1 S_{12} G_2^T \parallel^2 + \parallel R_{23} - G_2 S_{23} G_3^T \parallel^2 + \\ &+ \parallel R_{34} - G_3 S_{34} G_4^T \parallel^2 + tr(G_3^T L_3 G_3) + tr(G_4^T L_4 G_4)), \\ &s.t. G_1 \geq 0, G_2 \geq 0, G_3 \geq 0, G_4 \geq 0, \\ &G_1^T G_1 = I, G_2^T G_2 = I, G_3^T G_3 = I, G_4^T G_4 = I \end{aligned} \tag{3}$$

where matrices $G_1^{n_1 \times k_1}$, $G_2^{n_2 \times k_2}$, $G_3^{n_3 \times k_3}$ and $G_4^{n_4 \times k_4}$ indicate cluster memberships for categories, drugs, proteins and pathways, respectively; based on their entries, n_2 drugs are assigned to k_2 drug clusters, n_3 proteins are assigned to k_3 protein clusters and n_4 pathways are assigned to k_4 pathway clusters. The parameter k_1 is set equal to n_1 as the clustering of drug's categories is not needed for drug repositioning.

For the use of the NMTF graph regularized algorithm for drug repositioning, we need to estimate the rank parameters, i.e., the number of clusters k_2, k_3 and k_4. They are chosen to be $k_2 < n_2, k_3 < n_3, k_4 < n_4$, and can be estimated by computing the cluster stability [23], [24]. A well-known measure for cluster stability over the total number of NMTF runs is the *dispersion coefficient* ρ [24]. Its values range in $0 \leq \rho \leq 1$, where 1 denotes a stable cluster. We performed multiple factorization runs for different triplets of rank parameters and computed the dispersion coefficient for each triplet. Finally, we chose the values of k_2, k_3 and k_4 corresponding to the highest cluster stability. Namely, $\rho_2 = 0.9519$ for $k_2 = 200$, $\rho_3 = 0.9750$ for $k_3 = 300$ and $\rho_4 = 0.9855$ for $k_4 = 350$.

C. Evaluation metrics

We evaluated the performances of RF and NMTF methods using Precision-Recall curves, which are based on precision rather than the false positive rate; thus, they better reflect model performance when predicting from sparse datasets. To assess the performance of our method, we trained the NMTF model by removing randomly the $x\%$ of the category-drug connections, i.e., deleting $x\%$ of links in the R_{12} matrix. We performed several tests varying the percentage of removed links from 10% up to 80% of the total. For each percentage, we repeated the test 20 times and reported as precision and recall their means over the 20 iterations. Given n as the number of category-drug association deleted, TP indicates the true positive category-drug predictions (i.e., the correctly predicted drug's annotations), FP the false positive predictions and FN the false negative category-drug associations. Then, we used the following well-known precision and recall metrics:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Precision computes the portion of true predicted category-drug associations out of all predicted ones. Recall instead measures the fraction of all the actual category-drug annotations that are predicted. Results of the Precision-Recall curves can be summarized with the average precision score (APS) [25], which is the sum of precisions achieved at each threshold weighted for the difference between recalls at that threshold and recalls at the previous threshold:

$$APS = \sum_n (R_n - R_{n-1}) P_n \tag{6}$$

where P_n and R_n are the precision and recall at threshold n.

D. Comparative study

To evaluate the relevance of the NMTF graph regularized method for drug repositioning, we compared it with the state of art method proposed in [3]. Thus, we implemented the approach in [3], which is based on drug pairwise similarity, and evaluated it on our data. The drug pairwise similarity, $S(d_i, d_j)$, between two drugs, d_i and d_j, is computed as linear combination of the similarities of their molecular structures and their target profiles (i.e., the bipartite graph featured by d_i, d_j and their target proteins). It is a score ranging from 0 to 1, which has been defined as follows:

$$S(d_i, d_j) = (1 + \lambda) * S_{str}(d_i, d_j) + \lambda * S_{tar}(d_i, d_j) \qquad (7)$$

where $S_{str}(d_i, d_j)$ is the molecular structure similarity between d_i and d_j, $S_{tar}(d_i, d_j)$ is the similarity between the target proteins of d_i and d_j, and λ $(0 < \lambda < 1)$ is a constant for weighting the target similarity [3].

We computed the structure similarity $(S_{str}(d_i, d_j))$ as the Tanimoto coefficient [22] of the chemical structure data present in DrugBank and stored as chemical fingerprint of each drug, where available. To evaluate the target similarity, we implemented a new network based on drugs-proteins bipartite network. According to [3], nodes are all possible combinations of drug pairs and protein pairs, and edges between them exist only if the two drugs have at least one target protein in common from the protein pair. The target similarity $(S_{tar}(d_i, d_j))$ between d_i and d_j is the average similarity of protein pairs connecting the $(d_i$,$d_j)$ pair. Finally, we inferred the new indications of drug d_i by its similarity with d_j, i.e., if d_j belongs to a certain category c, then d_i can be repositioned to that category.

For the comparative study, we limited the analysis to the 784 drugs with chemical structure data available, and we computed the APS scores for the drug similarity method and NMTF method on this set of drugs (Figure 3).

IV. Results and Discussion

We tested our method on several datasets, randomly built by removing from the dataset respectively the 10%, the 30%, the 60%, the 70% and 80% of the category to drug links. For each dataset we then computed the average reconstructed $\bar{R}_{12}$ matrix of 20 runs (or randomization) of the NMTF method and applied various thresholds δ, spanning from 0 to 1, in order to draw the precision-recall curves in Figure 2. Results show that our method performs well till when less than 30% of the association were removed, while the performances decrease dramatically when considering less true links, as expected. However, as presented in Figure 2, precision-recall curves for NMTF decreasing training set show higher APS scores than the Random Classifier (the APS score of a Random Classifier in this case is 0.1). For the 10% dataset in Figure 2, the predictor scored high values of both precision and recall. For example with a threshold $\delta = 0.1$ $Precision = 0.7$ and $Recall = 0.6$; incrementally greater thresholds (from 0.1 to 1) have $0.7 < Precision < 1$ and a decreasing recall, meaning that the predicted category-drug links overlapping with the true links are almost the same as the predicted links and they significantly decrease over the total number of true links when the threshold increases (Table I). Instead, thresholds smaller than 0.1 lead to low values of precision and high values of recall. To clarify NMTF results, we reported in Table I the number of predicted links (third column of the table) compared to the ones that have been removed from the training set (i.e., the test set and about the 10% of the total links) for one run (or randomization) out of 20. True positive links are the intersections between the test set links and the predicted links. Thus, for thresholds from 0 to 0.4 the number of predicted links overcomes the number of true positive links, and 160 true positive links are lost in the same range of thresholds. Moreover, from the 0.5 to 0.7 threshold range, the number of predicted links is equal to the intersection between true and predicted links, i.e., $Precision = 1$.

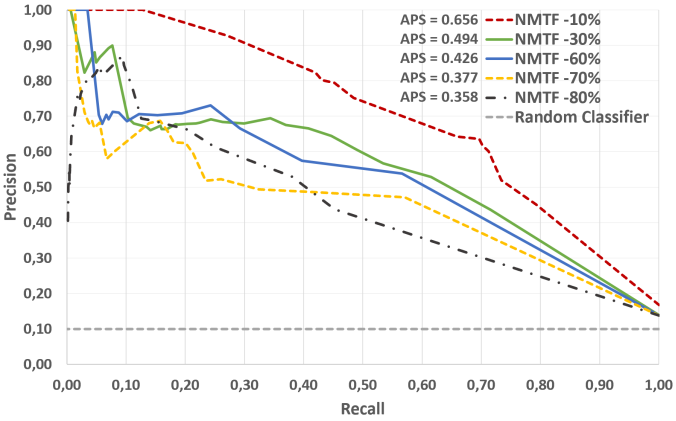

Fig. 2. Precision-Recall curves for NMTF different training sets. The curves are evaluated from the R_{12} matrix reconstruction, when the 10%, 30%, 60%, 70% and 80% of the links are kept out.

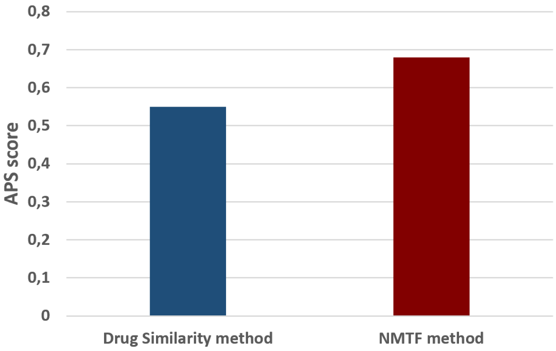

Fig. 3. Comparison of APS scores for NMTF method and drug similarity method [3]. The scores are computed considering 784 drugs, where chemical data were available.

We also compared NMTF with a predictor based on the drug similarity method presented in [3], as described in the previous Section. The APS scores for NMTF and the drug similarity method are reported in Figure 3. They show a better performance of the NMTF method compared to the [3] one. From the NMTF reconstruction of the partial R_{12} matrix (i.e., obtained removing the 10% of the category-drug links over 20 randomizations), we can retrieve the deleted links with higher precision and recall. Furthermore, our method can retrieve results also for drugs without a known molecular structure.

Finally, we decided to manually validate some predicted links that are not labeled as true links in the initial dataset. We considered predicted links with threshold greater than 0.3 in Table I and we validated them on the literature (II).

Firocoxib is categorized by DrugBank as a non-steroidal anti-inflammatory drug currently used in dogs and horses, however [26] has investigated its analgesic efficacy in mouse model with good results confirming our prediction. Mometasone is a synthetic corticosteroid with anti-inflammatory properties, but it is not labeled as analgesic in DrugBank. However, the Japan Standard Commodity Classification present in the KEGG Drug database [27] has annotated this drug as analgesic, corroborating our finding. Adenosine is a nucleoside that is composed of adenine and d-ribose; it has been classified as analgesic by DrugBank, but it also has anti-inflammatory properties found in [28]. Butorphanol, Levacetylmethadol, Levorphanol, Meperidine and Pholcodine are widely used opioid analgesics not having anti-inflammatory properties according to DrugBank, nevertheless their use as anti-inflammatory agents has been demonstrated especially in peripheral inflammatory pain [29], [30]. Lumiliximab is a chimeric monoclonal antibody that is used as an immunosuppressive drug; however, DrugBank database fails to annotate Lumiliximab as a immunosuppressor, whereas our method successfully labeled this drug. Ethanol is a liquid, rapidly absorbed from the gastrointestinal tract; it has bactericidal activity but DrugBank does not categorized its other implications. It has been reported its widely effects on the nervous system [31], confirming our categorization. Prasterone, also known as dehydroepiandrosterone (DHEA) is a steroid produced by the adrenal cortex. DrugBank omits its effects on the central nervous system [32], but our method suceeds to find them.

All these manually curated annotation predictions confirm the validity of NMTF approach for drug repositioning and gives space to more specific applications.

Experiments have been run on a MS-Windows machine equipped with an Intel i7-8750H processor and 16 GB of RAM. Every run of the NMTF method takes between 5 and 10 minutes depending on the number of iterations needed to reach the convergence, with a total use of 1.1 GB of RAM.

TABLE I
TOTAL NUMBER OF TRUE, PREDICTED AND TRUE PREDICTED CATEGORY-DRUG LINKS FOR A RANDOMIZATION OUT OF 20.

Threshold	True links to predict	Predicted Links	True Positive
0	188	1124	188
0.1	188	135	77
0.2	188	102	60
0.3	188	56	45
0.4	188	30	28
0.5	188	10	10
0.6	188	6	6
0.7	188	2	2
0.8	188	0	0
0.9	188	0	0
1	188	0	0

TABLE II
PREDICTED NEW LINKS FOR THRESHOLD GREATER THAN 0.3.

Drugs IDs	Drugs Names	Predicted Links	References
DB09217	Firocoxib	Analgesic	[26]
DB00764	Mometasone	Analgesic	[27]
DB00640	Adenosine	Anti-inflammatory	[28]
DB00611	Butorphanol	Anti-inflammatory	[27]
DB01227	Levacetylmethadol	Anti-inflammatory	[27]
DB00854	Levorphanol	Anti-inflammatory	[27]
DB00454	Meperidine	Anti-inflammatory	[27]
DB09209	Pholcodine	Anti-inflammatory	[27]
DB06162	Lumiliximab	Immunosuppressor	[19]
DB00898	Ethanol	Nervous System Agent	[31]
DB01708	Prasterone	Nervous System Agent	[32]

V. CONCLUSIONS

The demand for computational drug repositioning is increasing over the years, leading to low-priced methods for drug discovery compared to traditional methods. Moreover, drugs information is becoming more and more accessible through different sources. Thus, methods that can both predict and integrate heterogeneous information are the most likely to be used for this purpose.

In this study, we developed a network-based method for predicting potential new drug indications by exploring drug-target and target-pathways associations. Data from DrugBank, Reactome and Uniprot databases were obtained and merged in a large network. The entire framework contains drugs' categories, drugs, proteins and pathways as nodes. NMTF method learns from the existing relationships among nodes and reconstructs the category-drug network for predicting drug's annotations. The proposed NMTF quadripartite network model successfully finds novel uses for already approved drugs and it has better performance than the drug similarity approach. We also manually validated some new drug annotations based on the literature, demonstrating that our approach can complete missing information from DrugBank and it can predict new uses for approved drugs.

Future work will be focused on extending our approach to the drug-disease applications, i.e. adding disease-specific categories and drugs to the network. Drug discovering is particularly relevant in therapeutic areas, such as potential treatments for cancer, and our proposed method has showed very good preliminary results. Its applicability in cancer or other severe diseases may lead to interesting drug repurposing.

ACKNOWLEDGMENTS

This research is funded by the ERC Advanced Grant project 693174 GeCo (Data-Driven Genomic Computing), 2016- 2021. We thank the entire GeCo group and the Data Science group of Politecnico di Milano for useful discussions and insights. We would also like to show our gratitude to Andrea Beccari and Marica Gemei for attracting our attention to drug repurposing problems.

References

[1] T.L. Doan, M. Pollastri, M.A. Walters, G.I. Georg, “The Future of Drug Repositioning: Old Drugs, New Opportunities,” Annu Rep Med Chem, vol. 46, pp. 385–401, 2011.

[2] J.L. Medina-Franco, Y. Jakyung, A. Dueas-Gonzlez, “DNA Methyltransferase Inhibitors for Cancer Therapy,” Epigenetic Technological Applications, pp. 265–290, 2015.

[3] J. Li, Z. Lu, “A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity,” Proc. Int Conf Bioinformatics Biomed, pp. 1–4, 2014.

[4] M. Campillos, M. Kuhn, A.C. Gavin, L.J. Jensen, P. Bork, “Drug target identification using side-effect similarity,”, Science, vol. 321(5886), pp. 263–266, 2008.

[5] J. Li, X. Zhu, J.Y. Chen, “Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts,” PLoS Comput Biol, vol. 5(7), 2009.

[6] A.P. Chiang, A.J. Butte, “Systematic evaluation of drug-disease relationships to identify leads for novel drug uses,” Clin Pharmacol Ther, vol. 86(5), pp. 507–510, 2009.

[7] C. Ding, T. Li, W. Peng, H. Park, “Orthogonal nonnegative matrix trifactorizations for clustering”, Proc. ACM SIGKDD, pp. 126–135, 2006.

[8] V. Gligorijevic, N. Malod-Dognin, N. Przulj, “Patient-specific data fusion for cancer stratification and personalized treatment”, Pac. Symp. Biocomput., vol. 21, pp. 321–332, 2016.

[9] N. Del Buono, G. Pio, “Non-negative matrix tri-factorization for coclustering: an analysis of the block matrix”, Information Sciences, vol. 301, pp. 13-26, 2015.

[10] C. Ding, X He, H. Simon, “On the equivalence of nonnegative matrix factorization and spectral clustering”, Proc. SIAM Int. Conf. Data Min., pp. 1–5, 2005.

[11] R. Zass, A. Shashua, “A unifying approach to hard and probabilistic clustering”, Proc. ICCV, vol. 1, pp. 294–301, 2005.

[12] V. Gligorijevic and N. Przulj, “Methods for biological data integration: perspective and challenges”, J. R. Soc. Interface, vol. 12(112), pp. 1–20, 2015.

[13] S.A. Vavasis, “On the complexity of nonnegative matrix factorization”, Proc. SIAM J. Optim., vol. 20, pp. 1364–1377, 2009.

[14] F. Wang, T. Li, C. Zhang, “Semi-supervised clustering via matrix factorization”, Proc. SIAM Int. Conf. Data Min., pp. 1–12, 2008.

[15] D. Cai, X. He, J. Han, T.S. Huang, “Graph regularized non-negative matrix factorization for data representation”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, pp. 1548–1560, 2011.

[16] F. Shang, L. Jiao, F. Wang, “Graph dual regularization non-negative matrix factorization for co-clustering”, Pattern Recognit., vol. 45, pp. 2237–2250, 2012.

[17] M. Zitnik, V. Janji, C. Larminie, B. Zupan, N. Prulj, “Discovering disease-disease associations by fusing systems-level molecular data”, Sci. Rep., vol. 3(3202), pp. 1–9, 2013.

[18] H. Wang, H. Huang, C. Ding, F. Nie, “Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization”, J. Comput. Biol., vol. 20(4), pp. 344–358, 2013.

[19] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, et al., “DrugBank 5.0: a major update to the DrugBank database for 2018”, Nucleic Acids Res., vol. 46(D1), pp. D1074-D1082, 2018.

[20] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, et al., “The Reactome Pathway Knowledgebase”, Nucleic Acids Res., vol. 46(D1), pp. D649–D655, 2018.

[21] The UniProt Consortium, “UniProt: a hub for protein information”, Nucleic Acids Res., vol. 43(D1), pp. D204-D212, 2015.

[22] D. Bajusz, A. Rcz, K. Hberger, “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?”, J Cheminform., vol. 7(20), 2015.

[23] J.P. Brunet, P. Tamayo, T.R. Golub, J.P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization”, Proc. Natl. Acad. Sci. U.S.A., vol. 101(12), pp. 4164–4169, 2004.

[24] H. Kim, H. Park, “Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis”, Bioinformatics, vol. 23(12), pp. 1495–1502, 2007.

[25] S. Wanhua, Y. Yan, Z. Mu, “A Relationship Between the Average Precision and the Area Under the ROC Curve”, Proc. ACM ICTIR, pp. 349–352, 2015.

[26] B. Reddyjarugu, T. Pavek, T. Southard, J. Barr, B. Singh, “Analgesic Efficacy of Firocoxib, a Selective Inhibitor of Cyclooxygenase 2, in a Mouse Model of Incisional Pain,” J Am Assoc Lab Anim Sci, vol. 54(4), pp. 405–410, 2015.

[27] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs”, Nucleic Acids Res, vol. 45(D1), pp. D353-D361, 2017.

[28] G. Hask, B. Cronstein, “Regulation of inflammation by adenosine”, Front Immunol, 2013.

[29] J.S. Walker, “Anti-inflammatory effects of opioids, ” Adv Exp Med Biol, vol. 521, pp. 148–160, 2003.

[30] S. Katerina, J.J. Iwaszkiewicz, J. Schneider, S. Hua, “Targeting peripheral opioid receptors to promote analgesic and anti-inflammatory actions,” Front Pharmacol, 2013.

[31] H. Kalant, “Direct effects of ethanol on the nervous system,” Fed Proc, vol. 34(10), pp. 1930–1941, 1975.

[32] A.G. Gravanis, S.H. Mellon, “Hormones in Neurodegeneration, Neuroprotection, and Neurogenesis,” John Wiley & Sons, pp. 349, 2011.