Network-based Drug Ranking and Repositioning with respect to DrugBank Therapeutic Categories

Matteo Re, and Giorgio Valentini

Abstract—Drug repositioning is a challenging computational problem involving the integration of heterogeneous sources of biomolecular data and the design of label ranking algorithms able to exploit the overall topology of the underlying pharmacological network. In the context of the drug repositioning problem, the inference step is usually performed into an inhomogeneous similarity space induced by the relationships existing between drugs and a second type of entity (e.g. disease, target, ligand set), thus making unfeasible a drug ranking within a homogeneous pharmacological space. To deal with this challenging problem, we designed a general framework based on bipartite network projections by which homogeneous pharmacological networks can be constructed and integrated from heterogeneous and complementary sources of chemical, biomolecular and clinical information. Moreover, we present a novel algorithm based on kernelized score functions that adopts both local and global learning strategies to effectively rank drugs in the integrated pharmacological space. We applied the proposed methods to a novel semi-supervised drug ranking problem: prioritizing drugs in integrated bio-chemical networks according to specific DrugBank therapeutic categories. Detailed experiments with more than 80 DrugBank therapeutic categories involving about 1300 FDA approved drugs show the effectiveness of the proposed approach.

Index Terms—Drug ranking, drug repositioning, network integration, kernel functions, systems biology, graph nodes ranking

1 Introduction

The very conservative drug development strategy, which typically consists in the discovery of new therapeutics targets followed by a slow, costly and risky validation, results in a consistent increment in research and development spending [1]. On the contrary, repurposing already approved and marketed drugs can speed up their application to clinical practice, because in this way we can take advantage of existing rigorous testing required by the U.S. Food and Drug Administration (FDA) and other regulatory agencies. Drugs repurposing, also referred to as drugs repositioning, is less costly when compared to the results of traditional discovery efforts, which typically takes 10-15 years and upwards $1 billion, while revenues due to repurposed drugs can exceed billions [2].

Drug repositioning, i.e. the prediction of novel therapeutic indications for existing drugs, is a challenging problem in modern computational biology. Computational approaches for drug repositioning focused mainly on small-scale applications, such as the analysis of specific classes of drugs or drugs for specific diseases [3], [4], [5], [6]. Large-scale applications, involving a relatively large number of drugs and diseases, count only a few examples [7], [8], [9], [10].

Different computational tasks related to the the drug repositioning problem have been proposed, ranging from clustering drugs either considering their pharmacophore descriptors [3] or Connectivity Map-based networks [8], to prediction of drug-target interactions [11], [12], or drug-disease associations [13], [9] using supervised or semi-supervised approaches. While the clustering approach does not require "a priori" knowledge about drugs (but should in principle require the application of methods to assess the reliability of clustering results [14]), the latter approach requires that at least a partial labeling of the drugs is known in advance, but by exploiting the available "a priori" knowledge classical techniques to evaluate supervised algorithms can be applied to assess the prediction performances [15].

In the context of semi-supervised learning of network labeling, we propose a novel prediction task, i.e. the large-scale ranking of drugs with respect to DrugBank therapeutic categories[16]. We chose DrugBank categories since their associations to drugs are manually curated using medical literature such as PubMed, e-Therapeutics (http://www.e-therapeutics.ca) and STAT!Ref (AHFS) (http://online.statref.com), and because "at present, there is not a comprehensive and systematic representation of known drugs indications that would enable a fine-scale delineation of types of drug-disease relationships" [17]. The ranking of drugs for each DrugBank therapeutic category (TC) can allow the choice of top ranked "false positive" drugs as natural candidates for drug repositioning, while a pure classification approach cannot provide such preferential candidates.

Several works showed that network integration plays a central role in different molecular systems biology

• *Matteo Re, and Giorgio Valentini are with DI, Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39, Milano, Italy, e-mail: {re,valentini}@di.unimi.it*

problems [18], ranging from disease genes discovery [19] to gene function prediction [20] and drug repositioning [21]. Unfortunately, in the context of drug repositioning, the inference step is usually performed into an inhomogeneous similarity space induced by the relationships existing between drugs and a second type of entity (e.g. disease, target, ligand set), thus making unfeasible a drug ranking within homogeneous pharmacological spaces. To deal with this problem, we propose a general framework based on bipartite networks projections for the construction of homogeneous pharmacological spaces, by which, starting from heterogeneous networks of data involving interactions between two different sets of nodes (e.g. drug-protein targets, drug-pathways, drug-side effects), we can obtain homogeneous drug-drug networks that implicitly embed previous interactions into homogeneous pharmacological spaces. The nature of these network-structured projected spaces allows the application of prediction algorithms to homogeneous drug-drug networks that no longer represent a physical reality, but informational constructs related to the pharmacological similarity between drugs. Bipartite graphs have been just successfully applied to integrate different "omics" data in yeast molecular networks [22], and to identify global relationships between different diseases [23]: in this work we apply them in a more general framework to improve the integration of different sources of chemical, biomolecular and clinical sources of information in the context of the drug ranking and repositioning problem.

Most of the node label ranking algorithms proposed for the analysis of biomolecular networks exploit local or global learning strategies to properly rank nodes, according to the biological property under investigation [24], [25], [18]. In this work we propose a very fast semi-supervised network method that combines both local and global learning strategies to exploit both "local" similarities between drugs and "global" similarities embedded in the topology of the pharmacological network, following an approach that we very recently successfully applied to the gene function prediction problem [26] and to discover genes related to diseases [27]. Indeed our proposed *Score Functions* adopt both local learning strategies based on a generalized notion of distance in a universal reproducing kernel Hilbert space, and global learning strategies based on the choice of random walk kernels kernels to exploit the overall topology of the underlying pharmacological network.

We evaluated the proposed approach by integrating three pharmacological similarity spaces accounting, respectively, for chemical similarity, drug-targets interaction similarity and drug-chemicals similarity, in order to rank a curated set of U.S. Food and Drug Administration (FDA) approved drugs according to the DrugBank therapeutic categories.

The paper is structured as follows: in Section 2 we present ψ*NetPro*, Pharmacological Spaces Integration based on Networks Projections, a method to construct homogeneous pharmacological spaces from heterogeneous bipartite networks, and moreover we discuss how to construct an integrated pharmacological space by a progressive combination of different projected networks obtained from heterogeneous sources of data. In Section 3 we introduce the drug ranking methods applied in this work, including our proposed *Score Functions based on Kernelized Random Walks*. In the successive section we provide a large set of experiments involving 81 DrugBank therapeutic categories to show the effectiveness of ψ*NetPro* and of the proposed drug ranking methods. The conclusions summarize the main results and the possible developments of this work.

2 ψ*NetPro*, Pharmacological Spaces Integration based on Networks Projections

We propose ψ*NetPro*, Pharmacological Spaces Integration based on Networks Projections, a general approach to construct and integrate different pharmacological similarity spaces capturing different pharmacological characteristics of drugs. In Section 2.1 we introduce the bipartite network projection method to construct homogeneous spaces from inhomogeneous spaces represented though bipartite networks, and in Section 2.2 we show how to construct and integrate different pharmacological spaces using different sources of chemical, biomolecular and pharmacological data.

2.1 Bipartite networks projections

Many relationships naturally come in a bipartite setting. Common examples are authors that write articles, people that visit web pages and many others. In computational biology this kind of relationships can be used, just to cite a few, for the investigation of the interactions between proteins and genes or between enzymes and metabolites using networks composed by two types of nodes.

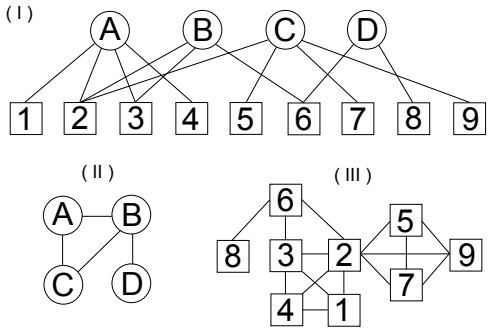


Fig. 1. A toy bipartite network and its unipartite projections. (I) Original bipartite network. Top ($\top$) nodes are labeled by letters and bottom ($\bot$) nodes are labeled by numbers. (II) Projection in the $\top$ domain. (III) Projection in the $\bot$ domain.

Bipartite or two-mode networks (Fig. 1 I) can be naturally modeled as bipartite graphs. A bipartite graph

is a triplet $B = (\top, \bot, E)$ where $\top$ is the set of top nodes, $\bot$ is the set of bottom nodes, $\top \cap \bot = \emptyset$ and $E \subseteq \top \times \bot$ is the set of edges. The difference with unipartite graphs consists in the fact that the nodes lie in two disjoint sets, and the edges are always between a node of one set and a node of the other set. Bipartite networks can be projected into one-mode networks (composed by a single type of nodes). More precisely the $\top$-projection of $B = (\top, \bot, E)$ is the graph $B_\top = (V^\top, E_\top)$ in which two nodes $u, v \in \top$ are connected if they share at least one neighbour $x \in \bot$ in the original bipartite graph B. The set of edges in the projected unipartite graph $B_\top$ is thus:

$$E_\top = \{(u, v), \exists x \in \bot : (u, x) \in E \wedge (v, x) \in E\} \quad (1)$$

The $\bot$-projection $G_\bot$ is defined dually (Fig. 1). This operation is commonly referred to as "binary mode projection" and is suitable for the induction of a homogeneous similarity space between vertices $v \in \top$ (or $\bot$) in the bipartite graph B (Fig. 1). In the following, for the sake of simplicity, we represent projected graph $B_\top = (V^\top, E_\top)$ as $G = (V, E)$ and its adjacency matrix as $\boldsymbol{W}$.

The binary mode projection produces one-mode networks containing binary edges, but more complex projection schemes can assign edge weights according to the edge weights in the bipartite two-mode network, or to the number of shared neighbors, or to the number of nodes which each shared neighbor is connected to [CITANEWMAN2001]. In our experiments we adopted the binary projection technique, since the bipartite drug-target data downloaded from the DrugBank database are unweighted, and for homogeneity we applied a binary projection also to the other considered data (see Section 2.2 for more details). The bipartite network projection scheme may induce different pharmacological similarity spaces depending on the nature of the bipartite network (e.g. drug-protein or drug-chemicals interaction bipartite networks), but the projected networks correspond to homogeneous pharmacological spaces representing different notions of induced pharmacological similarity between drugs.

2.2 Construction and integration of pharmacological networks

Once projected onto one-mode networks $G = (V, E)$, the drugs similarity spaces induced from the bipartite graphs can be combined using appropriate network integration methods and proper normalization techniques. For instance, we adopted the normalized graph Laplacian $\boldsymbol{L}$ [28] to make comparable the pharmacological networks $G =< V, E >$ represented through the corresponding symmetric adjacency matrices $\boldsymbol{W}$:

$$\boldsymbol{L} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}} \quad (2)$$

where $\boldsymbol{D}$ is a diagonal matrix with elements $d_{ii} = \sum_j w_{ij}$, $\boldsymbol{I}$ is the identity matrix and w_{ij} are the elements of the matrix $\boldsymbol{W}$.

In our setting we integrated multiple networks with a simple technique that assures a high coverage of the drugs included in the integrated pharmacological network, without penalizing drugs for which a specific source of data is unavailable. More precisely, given a set of n pharmacological networks $G^d =< V^d, E^d >, 1 \leq d \leq n$, constructed through appropriate bipartite graph projections, the integrated pharmacological network $\bar{G} =< \bar{V}, \bar{E} >$, with $\bar{V} = \bigcup_d V^d$ and $\bar{E} \subseteq \bigcup_d E^d$, can be derived by averaging the normalized edge weights only when data for the corresponding pair of drugs is actually available. In other words, if w_{ij}^d represents the weight of the edge $(v_i, v_j) \in E^d$, the weight $\bar{w}_{ij}$ of the edge $(v_i, v_j) \in \bar{E}$ is computed as follows:

$$\bar{w}_{ij} = \frac{1}{|D(i,j)|} \sum_{d \in D(i,j)} w_{ij}^d \quad (3)$$

where $D(i, j) = \{d | v_i \in V^d \wedge v_j \in V^d\}$.

It is worth noting that other network integration methods may lead to better results (e.g. weighted integrated networks that take into account the information content of each source of data), but we applied this simple approach only to show the feasibility and effectiveness of the proposed overall approach.

We constructed three pharmacological similarity networks reflecting the pairwise chemical structure similarity between drugs ($N_{structSim}$), the similarity between drugs derived from common protein targets ($N_{drugTarget}$) and the pairwise similarity from chemical-chemical interactions ($N_{drugChem}$) between the considered drugs and other chemicals involved in their pharmacological activity.

2.2.1 Chemical and pharmacological data bases.

We constructed three pharmacological similarity networks reflecting the pairwise chemical structure similarity between drugs ($N_{structSim}$), the similarity between drugs derived from common protein targets ($N_{drugTarget}$) and the pairwise similarity from chemical-chemical interactions ($N_{drugChem}$) between the considered drugs and other chemicals involved in their pharmacological activity. They have been constructed using data collected from the DrugBank [16] and STITCH [29] public databases.

DrugBank is a unique bioinformatics/chemoinformatics resource that combines detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information. In the current release DrugBank contains detailed information about 6707 drug entries including 1436 FDA-approved small molecule drugs. In order to construct a highly reliable drugs set we selected from DrugBank the largest set of FDA approved drugs targeting at least one FDA approved target. This led to the definition of a collection composed by 1253 drugs.

STITCH integrates data distributed over many databases. For instance, the chemical-chemical interaction networks stored in STITCH includes information

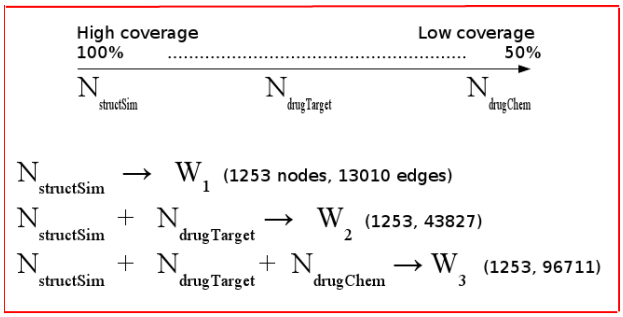


Fig. 2. Progressive integration of network data.

about the impact of genetic variation on drug response and from the Comparative Toxicogenomics Database (which contains more than 8500 direct chemical-disease relationships), thus ensuring the existence of drug-drug relationships induced by common genetics and/or toxicogenomics disease-association profiles [30], [31].

2.2.2 Constructing pharmacological spaces from different sources of data.

The construction of $N_{structSim}$ is based on the direct computation of the structural chemical similarities between each pair of drugs, while for the other pharmacological spaces we applied the projection techniques described in Section 2.1. More precisely it has been obtained by computing the Tanimoto similarity scores between each pair of drugs in the reference set [32]. The scores were obtained by comparing the simplified molecular input line entry specification (SMILES) annotations contained in DrugBank entries [33]. The obtained adjacency matrix was then converted into a binary matrix by thresholding the similarity scores according to the procedure reported in [11] (threshold $t = 0.5$).

The second considered similarity space, $N_{drugTarget}$, was obtained by creating a bipartite network between the drugs and all the FDA approved targets, according to the information stored in DrugBank. Once constructed, this network has been projected onto a one mode network and processed according to the procedures described in Section 2.1.

The third pharmacological similarity space ($N_{drugChem}$) has been constructed by processing the chemical-chemical interactions stored in the STITCH 2.0 database [34]. This dataset is expected to be informative because these interactions are obtained by considering many sources of information (i.e. metabolic pathways, binding experiments, phenotypic effects and drug-target relationships). In STITCH each predicted drug-chemical interaction is stored along with a quality score. The original bipartite graph encoding these interactions has been sparsified by removing all the interactions with score below 0.7. This threshold was empirically selected by testing all the values ranging from 0.5 to 0.9 at steps of 0.1 and searching for the larger value able to cover, after the binary mode projection, at least half of the drugs in our reference set (the vertices of the $N_{structSim}$ network). The thresholding led to a final coverage of 50% of the drugs in our reference set.

2.2.3 Progressive integration of pharmacological networks.

The computed pharmacological networks have been progressively integrated to enrich the encoded drug-drug relationships with different and complementary sources of information while preserving a high-coverage of drugs for large scale drugs repositioning. To this end we considered at first the $N_{structSim}$ space alone (that is the space with the highest drug coverage), then we progressively integrated the other two pharmacological spaces characterized by a lower coverage, that is respectively $N_{drugTarget}$ and $N_{drugChem}$. These progressively enriched pharmacological networks have been represented through the corresponding adjacency matrices $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{W}_3$, where the numeric index indicates the number of different integrated pharmacological networks (Fig. 2). Despite the number of nodes/drugs in the three networks is the same (1253), our "progressive integration" strategy yields to a significant increment in the number of the edges, that grow from 13010, to 43827 and 96711 respectively in $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{W}_3$. This correspond to a roughly 7.5 folds increment in the network density $\delta(G) = \frac{2m}{n(n-1)}$ where m is the number of existing edges and n is the number of nodes. The network densities of the pharmacological spaces involved in our experiments are 0.01658, 0.05587 and 0.12329 for $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{W}_3$ respectively. Fig. 3 provides a visual clue of the integrated $\boldsymbol{W}_3$ network.

3 Drug Ranking Methods

Drug ranking can be formalized as a semi-supervised node label ranking problem on a graph. Let $G =< V, E >$ be an undirected weighted graph, representing a pharmacological network $\boldsymbol{W}$, and let $V_C \subset V$ be a subset of drugs belonging to a priori known therapeutic category C, the *drug ranking problem* consists in finding a score function $S : V \longrightarrow \mathbb{R}^+$, by which we can directly rank vertices according to their likelihood to belong to a specific therapeutic category C: the higher the score, the higher the likelihood that a drug belongs to C. *Drug ranking* can be seen as a "one-class" semi-supervised learning problem on pharmacological networks $\boldsymbol{W}$, since we can exploit the labeling of the known positive vertices $v \in V_C$ belonging to the therapeutic category C, but also the similarity relationships between labeled or unlabeled vertices $v \in V \setminus V_C$.

In our experiments we compared results obtained with random walks and random walk with restart with our novel proposed method that can be interpreted as a kernelized extension of the classical random walks. As a baseline we applied a simple guilt-by-association-based method.

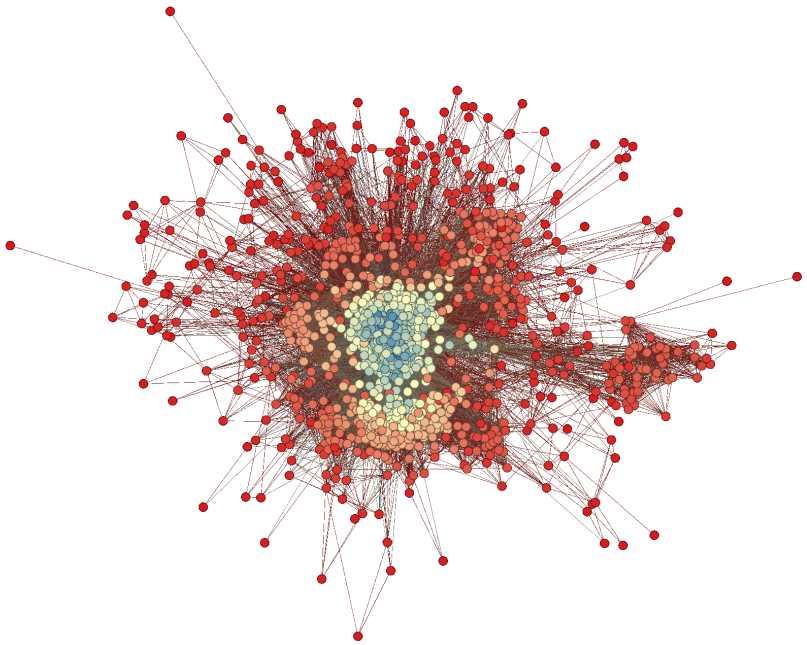


Fig. 3. Graph of the integrated $\boldsymbol{W}_3$ pharmacological network (1253 nodes and 96711 edges). Lighter nodes represent drugs with a higher number of connections (edges) with other drugs in the integrated pharmacological space.

3.1 Guilt by Association

Guilt by association (*GBA*) is a general biological principle by which a biomolecular entity that interacts or shares some features with another biomolecular entity can also shares some specific biological property. For instance, if a gene A shares an expression patterns or a genetic interaction with gene B and gene A is annotated for a given Gene Ontology (GO) term, it is likely that gene B can be annotated for the same term [35]. In computational biology this basic biological principle has been exploited to develop methods able to assign a given biological or molecular property on the basis of the labeling of neighborhoods in biomolecular networks [24], [36]. In the context of pharmacological networks (Section 2) we can assess the likelihood that a given drug belongs to a given Therapeutic Category C on the basis of the C-labeled drugs directly connected to the drug under study.

As a baseline, we implemented a simple version of the *GBA* approach, by which a score for each node/drug is computed by choosing the maximum of the weights $w_{ij} \in \boldsymbol{W}$ of the edges connecting the node v_i with positive labeled nodes $v_j \in V_C$ in the neighborhood $N(i)$ of v_i:

$$S(v_i, C) = max_{j \in N(i)} w_{ij} \tag{4}$$

where $N(i) = \{j | v_j \in V_C \wedge (v_i, v_j) \in E\}$.

3.2 Random Walks and Random Walks with Restart

Random walk (*RW*) algorithms [37] can capture not only relationships coming from direct neighborhoods between drugs, similarly to *guilt by association* methods, but also relationships coming from shared and more in general indirect neighbours between drugs. Indeed *RW* ranks drugs by exploring and exploiting the topology of the pharmacological network: random walks across the network are performed starting from a subset $V_C \subset V$ of drugs belonging to a specific therapeutic category C by using a transition probability matrix $\boldsymbol{Q} = \boldsymbol{D}^{-1}\boldsymbol{W}$, where $\boldsymbol{W}$ is the adjacency matrix, and $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $d_{ii} = \sum_j w_{ij}$. The elements q_{ij} of $\boldsymbol{Q}$ represent the probability of a random step from v_i to v_j. The initial probability of belonging to the set of drugs corresponding to a given therapeutic category can be set to $p_o = 1/|V_C|$ for the drugs $v \in V_C$ and to $p_o = 0$ for the drugs $v \in V \setminus V_C$: this represents the "a priori" knowledge about the membership of the drugs to a specific therapeutic category, and in principle these initial probabilities can be set to different values for each drug (if we dispose of "a priori' information detailed enough to justify this setting). Then *RW* adopts an iterative strategy to update the probability vector $\boldsymbol{p}_t$ of finding a "random walker" at step t in the nodes $v \in V$:

$$\boldsymbol{p}_{t+1} = \boldsymbol{Q}^T \boldsymbol{p}_t \tag{5}$$

The update (5) is iterated until convergence or can be stopped after a fixed number of steps if we would only like to partially explore the topology of the network. We could observe that the random walker could progressively "forget" the a priori information available for the therapeutic category C, by iteratively walking across the overall network. To avoid this problem, we can stop the *RW* algorithm after a few iterations, as outlined above, or we can apply the random walk with restart (*RWR*)

method: at each step the random walker can move to one of its neighbours or can restart from its initial condition with probability θ:

$$\boldsymbol{p}_{t+1} = (1-\theta)\boldsymbol{Q}^T\boldsymbol{p}_t + \theta\boldsymbol{p}_o \quad (6)$$

It can be shown that the stationary distribution of $\boldsymbol{p}$ in *RWR* is determined by the largest eigenvalue/eigenvector pair of the matrix $\boldsymbol{Q}' = [\theta\boldsymbol{I}+(1-\theta)\boldsymbol{Q}]$ obtained from (6), where $\boldsymbol{I}$ is the identity matrix, and values of $\boldsymbol{p}$ at convergence determine the ranking of the nodes [28]. With both *RW* and *RWR* methods at the steady state we can rank the vector $\boldsymbol{p}$ to prioritize drugs according to their likelihood to belong to the therapeutic category under study.

3.3 Score Functions based on Kernelized Random Walks

Random walks exploit the global topology of the network (Section 3.2), while *GBA* methods introduce simple, but effective local learning strategies to rank nodes according to the structure of their neighborhood. We propose a novel method that on the one hand generalizes the local learning strategy of *GBA* methods and on the other hand adopts a global learning strategy by embedding in a kernel function the random walking across the network.

More precisely, we can define a distance measure $D(v, V_C)$ between a drug $v \in V$ and the set of the drugs $x \in V_C$ in a reproducing kernel Hilbert space $\mathcal{H}$, according to a suitable mapping $\phi : V \to \mathcal{H}$. For instance, we can consider the minimum euclidean distance in the Hilbert space $\mathcal{H}$ between a drug $v \in V$ and the set of drugs V_C belonging to a specific therapeutic category:

$$D_{NN}(v, V_C) = \min_{x \in V_C} \left|\left|\phi(v) - \phi(x)\right|\right|^2 \quad (7)$$

By recalling that $< \phi(\cdot), \phi(\cdot) >= K(\cdot,\cdot)$, where $K : V \times V \to \mathbb{R}$ is a kernel function associated to the mapping ϕ, we can choose in principle any valid kernel, but in this context it is meaningful to use a *random walk kernel* [28] constructed from the adjacency matrices $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$, since it provides a similarity measure that takes into account direct and indirect relationships between drugs in the pharmacological space. The Gram matrix $\boldsymbol{K}$ associated to the one-step random walk kernel function $K(\cdot,\cdot)$ is obtained from the adjacency matrix $\boldsymbol{W}$ of the pharmacological network:

$$\boldsymbol{K} = (a-1)\boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}} \quad (8)$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{D}$ is the "degree" diagonal matrix with elements $d_{ii} = \sum_j w_{ij}$ and a is a value larger than 2. The *q-step random walk kernel* is a slight generalization of (8):

$$\boldsymbol{K}^q = [(a-1)\boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}]^q \quad (9)$$

where $q \geq 2$ is an integer representing the number of steps of the random walk across the graph and can be easily computed by adopting a recursive strategy:

$$\boldsymbol{K}^q = \boldsymbol{K}^{q-1}\boldsymbol{K} \quad (10)$$

When $q = 1$ it is simply the *one-step random walk kernel*, by which only the direct neighbours of each node are visited. By setting $q = 2$, the random walks consider also indirect neighbours, that is two nodes are similar if either they are directly connected or they share common nodes in their neighborhood. More in general, by setting $q > 2$ two vertices are considered similar if they are directly connected of if they are connected through a path including from 1 to $q-1$ intermediate vertices. In principle also very long paths could be considered, but this could introduce very remote similarities between genes, leading to behaviours similar to that of diffusion kernels [38]. The name of the kernel derives from the fact that (9) is up to scaling terms equivalent to a q-step random walk on the graph with random restarts, a well-known algorithm used for scoring web pages in the Google search engine [39].

By developing the square (7) we can derive the following similarity measure:

$$Sim_{NN}(v, V_C) = -\min_{x \in V_C} [K(v,v) - 2K(v,x) + K(x,x)] \quad (11)$$

By assuming an equal auto-similarity $K(x,x)$ for all $x \in V$, we can simplify (11), thus achieving the *nearest neighbours score* S_{NN}:

$$S_{NN}(v, V_C) = -\min_{x \in V_C} -2K(v,x) = 2\max_{x \in V_C} K(v,x) \quad (12)$$

It is easy to see that a different notion of distance based on the first k nearest-neighbours leads to the definition of the *k-nearest neighbours score* S_{kNN}:

$$S_{kNN}(v, V_C) = 2\sum_{x \in I_k(v)} K(v,x) \quad (13)$$

with $I_k(v) = \{x \in V_C | x \text{ is ranked in the first } k \text{ in } V_C\}$. In a similar way we can also derive the *average score* similarity measure S_{AV} based on the average distance D_{AV} with respect to to the set of drugs V_C belonging to the C therapeutic category:

$$S_{AV}(v, V_C) = \frac{2}{|V_C|}\sum_{x \in V_C} K(v,x) \quad (14)$$

It is worth noting that the S_{AV} score is similar to that recently proposed in the context of gene function prediction from synthetic lethality networks, and from this standpoint our approach can be viewed as an extension of the algorithm presented in [40]. By using the proposed kernelized score functions we can rank drugs with respect to their likelihood to belong to a given therapeutic category C simply by evaluating the random walk kernel. If the kernel matrix is computed in advance, the time complexity of the proposed algorithm is $\mathcal{O}(|V_C||V|)$, that is approximately linear with respect to the number of drugs when $|V_C| << |V|$.

4 Results and Discussion

We propose a novel learning problem in the context of drug ranking and repositioning: the prediction of the therapeutic category of drugs according to the annotations provided by DrugBank 3.0. The ψ*NetPro* construction and integration of the pharmacological networks $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{W}_3$ (Section 2) have been applied to predict the therapeutic category of drugs according to the annotations provided by DrugBank 3.0, by ranking nodes for each therapeutic category through the algorithms described in Section 3.

4.1 Experimental Setup

TABLE 1
DrugBank Therapeutic Categories (TC) with more than 15 drugs considered in the experiments. The first column reports the abbreviated name, the second the full DrugBank name and the third the cardinality of the TC.

Therapeutic categories with more than 15 drugs		
Abbreviated name	**Full DrugBank name name**	**Card.**
Adren.A.	Adrenergic_Agents	26
Adren.In.	Adrenergic_Uptake_Inhibitors	20
Adren.a.	Adrenergic_alpha.Agonists	23
Adren.b.	Adrenergic_beta.Antagonists	25
Analges.	Analgesics	40
Analg.Op.	Analgesics._Opioid	24
Anti.Aller.	Anti.Allergic_Agents	35
Anti.Arrh.	Anti.Arrhythmia_Agents	42
Anti.Bact.	Anti.Bacterial_Agents	103
Anti.HIV	Anti.HIV_Agents	22
Anti.Inf.A.	Anti.Infective_Agents	29
Anti.Inf.	Anti.Infectives	19
Anti.Ulcer	Anti.Ulcer_Agents	19
Anti.anx.	Anti.anxiety_Agents	35
Anti.infl.	Anti.inflammatory_Agents	49
Antiarr.A.	Antiarrhythmic_Agents	29
Anticonv.	Anticonvulsants	46
Antidysk.	Antidyskinetics	23
Antiemetics	Antiemetics	34
Antifungal	Antifungal_Agents	22
Antihist.	Antihistamines	24
Antihypert.	Antihypertensive_Agents	105
Antimetab.	Antimetabolites	26
Antineopl.	Antineoplastic_Agents	86
Antineopl.H.	Antineoplastic_Agents._Hormonal	18
Antipark.	Antiparkinson_Agents	27
Antipsyc.A.	Antipsychotic_Agents	39
Antipsyc.	Antipsychotics	27
Antiviral	Antiviral_Agents	25
Bronchodil.	Bronchodilator_Agents	33
Ca.Ch.Block.	Calcium_Channel_Blockers	22
Cephalosp.	Cephalosporins	30
Cycloox.Inh.	Cyclooxygenase_Inhibitors	24
Dietary.sup.	Dietary_supplement	47
Diuretics	Diuretics	23
Dopam..Ant.	Dopamine_Antagonists	29
Enzyme.Inh.	Enzyme_Inhibitors	64
GABA.Mod.	GABA_Modulators	26
Glucocort.	Glucocorticoids	21
Hist.H1.Ant.	Histamine_H1_Antagonists	28
Hypnot.Sed.	Hypnotics_and_Sedatives	41
Hypoglyc.	Hypoglycemic_Agents	22
Immunosup.	Immunosuppressive_Agents	20
Micronutr.	Micronutrient	45
Musc.Ant.	Muscarinic_Antagonists	23
Narcotics	Narcotics	22
Penicillins	Penicillins	20
Sympathol.	Sympatholytics	24
Sympathomim.	Sympathomimetics	32
Vasoconstr.	Vasoconstrictor_Agents	25
Vasodilator	Vasodilator_Agents	55

In order to obtain the therapeutic category labels we parsed the DrugBank entries belonging to our reference set (1253 FDA approved drugs, see Section 2.2.1) by extracting all the drug category annotations excluding the chemical categories (categories reflecting the chemical nature of the considered compounds). We then removed from our therapeutic categories set all the classes associated to less than 15 drugs obtaining 51 therapeutic classes, in order to exclude classes with too few positive examples to assure reliable predictions. The classes represented in this set are very broad in nature ranging, only to cite a few, from "Diuretics" to "Anti Bacterial Agents" and to "Antiparkinson Agents", and are characterized by a relatively high unbalance between labeled and unlabeled nodes (Table 1).

We compared the drug ranking methods described in Section 3, by using the pharmacological networks $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{W}_3$ (Section 2). While *GBA* and *RW* iterated till to convergence have no parameters, for *RWR* we run the algorithm with $\theta \in \{0.1, 0.3, 0.6, 0.9\}$, and we run also the *RW* algorithm with a limited number of iterations, by varying the number of steps $q \in \{1, 2, 3, 5, 10\}$. Also for the proposed score functions with random walk kernel we varied the number of steps $q \in \{1, 2, 3, 5, 10\}$. In our experiments we did not perform a fine tuning of the method's parameters for each class; we simply fixed the same parameters for all classes and chose the ones leading to the best results. It is worth noting that a fine tuning of the parameters for each class (e.g. by internal cross-validation) may lead to better overall results.

We evaluated the proposed ranking method by using a 5-folds cross validation scheme repeated 10 times. As the output of the proposed methods is a continuous score for each drug-therapeutic category pair, we computed the Area Under the ROC curve (AUC), and the precision at fixed recall levels by varying recall between 0.1 and 1 at 0.1 steps.

In Section 4.2 we present the compared AUC and precision at a given recall results averaged across the therapeutic classes, in Section 4.4 we report the AUC and precision at a fixed recall results for each therapeutic category, and Section 4.3 discusses the influence of the choice of the number of steps in Random Walk kernel score functions. Finally in Section 4.5 we discuss the effectiveness of the proposed methods when the cardinality of DrugBank therapeutic categories is very low, by presenting the drug ranking results for a subset of DrugBank classes with less than 15 annotated examples, and in Section 4.6 we report a preliminary analysis of the top ranked false positives as possible candidates for drug repositioning.

4.2 Average AUC and Precision at a Fixed Recall Results

Tab. 2 shows the AUC and precision at 40% recall averaged across the 51 DrugBank therapeutic classes with more than 15 drugs. For kernelized score functions, *RWR* and *RW* at fixed steps the parameters giving the best results are reported.

Independently of the considered methods, the average AUC and precision at 40% recall (P40R) increases as new pharmacological spaces are added: most of the increment is achieved when we integrate 2 pharmacological spaces ($\boldsymbol{W}_2$), but note that the apparently small increment

TABLE 2
Average AUC and precision at 40% recall across the DrugBank categories with more than 15 drugs.

Methods	**AUC**			**P40R**		
	W_1	W_2	W_3	W_1	W_2	W_3
S_{AV} 3 steps	0.8332	0.9233	**0.9372**	0.5330	**0.6497**	0.6931
S_{kNN} 2 steps k=31	**0.8373**	**0.9261**	0.9361	**0.5334**	0.6480	**0.7012**
S_{NN} 3 steps	0.8271	0.9067	0.9224	0.3803	0.4300	0.4653
$RWR\ \theta = 0.6$	0.8078	0.9203	0.9299	0.5238	0.6278	0.6839
RW 1 step	0.8175	0.9201	0.9272	0.4910	0.6240	0.6799
GBA	0.8027	0.9028	0.9095	0.3273	0.4127	0.4634
RW	0.6846	0.5780	0.5334	0.2224	0.0608	0.0366

obtained, e.g. by S_{kNN}, when we pass from 2 to 3 integrated pharmacological spaces is actually statistically significant according to the Wilcoxon paired signed rank test (*p-value*< 0.005). These results are also confirmed by the precision at fixed recall levels curves (Fig. 4): independently of the recall level and the ranking method considered precision with $\boldsymbol{W}_3$ is larger than precision with $\boldsymbol{W}_2$ and $\boldsymbol{W}_1$ pharmacological networks.

An exception is represented by the classical *RW* that deteriorates its performances when new sources of data are added (Tab. 2). Note that *RW* substantially fails in these ranking tasks, since just with $\boldsymbol{W}_1$ (i.e. considering only the raw chemical similarities between drugs) this method is significantly worse than all the other ones. This is likely to the fact that the random walk is performed until the convergence condition is reached, thus resulting in an exploration of too remote and not significant relationships between drugs, that introduce noise in the probabilities achieved at the steady state. Indeed both *RW* 1 step and *RWR* achieve largely better results, since they do not "forget" the initial conditions, by exploring only the direct neighborhood of each drug (*RW* 1 step) or restarting with a certain probability θ form the initial conditions (*RWR*).

The average AUC and P40R are always larger in S_{AV} and S_{kNN} with respect to the other compared methods (Tab. 2), and the differences across the therapeutic categories are always statistically significant ($p - value < 0.005$, Wilcoxon paired signed rank test) except for the AUC with $\boldsymbol{W}_1$ with respect to S_{NN}, and between S_{AV} and *RWR* and *RW* 1 step with $\boldsymbol{W}_2$. Quite surprisingly the simple *GBA* method achieves very good average results in terms of AUC, while with P40R (Tab. 2) and more in general with precision at fixed recall levels we observe a larger difference with respect to the other considered methods (Fig. 4). Note that a similar behaviour can be observed also in S_{NN}, even if quite always S_{NN} obtains significantly better results than *GBA* both in terms of AUC and P40R. This is not surprising since both the methods adopt a "nearest-neighbour" local learning strategy to compute the score associated to each gene (compare (4) and (12)), but S_{NN} embeds a random walk kernel that can exploit the overall topology of the network.

Summarizing, the integration of multiple sources of information into projected homogeneous pharmacological spaces plays a central role to significantly improve the ranking results. Moreover random walk kernel score functions and in particular S_{AV} and S_{kNN} achieve significantly better results than the other compared methods.

4.3 Influence of the Number of Steps in Random Walk Kernel Score Functions

Results of the previous section show that random walk kernel score functions and in particular S_{AV} and S_{kNN} achieve significantly better results than the other compared methods. To get more insights into the significance of the number of steps needed to effectively rank drugs in pharmacological networks, we compare in Tab. 3 the average AUC and P40R results of S_{AV} with 1, 2, 3, 5 and 10 steps random walk kernels. Values in boldface highlight the best average results in terms of AUC and P40R achieved with $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$. Interestingly enough, with most pharmacological networks, there is no a statistically significant difference between 3, 5 and 10 steps random walk kernels according to the Wilcoxon paired signed rank test, $p - value < 0.005$ (for instance, in terms of AUC with $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ and in terms of P40R with all the three pharmacological spaces). Recalling that 3 steps S_{AV} has been chosen as the best S_{AV} in terms of AUC (see Tab. 2), we can conclude that also increasing the number of steps, on the average, there is no decay of performance in terms of average AUC and P40R.

To gain more insights into the reasons underlying this learning behavior of kernelized score functions, we analyzed the number of wins for each therapeutic category both in terms of AUC and P40R among S_{AV} with 1, 2, 3, 5 and 10 steps random walk kernels, that is we counted how many times each k-steps random walk kernel achieved the maximal AUC or P40R (Fig. 5). We can observe that the "wins" are quite distributed across the random walk kernels with a different number of steps, especially if we consider the P40R (Fig. 5 (b)), while for the AUC (Fig. 5 (a)) we can observe a quite interesting "peak of wins" of the 10 steps random walk kernel with the full integrated $\boldsymbol{W}_3$ pharmacological space. These results show that, according to specific characteristics of each therapeutic class, different number of steps should

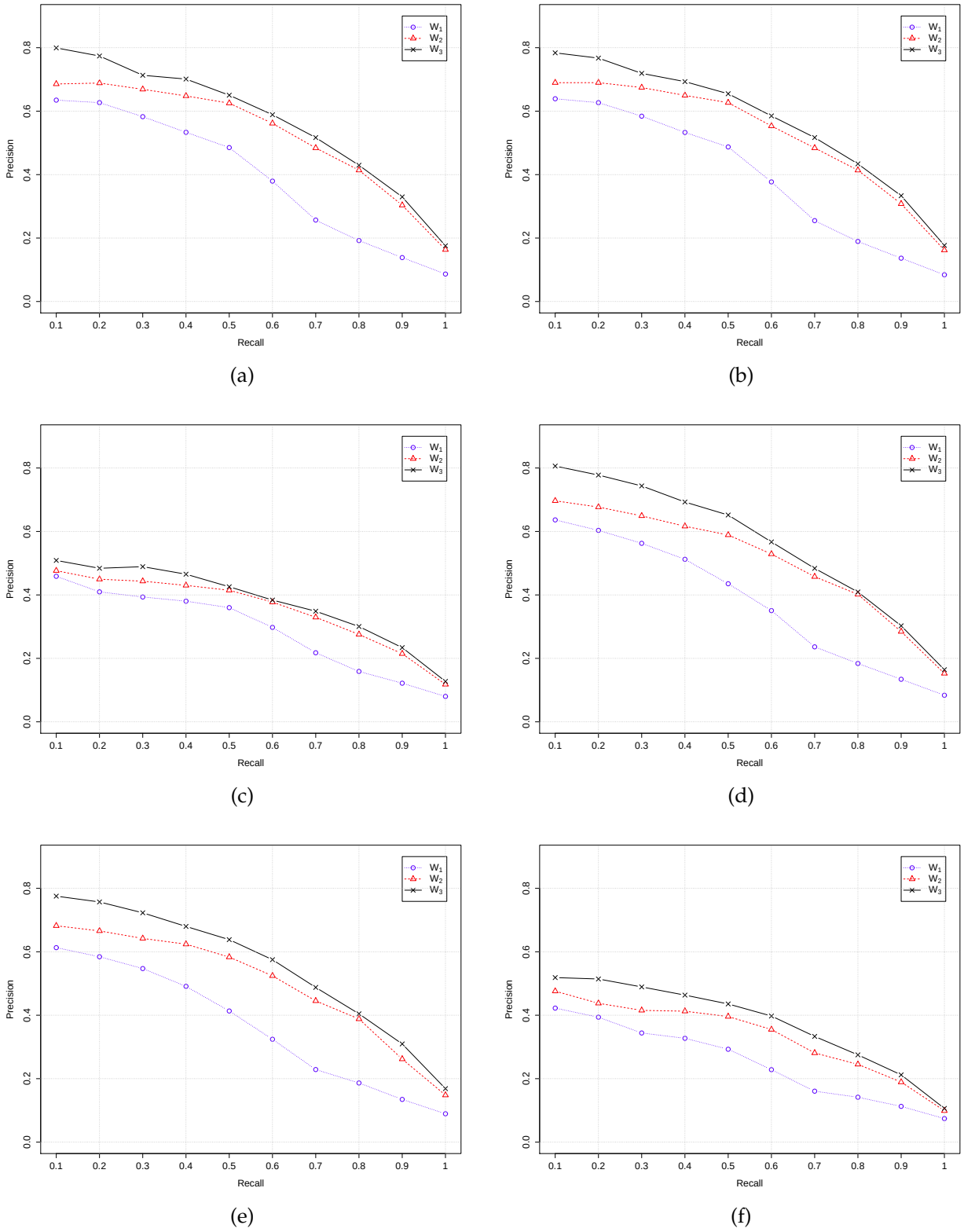
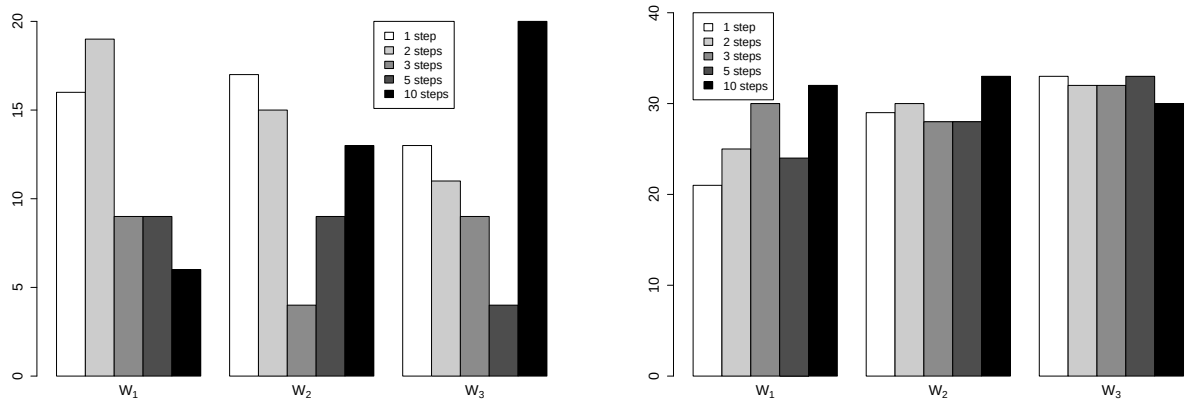


Fig. 4. Precisions at fixed recall levels, averaged across the 51 therapeutic DrugBank classes with more than 15 annotated drugs, with $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ pharmacological networks. (a) S_{kNN}; (b) S_{AV}; (c) S_{NN}; (d) *RWR*; (e) *RW* 1 step; (f) *GBA*.

be considered, in order to take into account, at least for some classes, also "indirect" similarities mediated through relatively long paths across the pharmacological space.

TABLE 3
Compared AUC and precision at 40% recall for S_{AV} with 1, 2, 3, 5 and 10 steps random walk kernels. Results are averaged across the DrugBank categories with more than 15 drugs.

N. of steps	AUC			P40R		
	W_1	W_2	W_3	W_1	W_2	W_3
1 step	0.8274	0.9252	0.9303	0.5206	0.6355	0.6996
2 steps	**0.8373**	**0.9261**	0.9360	0.5336	0.6482	**0.7005**
3 steps	0.8332	0.9233	**0.9372**	0.5330	**0.6497**	0.6931
5 steps	0.8226	0.9235	0.9365	0.5312	0.6452	**0.7005**
10 steps	0.8129	0.9239	0.9370	0.5319	0.6483	0.6955



(a) (b)

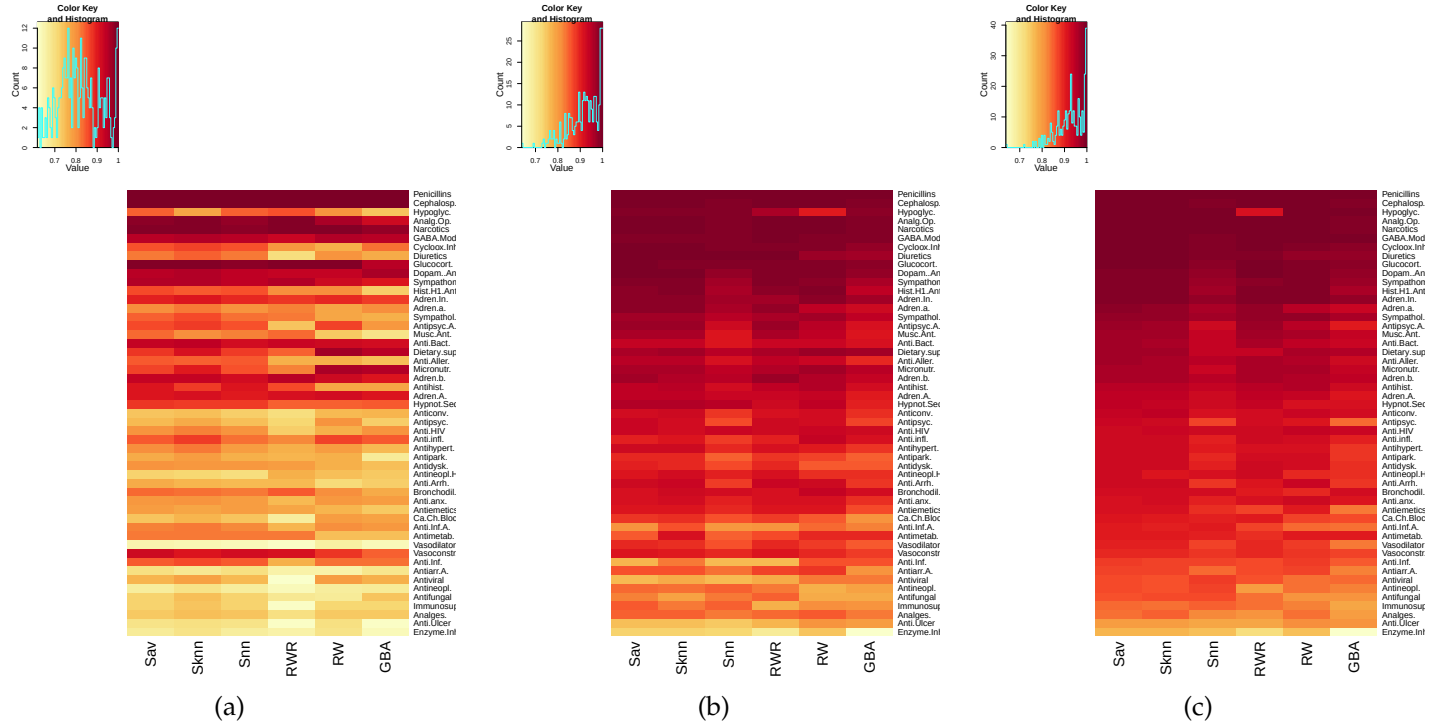
Fig. 5. Counts of the "wins" across the therapeutic classes for the S_{AV} score with 1, 2, 3, 5 and 10 steps random walk kernels, with $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ pharmacological networks: (a) Wins with respect to AUC; (b) Wins with respect to the precision at 40% recall.

It is also worth noting that with the classical *RW* algorithm, if we increase the number of steps, we incur in a progressive decrement of both AUC and P40R (data not shown): while the classical random walk tends to forget the initial "a priori" knowledge by iterating the update step (Section 3.2), the regularization properties of the random walk kernel and the possibility of staying at the same node at each step implicitly induced by the kernel itself (see (8)), allows to maintain the "a priori" knowledge represented by the core of "positive" drugs V_C and at the same to exploit the overall topology of the pharmacological space.

4.4 Per Class AUC and Precision at a Fixed Recall Results

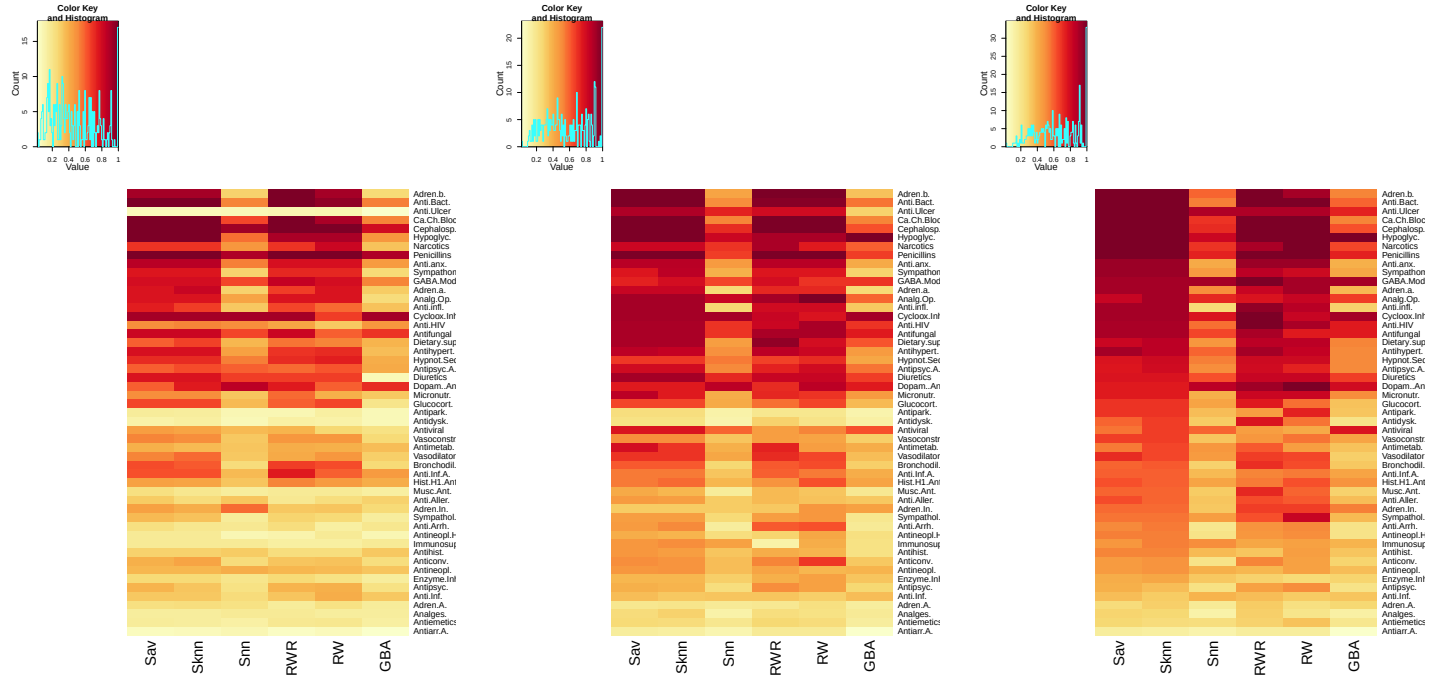
Fig. 6 shows the AUC results achieved by each drug ranking method for each DrugBank therapeutic category. The therapeutic categories are sorted according to the AUC values obtained by S_{AV} with the $\boldsymbol{W}_3$ pharmacological network. For *RW* we mean 1-step Random Walk (recall that by running classical *RW* till to convergence we obtain poor results). In Fig. 6 and 7 for each method we used the parameters listed in Tab. 2. For the correspondences between the therapeutic categories name abbreviations used in Fig. 6 and 7 and the full DrugBank names, please see Tab. 1.

By moving from $\boldsymbol{W}_1$ (Fig. 6 (a)) to $\boldsymbol{W}_2$ (Fig. 6 (b)) and $\boldsymbol{W}_3$ (Fig. 6 (c)) the heatmap "tones" to dark red, showing the effectiveness of the ψ*NetPro* approach, independently of the drug ranking method considered. The "color key" at the top left of each figure shows also an histogram of the distribution of AUC values across classes and across methods, showing a clear skewness towards high AUC values when we move from $\boldsymbol{W}_1$ to $\boldsymbol{W}_3$ (note the the "Count" ordinate scale doubles from Fig. 6 (a) to Fig. 6 (b) and from Fig. 6 (b) to Fig. 6 (c)). The same general trend can be also observed with the precision at 40% recall (Fig. 7), even if in this case the values are distributed across a wider range of values. Note that in Fig. 7 the therapeutic classes are sorted according to the results achieved by S_{kNN} with the $\boldsymbol{W}_3$ pharmacological network, the best performing method in terms of P40R (Tab. 2). Note also that the AUC values across classes are highly correlated between methods: this is more apparent with AUC, while in terms of P40R a very high correlation is maintained only between S_{kNN} and S_{AV} (the methods achieving the best results on the average) and partially between *RW* and *RWR*. *GBA* and S_{NN}



(a) (b) (c)

Fig. 6. TC with more than 15 drugs: per class AUC scores compared across methods. Yellow corresponds to the lowest AUC values, while red to the highest AUC values. (a) $\boldsymbol{W}_1$, (b) $\boldsymbol{W}_2$ and (c) $\boldsymbol{W}_3$ pharmacological networks.



(a) (b) (c)

Fig. 7. TC with more than 15 drugs: per class precision at 40% recall scores compared across methods. Yellow corresponds to the lowest precision values, while red to the highest values. (a) $\boldsymbol{W}_1$, (b) $\boldsymbol{W}_2$ and (c) $\boldsymbol{W}_3$ pharmacological networks.

show a high correlation both in terms of AUC and P40R: this fact confirm the considerations introduced in Section 4.2 about the similarity of the score functions characterizing these methods. The correlation between methods tend to increase when we use the integrated $\boldsymbol{W}_3$ pharmacological network, showing another time the key role of the projections and the integration to improve the overall ranking performances.

TABLE 4
The 10 best and the 10 worst AUC values of DrugBank Therapeutic Categories achieved by the S_{AV} score with 3 steps random walk kernel. The last three columns report the precision at 10, 20 and 40 % recall.

The 10 best classified TC				
TC	AUC	$P10R$	$P20R$	$P40R$
Penicillins	0.9999	1.0000	1.0000	1.0000
Cephalosporins	0.9995	1.0000	1.0000	1.0000
Hypoglycemic_Agents	0.9990	1.0000	1.0000	1.0000
Analgesics._Opioid	0.9979	1.0000	1.0000	0.8333
Narcotics	0.9977	1.0000	1.0000	1.0000
GABA_Modulators	0.9966	1.0000	1.0000	0.9166
Cyclooxygenase_Inhibitors	0.9950	0.7500	0.8333	0.9090
Diuretics	0.9948	0.6000	0.6250	0.6923
Glucocorticoids	0.9948	1.0000	1.0000	0.7692
Dopamine_Antagonists	0.9940	1.0000	1.0000	0.7500
The 10 worst classified TC				
TC	AUC	$P10R$	$P20R$	$P40R$
Vasoconstrictor_Agents	0.8975	0.7500	0.8333	0.6666
Anti.Infectives	0.8776	0.4000	0.5000	0.3478
Antiarrhythmic_Agents	0.8719	0.1363	0.0895	0.1428
Antiviral_Agents	0.8681	1.0000	1.0000	0.5555
Antineoplastic_Agents	0.8657	0.6923	0.6428	0.4605
Antifungal_Agents	0.8616	0.7500	0.8333	0.9000
Immunosuppressive_Agents	0.8428	0.6666	0.5714	0.4705
Analgesics	0.8402	0.1600	0.2285	0.2758
Anti.Ulcer_Agents	0.8001	1.0000	1.0000	1.0000
Enzyme_Inhibitors	0.7701	0.6363	0.5909	0.4000

TABLE 5
DrugBank Therapeutic Categories (TC) with less than 15 drugs considered in the experiments. The first column reports the abbreviated name, the second the full DrugBank name and the third the cardinality of the TC.

Therapeutic categories with less than 15 drugs		
Abbr. name	**Full DrugBank name**	**Card.**
Antipyr.	Antipyretics	7
Antigl.A.	Antiglaucomic_Agents	5
Antirh.A.	Antirheumatic_Agents	7
Antitub.A.	Antitubercular_Agents	7
Ang.Rec.Ant.	Angiotensin_II_Receptor_Antagonists	5
Antithr.A.	Antithrombotic_Agents	5
Osteopor.Pr.	Osteoporosis_Prophylactic	5
Antidotes	Antidotes	5
Nasal.Dec.	Nasal_Decongestants	7
Corticost.	Corticosteroids	5
Antichol.A.	Anticholesteremic_Agents	11
Dermat.A.	Dermatologic_Agents	11
Gastroint.A.	Gastrointestinal_Agents	10
Sulfonamides	Sulfonamides	10
Antid.Tric.	Antidepressive_Agents._Tricyclic	10
Muscle.Rela.	Muscle_Relaxants._Central	11
Anti.Asth.A.	Anti.Asthmatic_Agents	9
Fibrin.A.	Fibrinolytic_Agents	11
Nootropic.A.	Nootropic_Agents	9
Phenothiaz.	Phenothiazines	11
Antibiotics	Antibiotics	14
Anticoag.	Anticoagulants	15
Antid.IIGen.	Antidepressive_Agents._Second.Generation	14
Antihypoc.A.	Antihypocalcemic_Agents	13
AntiInf.Ur.	Anti.Infective_Agents._Urinary	14
CNS.Stim.	Central_Nervous_System_Stimulants	14
Neuroprot.A.	Neuroprotective_Agents	13
NSAIAs	Nonsteroidal_Anti.inflammatory_Agents_.NSAIAs.	14
Seroton.Inh.	Serotonin_Uptake_Inhibitors	13
Ang.Enz.In.	Angiotensin.converting_Enzyme_Inhibitors	13

While for some therapeutic categories such as "Penicillins" or "Cephalosporins" we can obtain high AUC values just with $\boldsymbol{W}_1$ (see the first two rows of the heatmap in Fig.Fig. 6 (a)), for other categories the integration of drug-target and drug-chemicals interaction information is of paramount importance to improve performances: consider, for instance, "Anticonvulsants" or "Anti-HIV Agents". This is not surprising since Penicillins are highly characterized from a chemical standpoint and hence can be effectively predicted by using the similarities between thier chemical structures, while if we consider other chemically more heteregeneous drugs such as Anti-HIV Agents, drug-target relationships play a central role to characterize this therapeutic category.

This is also more evident whe we consider the precision (Fig. 7). Several therapeutic categories need the ψ*NetPro* projection and integration to achieve an acceptable precision: for instance "Antiparkinson_Agents" and "Antidyskinetics" substantially increments their P40R values when the fully integrated $\boldsymbol{W}_3$ pharmacological network, by exploiting drug-drug relationships induced by common genetics and/or toxicogenomics disease-association profiles. Another case is represented by "Anti.Ulcer_Agents", for which Tanimoto coefficients are ineffective (Fig. 7, (a), third row of the heatmap), while with $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ integrated pharmacological spaces we can obtain a very significant P40R increment. Note that for most therapeutic classes we achive a substantial increase of P40R when we move from $\boldsymbol{W}_1$ to $\boldsymbol{W}_2$, while this is not always true when we move from $\boldsymbol{W}_2$ to $\boldsymbol{W}_3$: consider, for instance, "Diuretics" or "Antimetabolities".

For most classes S_{kNN} and S_{AV} achieve the best results, but in terms of P40R for some specific classes, *RWR* (e.g. with "Anti-HIV Agents" or "Cyclooxygenase_Inhibitors") and *RW* 1-step ("Dopamine_Antagonists" and "Sympatholytics") outperform kernelized score functions. We have not a clear explanation of this fact, but at least with respect to *RW* this could be the effect of the choice of the number of steps: indeed with both 1-step random walk kernel S_{kNN} and S_{AV} achieves significantly better results (recall that in Fig. 7 we reported results of 2 steps S_{kNN} and 3 steps S_{AV}). Score functions based on random walk kernels obtain high AUC values for most classes, but also with respect to the worst therapeutic categories (in terms of achieved AUC) we can obtain a reasonable ranking of the drugs (Tab. 4). These results show that for such classes we could obtain better results by integrative further informative sources of data projected into homogeneous pharmacological spaces through ψ*NetPro*.

4.5 Drug Ranking of Therapeutic Categories Characterized by Low Cardinality

Even if in our experiments we chose DrugBank therapeutic categories (TC) with more than 15 drugs in order to obtain more robust and reliable performance measures of the proposed methods, in this section we evaluate the performance of ψ*NetPro* and kernelized score functions using relatively small DrugBank therapeutic categories. More precisely, from the 131 therapeutic classes having from 5 to 15 drugs we randomly selected 10 TCs for each of the 5 to 8, 9 to 12 and 13 to 15 subgroups. The resulting TCs are listed in Tab. 5.

TABLE 6
Average AUC and precision at 40% recall across the DrugBank categories with less than 15 drugs.

Methods	**AUC**			**P40R**		
	W_1	W_2	W_3	W_1	W_2	W_3
S_{AV} 1 step	**0.6924**	0.8635	0.8984	0.2882	**0.4217**	**0.5082**
S_{AV} 10 steps	0.6455	**0.8650**	**0.9153**	0.2710	0.4048	0.4952
S_{kNN} 1 step k=9	**0.6924**	0.8635	0.8983	0.2878	0.4204	**0.5082**
S_{kNN} 10 steps k=9	0.6447	0.8640	0.9115	**0.2920**	0.4143	0.4958
S_{NN} 1 step	0.6916	0.8614	0.8959	0.2436	0.3399	0.4079
S_{NN} 10 steps	0.6447	0.8606	0.9116	0.2522	0.3741	0.4319
RW 1 step	0.6840	0.8620	0.9007	0.2606	0.3707	0.4818
RW 10 steps	0.6130	0.8206	0.8128	0.1605	0.2440	0.2840
RWR $\theta = 0.3$	0.6394	0.8601	0.9110	0.2360	0.3869	0.4915
GBA	0.6853	0.8598	0.8909	0.2146	0.3208	0.4105

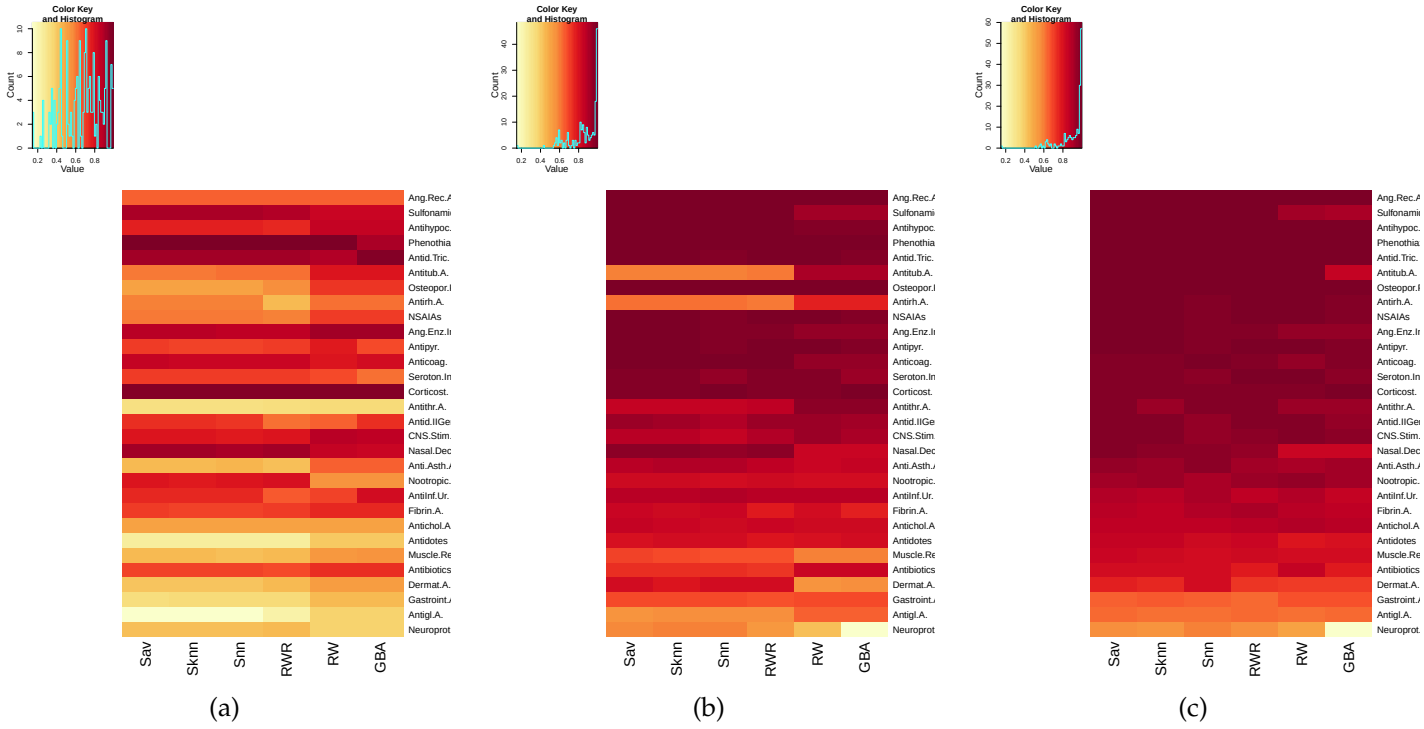


Fig. 8. TC with less than 15 drugs: per class AUC scores compared across methods. Yellow corresponds to the lowest AUC values, while red to the highest AUC values. (a) $\boldsymbol{W}_1$, (b) $\boldsymbol{W}_2$ and (c) $\boldsymbol{W}_3$ pharmacological networks.

As expected, average AUC and P40R results across TCs (Tab. 6) register a certain decrement with respect to the larger TCs analyzed in Section 4.2 (Tab. 2), more significant in terms of P40R, while the average AUC shows a less marked decrease. Tab. 6 shows that also with relatively small TCs the $\psi NetPro$ projection and integration of pharmacological space works nicely, leading to a significant increment of both AUC and P40R independently of the drug ranking method used. A visual clue of this fact is offered also by Fig. 8 and 9 that show respectively the per-class AUC and P40R results achieved by the ranking methods with $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ pharmacological spaces. Note however that P40R values are more scattered, revealing a significant decay in performance with respect to results realtive to the TCs with more than 15 drugs (Fig. 7 and 9).

Also with the "small" TCs S_{AV} and S_{kNN} achieve the best average results, but also *RWR* and *RW* 1 step obtain competitive results (Tab. 6). Interestingly enough, S_{AV} and S_{kNN} with 10 steps random walk kernel register the best AUC results, while this is not true for the classic *RW* 10 steps, as just observed with the "large" TCs analyzed in Section 4.3.

But the more significant fact with "small" TCs is that the integration introduces a more consistent advancement in both AUC and P40R: all the methods on the average approximately improve the AUC of about 20 percent points and double the precision passing from $\boldsymbol{W}_1$ to $\boldsymbol{W}_3$ (Tab. 6), while with "large" TCs the improvement is about 10 percent in terms of AUC and the P40R is augmented of at most one half by passing from $\boldsymbol{W}_1$ to $\boldsymbol{W}_3$ (Tab. 2). This fact is evident also

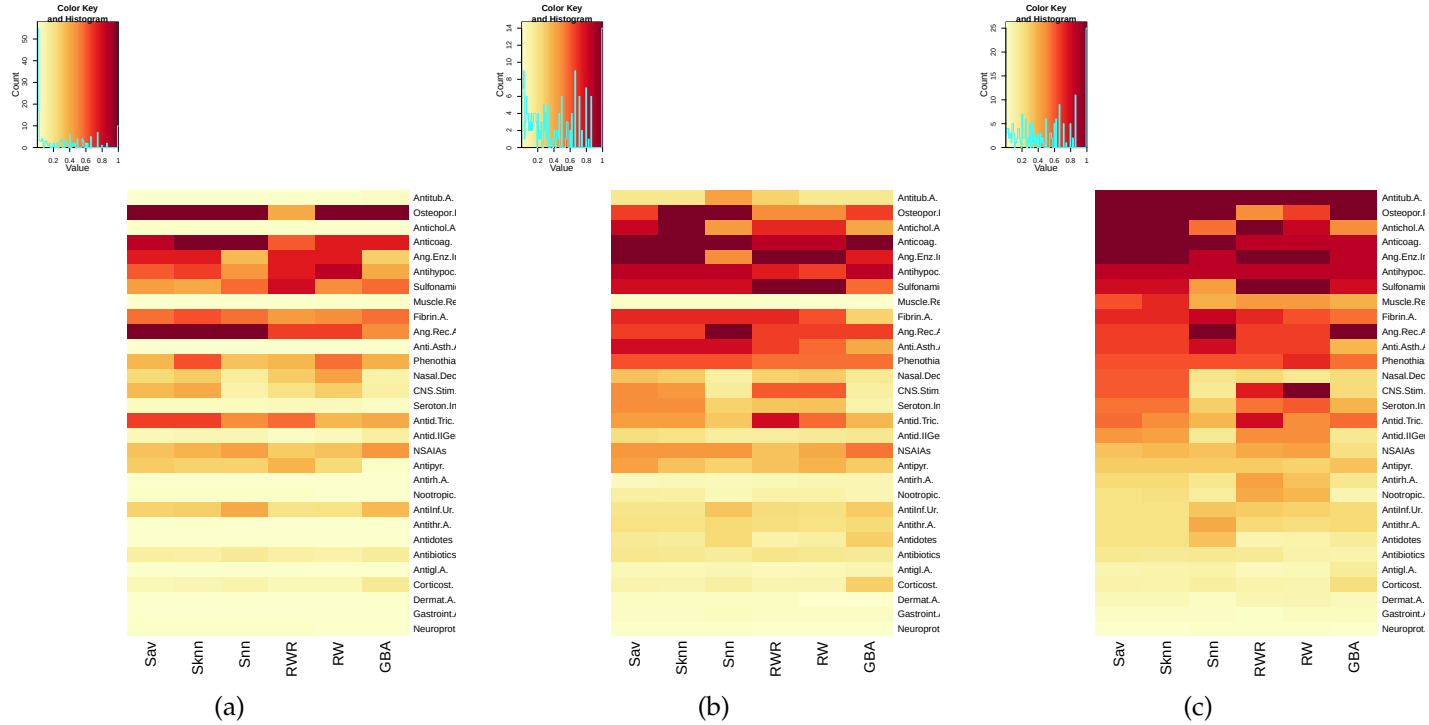


Fig. 9. TC with less than 15 drugs: per class precision at 40% recall scores compared across methods. Yellow corresponds to the lowest precision values, while red to the highest values. (a) $\boldsymbol{W}_1$, (b) $\boldsymbol{W}_2$ and (c) $\boldsymbol{W}_3$ pharmacological networks.

analyzing the per-class results (Fig. 8 and 9): some TCs such as "Angiotensin_II_Receptor_Antagonists" approximately double their average AUC by integrating more sources of data through ψ*NetPro*, independently of the ranking method used, and others, such as "Anticholesteremic_Agents" or "Antitubercular_Agents" pass from P40R values close to 0 with $\boldsymbol{W}_1$ to values close to $0.9 - 1.0$ with $\boldsymbol{W}_3$, with all the ranking methods or at least with the most performing S_{AV}, S_{kNN} or *RWR*.

4.6 Preliminary Analysis of Top Ranked False Positives

Cross-validated average results across classes show that our proposed methods are able to recover therapeutic classes of drugs. A thorough analysis of the results relative to each therapeutic category is out of the scope of this investigation, but in order to show the potential of the proposed method we report the analysis of the top ranked false positives predicted in three drug categories. All the ranking results show an AUC increment due to the progressive networks integration, and we chose among them three of the classes with the largest AUC improvement. "Antidyskinetics" drugs are used in the treatment of motor disorders. In this ranking task we obtained 0.730, 0.887 and 0.923 average AUC using the W_1, W_2 and W_3 networks respectively. The first top ranked negative (L-Tryptophan, DrugBank id: DB00150) was reported to be effective in preventing levodopa-induced motor complications in the treatment of patients affected by Parkinson disease [41], and hence could be associated to the "Antidyskinetics" category. In the ranking task associated with the "Anti HIV Agents" category we achieved respectively 0.753, 0.900 and 0.943 AUC results using our progressively integrated networks. The first top ranked negative was Darunavir (DB01264) and, according to the associated DrugBank entry, it is indicated in the treatment of HIV, but not annotated as "Anti HIV Agents", probably since just annotated as "HIV Protease Inhibitors". The top ranked false positive in the task associated with the "GABA Modulators" (AUC 0.941, 0.972 and 0.995) is Adinazolam (DB00546). This drug, and the four top ranked false positives in this task are benzodiazepines, a class of substances known to modulate the effect of GABA [42], [43].

5 Conclusions and Developments

The integration of multiple sources of information coded as heterogeneous bipartite networks into projected homogenous pharmacological spaces plays a key role to significantly improve the drug ranking results in DrugBank therapeutic categories. By constructing a network through Tanimoto coefficients computed from each pair of drug chemical fingerprints, we can obtain a large coverage initial pharmacological network including all the FDA approved drugs under study. By adding a novel pharmacological network constructed from known drug-target relationships through our proposed ψ*NetPro* procedure, we enrich our original large coverage network with highly informative novel edges, as witnessed by the very significant improvement in terms of both AUC

and precision at fixed recall achieved by each compared drug ranking method when we use $\boldsymbol{W}_2$ instead of $\boldsymbol{W}_1$ pharmacological network. Moreover, by adding further information, such as drug-drug relationships induced by common genetics and/or toxicogenomics disease-association profiles or from target chemicals belonging to the same pathway ($\boldsymbol{W}_3$) network), we can further obtain novel relationships that can signicantly improve performances for specific therapeutic categories (Fig. 6 and 7), but also overall AUC and P40R results (Tab. 2).

Our proposed kernelized scores S_{AV} and S_{kNN}, by introducing both local and global learning strategies for the semi-supervised ranking of drugs, achieve significantly better results than the other compared methods (Tab. 2), but also *RWR* and *RW* 1-step obtain sometimes comparable results (i.e. with the integrated $\boldsymbol{W}_2$ pharmacological network), showing that the construction and integration of informative pharmacological spaces is at least relevant as the design and the choice of proper label ranking algorithms.

The analysis of the performances of the score functions embedding random walk kernels with different numbers of steps (Section 4.3), shows that also indirect similarities mediated through relatively long paths across the pharmacolgical space can be relevant to correctly rank drugs with respect to DrugBank TCs. These results suggest that by tuning the number of steps for each TC or by adopting ensemble learning strategies [44] to include and combine random walk kernels with different number of steps may significantly improve the performances of the kerenelized score functions.

Results averaged across classes show that our proposed approach is able to correctly rank known drugs with respect to DrugBank TCs; moreover the analysis of the the results for each class reveals that for several TCs we can obtain AUC and P40R values that assure a highly reliable ranking and potential repositioning of drugs. Indeed a preliminary analysis of the top-ranked false positives shows that our proposed methods can discover potential drug candidates for novel therapeutic indications.

We would like also to outline that kernelized score ranking methods could be applied to significantly larger drug networks, due to their low computational complexity and scalability. Indeed the full ranking of drugs with 5 fold CV repeated 10 times with respect to the 81 considered TCs requires no more than 10 seconds on an Intel i7-860 2.80 GHz processor with 4 Gbytes of RAM. Hence, considering that in our experiments we analyzed about a thousand of FDA-approved drugs, we hypothesize that the same approach could be applied to thousands of investigational compounds, thus finding initial therapeutic indications for unknown drugs.

Moreover, we could apply the same network projection and integration approach to enrich the pharmacological space with new information coming, e.g., from annotated side-effects (as the one stored in public databases such as SIDER [45]), or from manually curated pathways databases such as Reactome [46], or from large collections of gene expression signatures as the ones included in the Connectivity Map public repository [7], or also from data obtained through Next Generation Sequencing techniques, one of the most promising biotechnologies for drug discovery and development [47].

Another possible development could consists in experimenting with real-valued network projections, to take into account the weights eventually associated to the edges of the bipartite network, or to expicitly consider multiple nodes of the "bottom" set shared by the same pair of vertices of the "top" set of nodes.

ACKNOWLEDGMENTS

The authors gratefully acknowledge partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views.

REFERENCES

[1] J. DiMasi *et al.*, "The price of innovation: new estimates of drug development costs," *Journal of health economics*, vol. 22, no. 2, pp. 151–185, 2003.

[2] T. Ashburn *et al.*, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews*, vol. 3, no. 8, pp. 28–55, 2004.

[3] T. Noeske, B. Sasse, H. Strak, *et al.*, "Predictiong compound selectivity by self-organizing maps: cross-activities of metabotropic glutamate receptor antagonists," *ChemMedChem*, vol. 1, pp. 1066–1068, 2006.

[4] G. Wei, D. Twomey, J. Lamb, *et al.*, "Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance," *Cancer Cell*, vol. 10, pp. 331–342, 2006.

[5] E. Kotelnikova, A. Yuryev, I. Mazo, and N. Daraselia, "Computational approaches for drug repositioning and combination therapy design," *Journal of Bioinformatics and Computational Biology*, vol. 8, pp. 593–606, 2010.

[6] J. Li, X. Zhu, and J. Chen, "Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts," *PLoS Computational Biology*, vol. 5, no. e1000450, 2009.

[7] J. Lamb *et al.*, "The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[8] F. Iorio, R. Bosotti, E. Scacheri, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. di Bernardo, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *PNAS*, vol. 107, no. 33, pp. 14 621–14 626, 2010.

[9] A. Gottlieb, G. Stein, E. Ruppin, and R. Sharan, "PREDICT, a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology*, vol. 7, no. 496, 2011.

[10] M. Sirota *et al.*, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Sci. Transl. Med.*, vol. 96, no. 3, pp. 96–77, 2011.

[11] M. Keiser, V. Setola, J. Irwin, *et al.*, "Predicting new molecular targets for known drugs," *Nature*, vol. 462, pp. 175–181, 2009.

[12] Y. Yamanishi, M. Kotera, M. Kaneisha, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. ISMB 2010, pp. i246–i254, 2010.

[13] A. Chiang and A. Butte, "Systematic evaluation of drug-disease relationships to identify leads for novel drug uses," *Clin. Pharmacol. Ther.*, vol. 86, pp. 507–510, 2009.

[14] A. Bertoni and G. Valentini, "Discovering multi-level structures in bio-molecular data through the Bernstein inequality," *BMC Bioinformatics*, vol. 9, no. S2, 2008.

[15] T. Hastie, R. Tibshirani, and R. Friedman, *The Elements of Statistical Learning, Second Edition*. New York: Springer, 2009.
[16] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. Guo, and D. Wishart, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res.*, vol. 39, no. Jan, pp. D1035–41, 2011.
[17] J. Dudley, T. Desphonde, and A. Butte, "Exploiting drug-disease relationships for computational drug repositioning," *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 303–311, 2011.
[18] A. Ma'ayan, "Network integration and graph analysis in mammalian molecular systems biology," *IET Syst. Biol.*, vol. 2, no. 5, pp. 206–221, 2008.
[19] W. Zhang, F. Sun, and R. Jiang, "Integrating multiple protein-protein interaction networks to prioritize disease genes: a Bayesian regression approach," *BMC Bioinformatics*, vol. 12, no. Suppl 1/S11, 2011.
[20] A. Fraser and E. Marcotte, "A probabilistic view of gene function," *Nature Genetics*, vol. 36, no. 6, pp. 559–564, 2004.
[21] H. Lee, T. Bae, J. Lee, D. Kim, Y. Oh, Y. Jang, J. Kim, J. Lee, A. Innocenti, C. Supuran, L. Chen, K. Rho, and S. Kim, "Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug," *BMC Syst Biol.*, vol. 6, no. 1:80, 2012.
[22] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *PNAS*, vol. 101, no. 9, pp. 2981–2986, 2004.
[23] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabasi, "The human disease network," *PNAS*, vol. 104, no. 21, pp. 8686–8690, 2007.
[24] S. Oliver, "Guilt-by-association goes global," *Nature*, vol. 403, pp. 601–603, 2000.
[25] A. Mitrofanova, V. Pavlovic, and B. Mishra, "Prediction of protein functions with gene ontology and interspecies protein homology data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 775–784, 2011.
[26] M. Re, M. Mesiti, and G. Valentini, "A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks," *IEEE ACM Transactions on Computational Biology and Bioinformatics*, (in press).
[27] M. Re and G. Valentini, "Cancer module genes ranking using kernelized score functions," *BMC Bioinformatics*, vol. 13, no. Suppl 14/S3, 2012.
[28] A. Smola and I. Kondor, "Kernel and regularization on graphs," in *Proc. of the Annual Conf. on Computational Learning Theory*, ser. Lecture Notes in Computer Science, B. Scholkopf and M. Warmuth, Eds. Springer, 2003, pp. 144–158.
[29] M. Kuhn, C. von Mering, M. Campillos, L. Jensen, and B. P., "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Res.*, vol. 36, no. Jan, pp. D684–8, 2008.
[30] L. Gong *et al.*, "PharmGKB: an integrated resource of pharmacogenomic data and knowledge," *Curr. protoc. Bioinformatics*, vol. 14, no. 17, 2008.
[31] A. Davis *et al.*, "The Comparative Toxicogenomics Database: update 2011," *Nucleic Acids Res.*, vol. 39, pp. D1067–D1072, 2011.
[32] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity - a review," *QSAR Comb. Sci.*, vol. 22, no. 9-10, pp. 1006–1026, 2003.
[33] D. Weininger, "Smiles, a chemical language and information system," *Journal of Chemical Information and Modeling*, vol. 28, no. 31, 1988.
[34] M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. Jensen, A. Beyer, and P. Bork, "STITCH 2: an interaction network database for small molecules and proteins," *Nucleic Acids Res.*, vol. 38, no. Jan, pp. D552–6, 2010.
[35] M. Mayer and P. Hieter, "Protein networks - built by association," *Nature Biotechnology*, vol. 18, no. 12, pp. 1242–1243, 2000.
[36] Arabidopsis Interactome Mapping Consortium, "Evidence for network evolution in an Arabidopsis interactome map," *Science*, vol. 333, pp. 601–607, 2011.
[37] L. Lovasz, "Random Walks on Graphs: a Survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, pp. 1–46, 1993.
[38] I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proceedings of the 19th International Conference on Machine Learning (ICML)*, 2002, pp. 315–322.
[39] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the seventh international conference on World Wide Web 7*. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998, pp. 107–117.
[40] G. Lippert, Z. Ghahramani, and K. Borgwardt, "Gene function prediction from synthetic lethality networks via ranking on demand," *Bioinformatics*, vol. 26, no. 7, pp. 912–918, 2010.
[41] R. Sandyk and H. Fisher, "L-tryptophan supplementation in parkinson's disease." *Int J Neurosci.*, vol. 45, no. (3-4), pp. 215–219, 1989.
[42] R. MacDonald and L. Jeffery, "Benzodiazepines specifically modulate GABA-mediated postsynaptic inhibition in cultured mammalian neurones," *Nature*, vol. 271, pp. 563–564, 1976.
[43] S. Hanson and C. Czajkowski, "Structural mechanisms underlying benzodiazepine modulation of the $GABA_A$ receptor," *The Journal of Neuroscience*, vol. 28, no. 13, pp. 3490–3499, 2008.
[44] M. Re and G. Valentini, "Ensemble methods: a review," in *Advances in Machine Learning and Data Mining for Astronomy*, ser. Data Mining and Knowledge Discovery. Chapman & Hall, 2012, pp. 563–594.
[45] M. Kuhn, M. Campillos, I. Letunic, L. Jensen, and B. P., "A side effect resource to capture phenotypic effects of drugs," *Mol Syst Biol*, vol. 6, no. 343, 2010.
[46] D. Croft, G. O'Kelly, G. Wu, M. Haw, R.and Gillespie, L. Matthews, M. Caudy, P. Garapati, *et al.*, "Reactome: A database of reactions, pathways and biological processes," *Nucleic Acids Res.*, vol. 39, no. Jan, pp. D691–D697, 2010.
[47] P. Woollard, N. Mehta, J. Vamathevan, S. Van Horn, B. Bonde, and D. Dow, "The application of next-generation sequencing technologies to drug discovery and development," *Drug Discovery Today*, vol. 16, no. 11-12, pp. 512–519, 2011.

PLACE PHOTO HERE

Matteo Re

Giorgio Valentini Giorgio Valentini received the "laurea" degree in Biological Science and in Computer Science from the University of Genova, and the Ph.D. in Computer Science from the same university. He is currently associate professor at DI, Computer Science Department of the University of Milano, where he attends to both teaching and research. His main research areas are computational biology and machine learning, with a special focus on biomolecular network analysis and gene function prediction. He is author of more than 100 papers published in international peer-reviewed journals, books and conference proceedings.