# Leveraging Non-negative Matrix Tri-Factorization and Knowledge-Based Embeddings for Drug Repurposing: an Application to Parkinson's Disease


Letizia Messa*
Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano
Milan, Italy

Carolina Testa*
Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano
Milan, Italy

Stephana Carelli
Center of Functional Genomics and Rare Diseases, Buzzi Children's Hospital
Milan, Italy

Federica Rey
Pediatric Clinical Research Center "Fondazione Romeo ed Enrica Invernizzi", DIBIC, Università degli Studi di Milano
Milan, Italy

Cristina Cereda
Center of Functional Genomics and Rare Diseases, Buzzi Children's Hospital
Milan, Italy

Manuela Teresa Raimondi
Department of Chemistry, Materials and Chemical Engineering "Giulio Natta", Politecnico di Milano
Milan, Italy

Stefano Ceri
Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano
Milan, Italy

Pietro Pinoli†
pietro.pinoli@polimi.it
Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano
Milano, Italy



## ABSTRACT

Drug repurposing, which involves using already approved drugs for new clinical targets, represents a cost-effective alternative to the development of new drugs. In this study, we introduce an innovative computational strategy, which uses Non-negative Matrix Tri-Factorization (NMTF) to generate vector embeddings of given sizes for drugs and drug targets; vector embeddings are then employed to generate predictions for drug repurposing using conventional classifiers, like random forest, logistic regression, and multi-layer perceptron.

Our approach leverages the NMTF method within a new approach to classification, named two-tower architecture, which is effective in solving complex tasks, such as the optimal prediction of targets for already approved drugs. This approach produces robust models, with AUROC reaching 0.90, which outperform traditional NMTF. We evaluate our method in the context of Parkinson's Disease; within the newly revealed drug-target predictions, we highlight compounds that exhibit potential in mitigating neurodegeneration, thereby revealing a potentially useful drug in relationships with a well-identified target.



## CCS CONCEPTS

• **Computing methodologies** → **Non-negative matrix factorization**; **Supervised learning by classification**; **Learning latent representations**; • **Applied computing** → **Computational biology**.

## KEYWORDS

Non-negative Matrix Tri-Factorization, Embeddings, Drug Repurposing, Predictors, Parkinson's Disease

**ACM Reference Format:**
Letizia Messa, Carolina Testa, Stephana Carelli, Federica Rey, Cristina Cereda, Manuela Teresa Raimondi, Stefano Ceri, and Pietro Pinoli. 2023. Leveraging Non-negative Matrix Tri-Factorization and Knowledge-Based Embeddings for Drug Repurposing: an Application to Parkinson's Disease. In *2023 10th International Conference on Biomedical and Bioinformatics Engineering (ICBBE 2023), November 09–12, 2023, Kyoto, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3637732.3637783


*Both authors contributed equally to this research.
†Corresponding author: pietro.pinoli@polimi.it




*ICBBE 2023, November 09–12, 2023, Kyoto, Japan*



ACM ISBN 979-8-4007-0834-3/23/11
https://doi.org/10.1145/3637732.3637783


## 1 INTRODUCTION

In the landscape of drug development, the quest for innovative and effective treatments has traditionally involved the painstaking process of discovering and developing entirely new compounds.

However, this approach is not only time-consuming but also financially burdensome, with the average cost of bringing a novel drug to market exceeding billions of dollars and spanning over a decade. Moreover, the vast majority of new molecules fail in early-stage clinical trials. Recent studies estimated that only 14% of compounds initially identified as potential drugs get approved by FDA [38].

Drug repurposing offers a novel and cost-effective alternative to the conventional drug development paradigm. It involves the identification of already approved drugs, initially developed for one medical condition, to treat different conditions, with significantly less time and investment. The repurposing of a compound may be performed either experimentally or computationally; however, computational tools can rapidly and efficiently test a large number of potential drugs and biological targets at a reasonable cost [24, 26].

We previously demonstrated the efficacy of the Non-negative Matrix Tri-Factorization (NMTF), a method that allows exploiting both data integration and machine learning, to various tasks in computational drug repurposing, including prediction of novel therapeutic indications [4], unknown drug targets [5], drug synergism [25], anticancer drug sensitivity [34], and synthetic lethal gene pairs [33]. NMTF takes a multipartite graph as input, which is represented as a set of association matrices between different node classes. By factorizing and reconstructing these matrices, NMTF can infer missing connections.

The capability of NMTF to process multipartite graphs enables the easy integration of multiple and heterogeneous data sources. Indeed, our previous research showed that the performance of NMTF in predicting new associations between a set of nodes X (e.g., drugs) and a set of nodes Y (e.g., targets) can be significantly improved when either X or Y is associated with side data (e.g., extracted from knowledge bases or derived from experiments). The association with side data, however, presents some limitations. First, as NMTF is typically trained using incremental update rules that consider all layers of the multipartite graph at each iteration, dealing with very large input graphs (in terms of number of layers) can dilute the information and lead to a significant drop in performance. Second, NMTF is a linear method - as it minimizes the Frobenius norm of the difference between the original and reconstructed matrices, which is linear in the model's parameters. However complex prediction tasks may benefit of non-linear classifiers.

In this study, we propose a novel approach, by embedding the NMTF method in a *two-tower architecture*, i.e. a novel machine learning model that learns two representations of the data: one for the drugs and one for the targets. The two representations are then merged and passed to a classifier, to generate a final prediction. In this setup, NMTF is used to create two embeddings, i.e. vector representations of drugs and targets, starting from two multipartite graph independently representing them, thus integrating and fusing information from heterogeneous sources. The embeddings are then used to train a common classifier to generate predictions. While NMTF has already been used for embedding generation [39], to the best of our knowledge, it has not been used before to generate embeddings from multipartite graphs.

## 2 RELATED WORK

Among most commonly used computational methods employed for drug repurposing, we find Machine Learning (ML) and Network Models. Machine Learning techniques applied in this context encompass logistic regression, support vector machine (SVM), random forest, neural networks (NN) and deep learning (DL) [12, 23]. There is a reported instance of two ML approaches based on similarity, both exploiting logistic regression. The first is PREDICT [9] wherein drug-drug similarity has been integrated with disease-disease similarity, serving as features in the prediction of drugs with similar properties for similar diseases using logistic regression. On the other hand, SPACE [18] also utilized logistic regression to forecast the therapeutic chemical class of a drug by integrating data from multiple sources. A SVM classifier has been employed by Napolitano et al. [22] to predict drug therapeutic categories of FDA-approved compounds. Particularly, they amalgamated three drug-drug similarity datasets, founded on gene expression signatures, chemical structures, and molecular targets, into a unified drug similarity matrix which served as kernel to train the multi-class SVM. Similarly, Wang et al. [36] defined a kernel function to correlate drugs with diseases integrating molecular structure, molecular activity, and phenotype data, then training the SVM classifier to computationally predict novel drug-disease interactions. The fusion of cell line genomics and drug chemical structures has been used to construct a feed-forward perceptron neural network model and a random forest regression model by Menden et al. [21] to predict cancer responses to drug treatments. Several other studies have explored the use of deep neural networks for drug repurposing and the identification of novel therapeutic indications [1, 2, 11, 29], collectively highlighting the potential of deep learning in this application.

In network models, entities (e.g drugs, genes, proteins, diseases, etc.) are represented by the nodes of the network while the edges symbolize the connections among them, representing a highly effective approach for modeling biological and biomedical objects, as well as for capturing their interactions and relationships. These methods allow the construction of heterogeneous networks by integrating entities and relationships from different data sources, unveiling previously unknown or concealed drug-disease connections [12, 23]. Yamanishi et al. [40] proposed a bipartite graph supervised learning model that leverages protein-protein interaction data, drug chemical structure information, and drug-target interaction networks to predict diverse drug-target interaction classes. Kinnings et al. [14] built a drug-drug network, representing drugs as nodes and connecting them based on drug chemical structure data and drug-target interaction similarity to uncover communities of drugs, ultimately unveiling therapeutic potentials and novel indications for existing drugs. On another front, Hu and Agarwal [10] used microarray gene expression profiles to construct a disease-drug network, coming up with a model which effectively identified drug repositioning opportunities and potential drug side effects. Lastly, Li and Lu [15] introduced a novel bipartite graph model that infers drug-target indications by considering drug pairwise similarity, combining drug chemical structure information with drug-target interactions.

## 3 METHODS

In this Section, we first introduce the Non-Negative Matrix Factorization (NMTF) method for predicting novel links in a bipartite graph. We then extend the method to multipartite graphs and show the update rules that are used to compute the NMTF in the most general case. Finally, we describe how we used NMTF to create embeddings for the two-tower architecture that we used for prediction.

### 3.1 Non-Negative Matrix Factorization (NMTF)

Let's define a bipartite graph as a graph $\mathcal{G} = \langle N, E \rangle$ where $N = L \cup R$, such that $L \cap R = \emptyset$, is a set of nodes and $E \subseteq L \times R$ is the a set of edges. Thus, a bipartite graph is a graph where the nodes can be divided into two disjoint groups, such that no nodes in the same group are connected by an edge. A bipartite graph graph can be encoded as its association matrix $X_{LR} \in \{0,1\}^{|L|\times|R|}$, where $X_{LR}[i,j] = 1$ if and only if the node $i \in L$ is connected to the node $j \in R$ (i.e., $(i,j) \in E$).

The NMTF can be applied to the $X_{LR}$ matrix in order to infer missing links between. Using iterative updated rules, NMTF decomposes the association matrix in three lower rank positive matrices $U \in \mathbb{R}_{\geq 0}^{|I|\times k_r}$, $S \in \mathbb{R}_{\geq 0}^{k_r \times k_l}$, and $V \in \mathbb{R}_{\geq 0}^{|J|\times k_l}$, with $k_r < L$ and $k_l < R$ such that the Frobenius norm of the difference between the original matrix and the product of the three

$$\mathcal{L}(\mathcal{G}|k_r, k_l) = |X_{LR} - USV^\top|^2_{Fro}$$

is minimized.

Then, the three matrices are used to compute the approximated matrix

$$\tilde{X}_{LR} = USV^\top.$$

Finally, fixed a threshold $\tau$, we predict a connection between the node $i \in L$ to the node $j \in R$ if and only if $\tilde{X}_{LR}[i,j] > \tau$. This technique, which is also known as *matrix completion*, is a popular choice for various fields including recommender systems, signal processing and image denoising.

### 3.2 NMTF for multipartite graphs

The NMTF method can be easily extended from bipartite graphs to multipartite graphs. For the sake of clarity, but without any loss of generality, we here present the extension in the context of predicting novel associations between nodes in two sets, L and R, where both L and R are connected to additional sets of nodes. We call this auxiliary set of nodes *side data*. An example of this graph topology is shown in Figure 1.

We can view a multipartite graph as a composition of bipartite graphs, which can be encoded as a set of association matrices, one for each bipartite graph. The extension of NMTF to multipartite graphs is then straightforward. For each association matrix $X_{IJ}$ connecting nodes of a set $I$ to nodes of a set $J$, NMTF finds three lower rank positive matrices $U_I$, $S_{IJ}$ and $V_J$ such that they minimize the objective function:

$$\mathcal{L}(\mathcal{G}|\Theta) = \sum_{X_{IJ}\in\mathcal{G}} |X_{IJ} - U_I S_{IJ} V_J^\top|^2_{Fro},$$

where $\Theta$ is the set of all the hyperparameters, i.e., the $k_I$ and $k_J$. It is clear that such a formulation would result in a set of independently factorized matrices, in which the data integration would play no role. In order to take advantage of all the information in the multipartite graph in a holistic way, we need to impose some additional constraints. In particular, whenever a chain is present in the graph such that the nodes of a set $I$ are connected to the nodes of a set $J$ and the nodes of $J$ are connected to the nodes of a third set $Z$ (e.g., the sub-graph connecting $L1$, $L$ and $R$ in Figure 1), then the $V_J$ factor matrix of $X_{IJ}$ has to be equal to the $U_J$ factor matrix of $X_{JZ}$.

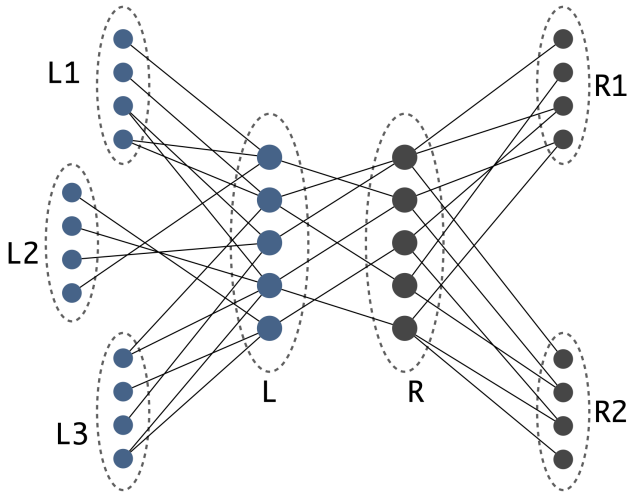


**Figure 1: Example of a multipartite graph. The bipartite graph connecting nodes in $L$ with nodes in $R$ is extended by associating side data (i.e., links to other node sets) to both sides. Notice that links between nodes of the same set are not allowed.**

Under this settings, it if possible to compute the set of factor matrices that minimize the objective function starting from a random initialization and iteratively applying the following update rules:

$$U_I \leftarrow U_I \odot \frac{\sum_Q X_{IQ} V_Q S_{IQ}^\top + \sum_Q X_{QI}^\top U_Q S_{QI}}{\sum_Q U_I S_{IQ} V_Q^\top V_Q S_{IQ}^\top + \sum_Q U_I S_{QI}^\top U_Q^\top U_Q S_{QI}}$$

$$V_J \leftarrow V_J \odot \frac{\sum_Q X_{QJ}^\top U_Q S_{QJ} + \sum_Q X_{JQ} V_Q S_{JQ}^\top}{\sum_Q V_J S_{QJ}^\top U_Q^\top U_Q S_{QJ} + \sum_Q V_J S_{JQ} V_Q^\top V_Q S_{JQ}^\top}$$

$$S_{IJ} \leftarrow S_{IJ} \odot \frac{U_I^\top X_{IJ} V_J}{U_I^\top U_I S_{IJ} V_J^\top V_J}$$

where $\odot$ and $\div$ stand for the element-wise multiplication and division, respectively. Usually those rules are iterated until the relative improvement of the objective function $\mathcal{L}$ between two consecutive iterations goes below a threshold $\alpha$.

### 3.3 NMTF for Embeddings Generation

Once we have presented the NMTF method for multipartite graph, we can exploit it to generate knowledge-based embeddings of elements of a set $D$ (e.g., genes) that are connected to side data from $N$ sets $S1, S2, \ldots, SN$ (e.g., terms in a curated vocabulary or ontology such as GO and KEGG).

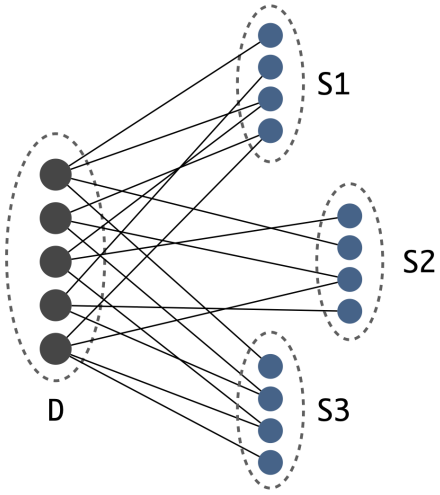


**Figure 2: Multipartite graph for knowledge-based embeddings. Each node of the set $D$ is enriched with connection to side data from 3 different datasets, $S1$, $S2$, and $S3$.**

The first step is to represent the information as a multipartite graph with the topology shown in Figure 2; the entities of interest $D$ are connected to the $N$ side dataset. No further connection are present (i.e., intra-set or between side datasets). Then, the graph is encoded as the corresponding set of $N$ association matrices $X_{D,S1}, X_{D,S2}, \ldots, X_{D,SN}$. Those matrices are jointly factorized by NMTF. For each $X_{D,Si}$ we obtain three matrices $U_D$, $S_{D,Si}$, and $V_{Si}$. Note that all the factorizations of the $N$ association matrices share the same left matrix $U_D$ due to the constraints introduced in Section 3.2. The $U_D$ matrix, which has a dimension of $|D| \times k_D$, represents the embeddings of the $D$ elements in the vector space $\mathbb{R}^{k_D}$.

### 3.4 Two-Tower Architecture

With reference to Figure 1, suppose our goal is to predict novel association between elements in $L$ and elements in $R$. To this end, we implemented a classification system based on a two-tower architecture, as depicted in Figure 3.

As mentioned in the Introduction a two-tower method is a machine learning model composed of two *towers* connected by a classifier. It is especially useful for tasks that involve predicting the interaction between two different types of entities. Each tower learns a feature representation of the entities it is responsible for. The two representations are then merged together (typically by simply concatenating them) and fed into a classifier.

In our proposal, we use NMTF to learn vectorial dense representations of the entities in each tower, namely the embeddings. Then, (s sub set of) the cross product of these two sets of embeddings

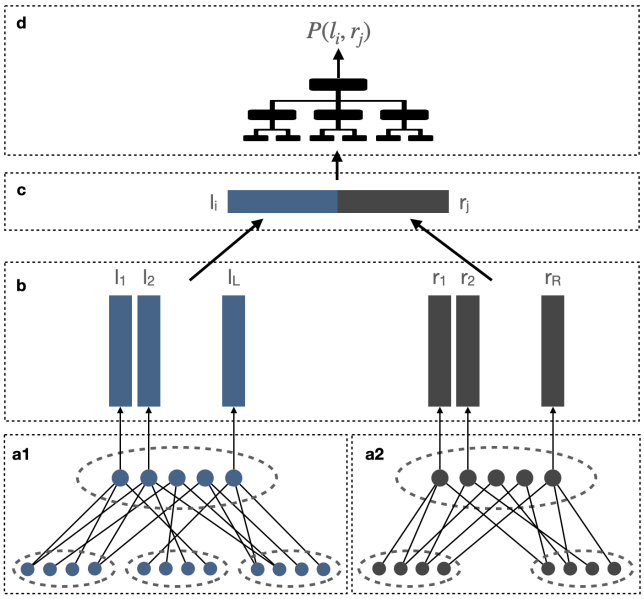


**Figure 3: Two-tower architecture. The following steps compose the pipeline: (a1, a2) NMTF is applied independently to the two entities; (b) the vectorial embeddings are produced; (c) the embeddings are concatenated; (d) finally, a classifier is applied.**

is then used as input to a classifier. The fact that only a subset of the cross product is needed for the classifier is significant. This implies that the latent encoding can be learned on a very large set of entities, even larger than the set of entities for which we want to infer novel associations.

We have experimented with three different classifiers: logistic regression, random forest, and multi-layer perceptron.

## 4 APPLICATION TO PARKINSON'S DISEASE

We applied our proposed method to the task of drug repurposing, specifically predicting novel potential targets for approved drugs. In this scenario, the two towers learn representations for drugs and genes, respectively.

We then produce a cross-product of the two sets of embeddings and concatenate the two vectors. This new set is used to train a classifier, which is then exploited to predict novel interactions.

In order to evaluate our approach in qualitative as well as quantitative terms, we used a dataset that is restricted to genes involved in the Parkinson's Disease.

### 4.1 Data extraction and integration

We first built two primary networks, namely *drug network* and *gene network*. Thereafter, to assess our method we compared it with the traditional NMTF, and thus we combined the two networks in a multipartite network. A detailed description of how the networks were built is given below.

*4.1.1* ***Drug network.*** The drug network (left part of Figure 4) was populated with data retrieved from DrugBank [37]; we considered a

set $D$ of 1,101 unique approved drugs. Each drug in $D$ is associated with terms in the set $A$ from the fourth level of the *Anatomical Therapeutic Chemical* (ATC) codes, a system that categorizes drugs into different groups at five different levels according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. The set $A$ comprises 530 different ATC codes. We also expanded the network with *categories* and *classification*, which were retrieved from DrugBank database. Thus, the drugs in $D$ are also associated with the 2,140 categories and with the 232 classifications in the two sets $C$ and $L$, respectively.

Thus, the drug multiparite network, shown on the left of Figure 4, can be encoded into the following three binary matrices:

- $X_{DA}$, such that $X_{DA}[d, a] = 1$ iff the drug $d$ is associated with the ATC code $a$;
- $X_{DC}$, such that $X_{DC}[d, c] = 1$ iff the drug $d$ is belongs to the category $c$;
- $X_{DL}$, such that $X_{DL}[d, l] = 1$ iff the drug $d$ is included the drug classification $l$;

*4.1.2* ***Gene network.*** The gene network was constructed by focusing on genes potentially associated with PD. We accomplished this by using publicly available RNA-Seq data from human post-mortem brain tissues of both PD and healthy control patients, sourced from the Gene Expression Omnibus (GEO) repository. Following the application of standardized data processing, we first conducted a differential expression analysis between the PD and healthy control groups, resulting in the identification of 3,670 genes with an adjusted p-value < 0.05. These genes intersected with Gene Ontology (GO) annotations retrieved from Molecular Signature Database (MSigDB) [16] formed the first matrix of this network. We obtained a matrix of $G$ = 3,049 unique genes and $GO$ = 4,047 unique GO terms. Then, we expanded the network and we considered also the Curated gene sets from MSigDB, which includes genes involved in the same pathways, curated from biomedical literature: BioCarta [27], KEGG [13], PID [28], Reactome [7] and WikiPathways [19] databases.

The gene multiparite graph, shown on the right of Figure 4, can be encoded as the following two binary matrices:

- $X_{GGO}$ such that $X_{GGO}[g, t] = 1$ iff the gene $g$ is annotated to the GO term $t$;
- $X_{GC}$ such that $X_{GC}[g, c] = 1$ iff the gene $g$ belongs to the curated set $c$.

The dimensions of the resulting binary matrices both for drug and gene networks are reported in Table 1. These matrices served as the input data for the two towers of our architecture which creates the embeddings, subsequently employed in the classification tasks.

To assess and compare our improved approach with the traditional NMTF method, we added the bipartite graph between the drugs and the genes; according to the information retrieved from DrugBank [1] we build the graph such that an edge exists between a drug $d$ and a gene $g$ only if $g$ is a target of $d$. Thus we finally obtained the whole multipartite network depicted in Figure 4. In DrugBank, a gene being a target of a drug means that the drug interacts with or affects the gene's protein product. This interaction can involve direct binding, indirect regulation, or modulation of gene expression.

[1] https://go.drugbank.com

**Table 1: Dimensions of the matrices that encode the multipartite graph used for testing the two-tower model against NMTF**

| | Two-tower | | NMTF | |
|---|---|---|---|---|
| **Matrix** | **#Row** | **#Columns** | **#Rows** | **#Columns** |
| $X_{DT}$ | | | 464 | 309 |
| $X_{DA}$ | 1,101 | 530 | 464 | 310 |
| $X_{DC}$ | 1,101 | 2,140 | 464 | 1,528 |
| $X_{DL}$ | 1,101 | 232 | 464 | 156 |
| $X_{GGO}$ | 3,049 | 4,047 | 309 | 1,299 |
| $X_{GC}$ | 3,049 | 3,798 | 309 | 1,181 |

Starting from this network we pruned it in order to make it ideal for NMTF; specifically, we only considered drugs that are linked to at least one target gene in the set. This resulted in the creation of the $X_{DT}$ matrix, comprising 464 drugs connected to 309 genes. On cascade, we also pruned the data from the side sets, both for genes and drugs. The dimensions of the matrices incorporated in this network and used to compare traditional NMTF with our novel approach are reported in Table 1.

To keep the comparison fair, the $X_{DT}$ matrix containing drugs and genes was used to filter the embeddings used in the classification tasks, allowing for a direct comparison of performance with the traditional NMTF method.

All these matrices were considered to produce the outcomes described in the following Section 4.2.

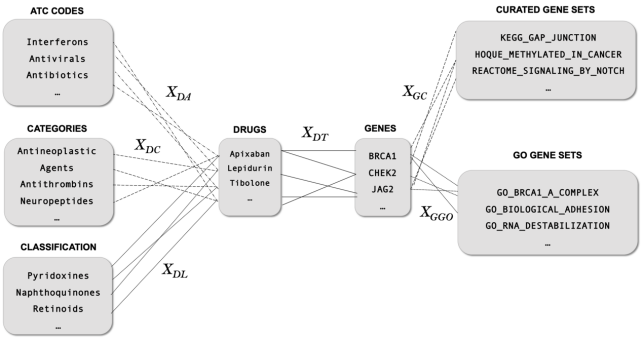


**Figure 4: Graphical representation of the multipartite graph used in our test. The edges between DRUGS and GENES are present only for the experiments with NMTF; however its information is used for training the classifier of the two-tower model. The embedding are computed on the two subgraph starting from DRUGS and GENES, respectively.**

## 4.2 Results

*4.2.1* ***Embeddings generation via NMTF and drug-target prediction.*** To derive embeddings for the drug and the gene networks, we exploited the NMTF decomposition step and we generated two factorized matrices, denoted as $U_{drugs}$ and $U_{genes}$. Moreover, to

comprehensively explore potential variations arising from changes in the embedding dimension, we executed NMTF considering different factorization ranks, $k$ = 10, 25 and 50. Then, we cross-matched the factorized matrices, combining all rows of the $U_{drugs}$ matrix with all rows of the $U_{genes}$ matrix, and subsequently we used the final cross-matched matrix as input matrix for RF, LR and MLPC across different values of k while employing two key techniques: K-Fold Cross Validation with 5 folds and hyper-parameter tuning. We evaluated the different predictors performances by calculating the Area Under the Receiver Operating Characteristic (AUROC). This approach led to interesting outcomes, underscoring its effectiveness. For K-Fold Cross Validation with $k$ = *10*, our analysis yielded intriguing findings. RF model demonstrated a solid AUROC of 0.87, indicating its ability to make robust predictions. In contrast, LR model achieved a lower AUROC of 0.64, suggesting a relatively weaker performance compared to RF. On the other hand, MLP model showcased impressive predictive power with an AUROC of 0.85, even under the constraint of a smaller k value. Moving on to Hyperparameter Tuning with $k$ = 10, RF and LR slightly improved with an AUROC of 0.90 and 0.87 respectively, while LR maintained an AUROC of 0.64.

For K-Fold Cross Validation with $k$ = 25, we observed promising results with RF achieving an AUROC of 0.88, while LR exhibited an AUROC of 0.71. Impressively, MLP also achieved an AUROC of 0.88. Moving on to Hyperparameter Tuning with $k$ = 25, RF displayed improved performance with an AUROC of 0.90, while LR maintained an AUROC of 0.71, and MLP slightly decreased to an AUROC of 0.87. An illustrative depiction of AUROC values for k=25 after Hyperparameter Tuning is presented in Figure 5 providing a visual representation of the achieved predictive performances.

Expanding our investigation to K-Fold Cross Validation with a larger $k$ = 50, we continued to witness strong predictive capabilities, with RF achieving an AUROC of 0.89, LR improving to an AUROC of 0.74, and MLP maintaining its high AUROC of 0.88. Lastly, under Hyperparameter Tuning with $k$ = 50, RF once again outperformed other models with an AUROC of 0.90, while LR matched its performance with an AUROC of 0.74, and MLP retained its AUROC of 0.87. These results underscore the robustness of this method which couples NMTF with predictors models to drug-target prediction.

*4.2.2* ***Traditional NMTF.*** To validate the effectiveness of our proposed approach, we conducted a comparative analysis with the traditional NMTF method. Specifically, we applied NMTF to the integrated network shown in Figure 4B and validated through a mask validation strategy: we defined a mask M of the same dimension of the input matrix, which randomly hides, i.e., set to 0 regardless of their real values, the 20% of the entire elements of the $X_{DT}$. Thus, we applied the NMTF decomposition substituting $X_{DT}$ with the masked matrix and we evaluated how well it could reconstruct the real matrix my measuring the AUROC. The same procedure was repeated 25 times, every time changing the mask. As explained in section 4.2.1, we tested different configurations with k = 25 and 50.

First, we applied NMTF to the solely input association matrix $X_{DT}$ obtaining an AUROC of 0.6756 for $k$ = 25 and 0.6803 for $k$ = 50. We subsequently expanded the network by adding first the additional information related to genes (i.e., $X_{GGO}$, $X_{GC}$) and then to drugs (i.e., $X_{DA}$, $X_{DC}$, $X_{DL}$). The inclusion of solely gene-related information had a negligible impact on the performance, with AUROC values of 0.6801 for $k$ = 25 and 0.6788 for $k$ = 50. The most significant improvement was observed when we introduced solely drug-related information, resulting in AUROC values of 0.7381 for $k$ = 25 and 0.758 for $k$ = 50. However, as expected, the integration of all information into the network, comprising the principal association matrix $X_{DT}$ along with both drug and gene information, did not lead to better predictive performance, as shown in Figure 6. This is proved by a decrease of almost the 12% in the value of the AUROC metric, from 0.6756 to 0.5526 for $k$ = 25 and from 0.6803 to 0.557 for $k$ = 50. This suggests that our implemented method, which incorporates all available information, succeeded in enhancing the model's predictive capabilities, unlike the standalone NMTF method.

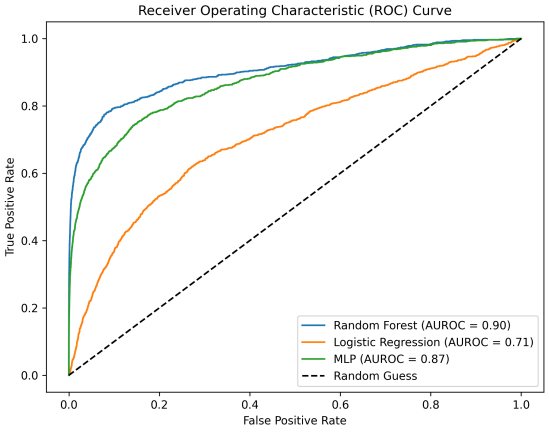


**Figure 5: AUROC representation of the implemented method for drug-target prediction with RF, LR and MLP after hyper-parameter tuning for $k$ = 25**

*4.2.3* ***Newly identified drug-target predictions.*** Given the remarkable performances demonstrated by RF and MLP models, we have expanded our investigation to evaluate the novel predictions generated by the classifiers with regard to potential treatments for PD. Among these predictions, we have identified several promising compounds: *Rutin*, *Fingolimod*, *Bexarotene*, *Statins* (*Pravastatin* and *Atorvastatin*), and *Capsaicin*. These already approved compounds were evaluated for their interesting mechanism of action and potential neuroprotective effects. Newly identified drugs, along with their associated targets, are reported in Table 2.

*Rutin*, a natural compound, has demonstrated potential in protecting the brain from neurodegenerative conditions by reducing neuroinflammation, enhancing antioxidant activity, and influencing gene expression related to PD [6]. *Fingolimod*, currently used in the treatment of Multiple Sclerosis, has shown neuroprotective effects in various animal models of neurodegenerative diseases, improving disease-related symptoms such as cognition and motor abilities, along with reduction of neuroinflammatory markers [3]. *Bexarotene*, originally used to treat cutaneous T-cell lymphoma, has displayed promise in the treatment of Alzheimer's disease (AD), since it has exerted different protective mechanisms, including inhibiting beta

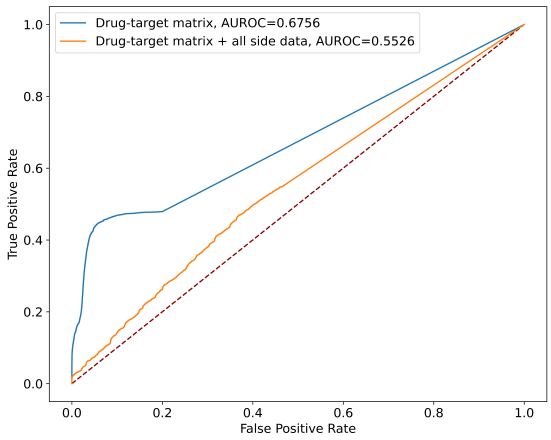


**Figure 6: AUROC representation of traditional NMTF drug-target prediction performances using a random mask on the solely association input matrix of $X_{DT}$ and on the integrated network for $k = 25$.**

**Table 2: New drug-target predictions made by our proposed approach**

| DrugBank ID | Drug Name | Target |
|---|---|---|
| DB01698 | Rutin | ALB |
| DB08868 | Fingolimod | SLC12A5 |
| DB00307 | Bexarotene | MAP3K3 |
| DB00175 | Pravastatin | CAMK2G |
| DB01076 | Atorvastatin | CS |
| DB06774 | Capsaicin | ALB |

amyloid production and aggregation, regulating neuroinflammation, and improving cognitive functions [17]. *Pravastatin* and *Atorvastatin* have demonstrated neuroprotective effects in an in vitro induced model of neuroinflammation and neurodegeneration [20]. Lastly, *Capsaicin*, studied in relation to AD, has shown potential in preventing the aggregation of beta amyloid [35]. Notably, while these drugs rank highly in our new predictions, their traditional NMTF scores are significantly low, making their identification unlikely without our enhanced approach.

Additionally, it is important to emphasize that some of the identified targets hold direct relevance to the disease or have previously been associated with the predicted drugs. For instance, *Rutin* and *Capsaicin* are predicted to interact with albumin (ALB), a protein whose increased serum levels have been associated with improved cognitive functions in PD [32]. Similarly, *Fingolimod* is predicted to interact with SLC12A5, a gene belonging to the SLC transporter family that plays a crucial role in the regulation of neurotransmission related to PD [8].

These findings further strenghten the validity of the proposed method in making predictions, suggesting significant connections between the predicted already approved drugs and the identified targets, with potential therapeutic implications for the disease.

## 5 DISCUSSION AND CONCLUSIONS

In this study, we introduced an innovative enhancement to the NMTF method to drive drug repurposing. Our approach involved leveraging both data integration and machine learning techniques, specifically the NMTF method and various machine learning predictors, to predict drug-target interactions. We designed a novel architecture for building predictions, named two-tower architecture, which employs the NMTF method in a new, unconventional way, and demonstrated the effectiveness of our approach. We showed that NMTF successfully extracted meaningful features and relationships from the data, highlighting the potential for drug repurposing.

By combining NMTF embeddings with machine learning predictors like random forest, logistic regression, and multi-layer perceptron, we achieved promising results. Our models exhibited strong predictive capabilities with AUROC values ranging from 0.71 to 0.90, depending on the specific machine learning algorithm and the choice of factorization rank $k$. We compared our approach to traditional NMTF method on the same network configuration. While traditional NMTF achieved AUROC values between 0.5526 and 0.6803, our enhanced method consistently outperformed it.

Among the newly identified drug-target predictions, we uncovered compounds that showed promise in protecting against neurodegeneration, including *Rutin*, *Fingolimod*, *Bexarotene*, *Pravastatin*, *Atorvastatin* and *Capsaicin* which exhibited the ability to reduce neuroinflammation and influence mechanisms protecting the brain [3, 6, 17, 20, 35]. A comment is required for statins since they are widely used in cardiovascular diseases and our study emphasized the importance of considering individual statins for their specific effects within the central nervous system in future studies.

Moreover, among the newly identified targets, particular interest goes to albumin, which has been predicted to interact with *Rutin* and *Capsaicin*. Indeed, higher levels of albumin have been associated with improvement in cognitive functions and protective role in motor impairments in PD patients [32]. Furthermore, recent studies shed a light on the association between *Rutin* and albumin as they showed that their interaction could improve *Rutin* efficacy [30, 31]. Nonetheless, it is essential to highlight that these results are based on computational predictions and require comprehensive both preclinical and clinical validation to confirm their efficacy in practice.

The primary drawbacks of our method involve handling incomplete and noisy data, which may adversely impact the generation of embeddings, thus reducing their accuracy and reliability. This may lead to potential misinterpretation or errors in downstream analyses and difficulty in achieving the desired results.

In conclusion, our study presents a novel and powerful approach to building predictions. This advancement unlocks numerous future extensions, including the potential for triplet-based predictions involving drug-protein-disease interactions or drug-disease-patient profiles, which are paramount for precision medicine. Furthermore, it allows the integration of complex data types that cannot be easily represented as graphs, like amino acid sequences of target proteins. In such cases, an embedding of the additional information may be effectively computed by means of Recurrent or Convolutional

Neural Network and then concatenated to the knowledge-based embedding obtained by NMTF. In summary, our tower-based architecture is a powerful and general concept, that adapts to many application domains.

## ACKNOWLEDGMENTS

This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program. C.T. and M.T.R. are funded by the ERC AdG 101053122, BEACONSANDEGG. L.M. is funded by GR-2019-12368701.

## REFERENCES

[1] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. 2016. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics* 13, 7 (2016), 2524–2530.

[2] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. 2017. Low data drug discovery with one-shot learning. *ACS central science* 3, 4 (2017), 283–293.

[3] Pablo Bascuñana, Luisa Möhle, Mirjam Brackhan, and Jens Pahnke. 2020. Fingolimod as a treatment in neurologic disorders beyond multiple sclerosis. *Drugs in R&D* 20 (2020), 197–207.

[4] Gaia Ceddia, Pietro Pinoli, Stefano Ceri, and Marco Masseroli. 2019. Non-negative matrix tri-factorization for data integration and network-based drug repositioning. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 1–7.

[5] Gaia Ceddia, Pietro Pinoli, Stefano Ceri, and Marco Masseroli. 2020. Matrix factorization-based technique for drug repurposing predictions. *IEEE journal of biomedical and health informatics* 24, 11 (2020), 3162–3172.

[6] Adaze Bijou Enogieru, William Haylett, Donavon Charles Hiss, Soraya Bardien, and Okobi Eko Ekpo. 2018. Rutin as a potent antioxidant: Implications for neurodegenerative disorders. *Oxidative Medicine and Cellular Longevity* (2018).

[7] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. 2018. The reactome pathway knowledgebase. *Nucleic acids research* 46, D1 (2018), D649–D655.

[8] Yajing Gan, Zihan Wei, Chao Liu, Guoyan Li, Yan Feng, and Yanchun Deng. 2022. Solute carrier transporter disease and developmental and epileptic encephalopathy. *Frontiers in Neurology* 13 (2022), 1013903.

[9] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* 7, 1 (2011), 496.

[10] Guanghui Hu and Pankaj Agarwal. 2009. Human disease-drug network based on genomic expression profiles. *PloS one* 4, 8 (2009), e6536.

[11] ShanShan Hu, Chenglin Zhang, Peng Chen, Pengying Gu, Jun Zhang, and Bing Wang. 2019. Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. *BMC bioinformatics* 20 (2019), 1–12.

[12] Tamer N Jarada, Jon G Rokne, and Reda Alhajj. 2020. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of cheminformatics* 12, 1 (2020), 1–23.

[13] Minoru Kanehisa and Susumu Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28, 1 (2000), 27–30.

[14] Sarah L Kinnings, Nina Liu, Nancy Buchmeier, Peter J Tonge, Lei Xie, and Philip E Bourne. 2009. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS computational biology* 5, 7 (2009), e1000423.

[15] Jiao Li and Zhiyong Lu. 2012. A new method for computational drug repositioning using drug pairwise similarity. In *2012 IEEE international conference on bioinformatics and biomedicine*. IEEE, 1–4.

[16] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 12 (2011), 1739–1740.

[17] Yangtao Liu, Pengwei Wang, Guofang Jin, Peijie Shi, Yonghui Zhao, Jiayi Guo, Yaling Yin, Qianhang Shao, Peng Li, and Pengfei Yang. 2023. The novel function of bexarotene for neurological diseases. *Ageing Research Reviews* (2023), 102021.

[18] Zhongyang Liu, Feifei Guo, Jiangyong Gu, Yong Wang, Yang Li, Dan Wang, Liang Lu, Dong Li, and Fuchu He. 2015. Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources. *Bioinformatics* 31, 11 (2015), 1788–1795.

[19] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, et al. 2021. WikiPathways: connecting communities. *Nucleic acids research* 49, D1 (2021), D613–D621.

[20] AJ McFarland, AK Davey, CM McDermott, GD Grant, J Lewohl, and S Anoopkumar-Dukie. 2018. Differences in statin associated neuroprotection corresponds with either decreased production of IL-1$\beta$ or TNF-$\alpha$ in an in vitro model of neuroinflammation-induced neurodegeneration. *Toxicology and Applied Pharmacology* 344 (2018), 56–73.

[21] Michael P Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H Benes, Pedro J Ballester, and Julio Saez-Rodriguez. 2013. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one* 8, 4 (2013), e61318.

[22] Francesco Napolitano, Yan Zhao, Vânia M Moreira, Roberto Tagliaferri, Juha Kere, Mauro D'Amato, and Dario Greco. 2013. Drug repositioning: a machine-learning approach through data integration. *Journal of cheminformatics* 5, 1 (2013), 1–9.

[23] Kyungsoo Park. 2019. A review of computational drug repurposing. *Translational and clinical pharmacology* 27, 2 (2019), 59–63.

[24] Vineela Parvathaneni, Nishant S Kulkarni, Aaron Muth, and Vivek Gupta. 2019. Drug repurposing: a promising tool to accelerate the drug discovery process. *Drug discovery today* 24, 10 (2019), 2076–2085.

[25] Pietro Pinoli, Gaia Ceddia, Stefano Ceri, and Marco Masseroli. 2021. Predicting drug synergism by means of non-negative matrix tri-factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19, 4 (2021), 1956–1967.

[26] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. 2019. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery* 18, 1 (2019), 41–58.

[27] Andrew D Rouillard, Gregory W Gundersen, Nicolas F Fernandez, Zichen Wang, Caroline D Monteiro, Michael G McDermott, and Avi Ma'ayan. 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016 (2016).

[28] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. 2009. PID: the pathway interaction database. *Nucleic acids research* 37, suppl_1 (2009), D674–D679.

[29] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. 2018. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* 4, 1 (2018), 120–131.

[30] Priti Sengupta, Prasenjit Mondal, Souryadeep Mukherjee, Sarmishtha Chanda, and Adity Bose. 2020. Rutin-serum albumin interaction in different media and its effective dose selection in radiation-induced cytotoxicity on human blood cells. *Journal of Herbal Medicine* 21 (2020), 100322.

[31] Priti Sengupta, Pinki Saha Sardar, Pritam Roy, Swagata Dasgupta, and Adity Bose. 2018. Investigation on the interaction of Rutin with serum albumins: Insights from spectroscopic and molecular docking techniques. *Journal of Photochemistry and Photobiology B: Biology* 183 (2018), 101–110.

[32] Shujun Sun, Yiyong Wen, and Yandeng Li. 2022. Serum albumin, cognitive function, motor impairment, and survival prognosis in Parkinson disease. *Medicine* 101, 37 (2022).

[33] Carolina Testa, Sara Pidò, Emanuela Jacchetti, Manuela T. Raimondi, Stefano Ceri, and Pietro Pinoli. 2023. Inference of Synthetically Lethal Pairs of Genes Involved in Metastatic Processes via Non-Negative Matrix Tri-Factorization. In *International Conference on Bioinformatics and Biomedical Technology*. ACM.

[34] Carolina Testa, Sara Pidò, and Pietro Pinoli. 2021. A Non-Negative Matrix Tri-Factorization Based Method for Predicting Antitumor Drug Sensitivity. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer, 94–104.

[35] Jun Wang, Bin-Lu Sun, Yang Xiang, Ding-Yuan Tian, Chi Zhu, Wei-Wei Li, Yu-Hui Liu, Xian-Le Bu, Lin-Lin Shen, Wang-Sheng Jin, et al. 2020. Capsaicin consumption reduces brain amyloid-beta generation and attenuates Alzheimer's disease-type pathology and cognitive deficits in APP/PS1 mice. *Translational psychiatry* 10, 1 (2020), 230.

[36] Yongcui Wang, Shilong Chen, Naiyang Deng, and Yong Wang. 2013. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PloS one* 8, 11 (2013), e78518.

[37] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* 34, suppl_1 (2006), D668–D672.

[38] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. 2018. Estimation of clinical trial success rates and related parameters. *Biostatistics* 20, 2 (01 2018), 273–286.

[39] A Xenos, N Malod-Dognin, S Milinković, and N Pržulj. 2021. Linear functional organization of the omic embedding space. *Bioinformatics* 37, 21 (07 2021), 3839–3847. arXiv:https://academic.oup.com/bioinformatics/article-pdf/37/21/3839/50337488/btab487.pdf https://doi.org/10.1093/bioinformatics/btab487

[40] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, 13 (2008), i232–i240.