Drug Repositioning Predictions by Non-Negative Matrix Tri-Factorization of Integrated Association Data

Gaëtan Dissez
gaetan.dissez@polytechnique.edu
École polytechnique
Palaiseau, France

Gaia Ceddia
gaia.ceddia@polimi.it
Politecnico di Milano
Milan, Italy

Pietro Pinoli
pietro.pinoli@polimi.it
Politecnico di Milano
Milan, Italy

Stefano Ceri
stefano.ceri@polimi.it
Politecnico di Milano
Milan, Italy

Marco Masseroli
marco.masseroli@polimi.it
Politecnico di Milano
Milan, Italy

ABSTRACT

Drugs repurposing (i.e., the reuse of existing drugs for new medical indications) is attracting the interest of pharmaceutical companies, as it speeds up the drug development process and substantially reduces the need for clinical trials. Thus, computational methods for drug repositioning are gaining increasing interest. In this work, we propose a drug repositioning algorithm based on the Non-Negative Matrix Tri-Factorization (NMTF) of integrated association data. We show how to construct a general-purpose graph encompassing the most relevant aspects in drug discovery and how to ensure fast convergence of the algorithm. In particular, we study how initialization and termination may significantly affect the outcome quality for the drug repurposing application. We also evaluate our computationally predicted repurposed drugs based on the literature and find confirmation for our prediction, proving that our method can successfully be applied to hypothesis generation for drug repurposing.

CCS CONCEPTS

• **Computing methodologies** → **Non-negative matrix factorization**; • **Applied computing** → **Biological networks**; *Systems biology*; *Health informatics.*

KEYWORDS

Drug repositioning; drug repurposing; data fusion; machine learning; biological networks; data integration

ACM Reference Format:
Gaëtan Dissez, Gaia Ceddia, Pietro Pinoli, Stefano Ceri, and Marco Masseroli. 2019. Drug Repositioning Predictions by Non-Negative Matrix Tri-Factorization of Integrated Association Data. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3307339.3342154

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342154

1 INTRODUCTION

Drug discovery, i.e., de novo identification of therapeutic drugs, has become an expensive task for pharmaceutical companies [8, 19]. The number of approved drugs compared to the investments done in this field has dramatically decreased over the years. Moreover, drug discovery takes about 9-12 years, meaning that a new drug may be placed in the market after several years and billions of dollars of investments [6]. This is due to the conservative protocols of drug discovery within the pharmaceutical companies, which concern the discovery of a new therapeutic target and compounds that can modulate its activity, followed by a long experimental and clinical validation process [8]. Drug repositioning or repurposing (i.e., finding new indications for known drugs) speeds up the drug development process and substantially reduces the need for clinical trials; this leads to substantial revenues for pharmaceutical companies and to important benefits for patients. Thus, computational methods for drug repositioning are gaining increasing interest.

In this work we propose a drug repositioning algorithm based on the Non-Negative Matrix Tri-Factorization (NMTF) of integrated association data. NMTF is a linear algebra algorithm, firstly proposed by Ding et al. [7], designed to factorize an input matrix in three matrices of non-negative elements. NMTF has been used for several biological applications, such as gene prioritization [33], finding patient-specific treatments [10], and disease association predictions [32]. When applied to an association matrix (i.e., a binary matrix representing associations between the components of two sets), the NMTF has been proven to be an effective method for simultaneously clustering the elements of the two sets and predicting missing associations. Furthermore, the enhancement of the original association matrix by associating the elements of the two sets with elements of external sets boosts the prediction capability of NMTF.

In order to take advantage of these NMTF features, first we created an extended graph including a set of approved drugs associated with their indications, expressed through a controlled vocabulary, and also associated both with the disease for which they are currently used and with target proteins, which in turn are associated with biological pathways. Then, we applied the NMTF to the matrix representing the drug-indication associations of the graph, so as to predict missing drug-indication links suggesting novel potential therapeutic drug usages.

The focus of this study is both on the construction of a general-purpose graph, encompassing the most relevant aspects in drug discovery, and also on how to ensure fast convergence of the algorithm. In particular, we assessed that the method is highly sensible to how factor matrices are initialized, and we designed and compared several initialization strategies. Moreover, we also understood that determining the termination conditions is not trivial due to the possibility that, by having too many iterations, reconstructed matrices may overfit their inputs. Thus, the major contribution of this work is to show how initialization and termination may significantly affect the quality of outcome when NMTF is used for solving a link prediction problem, in a way that can find other significant applications besides drug repurposing. With respect to our main objective, we biologically evaluated our results using the literature and found significant drug-indication associations, suggesting that the method could successfully be applied to hypothesis generation for drug repurposing.

2 RELATED WORK

Computational repositioning strategies have been categorized as *drug-based* or *disease-based*, where the former one indicates that a drug is re-purposed from the chemical or pharmaceutical perspective and the latter one means that the drug repositioning process starts from the symptomatology or pathology perspective [8]. Drug-based strategies may be useful in understanding the whole pharmacological spectrum of drugs, instead disease-based approaches are deployed when a disease-specific strategy is needed.

Drug-based computational methods include chemical similarity, molecular activity similarity and molecular docking approaches. The chemical similarity method hypothesizes that the structure and the chemical properties of a drug are associated with its effectiveness during a treatment [19]. Thus, drugs sharing similar molecular structure may have analogous therapeutic effects. Moreover, the increase of publicly available databases of chemical structures made the implementation of chemical similarity methods straightforward. Drug chemical similarity is mostly measured by 2D topological fingerprints or 3D conformations, and it is an area currently under development [19]. Both *Keiser et al.* [13] and *Li et al.* [17] integrated drug structure similarity and drug-target associations in their work. They both computed pairwise similarity scores using the Tanimoto coefficient between 2D chemical fingerprints. However, [13] uses a statistical model to infer significant drug-target associations followed by a repurposing assumption. Conversely, [17] uses a linear combination of chemical and target profile similarities to compute drug similarity scores for repositioning. Limitations of the chemical similarity approach for drug repositioning lie in the fact that many structures contain errors and that drugs go through several metabolic transformations, making their effects unpredictable using chemical properties alone [9].

The molecular activity (MOA) of a drug is the perturbation held by that active drug in a biological system and it can be measured using gene expression microarrays [8]. This leads to the construction of molecular activity profiles for each drug, which can be used to determine drug pairwise similarity. One of the most comprehensive source of molecular activities of drugs is the Connectivity Map project [26]; it currently contains information about 6,125 drugs and their MOA profiles. Limitations of this approach are mainly due to the quality of the technology and the biological models used for MOA measurement (i.e., in vitro models vs. in vivo metabolic transformations).

Molecular docking methods are based on the evaluation of new physical interactions between drugs and targets using the 3D molecular structure of both. If a drug is predicted to physically interact with a new target, known to be important in a particular disease, then that drug can be considered as a repositioning candidate for the disease. Molecular docking limitations include the lacking of 3D complete structures for proteins and the great amount of false-positive interactions found [8].

Disease-based approaches leverage on drug indications about diseases to repurpose the drugs. For example, *Chiang et al.* [2] used a 'guilt by association' approach in which diseases sharing a significant number of therapies are considered to be similar; consequently, the drugs used for a disease may be repositioned for other similar diseases. Another strategy is to search for disease similarity at the molecular level as done in [12]. *Campillos et al.* [1], instead, used a side effect similarity approach for drug repositioning. Limitations of disease-based approaches include the ability to measure disease molecular profiles and to collect comprehensive side-effect information [8].

With the increase of drug related data and the development of open source platforms, new approaches for drug repositioning integrating heterogeneous data types, such as chemical, clinical or MOA data, have been considered [19]. For example, *Napolitano et al.* [24] proposed a Support Vector Machine classification method to predict drug therapeutic class using chemical structure similarity, molecular target similarity and drug gene expression similarity. Additionally, with the aim of identifying new indications for existing drugs, *Li et al.* [18] developed a causal network (CauseNet) made of 5 layers: drugs, protein targets, pathways, genes and diseases, and integrating 4 different open databases. More recently, *Luo et al.* [21] developed a computational framework to predict novel drug-target interactions from drug-related heterogeneous information, assuming that novel drug-target associations lead to drug repositioning. These examples are more successful than methods that use drug-based and disease-based approaches individually, showing better perfomance in both sensitivity and specificity [8, 19].

3 METHODS

In this Section we first describe the considered datasets and how we combined them (3.1); then, we illustrate the Non-Negative Matrix Tri-Factorization method. About the latter one, after a brief introduction to the method (3.2), we illustrate how it is adapted to the drug repositioning problem, both in terms of model (3.3) and update rules for driving the method's iterations (3.4); we also discuss the inclusion of intra-data type relations (3.5) and an evaluation metric for validation (3.6). We then focus on the three most innovative aspects of our contribution from a computational perspective: how we initialize the matrices (3.7), how we perform the hyperparameter tuning (3.8) and how we define stopping criteria (3.9).

3.1 Datasets and their Use

For drug repositioning predictions, besides considering a drug-indication dataset, we enriched it with three additional datasets in order to provide as much information as possible as input to the NMTF prediction. In particular, we associated the drugs with the diseases for which they are commonly used as treatment, and we associated each drug with its targeted proteins, and each protein with the biological pathways it contributes to. As depicted in Figure 1, the elements of these five data types (indications, drugs, diseases, proteins and pathways) are connected by four bipartite graphs. We represent each bipartite graph as a binary association matrix. In particular, R_{12} associates therapeutical indications with drugs, R_{23} associates drugs with proteins, R_{25} associates drugs with diseases and R_{34} associates proteins with biological pathways. Moreover, for the proteins and pathways sets we also considered self intra-connections; in the former case the links represent protein-protein interactions, in the latter case they represent the hierarchical structure of pathway ontology.

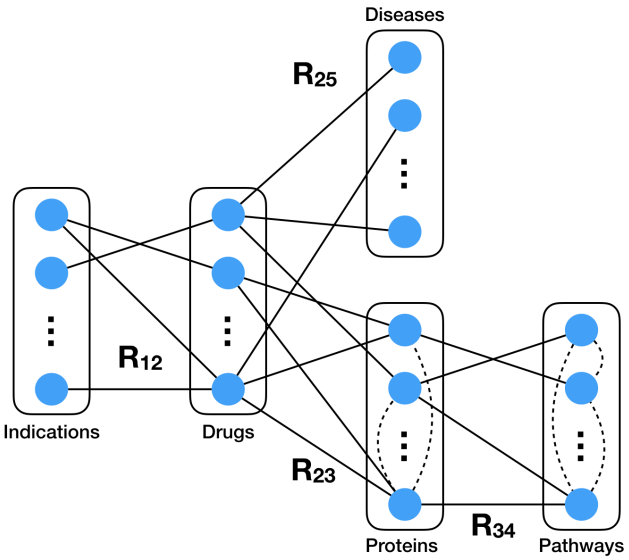

Figure 1: Network illustration of integrated association data. R_{12}, R_{23}, R_{34} and R_{25} are binary association matrices between indications and drugs, drugs and proteins, proteins and pathways, and drugs and diseases, respectively.

For constructing the network presented in Figure 1, we extracted approved drugs, their targets and current indications from DrugBank [30], pathways of the involved targets from Reactome [4] and drug-treated diseases from Therapeutic Target Database (TTD) [20]. We added protein-protein interanctions from UniProt [3] and pathway hierarchical relationships from Reactome. The used datasets include 141 indications, 3261 drugs, 3691 proteins, 1914 pathways and 841 diseases; R_{12}, R_{23}, R_{34} and R_{25} contain 23517, 13432, 28345 and 2417 links, respectively.

3.2 Non-Negative Matrix Tri-Factorization

Consider a matrix $R \in \mathbb{R}^{(N \times M)}$ and let us define $\mathbb{R}_+ = \{x : x \in \mathbb{R} \wedge x \geq 0\}$ the set of non-negative real numbers. The NMTF method factorizes R in three components $G_1 \in \mathbb{R}_+^{(N \times k_1)}$, $S_{12} \in \mathbb{R}_+^{(k_1 \times k_2)}$ and $G_2 \in \mathbb{R}_+^{(M \times k_2)}$, such that:

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

and the three matrices G_1, S_{12} and G_2 minimize the cost function J, in this case simply defined as the Frobenius norm:

$$J(G_1, S_{12}, G_2) = \|R - G_1 S_{12} G_2^T\|^2.$$

The additional constraint on the orthogonality of both G_1 and G_2 (i.e., $G_1 G_1^T = I$ and $G_2 G_2^T = I$) guarantees the uniqueness of the solution.

Although the NMTF is clearly applicable to any matrix, for the sake of simplicity in this paper we consider the input matrix to be the association matrix of a bipartite graph between elements of two distinct sets. This being stated, the NMTF method is easily extensible to more complex graph topology. For instance, consider two association matrices R_{AB} and R_{BC}, where the former one connects elements of the set A to elements of the set B and the latter one connects elements of the set B to elements of a third set C. One can use the NMTF to compute a set of non-negative matrices G_A, G_B, G_C, S_{AB} and S_{BC} that minimizes

$$\|R_{AB} - G_A S_{AB} G_B^T\|^2 + \|R_{BC} - G_B S_{BC} G_C^T\|^2$$

In such cases, the factorization of each matrix influences the factorization of the other one. This feature, combined with the fact that we only factorize association matrices regardless of the actual type of the elements within the connected sets, makes the NMTF a valuable tool for data integration in the context of many data analysis tasks, such as clustering, co-clustering, link prediction and anomaly detection.

Typically, the NMTF problem is solved by randomly initializing the elements of the matrices and then iterating a set of matrix updating rules until convergence. Hereafter, we describe how we make use of the NMTF output to predict novel drug indications.

3.3 Using NMTF for Drug Repositioning

In the network scenario described in Figure 1, predicting novel potential indications for the set of considered drugs means to infer new links between the indication and the drug sets (i.e., within the R_{12} matrix). In order to do so, we apply the NMTF simultaneously on the four input matrices (R_{12}, R_{23}, R_{34}, R_{25}); thus, we factorize them all in such a way that the final result is influenced by the information in all the four matrices. More formally, we adopt the NMTF method in order to jointly factorize all the four association matrices of the graph in Figure 1.

We first select a set of parameters $k_1, k_2, k_3, k_4, k_5 \in \mathbb{N}$ that allows us to compute the matrices G_1, G_2, G_3, G_4, G_5, S_{12}, S_{23}, S_{34} and S_{25}, which respect non-negativeness and orthogonality constraints and minimize the sum of the Frobenius norm distances. In particular, the matrix R_{12} is factorized in the three matrices G_1, S_{12} and G_2. We multiply the three factor matrices to build a non-negative matrix $\hat{R}_{12}$ that approximates R_{12}:

$$\hat{R}_{12} = G_1 S_{12} G_2 \approx R_{12}$$

Then, we *binarize* the $\hat{R}_{12}$ by setting a threshold $\delta \in [0,1]$ and producing a third matrix $\hat{R}_{12}^{(\delta)}$ such that $\hat{R}_{12}^{(\delta)}[i,j] = 1$ if $\hat{R}_{12} > (\delta)$, and $\hat{R}_{12}^{(\delta)}[i,j] = 0$ otherwise. Thus, given a drug d and an indication i, four disjoint events may occur:

- $R_{12}[i,d] = \hat{R}_{12}^{(\delta)}[i,d] = 0$: In this case, d is not associated with i, either in the input graph or in the predicted one;
- $R_{12}[i,d] = \hat{R}_{12}^{(\delta)}[i,d] = 1$: In this case, the drug d is associated with the indication i in both graphs, in other words the prediction method confirms the association;
- $R_{12}[i,d] = 1$ and $\hat{R}_{12}^{(\delta)}[i,d] = 0$: In this case, in the input matrix an association is present between d and i that the method was not able to recall; this can be interpreted as an hint to revise such association, which may also be due to an incorrect annotation in the analyzed data;
- $R_{12}[i,d] = 0$ and $\hat{R}_{12}^{(\delta)}[i,d] = 1$: This is the most interesting case, as in the reconstructed association matrix there is a new d-i link that was not present in the input; it predicts that the drug d may be used for the indication i.

3.4 Update Rules

Considering all inter-data type links in Figure 1 (i.e., between indications and drugs, diseases and drugs, drugs and proteins, proteins and pathways), the NMTF objective function is defined as:

$$J(G_1,G_2,G_3,G_4,G_5,S_{12},S_{23},S_{34},S_{25}) = \|R_{12} - G_1 S_{12} G_2^T\|^2$$
$$+\|R_{23} - G_2 S_{23} G_3^T\|^2 + \|R_{34} - G_3 S_{34} G_4^T\|^2 + \|R_{25} - G_2 S_{25} G_5^T\|^2$$

where $\|\|$ is the Frobenius norm. As explained in Section 3.2, this objective function needs to be minimized under the constraints:

$$G_1 \geq 0, G_2 \geq 0, G_3 \geq 0, G_4 \geq 0, G_5 \geq 0$$
$$G_1^T G_1 = \mathbf{I}, G_2^T G_2 = \mathbf{I}, G_3^T G_3 = \mathbf{I}, G_4^T G_4 = \mathbf{I}, G_5^T G_5 = \mathbf{I}$$
$$S_{12} \geq 0, S_{23} \geq 0, S_{34} \geq 0, S_{25} \geq 0$$

which means that G_1, G_2, G_3, G_4 and G_5 must be non-negative and orthogonal matrices, and $S_{12}, S_{23}, S_{34}, S_{25}$ must be non-negative matrices. Let $\mathbf{G} = diag(G_1, G_2, G_3, G_4, G_5)$, $\mathbf{S} = diag(S_{12}, S_{23}, S_{34}, S_{25})$; thus, the former constraints can be summarized as: $\mathbf{G} \geq 0, \mathbf{G}^T\mathbf{G} = \mathbf{I}, \mathbf{S} \geq 0$. Then, considering (λ, μ_G, μ_S) the Lagrangian multiplier matrices, the Lagrangian $\mathcal{L}$ associated with the problem is:

$$\mathcal{L}(\mathbf{G},\mathbf{S},\lambda,\mu_G,\mu_S) = J(\mathbf{G},\mathbf{S}) + tr(\lambda(\mathbf{G}^T\mathbf{G} - \mathbf{I})) - tr(\mu_\mathbf{G}\mathbf{G}^T) - tr(\mu_\mathbf{S}\mathbf{S}^T)$$

with *tr* the trace function. From the Karush-Kuhn-Tucker (KKT) theorem applied to this Lagrangian it is possible to verify that the solution of our problem follows the so-called *KKT conditions*:

- *Stationarity*: $\nabla_\mathbf{G}\mathcal{L} = 0$ and $\nabla_\mathbf{S}\mathcal{L} = 0$
- *Primal admissibility*: $\mathbf{G} \geq 0$, $\mathbf{S} \geq 0$ and $\mathbf{G}^T\mathbf{G} = \mathbf{I}$
- *Dual admissibility*: $\mu_G \geq 0$ and $\mu_S \geq 0$
- *Complementary slackness*: $\mu_{G_{ij}}\mathbf{G}ij = 0, \mu_{S_{ij}}\mathbf{S}ij = 0, \forall(i,j)$

As explained and proved in [7] and [11], it is possible to find update rules for $(G_1, G_2, G_3, G_4, G_5)$ and $(S_{12}, S_{23}, S_{34}, S_{25})$ that make them converging toward a solution that verifies those KKT conditions. For our drug repositioning task, as such update rules we used the following ones:

$$G_{1_{(i,j)}} \leftarrow G_{1_{(i,j)}} \sqrt{\frac{(R_{12}G_2S_{12}^T)_{i,j}}{(G_{11}R_{12}G_2S_{12}^T)_{i,j}}}$$

$$G_{2_{(i,j)}} \leftarrow G_{2_{(i,j)}} \sqrt{\frac{(R_{12}^TG_1S_{12} + R_{23}G_3S_{23}^T + R_{25}G_5S_{25}^T)_{i,j}}{(G_{22}R_{12}^TG_1S_{12} + G_{22}R_{23}G_3S_{23}^T + G_{22}R_{25}G_5S_{25}^T)_{i,j}}}$$

$$G_{3_{(i,j)}} \leftarrow G_{3_{(i,j)}} \sqrt{\frac{(R_{23}^TG_2S_{23} + R_{34}G_4S_{34}^T)_{i,j}}{(G_{33}R_{23}^TG_3S_{23} + G_{33}R_{34}G_4S_{34}^T)_{i,j}}}$$

$$G_{4_{(i,j)}} \leftarrow G_{4_{(i,j)}} \sqrt{\frac{(R_{34}^TG_3S_{34})_{i,j}}{(G_{44}R_{34}^TG_4S_{34})_{i,j}}}$$

$$G_{5_{(i,j)}} \leftarrow G_{5_{(i,j)}} \sqrt{\frac{(R_{25}^TG_2S_{25})_{i,j}}{(G_{55}R_{25}^TG_2S_{25})_{i,j}}}$$

$$S_{12_{(i,j)}} \leftarrow S_{12_{(i,j)}} \sqrt{\frac{(G_1^TR_{12}G_2)_{i,j}}{(G_1^TG_1S_{12}G_2^TG_2)_{i,j}}}$$

$$S_{23_{(i,j)}} \leftarrow S_{23_{(i,j)}} \sqrt{\frac{(G_2^TR_{23}G_3)_{i,j}}{(G_2^TG_2S_{23}G_3^TG_3)_{i,j}}}$$

$$S_{34_{(i,j)}} \leftarrow S_{34_{(i,j)}} \sqrt{\frac{(G_3^TR_{34}G_4)_{i,j}}{(G_3^TG_3S_{34}G_4^TG_4)_{i,j}}}$$

$$S_{25_{(i,j)}} \leftarrow S_{25_{(i,j)}} \sqrt{\frac{(G_2^TR_{25}G_5)_{i,j}}{(G_2^TG_2S_{25}G_5^TG_5)_{i,j}}}$$

where $G_{11} = G_1G_1^T$, $G_{22} = G_2G_2^T$, $G_{33} = G_3G_3^T$, $G_{44} = G_4G_4^T$ and $G_{55} = G_5G_5^T$ within the update rules. It is necessary to note that using multiplicative rules, if an element of any matrix equals zero, it will always remain equal to zero over iterations. Thus, the choice of the initialization is crucial.

3.5 Intra-Data Type Relations

We also considered intra-data type relations, specifically between proteins and between pathways, which can be described through two adjacency matrices W_3 and W_4, respectively. Given these two matrices, it is easy to compute the Laplacian matrices L_3 and L_4 such that:

$$L_3 = D_3 - W_3$$
$$L_4 = D_4 - W_4$$

where D_3 and D_4 are the degree matrices associated with W_3 and W_4, respectively. The *degree matrix* associated with a graph is a diagonal matrix in which the *i-th* coefficient corresponds to the degree of the node i in the graph. In other words, the *i-th* coefficient of D is the sum of the *i-th* row elements of the associated W matrix. In order consider the intra-data type relations between proteins (and between pathways), our objective function needs to take into account that proteins (and pathways) that are linked together are more likely to share the same behavior with regards to drugs (and to proteins, respectively). Thus, we added two new terms in the

objective function:

$$J(\mathbf{G},\mathbf{S}) = \|R_{12} - G_1 S_{12} G_2^T\|^2 + \|R_{23} - G_2 S_{23} G_3^T\|^2 + \|R_{34} - G_3 S_{34} G_4^T\|^2 + \|R_{25} - G_2 S_{25} G_5^T\|^2 + tr(G_3^T L_3 G_3) + tr(G_4^T L_4 G_4)$$

with tr the trace function. From this refined objective function, update rules were extended as explained in Section 3.4.

3.6 Evaluation Metric

We anticipate here a discussion of the evaluation metric we use to computationally validate our method, as it is necessary for explaining and justifying how we selected the initialization strategy, hyperparameters and stop criterion. Specifically, to validate our model, we build a mask M of the same dimension of R_{12}; such mask is a binary matrix where only 10% of the elements are ones and their position is selected randomly. We create a matrix R'_{12} such that:

$$R'_{12}[i,j] = \begin{cases} R_{12}[i,j], & \text{if } M[i,j] = 0 \\ 0, & \text{otherwise} \end{cases}$$

Then, in the global graph we substitute R_{12} with the masked matrix R'_{12}, apply the method and construct the matrix $\hat{R'}_{12}$. Finally, to assess the goodness of the method and the configuration used, we evaluate how well $\hat{R'}_{12}$ resembles R_{12}, by focusing only on those elements that were masked (i.e., pairs of a drug d and an indication i such that $M[d,i] = 1$). In order to do that, we set a threshold $0 \leq \delta \leq 1$ and create the binary matrix $\hat{R'}^{(\delta)}_{12}$. For the elements of such matrix that overlap the elements equal to one in the mask, we can distinguish the classical four classes:

- (TP) true positives, equal to one in both R_{12} and $\hat{R'}^{(\delta)}_{12}$;
- (FP) false positives, equal to one only in $\hat{R'}^{(\delta)}_{12}$;
- (TN) true negatives, equal to zero in both R_{12} and $\hat{R'}^{(\delta)}_{12}$;
- (FN) false negatives, equal to one only in R_{12}.

Depending on the chosen value of the threshold δ, we can compute the *recall*:

$$Recall_\delta = \frac{TP}{TP + FN}$$

as well as the *precision*:

$$Precision_\delta = \frac{TP}{TP + FP}.$$

From these two measures, letting the threshold δ span between 0 and 1, we can draw precision-recall (PR) curves, which capture the dependency between the two metrics and provide an intuitive visual representation of it. Comparing such curves is not trivial; thus, we consider the *Average Precision Score* (APS), which summarizes in a single number the information contained in a PR curve. It can be defined as:

$$APS = \sum_{\delta_1, \delta_2, \ldots, \delta_n} (Recall_{\delta_i} - Recall_{\delta_{i-1}}) Precision_{\delta_i}$$

and can be interpreted as an approximation of the area under the PR curve. Higher values of APS correspond to better performance of the model.

3.7 Initialization

The first step of the NMTF algorithm is the initialization of the factor matrices. Although [7] highlighted that the NMTF algorithm converges to a minimum for the loss function, it is fundamental to understand that this minimum does not necessarily correspond to the maximum Average Precision Score (APS) that can be obtained. For instance, in an hypothetical scenario where the NMTF algorithm perfectly fits the initial matrices, it would be impossible to find new links between indications and drugs, resulting in a method not useful for the task of drug repositioning. Therefore, the initialization is crucial and the performance of our prediction method deeply relies on it.

In this Section we describe several alternatives for the initialization task, starting from the most trivial ones to the more sophisticated ones. In general, we are required to initialize all nine factor matrices G_1, G_2, G_3, G_4, G_5, S_{12}, S_{23}, S_{34} and S_{25}. The most trivial initialization consists of drawing random values for the elements of the factor matrices $G_1, G_2, G_3, G_4, G_5, S_{12}, S_{23}, S_{34}$ and S_{25} from a uniform distribution in the range $[0, 1]$. This task can be simplified by initializing solely the five G_1, G_2, G_3, G_4 and G_5 matrices, and then computing the values of the S_{12}, S_{23}, S_{34} and S_{25} matrices as follows:

$$S_{12} = G_1^T R_{12} G_2$$
$$S_{23} = G_2^T R_{23} G_3$$
$$S_{34} = G_3^T R_{34} G_4$$
$$S_{25} = G_2^T R_{25} G_5$$

However, using this initialization, the final results vary a lot, as shown in Figure 2 where we present the results obtained on 10 different runs of the algorithm for different initialization methods.

Also another method, named *random ACOL*, firstly introduced by [16], uses some randomization to initialize the matrices, but besides it takes into account the original data held in R_{12}, R_{23}, R_{34} and R_{25}. For instance, to initialize G_1, the columns of R_{12} are partitioned into k_1 random groups and each column of G_1 is initialized by averaging the columns in one of such groups. Despite being computationally inexpensive, the random ACOL initialization provides better results than the uniform distribution. However, the initialization can be improved to reach a better Average Precision Score.

Another initialization method, suggested by [7], [28] and [31], performs a *k-means clustering* on the columns (or rows) of the association matrices and initializes G_1, G_2, G_3, G_4 and G_5 with the centroids of the clusters.

A last and more promising algorithm, namely the *spherical k-means initialization* [29], consists of an improvement of the previous one. In this case, before the clustering is performed, all elements are projected on a unity radius multidimensional sphere and therefore considered as a high-dimensional unit-length vector. The k-means clustering is then performed, but considering the cosine rather than the euclidean distance as measure of similarity between any pair of elements. As a result, the centroid of each cluster is a unit-length vector that maximizes the cosine similarity with respect to the members of the cluster. Each of these centroids is then used to initialize the G_x matrices. This method, which only takes into account angles and not distances, is relevant for our problem. Indeed,

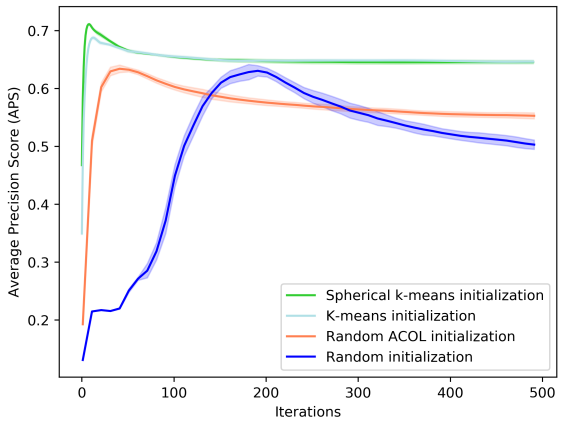

Figure 2: Average Precision Score (APS) over iterations for four different initialization methods. Each initialization is repeated 10 times; so, each curve shows the mean and the standard deviation of the corresponding method. The spherical k-means initialization has the smallest standard deviation and the highest Average Precision Score, which is reached in less iterations.

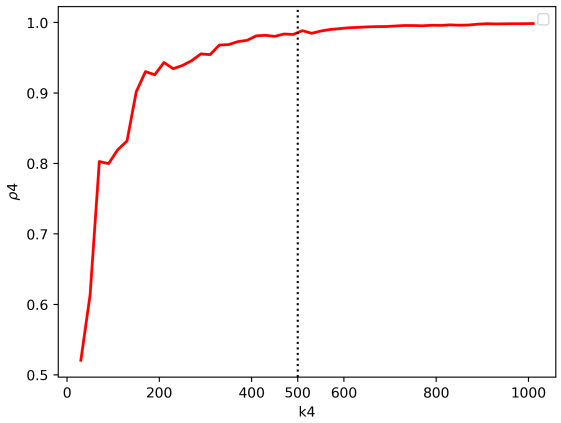

Figure 3: Evolution of the dispersion coefficient ρ_4 for the matrix $\hat{R}_{34}$, computed by our NMTF algorithm for different k_4 values. The other hyperparameters are fixed as $(k_1, k_2, k_3, k_5) = (500, 141, 500, 300)$.

for example, it gives the same weight to all drugs, no matter the number of indications they are associated with.

In Figure 2 we compare the four types of initialization that we tested. As it can be seen, the trivial *random initialization* scores worst than all the others, since it takes much more iterations to get to the peak of APS, and also has more variability in the results. The *random ACOL* is as good as the *random initialization* in term of APS, but it reaches the maximum score in a fourth of the iterations. Finally, the two initializations based on *k-means* score better APSs in very few iterations, with the *spherical k-means* being slightly better. Moreover, the NMTF initialized with any of the last two methods shows to be robust with respect to the number of iterations. Thus, we used the spherical k-means initialization for our tests.

3.8 Hyperparameter Tuning

For our application, we need to select a set of hyperparameters k_1, k_2, k_3, k_4 and $k_5 \in \mathbb{N}$. We select the best value for every hyperparameter based on the associated *dispersion coefficient* ρ [14]. The dispersion coefficients ρ_1, ρ_2, ρ_3, ρ_4 and ρ_5 for k_1, k_2, k_3, k_4 and k_5 are respectively computed on the matrix $\hat{R}_{12}, \hat{R}_{12}, \hat{R}_{23}, \hat{R}_{34}$ or $\hat{R}_{25}$; such coefficients range between 0 and 1, with higher values indicating more stable solutions.

The relationship between the dispersion coefficient and the associated hyperparameter is similar in all situations; after a strong increase for small values of the hyperparameter, the dispersion coefficient seems to converge slowly towards a value close to 1 (Figure 3). This behavior means that the higher is the hyperparameter, the more stable are the results given by the algorithm. Therefore, we fix each hyperparameter to the value at which the associated dispersion coefficient stabilizes, as a satisfying compromise between robustness of our method and computational costs. As an example, Figure 3 shows the dispersion coefficient ρ_4 for the hyperparameter k_4. This dispersion coefficient stabilizes when k_4 is higher than 500, which leads us to choose this value for the hyperparameter k_4.

Notice that, using the spherical k-means initialization, the hyperparameter values correspond to the number of clusters in which the data are partitioned. Therefore, these hyperparameters cannot have values higher than the number of items in the associated dataset. For instance, with our used dataset, k_2 have to be smaller or equal than 141, the number of indications in the dataset. According to the dispersion coefficients of all obtained results, we chose k_1, k_2, k_3, k_4 and k_5 equal to 500, 141, 500, 500 and 300, respectively. For such hyperparameter values, $\rho_1 = 0.990$, $\rho_2 = 0.990$, $\rho_3 = 0.981$, $\rho_4 = 0.983$ and $\rho_5 = 0.999$.

3.9 Stop Criterion

Update rules guarantee that the objective function J, also named *loss function*, is monotonically decreasing over iterations [7]. Thus, the NMTF algorithm keeps improving the approximation of the given matrices over the iterations. However, the capability of our model to predict missing links between drugs and indications could decrease as the differences between the initial drug-indication matrix and the reconstructed one reduce, since we are *overfitting* to the input. Therefore, finding a proper stop criterion for the training phase of our model is paramount and not trivial.

Several methods are worth considering to solve this problem. The easiest would be to define a maximum number of iterations. However, this technique is not appealing as it does not give any semantic to the stop criterion.

Instead, we implemented a stop criterion based on the evolution of the loss function. Since this function is strictly decreasing under the update rules we use, we stop the iterations of the NMTF training when the change of the function value between two consecutive

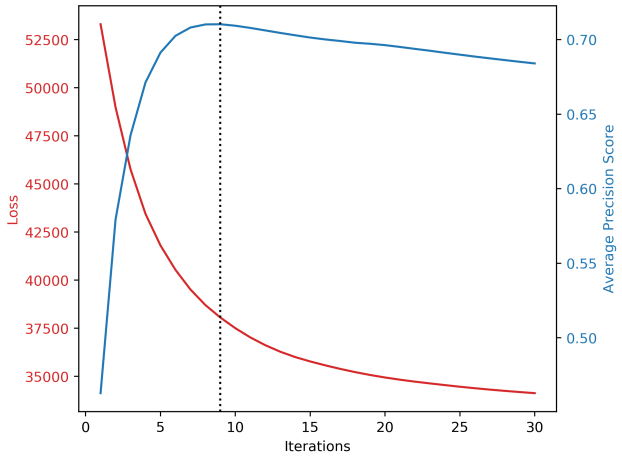

Figure 4: Application of the stop criterion on the loss (objective function) for $\epsilon = 0.02$. In this situation, with $(k_1, k_2, k_3, k_4, k_5) = (500, 141, 500, 500, 300)$ the algorithm would stop after 9 iterations, as represented by the black dashed vertical line, which also corresponds to the maximum of the Average Precision Score (APS).

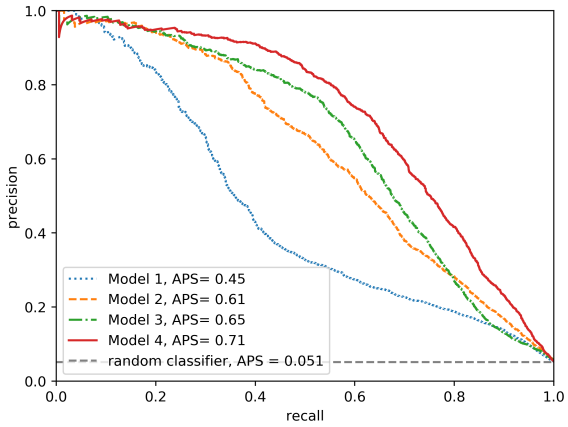

Figure 5: Precision-Recall curves of four progressively improved versions (models 1 to 4) of the NMTF method.

iterations is smaller than a given threshold ϵ, i.e,:

$$\frac{J(\mathbf{G}^{(n)}) - J(\mathbf{G}^{(n+1)})}{J(\mathbf{G}^{(n+1)})} < \epsilon.$$

Experimentally, this method proved particularly relevant: Indeed, for a proper ϵ, the algorithm stops when the Average Precision Score is maximum on the validation set. It means that this stop criterion is able to stop the algorithm when the model performs best at finding the missing links. Moreover, this characteristic of ϵ proves its effectiveness independently from the parameters' choice. Figure 4 shows that in our case this stop criterion permits to reach the higher APS score after 9 iterations (with $\epsilon = 0.02$).

4 RESULTS

In this Section we highlight the most relevant considerations derived from the results that we obtained in the experiments performed on the considered data with our method and in comparison with a state of the art method. We also report about the results of the scientific evaluation performed on the new drug-indication and drug-disease associations predicted by our method.

4.1 Benefits of Adding more Data Types

Adding data types and intra-data type relationships (protein-protein similarities and pathway hierarchical relationships) improves the prediction performance. Table 1 shows that the APS scored by our NMTF method slowly, but steadily, increases as data types and intra-data type associations are added.

4.2 Benefits of Optimizing the Model

Taking into account all available data, we optimized the initial model in order to get the best APS value. In Figure 5, we report the Precision-Recall curve of four different models increasingly optimized; the Figure illustrates the relevant contribution of our work to optimizing the NMTF method for drug repositioning. The four models present successive improvements to the method:

- **Model 1** is our baseline; it uses random initialization and uniformed hyperparameters, and stops after 500 iterations;
- **Model 2** adds spherical k-means initialization to Model 1, as discussed in Section 3.7;
- **Model 3** adds tuned hyperparameters to Model 2, as discussed in Section 3.8;
- **Model 4** adds the new stop criterion to Model 3, as discussed in Section 3.9 (while Models 1 to 3 use 500 iterations).

4.3 Comparative Study

We compared our results with the state of the art method proposed in [21]. The computational pipeline developed in [21], called DTINet, integrates multiple information about drugs from different sources to construct a heterogeneous network. DTINet uses drug-target, drug chemical similarity, side effect and drug-disease data from DrugBank, UniProt and CTD [5] databases. It first extracts significant features for each drug and protein by means of a compact feature learning algorithm. Then, it projects feature vectors of drugs onto the feature vector space of proteins, selecting as best projection the one corresponding to when drugs are geometrically close to the feature vectors of their known interacting proteins [21].

Table 1: Data integration benefits on the APS

Matrices	APS	Improvement
R_{12}	0.698	——
R_{12}, R_{23}	0.707	1.23%
R_{12}, R_{23}, R_{34}	0.709	1.57%
$R_{12}, R_{23}, R_{34}, W_3, W_4$	0.711	1.86%
$R_{12}, R_{23}, R_{34}, W_3, W_4, R_{25}$	0.714	2.30%

Finally, it extracts new significant drug-target interactions according to the geometric distance of the feature vectors: specifically, if the feature vector of a drug is close to the feature vector of a protein in the projected space, then that protein is a new candidate target for the drug.

We used predicted drug-target interactions from DTINet to infer new indications for drugs by their target profile similarity, i.e., if two drugs have similar predicted targets, they can share their uses (or indications). For the comparative study, we limited the analysis to the 546 drugs with available drug-related information used by DTINet, and we computed the APS values for the DTINet and NMTF methods on this set of drugs (Figure 6); the NMTF method has a greater score compared to DTINet, meaning that NMTF can retrieve missing indications of drugs with higher precision and recall. Furthermore, our NMTF method can compute drug-indication associations even when data about the drug chemical structure is not available.

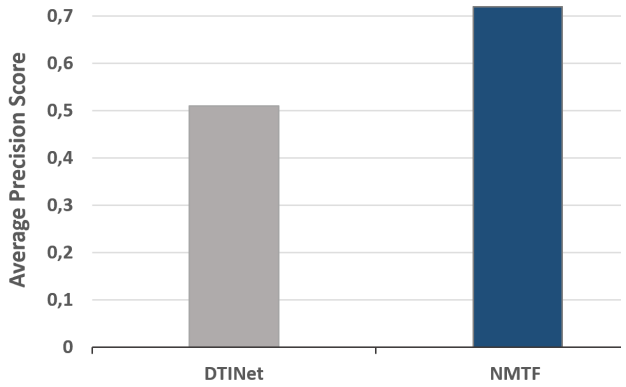

Figure 6: Comparison of Average Precision Scores for DTINet [21] and NMTF methods: 0.51 vs. 0.72, respectively. The scores are computed considering the 546 drugs evaluated in [21].

4.4 NMTF Predictions

We further evaluated the effectiveness of our new drug-indication and drug-disease predictions by searching the predicted pairs against the scientific literature and databases available.

Enoxacin and Pefloxacin are antibacterial agents that in the TTD database were not assigned to the treatment of bacterial infections, but our method associated with this disease. In other words, our method provides information that was absent in the TTD database used as input data source, despite being confirmed by the literature. Moreover, Enoxacin and Pefloxacin have been linked to the *moderate risk QTc-Prolonging agents* indication by our NMTF method; conversely, DrugBank reported these drugs only as *potential QTc-Prolonging agents*. Prolongation of the QT interval, i.e., the time between the first contraction of cardiac ventricles and the relaxation of the ventricles, may be useful for short QT syndrome, that is a genetic disease causing abnormal heart rhythms and sudden cardiac death [23]. Thus, Enoxacin and Pefloxacin could be repositioned for this syndrome.

Similarly, our NMTF method linked Aripiprazole lauroxil to schizophrenia; this drug is well-known for its use in the treatment of schizophrenia, but this information was not in the input TTD database. From a pharmacological perspective, Aripiprazole is the only approved antipsychotic that reduces dopaminergic neurotransmission through partial agonism, not antagonism [22]. Thus, it can be considered as a *neurotransmitter agent*, and the NMTF method correctly classified it. In other words, the NMTF method associates Aripiprazole lauroxil to schizophrenia and to the *neurotransmitter agent* indication.

Likewise, we discovered a link between Azidocillin and bacterial infection disease; this is another annotation that was not present in TTD, although Azidocillin is a penicillin antibiotic with antibacterial properties. Moreover, it has been indicated as an *anti-infective agent* according to the NMTF method, meaning that its anti-infection characteristics are particularly relevant.

One of the highest scored drug-indication associations is the one between Acamprosate and *central nervous system depressants* category. Acamprosate is a drug used for treating alcohol dependence, whose side effects include depression according to the SIDER database [15]. Other high ranked drug-disease associations predicted (e.g., Tiapride - Schizophrenia, Perospirone - Schizophrenia, Cilazapril - Hypertension, Olmesartan - Hypertension, Ceftibuten - Bacterial infections, Mezlocillin - Bacterial infections, Nafcillin - Bacterial infections, Candesartan cilexetil - Hypertension, Cefapirin - Bacterial infections, Cefotiam - Bacterial infections) can be found in the literature [30], but they are not provided by the TTD database.

Furthermore, our method was able to find interesting novel drug-disease associations; for example, Isoflurane was suggested to be repositioned for the treatment of human epilepsy. This finding is corroborated by [27], where an initial study in done on rat model. Another novel drug-disease association is the one between Vofopitant (currently under investigation) and nausea; this is also considered in [25] during the development of Phase II of this drug.

All these manually curated studies corroborate the use of our NMTF approach for drug repositioning, allowing correct identification of both drug indications and drug mappings to diseases.

4.5 Implementation Details

The entire framework for the analyses and the predictions presented in this manuscript has been developed in Python. The most computationally intensive tasks of the software have been parallelized to speed up the execution. Using 10 threads, the full prediction pipeline takes 3 minutes and 10 seconds on a Dell PowerEdge R730xd workstation equipped with two Intel Xeon E-2660 CPUs and requires less than 1.5 GB of RAM. The source code is available at *https://github.com/DEIB-GECO/NMTF-DrugRepositioning* under Apache2 license.

5 CONCLUSIONS

Drug repositioning has become an important task to reduce the costs and timing of drug discovery. Thus, the number of computational methods addressing this task, mainly taking advantage of public heterogeneous databases increasingly available, has increased over the years. In this study, we discussed the applicability, adaptation and extension of the NMTF to drug repositioning,

both in terms of model and computational aspects influencing the method performance. Starting from indication-drug, drug-target, drug-disease, and target-pathway associations, and leveraging the NMTF method, we can reliably predict novel potential drug indications or drug treatment usages.

We showed that the NMTF-based method we propose can effectively support multiple heterogeneous inputs, and we provided several optimizations, which improve the quality of the results w.r.t. standard method application. Such optimizations are innovative aspects of our contribution to the computational method, ranging from initialization and hyperparameter tuning, to update rules and stopping criteria.

Our NMTF model effectively finds novel indications and usages for drugs, and performs comparatively better than the DTINet state of the art approach in terms of Average Precision Score. We also validated some predicted new drug-indication and drug-disease associations based on the literature, demonstrating that our approach can correctly complete missing information in DrugBank and TTD databases.

ACKNOWLEDGMENTS

This research is funded by the ERC Advanced Grant project 693174 “GeCo” (Data-Driven Genomic Computing), 2016-2021.

REFERENCES

[1] M Campillos, M Kuhn, and AC Gavin. 2008. Drug target identification using side-effect similarity. *Science* 321 (2008), 263–266.
[2] AP Chiang and Atul J Butte. 2009. Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics* 86, 5 (2009), 507–510.
[3] UniProt Consortium. 2016. UniProt: the universal protein knowledgebase. *Nucleic acids research* 45, D1 (2016), D158–D169.
[4] D Croft, G OâĂŹkelly, G Wu, R Haw, M Gillespie, L Matthews, M Caudy, P Garapati, G Gopinath, B Jassal, et al. 2010. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* 39, Suppl1 (2010), D691–D697.
[5] AP Davis, CJ Grondin, RJ Johnson, D Sciaky, R McMorran, J Wiegers, et al. 2018. The comparative toxicogenomics database: Update 2019. *Nucleic acids research* 47, D1 (2018), D948–D954.
[6] M Dickson and JP Gagnon. 2004. The cost of new drug discovery and development. *Discov Med.* 4, 22 (2004), 172–179.
[7] C Ding, T Li, W Peng, and H Park. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proc SIGKDD*. ACM, NY, USA, 126–135.
[8] JT Dudley, T Deshpande, and AJ Butte. 2011. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 12, 4 (2011), 303–311.
[9] D Fourches, E Muratov, and A Tropsha. 2010. Trust but verify: on the importance of chemical structure curation in cheminformatics ans QSAR modeling research. *J Chem Inf Model* 50 (2010), 1189–1204.
[10] V GLIGORIJEVIĆ, N Malod-Dognin, and N PRŽULJ. 2016. Patient-specific data fusion for cancer stratification and personalised treatment. In *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, Pac Symp Biocomput, Fairmont Orchid, Hawaii(US), 321–332.
[11] Q Gu and J Zhou. 2009. Co-clustering on manifolds. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Paris, France, 359–368.
[12] G Hu and P Agarwal. 2009. Human disease-drug network based on genomic expression profiles. *PLoS One* 4 (2009), 43–46.
[13] MJ Keiser, V Setola, and JJ Irwin. 2009. Predicting new molecular targets for known drugs. *Nature* 462 (2009), 175–181.
[14] H Kim and H Park. 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23, 12 (2007), 1495–1502.
[15] M Kuhn, I Letunic, LJ Jensen, and P Bork. 2015. The SIDER database of drugs and side effects. *Nucleic acids research* 44, D1 (2015), D1075–D1079.
[16] AN Langville, CD Meyer, R Albright, J Cox, and D Duling. 2014. Algorithms, initializations, and convergence for the nonnegative matrix factorization. *arXiv preprint arXiv* 1407, 7299 (2014), 1–18.
[17] J Li and Z Lu. 2012. A new method for computational drug repositioning using drug pairwise similarity. In *Proc Int Conf Bioinformatics Biomed*, Vol. 2012. IEEE, PA, USA, 1119–1126.
[18] J Li and Z Lu. 2013. Pathway-based drug repositioning using causal inference. *BMC Bioinformatics* 14, Suppl16 (2013), S3.
[19] J Li, S Zheng, B Chen, AJ Butte, SJ Swamidass, and Z Lu. 2015. A survey of current trends in computational drug repositioning. *Brief Bioinform* 17, 1 (2015), 2–12.
[20] YH Li, CY Yu, XX Li, P Zhang, J Tang, Q Yang, T Fu, X Zhang, X Cui, G Tu, et al. 2017. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic acids research* 46, D1 (2017), D1121–D1127.
[21] Y Luo, X Zhao, J Zhou, J Yang, Y Zhang, W Kuang, J Peng, L Chen, and J Zeng. 2017. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications* 8, 1 (2017), 573.
[22] RB Mailman and V Murthy. 2010. Third generation antipsychotic drugs: partial agonism or receptor functional selectivity? *Current pharmaceutical design* 16, 5 (2010), 488–501.
[23] A Mazzanti, A Kanthan, N Monteforte, M Memmi, R Bloise, V Novelli, C Miceli, S O'Rourke, G Borio, A Zienciuk-Krajka, et al. 2014. Novel insight into the natural history of short QT syndrome. *Journal of the American College of Cardiology* 63, 13 (2014), 1300–1308.
[24] F Napolitano, Y Zhao, and VM Moreira. 2013. Drug repositioning: a macine-learning approach trough data integration. *J Cheminform* 5 (2013), 30.
[25] Howard S, Eric J, and Benjamin R. 2012. Postoperative nausea and vomiting. *Annals of Palliative Medicine* 1, 2 (2012), 1–9.
[26] A Subramanian et al. 2017. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. *Cell* 171, 6 (2017), 1437–1452.
[27] MC Veronesi, DJ Kubek, and MJ Kubek. 2008. Isoflurane exacerbates electrically evoked seizures in amygdala-kindled rats during recovery. *Epilepsy research* 82, 1 (2008), 15–20.
[28] S Wild, J Curry, and A Dougherty. 2004. Improving non-negative matrix factorizations through structured initialization. *Pattern recognition* 37, 11 (2004), 2217–2232.
[29] S Wild, WS Wild, J Curry, A Dougherty, and M Betterton. 2003. *Seeding non-negative matrix factorizations with the spherical k-means clustering*. Ph.D. Dissertation. University of Colorado.
[30] DS Wishart, C Knox, AC Guo, D Cheng, S Shrivastava, D Tzur, B Gautam, and M Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36, Database (2008), D901–D906.
[31] Y Xue, CS Tong, Y Chen, and W-S Chen. 2008. Clustering-based initialization for non-negative matrix factorization. *Appl. Math. Comput.* 205, 2 (2008), 525–536.
[32] M Žitnik, V Janjić, C Larminie, B Zupan, and N Pržulj. 2013. Discovering disease-disease associations by fusing systems-level molecular data. *Scientific reports* 3 (2013), 3202.
[33] M Žitnik, EA Nam, C Dinh, A Kuspa, G Shaulsky, and B Zupan. 2015. Gene prioritization by compressive data fusion and chaining. *PLoS computational biology* 11, 10 (2015), e1004552.