Matrix Factorization-based Technique for Drug Repurposing Predictions

G. Ceddia, P. Pinoli, S. Ceri, and M. Masseroli

***Abstract*— Classical drug design methodologies are hugely costly and time-consuming, with approximately 85% of the new proposed molecules failing in the first three phases of the FDA drug approval process. Thus, strategies to find alternative indications for already approved drugs that leverage computational methods are of crucial relevance. We previously demonstrated the efficacy of the Non-negative Matrix Tri-Factorization, a method that allows exploiting both data integration and machine learning, to infer novel indications for approved drugs. In this work, we present an innovative enhancement of the NMTF method that consists of a shortest-path evaluation of drug-protein pairs using the protein-to-protein interaction network. This approach allows inferring novel protein targets that were never considered as drug targets before, increasing the information fed to the NMTF method. Indeed, this novel advance enables the investigation of drug-centric predictions, simultaneously identifying therapeutic classes, protein targets and diseases associated with a particular drug. To test our methodology, we applied the NMTF and shortest-path enhancement methods to an outdated collection of data and compared the predictions against the most updated version, obtaining very good performance, with an Average Precision Score of 0.82. The data enhancement strategy allowed increasing the number of putative protein targets from 3,691 to 15,295, while the predictive performance of the method is slightly increased. Finally, we also validated our top-scored predictions according to the literature, finding relevant confirmation of predicted interactions between drugs and protein targets, as well as of predicted annotations between drugs and both therapeutic classes and diseases.**

***Index Terms*— Non-negative Matrix Tri-Factorization, drug repurposing, drug repositioning, shortest-paths, link prediction, complex networks**

I. Introduction

The process of drug development has become more and more time-consuming and expensive during the years. Recent studies report an average of 15-years period and over $2 billion to bring a new drug to the market [1], [2]. The challenge for new compounds is to make it to the phase II of clinical trials; only 10% pass the phase I due to safety concerns and ineffectiveness [3]–[5]. This leads to a steady drop in productivity; thus, the drug discovery process can no longer satisfy the market demand for new treatments, and rare disorders that need *de novo* therapies are completely neglected [3], [6]. As a result, big data analysis of biological and medical data has been considered, both to investigate new therapeutic opportunities for already approved drugs (namely drug repurposing), as well as to study their actions and adverse effects.

Drug repurposing (also called drug repositioning) is a strategy that deals with the limits of drug discovery by finding new purposes for known drugs. This approach leverages deep knowledge of drugs and gives several advantages over developing a completely new compound [1]. The percentage of success is higher: as repurposed drugs have already passed safety controls in different biological models and humans (when phase I of clinical trials is completed), the only critical step for these drugs to be approved for the new purpose is the efficacy during trial phases [1]. Therefore, the time spent on the development of a repurposed drug is extremely shorter than that needed for a new drug [7]. Furthermore, the repurposing strategy takes fewer investments: while the costs of phase III for a repurposed drug are similar as for a new drug, those of phase I and phase II are substantially reduced [1]. Overall, drug repositioning is a less risky, more rapid and cheaper approach than the traditional drug discovery, and it may also lead to interesting new findings associated with known drugs. It may be performed either experimentally or computationally; however, computational algorithms that can give indications on the experimental repurposing process are of pivotal importance to narrow down the repositioning search.

In this perspective, we previously proposed a novel computational method for drug repurposing prediction, which takes advantage of multiple data types and sources available [8], [9]. Specifically, we showed how the integration of several heterogeneous data types from different sources with the network-based Non-negative Matrix Tri-Factorization (NMTF) approach enables drug-therapeutic class, drug-protein target, and drug-disease link predictions. In this paper, we introduce an important innovation of the method, by adopting shortest paths as a means to infer more connections among nodes than those explicitly included in the integrated networks. Thanks to this innovation, our shortest-path enhanced NMTF method may lead to novel drug-protein target interaction discoveries, new drug annotations and new drug-disease associations. In particular, the method allows inferring as drug targets proteins that are not directly associated with known drugs. To the best of our knowledge, the use of shortest path in NMTF is original, although shortest paths have been successfully injected in many other computational methods.

II. State-of-the-Art Computational Approaches for Drug Repurposing

Computational approaches for drug repurposing can be classified into five major groups, which are: signature matching, molecular docking, genetic association, pathway mapping and backward clinical analysis techniques [1].

Signature matching is based on the comparison of actions/effects between a drug and another drug, disease, or phenotype [10]. Comparisons can be computed by matching transcriptomic signatures, chemical structures, or adverse effect profiles [1]. Thus, drug-drug comparisons can be done by calculating the differential gene expression profiles of a biological specimen before and after treatment with each drug and then comparing them [11]. We can also match the transcriptomic signatures of an untreated disease and a drug-treated biological specimen; if the two signatures are negatively correlated, then the drug may have a potential effect on that disease [12]. Drug-drug similarity approaches for drug repurposing state that if two drugs are significantly similar (according to their induced transcriptomic signatures, chemical structures, or adverse effect profiles), then they could share therapeutic applications [13], as presented in [14]. Successful examples employing the signature matching techniques include [15]–[18]. So et al. [15] proposed a framework to compute similarities between transcriptomic signatures from genome-wide association studies (GWAS) and the Connectivity Map [19]. Wang et al. [16] extracted similar signatures from the entire GEO repository [20] by using a particular training set. Iorio et al. [17]

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

G. Ceddia, P. Pinoli, S. Ceri, and M. Masseroli are with the Department of Electronics, Information and Bioengineering at Politecnico di Milano, via Ponzio 34/5, 20133, Milan, Italy. (e-mail: first.last@polimi.it).

refined the CMap signatures using drug secondary effects. Aliper at al. [18] integrated a large amount of transcriptomic signatures with MeSH therapeutic use data to classify drugs by therapeutic categories using deep neural networks. However, these approaches have limitations associated with signature data, including the quality of measurements and the different dosages from in vitro to in vivo models.

Molecular docking is based on 3D structure similarity between drugs and therapeutic targets, and it is used to predict new binding opportunities for drugs [21]. Although this technique has several issues, including data availability, its applicability has improved over the years [1].

Genetic association and pathway mapping techniques are strictly correlated. GWAS identify novel genetic variants associated with common diseases, i.e., they provide information about possible new therapeutic targets. From the drug repositioning perspective, drugs that are already used for other diseases could be repurposed to address the novel targets. GWAS data are also adopted for pathway-based or network-based approaches, where comparative network analyses of drug-treated versus untreated disease phenotypes provide repurposing candidate, as presented in [22], [23].

In addition, post-marketing surveillance and clinical trial data can lead to drug repurposing by methodical analysis [1]. Limitations of this approach lie on the accessibility of data, due to commercial and confidentiality reasons [1].

Anyway, computational methods that can integrate more data types from different sources have provided better predictions and wider analyses of the effects of drugs [3], [24]. Network-based drug repositioning is the baseline to easily integrate several data sources summarizing genome-wide data for drugs, biomolecules, and diseases. Each source can be interpreted as a network and the integrated network can be used to repurpose a single drug or a combination of drugs [24]. For example, Luo et al. [25] proved that integrating heterogeneous data sources with their network-based method leads to better performances of the drug repositioning model than using similarity-based methods. Other network-based methods include [26]–[29]. Wang et al. [26] used chemical structure, target protein, and side-effect data to correlate drugs with diseases; the correlations are used to predict novel drug-disease interactions using Support Vector Machines (SVM). Napolitano et al. [27] integrated gene expression, chemical structure, and target interaction profile data into a drug-similarity matrix used as input for a multi-class SVM classifier. Performance comparisons between drug-similarity matrices from a single source or from integrated sources as input for the SVM classifier showed better results for the integrated source ones [27]. Zitnic et al. [28] instead used a matrix factorization approach to classify diseases by combining several data types about drugs, diseases, and genes. The results of this approach demonstrated that each added data type improves the classification performance. Alaimo et al. [29] used a tripartite graph to infer associations between molecules and diseases. All these examples together prove that drug integrated data and network-based methods lead to more powerful predictive models than other approaches [3].

Our study aims at providing a novel network-based computational method that easily integrates heterogeneous data from different sources, and predicts drug annotations (therapeutic classes and drug categories), drug-protein target and drug-disease associations with good reliability. The baseline of our approach is the plain Non-negative Matrix Tri-Factorization method applied to drug repositioning [8]; for it, we previously proposed some computational optimizations regarding initialization techniques and parameter choice [9]. We also proved that integrating multiple data types improves its performance with respect to considering a single data type [9].

Here, we significantly extend our previous work by proposing a novel enhanced computational framework innovatively using a shortest-path based approach; the most significant innovation applies to protein networks, as it enables to build drug-protein target predictions for proteins that are not known to be targeted by any drug, through their indirect associations with other targets of known drugs. Thus, our new framework leads to novel drug-target interaction discoveries. Comparisons with state-of-the-art methods show that our shortest-path enhanced NMTF-based approach outperforms the others. We report top-scored predictions obtained with our method and validate them with biological findings from the literature and wet lab experiments, demonstrating the huge potential of our method. For example, we predict interesting drugs as estrogen receptor interactors and as associated with some cancer types, and then find experimental validation in the literature of both their activity as estrogen receptor ligands and their anti-proliferative effects that may play a role in cancer treatment. We also find drug-therapeutic class predictions that give indications on the possible effects of drugs never annotated before.

III. Methods

Our novel enhanced method for drug repurposing, aiming at predicting both new therapeutic indications and protein targets for known drugs, comprises two macro steps: (a) the integration of heterogeneous data about drugs and proteins, which might be possible drug targets, and (b) the prediction of novel information, i.e., associations between data entities. The first task is achieved by modeling different entities and their relationships as a multi-partite graph, while the second one by means of a matrix factorization technique, namely the Non-negative Matrix Tri-Factorization [30], applied on the association matrices of the multi-partite graph. In this Section, we describe these two steps on a generic multi-partite graph.

A. Baseline: Non-negative Matrix Tri-Factorization

Consider a matrix $R_{AB} \in \mathbb{R}_+^{(|A|\times|B|)}$ to be the association matrix between the nodes of a set A and the nodes of a set B of a graph, where $\mathbb{R}_+ = \{x : x \in \mathbb{R} \wedge x \geq 0\}$ denotes the set of positive real numbers and, for a given pair of nodes $a \in A$ and $b \in B$, the element $[a, b] \in R_{AB}$ is $R_{AB}[a, b] = \omega > 0$ if a and b are connected by an edge of the graph with weight ω, $R_{AB}[a, b] = 0$ otherwise. While the Non-negative Matrix Tri-Factorization is a general purpose method that can be applied to factorize any matrix of non-negative elements [31], for the sake of simplicity and conciseness we restrict our discussion to the aforementioned association matrix R_{AB}. Thus, by means of NMTF, for a given pair of parameters $0 < k_1 < |A|$ and $0 < k_2 < |B|$, we can approximate the matrix R_{AB} with a matrix $\hat{R}_{AB}$ computed as the product of three factors:

$$R_{AB} \approx \hat{R}_{AB} = G_A S_{AB} G_B^T$$

such that $G_A \in \mathbb{R}_+^{(|A|\times k_1)}$, $S_{AB} \in \mathbb{R}_+^{(k_1\times k_2)}$ and $G_B \in \mathbb{R}_+^{(|B|\times k_2)}$ are the three matrices of constrained size that minimize the Frobenius norm:

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

The matrices G_A and G_B are constrained to be orthogonal, i.e., $G_A G_A^T = I$ and $G_B G_B^T = I$, with I the unitary matrix; this guarantees the uniqueness of the factorization solution.

Given a non-negative matrix, its NMTF decomposition is computed, starting from an initialization of the three matrices G_A, S_{AB} and G_B, by iterating the application of a set of multiplicative update rules that compute the values of the three matrices at the current iteration from the values of the same matrices at the previous

iteration. Such multiplicative update rules, as first proposed in [30] and described in detail in [9] for the drug repurposing application, guarantee that the Frobenius norm between the input matrix and its NMTF reconstruction (i.e., $\hat{R}_{AB}$) monotonically decreases, with a better approximation being computed at each iteration. Thus, the multiplicative update rules are carried on until a stop criterion is met (e.g., the improvement of the Frobenius norm between the input matrix and its approximated decomposition in two subsequent iterations is below a user-specified threshold [32]).

One of the most interesting properties of the NMTF is the possibility to easily extend the decomposition to the association matrices of a network of two or more layers (i.e., sub-graphs). Suppose that, in addition to the two sets A and B of nodes, in the network we have a third set C of nodes, and let R_{AB} and R_{BC} be the association matrices between the elements of A and B and between the elements of B and C, respectively. 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 the error:

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

Notice that in such formulation the matrix G_B belongs to the factorization of both the R_{AB} and R_{BC} association matrices; thus, its values are influenced by the information in both the sub-graphs these matrices represent. Using this *chaining rule*, we can compute simultaneously the NMTF decomposition of all the association matrices in a multi-partite network.

B. Optimizations: Initialization, Stop Criterion and Parameters

Different initialization strategies for the NMTF can be adopted; indeed, the selection of a better initialization may lead to some advantages, such as avoiding local optima or requiring less iterations to reach the convergence [33]. In our previous work [9], aside of the trivial random initialization, we investigated several alternatives, namely the *random ACOL*, the *k-means clustering* and the *spherical k-means clustering* initializations. The latter one showed better results than other techniques for the drug repurposing application. Thus, we adopt the *spherical k-means* also in this work, particularly for the results that we present in the next Section III-C. We then verified that it gives better results also in this case (data not shown).

A second choice in the scientist's hand is the stop criterion used to decide when terminating the iteration of the multiplicative update rules. The convergence of the algorithm to a local stationary point has been proven by means of the minimization of the objective function (i.e., the approximation error of the NMTF decomposition) and the use of the multiplicative update rules [34]. The convergence is measured as the relative difference between the objective function values at two successive iterations of the algorithm [32], [35]. The optimization is hence stopped when the relative difference of the objective function reaches a user-specified threshold [35]:

$$|J_{i+1} - J_i|/J_i < \epsilon$$

where J is the objective function, i is the iteration, and $\epsilon = 10^{-3}$ is used in our experiments (following the results presented in [9] for the drug repurposing application).

A final important aspect of the NMTF is the choice of the dimensions of the decomposition factor matrices. In our previous studies regarding the drug repurposing application, we also tested various methods to identify the best set of such parameters, and found that the dispersion coefficient metric [31] is the most suitable choice [8], [9]; this coefficient ranges between 0 and 1, with higher values indicating more stable solutions. Thus, we select the set of parameters that maximizes the dispersion coefficient.

C. Shortest-Paths Enhanced Matrix

Shortest-path analysis has several biological applications. For example, Zhou et al. [36] used the shortest-paths to find transitive functional annotations of genes in a network based on expression data. The aim of our application is different: we want to quantify the minimum distance between two elements of different type in a bipartite network, as a weight of their possible interaction. When additional information about relationships between elements of the same type is available, we innovatively propose to take advantage of it in order to infer indirect unknown relationships between these elements and those of another type, which some of these elements are known to be directly associated with. This is particularly relevant since it also allows increasing (typically greatly) the amount of these elements that can be modeled as nodes of a multi-partite graph, i.e., beyond the only nodes with known relationships with other nodes of the graph. More specifically, we propose a shortest-paths-based method to infer unknown relationships between the nodes of a multi-partite graph, as well as between them and other nodes that were not originally linked in the graph, thus largely increasing the number of both the graph edges and the nodes of a sub-graph. We show how to use such a method to enhance a graph association matrix in order to improve NMTF predictions and infer novel putative associations between two sets of the graph nodes.

Consider two sets A and B of nodes connected by a bipartite graph, and suppose that intra-set node relationships are also available, e.g., for the nodes in B (if such relationships are available also for the nodes in A, the following is straightforwardly extended). In this case, given an association matrix R_{AB} with the known relationships between pairs of nodes $a \in A$ and $b \in B$, we can construct a matrix R'_{AB} by substituting the elements $R_{AB}[a, b]$, which typically are equal to 1 if a is known to be associated with b or 0 otherwise, with a function of the shortest path between a and b. More in detail, we search, if it exists, a path $P_{ab} = \langle b, b_1, \ldots, b_i, \ldots, b_n, a \rangle$ connecting the node a with the node b through a sequence of nodes b_i. Notice that a path between a and b can generally comprise many other nodes of A; by design, we decided to restrict our exploration only to those paths that traverse as many edges as needed in the intra-nodes relationship but only one link, specifically the last, from the R_{AB} association matrix. If at least one of such paths exists, we select one of the shortest length. Notice that many of such shortest paths may exist, but this is not relevant since we are only interested in the length of the path and not in the actual traversed nodes. Finally, for each pair of nodes a and b, we set the value of the matrix $R'_{AB}[a, b]$ as:

$$R'_{AB}[a, b] = \begin{cases} \alpha^{|P_{ab}|-1}, & \text{if } P_{ab} \text{ exists} \\ 0, & \text{otherwise} \end{cases}$$

where $|P_{ab}|$ denotes the length of any of the shortest paths between a and b, and $0 < \alpha < 1$ is a parameter meant to reduce the importance of the association between a node a and a node b that is mediated by many other nodes in the set B. Notice that, if a and b are directly (i.e., known to be) connected then $|P_{ab}| = 1$ and $R'_{AB}[a, b] = 1$, while the value of the inferred $R'_{AB}[a, b]$ weight decreases exponentially with the length of the shortest P_{ab} path, i.e., with $|P_{ab}|$; moreover, the lower is the value of the parameter α, the lower are the weights that the method computes for the associations between the elements of A and B connected through longer paths. In this way, we can enhance an association matrix by adding to the included known associations other weaker associations, inferred based on available indirect relationships and as weak as the distance of the indirect relation.

D. Prediction of Novel Associations

Consider an association matrix R_{AB} of (a layer of) a graph, its enhanced version R'_{AB} and its NMTF approximation $\hat{R}'_{AB}$. One can use the latter matrix to infer novel associations between the elements of the two sets A and B represented in the matrix. This can be done by varying a threshold $\delta \in [0,1] \subset \mathbb{R}_+$ and constructing the matrix $\hat{R'}^{\delta}_{AB}$ such that $\hat{R'}^{\delta}_{AB}[i,j] = 1$ if and only if $\hat{R'}_{AB}[i,j] > \delta$, otherwise $\hat{R'}^{\delta}_{AB}[i,j] = 0$. As a common practice in machine learning applications, we can define the four classes of predicted elements, namely True Positives (TP_δ), False Positives (FP_δ), True Negatives (TN_δ) and False Negatives (FN_δ) for each $\delta \in [0,1]$. However, differently from one would expect such classification to be performed, in our approach we compare $\hat{R}'_{AB}$ directly against R_{AB}, rather than R'_{AB}. This implies that the enhancement of the association matrix by means of the shortest-paths method is an influential component of the overall prediction framework. Therefore, for a given δ we call TP_δ the pairs $[i,j]$ for which $\hat{R'}^{\delta}_{AB}[i,j] = R_{AB}[i,j] = 1$, FP_δ the ones for which $\hat{R'}^{\delta}_{AB}[i,j] = 1$ and $R_{AB}[i,j] = 0$, TN_δ the ones for which $\hat{R'}^{\delta}_{AB}[i,j] = R_{AB}[i,j] = 0$, and FN_δ the ones for which $\hat{R'}^{\delta}_{AB}[i,j] = 0$ and $R_{AB}[i,j] = 1$. We are particularly interested in the False Positives, since they may indicate novel associations inferred by the method between the elements of the two sets A and B, i.e., which are not in the original matrix $R_{AB}[i,j]$ since still unknown and not discovered yet by traditional methods.

E. Validation Metrics

Once the $\hat{R'}_{AB}$ matrix has been computed, we can evaluate the *Recall*, or *True Positive Rate* (TPR), and the *Precision* of the method as a function of δ:

$$Recall_\delta = \frac{TP_\delta}{TP_\delta + FN_\delta} \qquad Precision_\delta = \frac{TP_\delta}{TP_\delta + FP_\delta}$$

where TP_δ, FN_δ and FP_δ are as defined in Section III-D. By varying the δ value in the range from 0 to 1, we can compute the *Average Precision Score* (APS), which corresponds to the area under the Precision-Recall curve:

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

For the performance evaluation we also use the *False Positive Rate* (FPR) and the *Area Under the Curve* (AUC) of the FPR-TPR curve, or *Receiver Operating Characteristic* (ROC) curve, with:

$$FPR_\delta = \frac{FP_\delta}{TN_\delta + FP_\delta}$$

F. Implementation and Availability

The method was developed in Python programming language. For the enrichment of the drug-to-protein matrix by means of shortest paths and protein-to-protein interactions we leveraged on the functionalities of the NetworkX[1] library. The software is open source and publicly available at https://github.com/DEIB-GECO/NMTF-DrugRepositioning.

IV. Drug-related Data and Sources

We integrate data about drugs and their therapeutic classes, protein targets, and application in disease cares, as well as about protein-protein interactions and associations between proteins and biological pathways. In order to do it, we represent these heterogeneous data as a multi-partite network with five classes of nodes that represent therapeutic classes, drugs, diseases, proteins, and pathways, respectively. Network edges connect drugs to their recognized therapeutic classes, to their known protein targets and to the diseases for whose care the drugs are used, while proteins are connected to the biological pathways in which they are involved. Furthermore, for the protein and pathway nodes, we represent intra-class connections as well. The former ones represent protein to protein interactions, while the latter ones capture the existing hierarchy of biological pathways, from the most generic to the most specific one. The full multi-partite graph is schematized in Figure 1.

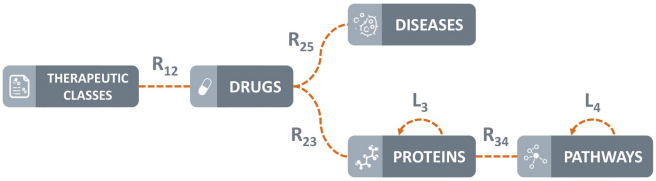

Fig. 1. Schematic representation of the multi-partite graph that integrates the variety of data types and their relationships comprehensively considered by our method. The labels on the connection lines represent the name of the association matrices between the components of the connected classes. Notice that proteins and pathways also have intra-class relationships.

To build the multi-partite graph in Figure 1, we use multiple heterogeneous data integrated from four different sources. Approved drugs, their current indications (therapeutic classes) and protein targets are taken from DrugBank [37] (version 5.1.2), pathways in which the protein targets are involved are from Reactome [38] (version 70), and drug-disease annotations from Therapeutic Target Database (TTD) [39] (version of July 2019); we also consider protein-protein interactions from BioGrid [40] (version 3.578) and pathway hierarchical relationships from Reactome.

For the experiments described in Section V, we use 141 therapeutic indications, 3,261 drugs, 3,691 protein targets, 1,914 pathways, 844 diseases and their relationships, with matrices R_{12}, R_{23}, R_{34}, R_{25}, L_3 and L_4 containing 23,517, 13,433, 28,345, 2,406, 39,756 and 3,858 links, respectively.

In Section VI we extend the number of considered proteins to 15,295, by adding all the proteins that directly interact with at least one of the 3,691 protein targets and using the shortest-path technique to estimate their relationship weight with the considered drugs. This allows the evaluation of a much greater number of novel potential drug-target interactions.

V. Computational Validation Methods and Prediction Performances

To computationally evaluate the performance of our method, we use two different strategies and measures, which explain how the shortest-path enhanced matrix and the choice of the α and path length parameters influence the NMTF-based method performance:

A. We measure the ability of our method to predict novel associations between drugs and therapeutic classes by computing the APS on the R_{12} matrix for different values of the α and path length parameters. To this aim, first, we run our method on an outdated version of the data collection, including associations between drugs and therapeutic classes retrieved from DrugBank version 5.1.2 (released in December 2018). Then, we check the obtained predictions on the latest DrugBank version (5.1.4, released in July 2019), produced 7 months later. Such an updated

[1]https://networkx.github.io

version contains 508 novel therapeutic class-drug annotations, which we use as validation set. Here, we evaluate our method using the APS score since we are interested in evaluating how the method is able to precisely find the missing associations.

B. We also measure the performance of our method in inferring the correct drug-protein target interactions for proteins that are not yet associated with any drug. In order to do so, we remove all the edges of a set of randomly selected protein target nodes in R_{23} and compute the AUC ROC of the computationally predicted interactions after using the shortest-path enhancement method with different values of the α and path length parameters. For this setting, we evaluate our method using the AUC ROC score since we want the method to retrieve the complete interaction profiles of the missing proteins in the best way.

Regarding the parameters related to the dimensions of the decomposition factor matrices (Section III-B), we chose k_1, k_2, k_3, k_4 and k_5 equal to 500, 141, 500, 500 and 300, respectively. For such parameter values, the dispersion coefficients ρ_1, ρ_2, ρ_3, ρ_4 and ρ_5 are equal to 0.998, 0.999, 0.985, 0.984 and 0.998, respectively. The greater the dispersion coefficients, the greater is the performance of the associated model [31]; for example, in our case for a bad choice of parameter values (e.g., $k_1 = k_2 = k_3 = k_4 = k_5 = 10$, whose related dispersion coefficients are $\rho_1 = 0.931$, $\rho_2 = 0.859$, $\rho_3 = 0.703$, $\rho_4 = 0.628$ and $\rho_5 = 0.821$) the APS score and AUC ROC values are 0.517 and 0.849, respectively, whereas a good choice of parameter values as the ones we chose has the APS score and AUC ROC values equal to 0.817 and 0.952, respectively.

A. Therapeutic Class-Drug Prediction Validation

Comparing DrugBank database versions 5.1.2 and 5.1.4, the newest release includes 508 novel therapeutic class-drug annotations; we use them as validation set for the evaluation of our method performance in predicting new links in the R_{12} association matrix.

After applying the NMTF method to all the R_{12}, R_{25}, R'_{23} and R_{34} matrices simultaneously (with DrugBank data from its database version 5.1.2 and the R'_{23} matrix enhanced with shortest-path data), we construct the $\hat{R}_{12}$ matrix by multiplying the computed factors G_1, G_2 and S_{12}. Then, we evaluate how well the validation set links are estimated in the $\hat{R}_{12}$ matrix. In order to do so, we set a threshold $0 \leq \delta \leq 1$ and create the binary matrix $\hat{R}^{\delta}_{12}$ by setting to 1 the $\hat{R}^{\delta}_{12}[i,j]$ elements corresponding to $\hat{R}_{12}[i,j] \geq \delta$ elements, 0 otherwise. Comparing the $\hat{R}^{\delta}_{12}$ matrix with the R_{12} one derived from the newest DrugBank database version 5.1.4, we can distinguish the matrix elements in the classical four classes: True Positives, equal to 1 in both R_{12} and $\hat{R}^{\delta}_{12}$, False Positives, equal to 1 only in $\hat{R}^{\delta}_{12}$, True Negatives, equal to 0 in both R_{12} and $\hat{R}^{\delta}_{12}$, and False Negatives, equal to 1 only in R_{12}.

Figure 2 shows the goodness of the therapeutic class-drug association prediction for different values of the α coefficient (as defined in Section III-C) and of the maximum path length considered in R'_{23}. To obtain each point in Figure 2 we modified the R'_{23} matrix such that the maximum path length considered in R'_{23}, for 10 repetitions of the NMTF algorithm, was equal to the x-axis value of the point. Thus, Figure 2 shows the relevance of adding more and more information in R'_{23} from the therapeutic class-drug prediction point of view.

As it can be seen, setting α equal to 0.2 provides the best performances over path lengths; indeed, it gives a very high average APS, equal to 0.82, better than the APS score (equal to 0.79) for unitary path length (i.e., for the R_{23} matrix without the shortest-path enhancement). Thus, the NMTF method applied after shortest-path enhancement can predict missing links in the R_{12} association matrix better than when the NMTF method is applied to binary association

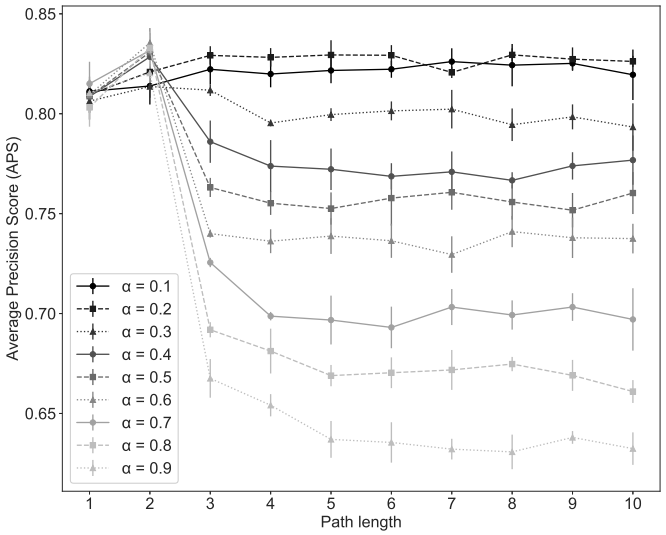

Fig. 2. Performance representation of the therapeutic class-drug association prediction method (Section V-A). Average Precision Score (APS) over the maximum path length of the associations included in the $\boldsymbol{R'_{23}}$ matrix. Each curve represents a different $\boldsymbol{\alpha}$ parameter value, as reported in the figure legend; each point is the mean value over 10 repetitions of the NMTF method and the error bars show the standard deviation across repetitions.

matrices only. Interestingly top score predictions are confirmed by both the plain NMTF (without shortest-path) and the shortest-path enhanced one; however, shortest-path addition allows us to better differentiate lower score predictions. This in turn has a positive effect on the overall performance of the prediction method. Moreover, the curve that corresponds to an α coefficient equal to 0.2 has very stable performances across path lengths, much more than the curves corresponding to α coefficients greater than 0.3, proving the robustness of our parameter choice.

Furthermore, for more a systematic validation we compared the distributions of the predicted probabilities of all the 508 novel therapeutic class-drug annotations included in the validation set with the distribution of the predicted probabilities of all the other such associations considered. The mean and median of the former ones resulted 0.179 and 0.117, respectively, whereas those of the latter ones were 0.154 and 0.098, respectively; the two distributions resulted statistically significantly different according to the Wilcoxon-Mann-Whitney test (p-value = $1.65 \cdot 10^{-4}$).

B. Drug-Protein Target Prediction Validation

To demonstrate that our method can properly predict novel drug-protein target interactions also for proteins that are not known as targets of any drug, for 300 protein targets randomly selected we remove all their drug-target binary associations (1,555 associations, regarding 898 drugs); then, we compute the shortest-path enhanced matrix R'_{23}, with the missing direct associations removed but leveraging on protein-protein interaction information, and we apply the NMTF approach. Finally, we evaluate if the computed $\hat{R'}_{23}$ matrix properly reconstructs the original drug-protein target interactions; to do so, after setting a value for the threshold $0 \leq \delta \leq 1$, we create the binary $\hat{R'}^{\delta}_{23}$ matrix, where TP, FP, TN and FN are as defined in Section III-D.

In Figure 3 we evaluate the shortest-path-based enhancement method for adding proteins in the prediction process that are not yet annotated as targets of drugs. The NMTF method, applied considering the binary association matrix R_{23}, cannot include proteins that are

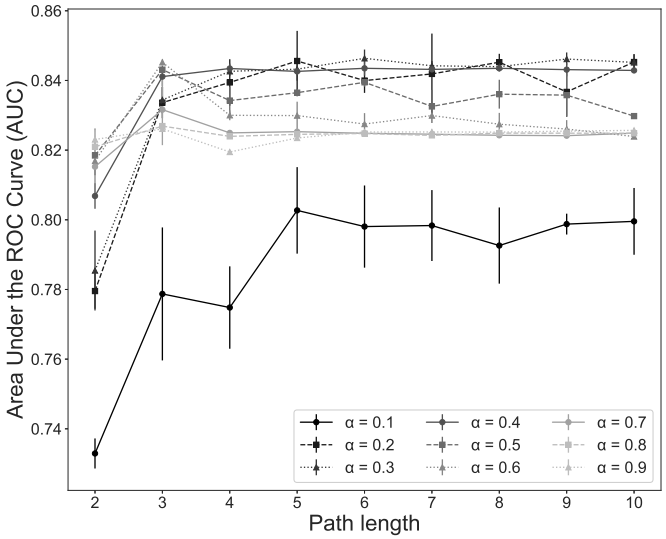

Fig. 3. Performance representation of the drug-protein target association prediction method (Section V-B). Area Under the ROC Curve (AUC) over the maximum path length of the associations included in the $\boldsymbol{R'_{23}}$ matrix. Each curve represents a different $\boldsymbol{\alpha}$ parameter value, as reported in the figure legend; each point is the mean value over 10 repetitions of the NMTF method executed considering increasingly longer path lengths, and the error bars show the standard deviation across repetitions.

unknown as drug targets. However, the NMTF method applied after shortest-path enhancement of the R_{23} matrix (i.e., considering the R'_{23} matrix instead) is able to overcome this issue. Figure 3 shows the AUC ROC scores for the drug-protein target associations retrieved with the NMTF method after shortest-path enhancement, computed only for the random proteins whose all drug associations were removed and for the related drugs. In a real case scenario, this method may lead to discover novel drug-protein target interactions, where such targets can be also proteins that were never considered as drug targets.

Each point in Figure 3 represents the mean value of the performance over 10 repetitions, in which the maximum path length considered in R'_{23} is equal to the point x-axis value. In the x-axis the value 1 is missing since binary associations (i.e., those with path length equal to 1) were removed for the validation process. Figure 3 shows that an α coefficient equal to 0.2 gives one of the best performances for path lengths greater than 3, reaching a mean AUC ROC value equal to 0.845, which corresponds to a remarkably high performance. Also, curves corresponding to α coefficients equal to 0.3 and 0.4 reach similar mean values of performances, i.e., 0.845 and 0.844 respectively; yet, their performance is not so good for the validation reported in Figure 2. Instead, the NMTF method with shortest-path enhancement and an α coefficient equal to 0.1 has the worst performance over path lengths, conversely to what occurs in the validation results illustrated in Figure 2; this makes such value not a good choice for the α coefficient. Together Figures 2 and 3 clearly show that shortest paths with path length greater than 3 do not carry useful information to improve the performance of our method.

C. Evaluation of Data Integration Improvement

Similarly to what done in [9], we assessed the benefits of the integration of data from heterogeneous sources, with a particular focus on the improvement brought by the integration of protein targets by the novel shortest-path method. We evaluated how the APS of the therapeutic class-drug annotation predictions (in the R_{12} matrix) varies when incrementally adding new layers in the multi-partite network. We obtained incremental APS values equal to 0.788, 0.792 and 0.804 when considering the $\langle R_{12}, R_{23}\rangle$, $\langle R_{12}, R_{23}, R_{34}\rangle$ and $\langle R_{12}, R_{23}, R_{34}, R_{25}\rangle$ networks, respectively, thus proving the benefits of integrating different heterogeneous datasets. Those APS values further increase to 0.808, 0.808 and 0.817 when the drug-protein target matrix R_{23} is enriched by using the shortest-path method with the optimal parameters $\alpha = 0.2$ and $1 \leq$ path length ≤ 3, demonstrating the gain of the shortest-path enhancement.

D. Comparative Study

We compared our drug repurposing results with those of two state-of-the-art methods proposed in [14] and [25]. Luo et al. [25] developed a network-based method, called DTINet, which integrates heterogeneous data sources for drug repositioning. DTINet uses a feature extraction method that extracts low-dimensional vector representations from the topology of the nodes in the integrated network. Then, it predicts and computes drug-protein target interactions and drug similarity measures through these representations. Li et al. [14] instead presented an approach based on drug similarity (DS), where a new indication for a drug is identified by its relations with other drugs. For this aim they compute a linear combination of chemical structure similarity and target similarity; the former one is measured by the Tanimoto coefficient of the 2D chemical fingerprints, while the latter one is measured using the drug-protein bipartite-graph considering common drug targets and their interaction information.

The predicted drug similarities from DTINet are used to infer new therapeutic indications for drugs by means of their interactions in the integrated network, i.e., if two drugs are similar according to DTINet, they can share their uses. An equal hypothesis is done on the drug similarities computed with the DS method. For the comparative study, we limited the analysis to the 607 drugs whose data can be retrieve for both approaches. We then computed the APS score and AUC ROC values for the NMTF, DTINet, and DS methods. As it can be seen in Figure 4, our method achieved higher APS score and AUC ROC values (0.863 and 0.931, respectively) than using DTINet (0.525 and 0.839, respectively), or DS (0.511 and 0.812, respectively) methods.

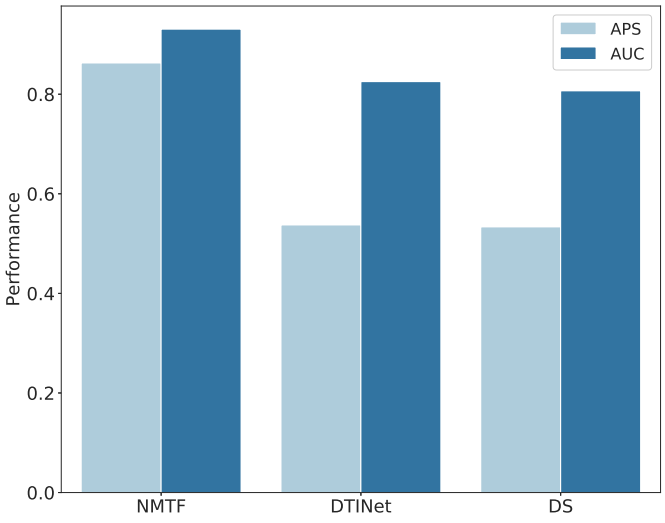

Fig. 4. Comparison of APS scores and AUC ROC values for NMTF, DTINet [25] and Drug Similarity (DS) [14] methods. The APS scores are equal to 0.863, 0.525 and 0.511, respectively. The AUC ROC values are equal to 0.931, 0.839 and 0.812, respectively. The scores are computed considering the 670 drugs evaluated in both [14] and [25].

VI. Literature Validation

Taking advantage of the shortest-path enhancement and of the good performances it provides as demonstrated in Section V, in what follow we extend the number of proteins in the modeled network, by including also those proteins that are not target of any drug according to the last version of the DrugBank database. In the network, the newly added proteins do not have binary associations with drugs, but their associations derive from shortest path lengths greater than 1. After applying our innovative computational framework to the new extended network, we manually validate the top new associations predicted for the R_{12}, R_{23} and R_{25} matrix, respectively. Towards this aim, we extract the high-score "False Positives" for each matrix and we validate them based on the associated literature findings. The hallmark of our method is to simultaneously get drug-centric predictions from the therapeutic class, the protein target, and the disease perspectives. The obtained results are following discussed.

A. Therapeutic Class Predictions

After applying the NMTF method with the shortest-path enhancement, we reconstruct the $\hat{R}_{12}$ matrix by multiplying the G_1, G_2 and S_{12} computed factors, and we normalize the values of $\hat{R}_{12}$ to have scores ranging from 0 to 1. Some of the obtained high-score therapeutic class predictions ("False Positives") are reported in Table I (prediction score greater than 0.85).

TABLE I
Top Therapeutic Class Predictions

Therapeutic Class	Drug Name	Drug ID	Score
Immunologic Factor	Emapalumab	DB14724	0.930
Cardiovascular Agent	Cacodylic Acid	DB02994	0.928
Immunoprotein	Lorlatinib	DB12130	0.926
Neurotransmitter Agent	Tropicamide	DB00809	0.881
Neurotransmitter Agent	Homatropine	DB11181	0.853

Emapalumab is an interferon-gamma blocking antibody. It is classified as a monoclonal antibody used for the treatment of patients with primary hemophagocytic lymphohistiocytosis (HLH) [37]. According to the experimental assays in [41], it affects the immune system by controlling the immune hyperactivation associated with the HLH disease. Thus, our algorithm positively associates this drug to the therapeutic class known as *Immunologic Factors* (with a 0.930 score), whereas DrugBank fails to report it.

Cacodylic acid is an organoarsenic compound containing arsenic and organic groups. It has been experimentally shown that Cacodylic acid can induce procoagulant activity and apoptosis in specific blood cells that play key roles in the development of various cardiovascular diseases [42], [43]. Our method successfully labels Cacodylic acid as a *Cardiovascular Agent* with a 0.928 score.

Lorlatinib is a small molecule used for the treatment of Non-small Cell Lung Cancer. It acts as an antineoplastic and immunomodulating agent by inhibiting the tyrosine kinase. Although this drug is classified as a small molecule, from the therapeutic class perspective Lorlatinib functions as an *Immunoprotein* [44], as our method predicts (prediction score equal to 0.926).

Tropicamide is an antimuscarinic drug that produces short-term dilation of the pupil. Tropicamide acts as a *Neurotransmitter Agent*; indeed it is a muscarinic receptor antagonist that blocks the activity of the muscarinic acetylcholine receptor (i.e., the ACh neurotransmitter) [45]. This validates the association that our algorithm predicts for Tropicamide with a 0.881 score. Also Homatropine is know to act as an antagonist of muscarinic acetylcholine receptors [46], confirming its classification as *Neurotransmitter Agent* provided by our NMTF-based method (with a 0.853 score).

B. New Drug-Target Interactions

The shortest-path enhancement method allowed us to include much more proteins into the considered network, i.e., proteins that were never considered as targets of drugs. This makes the predicted new drug-protein target interactions even of greater interest. Table II reports some of the highest-score drug-target predictions of our enhanced NMTF method, computed by reconstructing the $\hat{R}'_{23}$ matrix and normalizing the scores of its predicted links. They include Adinazolam and Clotiazepam drug predictions.

TABLE II
Top-Scored Predicted new Drug-Target Interactions

Uniprot Name	Uniprot ID	Drug Name	Drug ID	Score
GBRA4	P48169	Clotiazepam	DB01559	0.782
GBRA6	Q16445	Clotiazepam	DB01559	0.779
GBRA4	P48169	Adinazolam	DB00546	0.760
GBRA6	Q16445	Adinazolam	DB00546	0.757
GABRQ	Q9UN88	Clotiazepam	DB01559	0.635

According to our method, the Adinazolam drug interacts with the GBRA4 protein. Adinazolam derives from benzodiazepine, a class of psychoactive drugs; it has anxiolytic, anticonvulsant, sedative, and antidepressant properties [47], [48]. GBRA4 is the alpha-4 subunit of the GABA neurotransmitter. Benzodiazepines and GABA-A receptor are known to bind for the modulation of the chloride channel in cell membranes [49]. This provides a very strong biomolecular support to our predicted Adinazolam-GBRA4 interaction (0.760 score), as well as to the computationally predicted Adinazolam-GBRA6 interaction (0.757 score), where GBRA6 is the alpha-6 subunit of the GABA neurotransmitter.

Clotiazepam is another benzodiazepine derivative, which our method predicts to interact with the GBRA4, GBRA6 and GABRQ proteins (with score 0.782, 0.779 and 0.635, respectively). Each of these proteins is involved with the *chloride channel activity*; they all are subunits of the GABA-A receptor [49], which strongly supports their interaction with Clotiazepam as a mechanism to modulate the chloride channel activity.

Some of our computationally predicted drug-target interactions are also confirmed by the experimental results in [50]; in particular, Table III reports the ones regarding Simvastatin, Ketoconazole, Diclofenac and Itraconazole. These drugs were experimentally validated by Cheng et al. [50] for their possible interactions with two estrogen receptors, namely $ER\alpha$ and $ER\beta$. In vitro assays confirmed that Simvastatin, Ketoconazole, Diclofenac and Itraconazole act as novel estrogen receptor ligands, which confirms our predictions reported in Table III.

TABLE III
Experimentally-Validated new Drug-Target Interactions

Protein Name	Uniprot ID	Drug Name	Drug ID	Score
$ER\beta$	Q92731	Simvastatin	DB00641	0.727
$ER\beta$	Q92731	Ketoconazole	DB01026	0.722
$ER\alpha$	P03372	Diclofenac	DB00586	0.688
$ER\beta$	Q92731	Diclofenac	DB00586	0.572
$ER\alpha$	P03372	Itraconazole	DB01167	0.385
$ER\beta$	Q92731	Itraconazole	DB01167	0.316

Furthermore, we systematically tested our drug-protein target predictions against the available literature, automatically retrieving scientific publications from PubMed[2]. For each predicted drug-protein target pair, we evaluated if at least one paper mentioning it exists. We tested the 50,000 top scored predictions and we found that the

[2]https://www.ncbi.nlm.nih.gov/pubmed/

probability for a predicted drug-protein target pair to be mentioned in a scientific paper decreases with the prediction score given by our method. In particular, 4.9% of the top 5,000 predicted pairs are mentioned in a paper, while this percentage decreases to only 3.6% for the bottom 5,000 considered predictions.

C. Drug-Disease Novel Associations

To complete our drug-centric evaluation we consider the reconstructed $\hat{R}_{25}$ matrix, which gives novel drug-disease associations. In particular, we focus on the Itraconazole, Diclofenac, Simvastatin, and Ketoconazole drugs, which respectively have a total of 13, 22, 20 and 19 predicted disease associations with a score greater than 0.3. They include associations with specific types of cancer, as reported in Table IV, as well as with several infectious diseases. Itraconazole and Ketoconazole are antifungal medications, indeed used to treat fungal infections; Simvastatin is a lipid-lowering drug and Diclofenac is a non-steroidal anti-inflammatory compound [37]. They all show particularly interesting drug-disease associations. As a confirmation of their cancer association, in [50] and [51] Simvastatin and Ketoconazole showed anti-proliferative activities on breast cancer cell experiments, suggesting that these antifungal agents may have therapeutic effects also on breast cancer. Itraconazole and Diclofenac, instead, have already been considered as anti-cancer agents in recent case studies [52], [53].

TABLE IV
SOME DRUG-DISEASE NOVEL ASSOCIATIONS

Type of Disease	Drug Name	Drug ID	Score
Breast Cancer	Simvastatin	DB00641	0.984
Breast Cancer	Ketoconazole	DB01026	0.969
Prostate Cancer	Simvastatin	DB00641	0.555
Breast Cancer	Itraconazole	DB01167	0.541
Colorectal Cancer	Diclofenac	DB00586	0.447
Renal Cell Carcinoma	Simvastatin	DB00641	0.445
Prostate Cancer	Ketoconazole	DB01026	0.424
Ovarian Cancer	Ketoconazole	DB01026	0.327
Head and Neck Cancer	Ketoconazole	DB01026	0.320

Furthermore, as in [54] we compared our drug-disease predicted associations with available information about clinical trials[3]. Indeed, 181 of our 1,000 top predictions (18.1%) have at least one ongoing clinical trial reported, while on average only 7.1% of all predicted drug-disease pairs are associated with an ongoing trial. Additionally, we considered the scores that our method assigns to the drug-disease pairs and tested the difference between the scores of those pairs confirmed by a clinical trial and the others. Confirmed pairs have on average a higher associated score (0.0586) compared to the non-confirmed ones (0.0457); such difference is statistically significant accordingly to the t-test for the means (p-value = $5.0 \cdot 10^{-34}$). This confirms the validity of our top scoring drug-disease predictions.

VII. CONCLUSIONS

Computational drug repurposing has several advantages compared to the traditional drug discovery, including time-saving suggestions for the experimental repositioning of known drugs and cost-cutting opportunities for the clinical validations. The most promising approaches for computational drug repurposing are the ones that integrate heterogeneous information from different data types.

In our work, we developed a NMTF-based approach that integrates several data types by representing them as a multi-partite graph and predicts novel drug-centric annotations by graph matrix factorization. We also implemented the shortest-path enhancement method that exponentially increases the number of drug-protein target predictions, enabling the evaluation of unknown protein targets. To validate our method, we used two techniques based on therapeutic class and protein target predictions, respectively. The tests we performed demonstrated the validity of our method; the shortest-path enhanced NMTF-based approach that we propose scores 0.82 of APS for drug-to-therapeutic class predictions and 0.85 of AUC ROC for drug-to-target predictions. Such performances prove that our method is valuable for the repurposing of approved drugs, as well as for the prediction of novel protein targets for drugs. For both cases, in Figures 2 and 3 we reported our method performances while including increasing information (i.e., increasing shortest path lengths) and using different values of the α parameter (as defined in Section III-C). Both Figures show that the best choice for the α parameter is 0.2. Indeed, the curves corresponding to $\alpha = 0.2$ have the best performance for the therapeutic class-drug validation (Section V-A) and maintain one of the best scores for the drug-protein target validation (Section V-B). Moreover, the joint use of the NMTF and the shortest-path enhancement methods has shown better performances than the NMTF method applied only to binary association matrices.

We also validated some of our top-scored predictions with literature findings, demonstrating that our approach successfully annotates novel drug-centric predictions. We also reported experimentally validated examples of drug-target and drug-disease associations to confirm the validity and relevance of our predictions, and to prove that our results may lead to interesting repurposing opportunities.

[3]at https://www.clinicaltrials.gov/

ACKNOWLEDGMENT

We thank Andrea Gulino from the GeCo group of Politecnico di Milano for the useful discussions and insights.

REFERENCES

[1] S. Pushpakom, F. Iorio, P. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doing, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, M. Pirmohamed, "Drug repurposing: progress, challenges and recommendations", *Nat Rev Drug Discov*, vol. 18, no. 1, pp. 41–58, 2019.

[2] N. Nosengo, "Can you teach old drugs new tricks?", *Nature News*, vol. 534, no. 7607, pp. 314–316, 2016.

[3] R. A. Hodos, B. A. Kidd, S. Khader, B. P. Readhead, J. T. Dudley, "Computational Approaches to Drug Repurposing and Pharmacology", *Wiley Interdiscip Rev Syst Biol Med*, vol. 8, no. 3, pp. 186–210, 2017.

[4] J. Arrowsmith, P. Miller, "Trial watch: phase II and phase III attrition rates 2011-2012", *Nat Rev Drug Discov*, vol. 12, no. 8, pp. 569, 2013.

[5] I. Kola, J. Landis, "Can the pharmaceutical industry reduce attrition rates?", *Nat Rev Drug Discov*, vol. 3, pp. 711–716, 2004.

[6] A. Denis, L. Mergaert, C. Fostier, I. Cleemput, S. Simoens, "A comparative study of European rare diseases and orphan drug markets", *Health Policy*, vol. 97, pp. 173–179, 2010.

[7] A. Breckenridge, R. Jacob, "Overcoming the legal and regulatory barriers to drug repurposing", *Nat Rev Drug Discov*, vol. 18, pp. 1–2, 2018.

[8] G. Ceddia, P. Pinoli, S. Ceri, M. Masseroli, "Non-negative Matrix Tri-Factorization for data integration and network-based drug repositioning", *Proc IEEE Symp Comput Intell Bioinforma Comput Biol*, pp. 255-261, 2019.

[9] G. Dissez, G. Ceddia, P. Pinoli, S. Ceri, M. Masseroli, "Drug repositioning predictions by Non-negative Matrix Tri-Factorization of integrated association data", *Proc ACM Int Conf Bioinform Comput Biol*, pp. 25-33, 2019.

[10] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, B. L. Roth "Predicting new molecular targets for known drugs", *Nature*, vol. 462, pp. 175–181, 2009.

[11] F. Iorio, T. Rittman, H. Ge, M. Menden, J. Saez-Rodriguez, "Transcriptional data: a new gateaway to drug repositioning?", *Drug Discov Today*, vol. 18, pp. 350–357, 2013.
[12] J. T. Dudley, T. Deshpande, A. J. Butte, "Exploiting drug-disease relationships for computational drug repositioning", *Brief Bioinform*, vol. 12, pp. 303–311, 2011.
[13] A. P. Chiang, A. J. Butte, "Systematic evaluation of drug-disease relationships to identify leads for novel drug uses", *Clin Pharmacol Ther*, vol. 86, pp. 507–510, 2009.
[14] J. Li, Z. Lu, "A new method for computational drug repositioning using drug pairwise similarity", *Proc Int Conf Bioinformatics Biomed*, vol. 2012, pp. 1119–1126, 2012.
[15] H.C So, C.K.L Chau, W.T. Chiu, K.S. Ho, C.P. Lo, S.H.Y. Yim, P.C. Sham, "Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry", *Nat Neurosci*, vol. 20, no. 10, pp. 1342-1349, 2017.
[16] Z. Wang, C.D. Monteiro, K.M. Jagodnik, N.F. Fernandez, G.W. Gundersen, A.D. Rouillard, Q. Duan, "Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd", *Nat Commun*, vol. 7, no. 1, pp. 1–11, 2016.
[17] F. Iorio, R.L. Shrestha, N. Levin, V. Boilot, M.J. Garnett, J. Saez-Rodriguez, V.M. Draviam, "A semi-supervised approach for refining transcriptional signatures of drug response and repositioning predictions", *PLoS One*, vol. 10, no. 10, pp. e0139446, 2015.
[18] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, A. Zhavoronkov, "Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data" *Mol Pharm*, vol. 13, no. 7, pp. 2524–2530, 2016.
[19] A. Subramanian, R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, D.L. Lahr, "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles", *Cell*, vol. 171, no. 6, pp. 1437–1452, 2017.
[20] R. Edgar, M. Domrachev, A.E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Res*, vol. 30, no. 1, pp. 207–210, 2002.
[21] D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath, "Docking and scoring in virtual screening for drug discovery", *Nat Rev Drug Discov*, vol. 3, pp. 935–949, 2004.
[22] S. B. Smith, W. Dampier, A. Tozeren, J. R. Brown, M. Magid-Slav, "Identification of common biological pathaways and drug targets across multiple respiratory viruses based on human host gene expression analysis", *Plos One*, vol. 7, pp. e33174, 2012.
[23] C. S. Greene, B. F. Voight, "Pathway and network-based strategies to translate genetic dicoveries into effective therapies", *Hum Mol Genet*, vol. 25, pp. R94–R98, 2016.
[24] Z. Wu, Y. Wang, L. Chen, "Network-based drug repositioning", *Mol Biosyst*, vol. 9, no. 6, pp. 1268–1281, 2013.
[25] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, J, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information", *Nat Commun*, vol. 8, no. 1, pp. 573, 2017.
[26] Y. Wang, S. Chen, N. Deng, Y. Wang, "Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data", *Plos One*, vol. 8, no. 11, pp. e78518, 2013.
[27] F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato, D. Greco, "Drug repositioning: a machine-learning approach through data integration.", *J Cheminform*, vol. 5, no. 1, pp. 30, 2013.
[28] M. Zitnik, V. Janjic, C. Larminie, B. Zupan, N. Przulj, "Discovering disease-disease associations by fusing systems-level molecular data", *Sci Rep*, vol. 3, pp. 3202, 2013.
[29] S. Alaimo, R. Giugno, A. Pulvirenti, "ncPred: ncRNA-disease association prediction through tripartite network-based inference", *Frontiers in Bioengineering and Biotechnology*, vol 2, pp. 71, 2014.
[30] C. Ding, T. Li, W. Peng, H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering", *Proc SIGKDD ACM*, pp. 126-135, 2006.
[31] H. Kim, H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis", *Bioinformatics*, vol. 23, no. 12, pp. 1495-1502, 2007.
[32] F.G. Germain, G.J. Mysore, "Stopping criteria for non-negative matrix factorization based supervised and semi-supervised source separation", *IEEE Signal Process Lett*, vol. 21, no. 10, pp. 1284–1288, 2014.
[33] S. Wild, J. Curry, A. Dougherty, "Improving non-negative matrix factorizations through structured initialization", *Pattern Recognit*, vol. 37, no. 11, pp. 2217–2232, 2004.
[34] C.J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization", *IEEE Trans Neural Netw*, vol. 18, no. 6, pp. 1589–1596, 2007.
[35] A. Čopar, B. Zupan, M. Zitnik, "Fast optimization of non-negative matrix tri-factorization", *PLoS One*, vol. 14, no. 6, pp. e0217994, 2019.
[36] X. Zhou, M.C.J. Kao, W.H. Wong, "Transitive functional annotation by shortest-path analysis of gene expression data," *Proc Natl Acad Sci*, vol. 99, no. 20, pp. 12783–12788, 2002.
[37] D.S Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets", *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D901–D906, 2008.
[38] D. Croft, G. O'kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, "Reactome: a database of reactions, pathways and biological processes", *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D691–D697, 2010.
[39] Y.H. Li, C.Y. Yu, X.X. Li, P. Zhang, J. Tang, Q. Yang, T. Fu, X. Zhang, X. Cui, G. Tu, "Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics", *Nucleic Acids Res*, vol. 46, no. D1, pp. D1121–D1127, 2017.
[40] R. Oughtred, C. Stark, B. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, "The BioGRID interaction database: 2019 update", *Nucleic Acids Res*, vol. 47, no. D1, pp. D529–D541, 2019.
[41] M.T. Lam, S. Coppola, O.H. Krumbach, G. Prencipe, A. Insalaco, C. Cifaldi, S. Martinelli, "A novel disorder involving dyshematopoiesis, inflammation, and HLH due to aberrant CDC42 function", *J Exp Med*, vol. 216, no. 12, pp. 2778–2799, 2019.
[42] M. Carmignani, P. Boscolo, N. Castellino, "Metabolic fate and cardiovascular effects of arsenic in rats and rabbits chronically exposed to trivalent and pentavalent arsenic", *Arch Toxicol Suppl*, pp. 452-455, 1985.
[43] O.N Bae, K.M Lim, J.Y Noh, S.M Chung, S.H. Kim, J.H. Chung, "Trivalent methylated arsenical-induced phosphatidylserine exposure and apoptosis in platelets may lead to increased thrombus formation", *Toxicol Appl Pharmacol*, vol. 239, no. 2, pp. 144–153, 2009.
[44] G.M. O'Kane, N.B. Leighl, "Systemic therapy of lung cancer CNS metastases using molecularly targeted agents and immune checkpoint inhibitors", *CNS Drugs*, vol. 32, no. 6, pp. 527–542, 2018.
[45] N. M. Farber, S. Perez-Lloret, E. R. Gamzu, "Design and development of a novel supportive care product for the treatment of sialorrhea in Parkinson's disease", *Curr Top Med Chem*, vol. 15, no. 10, pp. 939–954, 2015.
[46] E. Leung, F. Mitchelson, "Modification by hexamethonium of the muscarinic receptor blocking activity of pancuronium and homatropine in isolated tissues of the guinea-pig", *Eur J Pharmacol*, vol. 80, no. 1, pp. 11–17, 1982.
[47] D. Dunner, J. Myers, A. Khan, D. Avery, D. Ishiki, R. Pyke, "Adinazolam—a new antidepressant: findings of a placebo-controlled, double-blind study in outpatients with major depression", *J Clin Psychopharmacol*, vol. 7, no. 3, pp. 170–172, 1987.
[48] K. Venkatakrishnan, L.L. Von Moltke, S.X. Duan, J.C. Fleishaker, R.I. Shader, D.J. Greenblatt, "Kinetic characterization and identification of the enzymes responsible for the hepatic biotransformation of adinazolam and N-desmethyladinazolam in man", *J Pharm Pharmacol*, vol. 50, no. 3, pp. 265–274, 1998.
[49] M. Chebib, G.A. Johnston, "The 'ABC'of GABA receptors: a brief review", *Clin Exp Pharmacol Physiol*, vol. 26, no. 11, pp. 937–940, 1999.
[50] F.X. Cheng, C. Liu, J. Jiang , W. Q. Lu, W.H. Li, G.X. Liu, "Prediction of drug-target interactions and drug repositioning via network-based inference", *PLoS Comput Biol*, vol. 8, no. 5, pp. 1–12, 2012.
[51] Z. Wu, W. Li, G. Liu, Y. Tang, "Network-Based Methods for Prediction of Drug-Target Interactions", *Front in Pharmacol*, vol. 9, no. 1134, pp. 1–14, 2018.
[52] P. Pantziarka, V. Sukhatme, G. Bouche, L. Meheus, V.P. Sukhatme, "Repurposing Drugs in Oncology (ReDO)—itraconazole as an anti-cancer agent", *Ecancermedicalscience*, vol. 9, pp. 521, 2015.
[53] P. Pantziarka, V. Sukhatme, G. Bouche, L. Meheus, V.P. Sukhatme,"Repurposing Drugs in Oncology (ReDO)—diclofenac as an anti-cancer agent", *Ecancermedicalscience*, vol. 10, pp. 610, 2016.
[54] K.Zhao, H.C. So, "Drug repositioning for schizophrenia and depression/anxiety disorders: A machine learning approach leveraging expression data", *IEEE J Biomed Health Inform*, vol. 23, no. 3, pp. 1304–1315, 2018.