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 this context we propose a novel semi-supervised drug ranking problem: prioritizing drugs in integrated bio-chemical networks according to specific DrugBank therapeutic categories. Algorithms for drug repositioning usually perform the inference step 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 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 algorithmic scheme based on kernelized score functions that adopts both local and global learning strategies to effectively rank drugs in the integrated pharmacological space using different network combination methods. 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

DRUG repositioning, i.e. the prediction of novel therapeutic indications for existing drugs, has recently raised the attention of the research community and of the big pharma companies, since it allows substantial savings in research and development spending with respect to traditional drug development strategies [1].

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 [2], [3], [4]. Large-scale applications, involving a relatively large number of drugs and diseases, count only a few examples [5], [6], [7], [8].

Different computational tasks related to the drug repositioning problem have been proposed, ranging from clustering drugs either considering their pharmacophore descriptors [2] or Connectivity Map-based networks [6], to prediction of drug-target interactions [9], [10], or drug-disease associations [11], [7] 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 [12]), the latter approach requires that at least a partial labeling of the drugs is known in advance, and by exploiting the available "a priori" knowledge, classical techniques to evaluate supervised algorithms can be applied to assess the prediction performances [13]. For more details about computational methods for drug repositioning based on chemical similarity [2], molecular activity similarity [6], [14], molecular docking [15], shared molecular patology [16], and side effect similarities [17], we refer the reader to the Dudley et al. comprehensive review [18].

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 (TCs) [19]. 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" [18]. The ranking of drugs for each DrugBank 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 problems [20], ranging from gene prioritization [21] to gene function prediction [22] and drug repositioning [23]. 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

• *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*

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.

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], [20]. 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 *Kernelized Score Functions* can be considered a generalization of both guilt-by-association methods [28], and kernel based algorithms for semi-supervised network analysis [29]. More precisely, we propose an algorithmic scheme from which we can derive different node/drug ranking algorithms by choosing or designing a specific distance and/or a specific kernel well-suited to capture the similarity between two nodes by possibly exploiting the overall topology of the network.

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

A preliminary version of this paper has been presented at the ISBRA conference [30]. This enhanced version adds more details about the proposed methods, novel experiments, including the comparison of different network integration strategies, and an extended presentation and discussion of the results.

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. Then we propose and discuss different network combination methods to integrate projected networks obtained from heterogeneous sources of "omic" 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 TCs to show the effectiveness of the proposed drug ranking and network construction and integration methods. The conclusions summarize the main results and 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. 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.

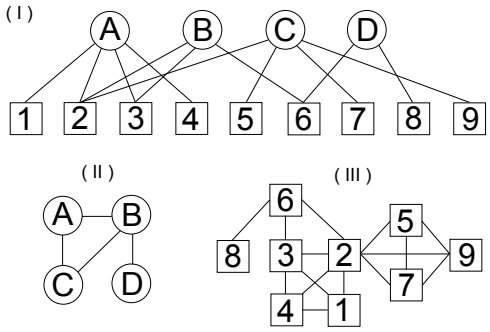


Fig. 1. A toy bipartite network and its unipartite projections. (I) Original bipartite network. Top ($\top$) nodes are labeled by letters and bottom ($\bot$) nodes are labeled by numbers. (II) Projection in the $\top$ domain. (III) Projection in the $\bot$ domain.

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 $x \in \perp$ in the original bipartite graph B. The set of edges in the projected unipartite graph $B_\top$ is thus:

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

The $\perp$-projection $B_\perp$ 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 $v \in \perp$) in the bipartite graph B (Fig. 1). In the following sections, for the sake of simplicity, we name the projected graph $B_\top = (V^\top, E_\top)$ $G = (V, E)$ and its adjacency matrix $\boldsymbol{W}$.

The binary mode projection produces one-mode networks containing binary edges, but more complex projection schemes can generate real-valued edges 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 [31]. 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 interactions), and 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 drug similarity spaces induced from the bipartite graphs can be combined using appropriate network integration methods and proper normalization techniques.

We adopted the normalized graph Laplacian $\boldsymbol{L}$ [32] to make comparable the pharmacological networks G 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 $\boldsymbol{W}$.

To combine the networks, we firstly adopted a progressive integration strategy, and secondly we experimented with different network integration methods, ranging form unweighted to weighted integration of the pharmacological spaces.

In the rest of this section we first briefly introduce the chemical and pharmacological data bases we used to construct pharmacological spaces from different sources of data (Section 2.2.1). Then we construct homogeneous pharmacological networks both directly considering the structural similarity between drug compounds and by exploiting network projections from heterogeneous drug-target and drug-chemical spaces into homogeneous pharmacological networks (Section 2.2.2). In Section 2.2.3 we present a progressive integration strategy to efficiently exploit the different drug coverage provided by each type of constructed pharmacological network. Finally in Section 2.2.4 we describe different network integration methods that we experimentally compared in Section 4.5.

2.2.1 Chemical and pharmacological data bases.

We constructed three pharmacological similarity networks reflecting different notions of similarity between drugs. The first ($N_{structSim}$), is based on edges encoding the similarity of the chemical structures of the drugs. This is the largest network considered in our experiments and the only one that is not computed through bipartite network projections. $N_{structSim}$ is expected to be the least informative pharmacological network, but its usage is motivated by its full coverage of our reference set composed by 1253 drugs. The second one ($N_{drugTarget}$) encodes a notion of drug similarity based on common targets shared by different drugs. The last network ($N_{drugChem}$) exploits information stored in the STITCH database (in the form of precomputed scores) in order to encode similarities between drugs based on their shared interactions between the considered drugs and other chemicals involved in their pharmacological activity (this goes beyond the notion of similarity due to shared protein-targets). Both $N_{drugTarget}$ and $N_{drugChem}$ have been constructed from bipartite heterogeneous networks projected into homogeneous pharmacological spaces (that is networks having only drugs as their nodes). All the aforementioned networks have been constructed using data collected from the DrugBank [19] and STITCH [33] public databases.

DrugBank is a unique bioinformatics and 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 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 [34], [35].

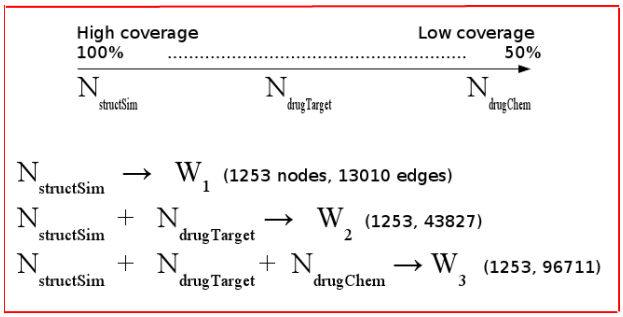


Fig. 2. Progressive integration of network data.

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 $N_{structSim}$ has been obtained by computing the Tanimoto similarity scores between each pair of drugs in the reference set [36]. The scores have been computed by comparing the molecular fingerprints of the simplified molecular input line entry specification (SMILES) annotations contained in DrugBank entries [37]. The chemical fingerprints have been constructed using the *get.fingerprint* function of the *rcdk* R package using, for the *type* parameter the value *extended* because this type of fingerprint take rings structures and atomic properties into account and is thus expected to be more informative than the basic fingerprint type computed by the *get.fingerprint* function. The obtained adjacency matrix was then converted into a binary matrix by thresholding the similarity scores. Instead of tuning the threshold value we arbitrarily adopted a threshold $t = 0.5$ because this is a commonly adopted threshold as stated, among the others, in [9].

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 [38]. A unique feature of STITCH is its ability to establish relationships between chemicals based on their shared relationships with other types of entities such as phenotypic effects, interference with metabolic pathways or co-occurrence in PUBMED abstracts. These very different sources of information are used to compute a chemical-chemical interaction score. The amount of chemicals contained in the STITCH database is larger than our reference set of 1253 drugs. We constructed a bipartite network connecting the drugs in our reference set with all the chemicals contained in STITCH using the STITCH chemical-chemical similarity scores. We then projected the resulting network onto a homogeneous network composed only by the drugs included in our reference set. 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 0.1 steps, 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 about 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. 5 (Supplemental Material) provides a visual clue of the integrated $\boldsymbol{W}_3$ network.

2.2.4 Network integration methods

To get more insights into the role of network integration methods in drug ranking, besides the progressive integration described above, we considered also the following methods:

- Unweighted Average (*UA*)
- Per-edge Unweighted Average (*PUA*)
- Max integration (*MAX*)
- Min integration (*MIN*)
- Weighted Average Per class (*WAP*)
- Weighted Average (*WA*)

In the *UA* method the weight of each edge is computed by simply averaging across the available n networks, and "missing data", i.e. pair of vertices v_i, v_j not present in a given network result in a weight $w_{ij} = 0$:

$$\bar{w}_{ij} = \frac{1}{n} \sum_{d=1}^{n} w_{ij}^d \tag{3}$$

The (*Per-edge Unweighted Average - PUA*) 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:

$$\bar{w}_{ij} = \frac{1}{|D(i,j)|} \sum_{d \in D(i,j)} w_{ij}^d \tag{4}$$

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

The *MAX* integration selects the largest weight among the available sources of data:

$$\bar{w}_{ij} = \max_d w_{ij}^d \tag{5}$$

Analogously, the *MIN* integration selects the minimum weight:

$$\bar{w}_{ij} = \min_d w_{ij}^d \tag{6}$$

The above methods do not require to learn parameters from the data, while the last two learn the "weights" α associated to each type of networks. The α parameter is associated to the "predictiveness strength" of each type of network. For instance, it could be related to the Area Under the Curve (AUC) or the precision at a given recall obtained for a given therapeutic category (TC). More precisely, having n networks and c TCs, we can compute the weight $\alpha^d(k)$ for the d^{th} network and for the k^{th} TC in the following way:

$$\alpha^d(k) = \frac{1}{\sum_{j=1}^{n} F^j(k)} F^d(k) \tag{7}$$

where $F^d(k)$ represents the metric applied to measure the accuracy of the prediction (e.g. the AUC or the precision at a fixed recall) with respect to k^{th} TC and the d^{th} network. The denominator in (7) simply assures that $\sum_{d=1}^{n} \alpha^d(k) = 1$. The $\alpha^d(k)$ can be computed for each TC k by estimating, e.g., the corresponding AUC by leave-one-out on the training data.

Once obtained the $\alpha^d(k)$, the *WAP* method integrates the networks in the following way:

$$\bar{w}_{ij}(k) = \sum_{d=1}^{n} \alpha^d(k) w_{ij}^d \tag{8}$$

It is worth noting that in this way we construct a different weighted integrated network for each TC.

We can also easily compute a "regularized" weight α^d, by averaging across classes. In this way we obtain a unique weight α^d for each network:

$$\alpha^d = \frac{1}{c} \sum_{k=1}^{c} \alpha^d(k) \tag{9}$$

The *WA* method, using the parameters (9) construct an unique integrated network, independently of the TC considered:

$$\bar{w}_{ij} = \sum_{d=1}^{n} w_{ij}^d \sum_{k=1}^{c} \frac{\alpha^d(k)}{c} = \sum_{d=1}^{n} \alpha^d w_{ij}^d \tag{10}$$

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 $v \in V$.

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.

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 share 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 [28]. 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], [39]. 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:

$$S(v_i, C) = max_{j \in N(i)} w_{ij} \tag{11}$$

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 [40] 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 TC 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 TC, 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 $v \in V$:

$$\boldsymbol{p}_{t+1} = \boldsymbol{Q}^T \boldsymbol{p}_t \tag{12}$$

The update (12) 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:

$$\boldsymbol{p}_{t+1} = (1-\theta)\boldsymbol{Q}^T \boldsymbol{p}_t + \theta \boldsymbol{p}_o \tag{13}$$

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 (13), where $\boldsymbol{I}$ is the identity matrix, and values of $\boldsymbol{p}$ at convergence determine the ranking of the nodes [32]. 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 TC:

$$D_{NN}(v, V_C) = \min_{x \in V_C} \left|\left| \phi(v) - \phi(x) \right|\right|^2 \tag{14}$$

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* [32] 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:

$$\boldsymbol{K} = (a-1)\boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}} \tag{15}$$

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 (15):

$$\boldsymbol{K}^q = [(a-1)\boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}]^q \tag{16}$$

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:

$$\boldsymbol{K}^q = \boldsymbol{K}^{q-1}\boldsymbol{K} \tag{17}$$

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 [41]. The name of the kernel derives from the fact that (16) 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 [42].

By developing the square (14) we can derive the following similarity measure:

$$Sim_{NN}(v, V_C) = -\min_{x \in V_C} [K(v,v) - 2K(v,x) + K(x,x)] \quad (18)$$

By assuming an equal auto-similarity $K(x,x)$ for all $x \in V$, we can simplify (18), thus achieving the *nearest neighbours score* S_{NN}:

$$S_{NN}(v, V_C) = -\min_{x \in V_C} -2K(v,x) = 2\max_{x \in V_C} K(v,x) \quad (19)$$

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}:

$$S_{kNN}(v, V_C) = 2\sum_{x \in I_k(v)} K(v,x) \quad (20)$$

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:

$$S_{AV}(v, V_C) = \frac{2}{|V_C|}\sum_{x \in V_C} K(v,x) \quad (21)$$

The S_{AV} score can be viewed as an extension of the algorithm recently proposed in the context of gene function prediction from synthetic lethality networks [43]. Indeed by choosing different local learning strategies and/or specific kernels well-suited to capture the global topology of the network, we can derive different ranking algorithms, including those proposed in [43]. 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 selected kernel function. 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 ranking algorithms described in Section 3 and 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 DrugBank TCs of drugs.

Binary network matrices of the constructed pharmacological spaces and DrugBank labels used in the experiments are available from: http://homes.di.unimi.it/∼re/DATA/ISBRA-DrugData.

4.1 Experimental Setup

In order to obtain the TC 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 firstly analyzed TCs associated to more 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 of Supplemental Material). Then, to test the effectiveness of our methods with TCs characterized by a small number of known associated drugs, we analyzed a set of randomly selected TCs with less than 15 annotated drugs.

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 version of 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 in the same range ($q \in \{1, 2, 3, 5, 10\}$).

We also evaluated the impact of different network integration strategies on the overall ranking performances. More precisely we compared the results obtained with the network integration methods described in Sect. 2.2.4. The α parameters (weights associated to each pharmacological network) have been computed using the AUC metric (see (7) in Sect. 2.2.4) estimated through leave-one-out techniques on the training data.

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 we are interested in evaluating the ranking of the drugs with respect to the TCs, 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 averaged across the therapeutic classes, while results obtained with DrugBank TCs characterized by a very low cardinality are available in

TABLE 1
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

the Supplemental Material. In Section 4.3 we discuss the influence of the choice of the number of steps in random walk kernel score functions, and in Section 4.4 we report the AUC and precision at a fixed recall results for each TC. In Section 4.5 we compare different network integration methods to evaluate their impact on on the DrugBank ranking task, 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. 1 shows the AUC and precision at 40% recall (P40R) averaged across the 51 DrugBank therapeutic classes with more than 15 drugs, using the progressive integration of pharmacological networks (Sect. 2.2.3). For kernelized score functions, *RWR* and *RW* at fixed steps the parameters giving the best results are highlighted in bold.

Independently of the considered methods, the average AUC and P40R increases as new pharmacological spaces are added: most of the AUC increment is achieved when we integrate 2 pharmacological spaces ($\boldsymbol{W}_2$), but note that the relatively small increment 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). With P40R results the increment is large also when we pass from $\boldsymbol{W}_2$ to $\boldsymbol{W}_3$. These results are enforced by the precision at fixed recall levels curves (Fig. 3): independently of the recall level and the considered ranking methods, 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* iterated till to convergence: it deteriorates its performances when new sources of data are added (Tab. 1). 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 due 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. Indeed both *RW* 1 step and *RWR* achieve significantly better results, since they do not "forget" the initial conditions, by exploring only the direct neighborhood of each drug (*RW* 1 step) or by restarting with a certain probability θ from the initial conditions (*RWR*).

The average AUC and P40R are always higher in S_{AV} and S_{kNN} with respect to the other compared methods (Tab. 1), and the differences across the TCs 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. 1) and more in general with precision at fixed recall levels we observe a larger decay with respect to the other considered methods (Fig. 3). Note that a similar behaviour can be observed also in S_{NN}, even if S_{NN} often obtains significantly better results than *GBA* both in terms of AUC and P40R: both methods adopt a "nearest-neighbour" local learning strategy to compute the score associated to each drug (see (11) and (19)), 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. According to these results, as a "rule of thumb" we suggest to apply S_{AV} and S_{kNN} score functions with 2 or 3 steps random walk kernels to rank drugs in the constructed pharmacological space. This recommendation comes from the analysis of the experimental results, and can be explained by the fact that usually "relatively close" neighbours are highly informative: by exploring long paths across the pharmacological space (i.e. by allowing a too large number of steps) we consider similarities mediated through multiple drugs in the network, thus introducing in several cases noise into the prediction score. Of course this depends also on the nature of the data and on the characteristics of the considered TC: for instance if we consider networks

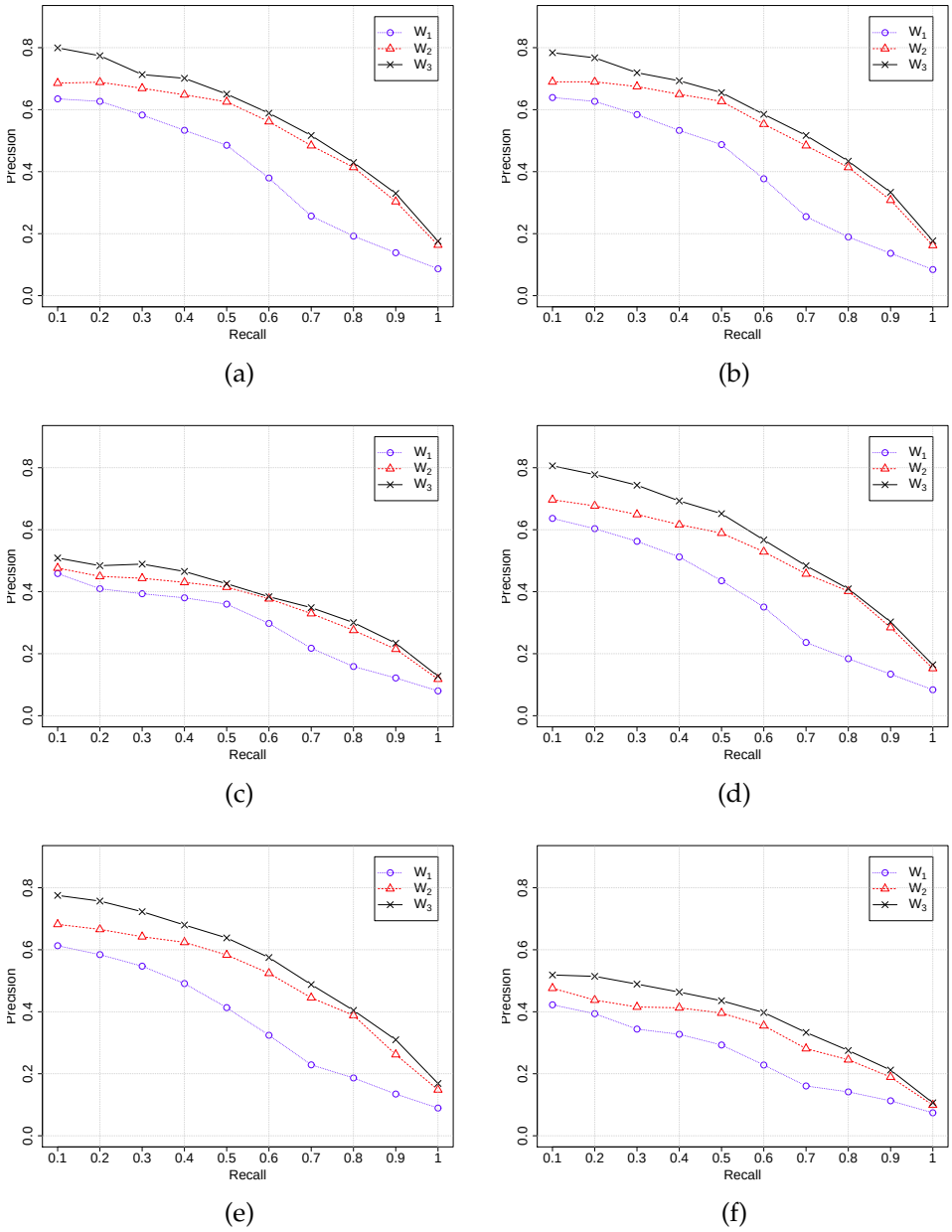
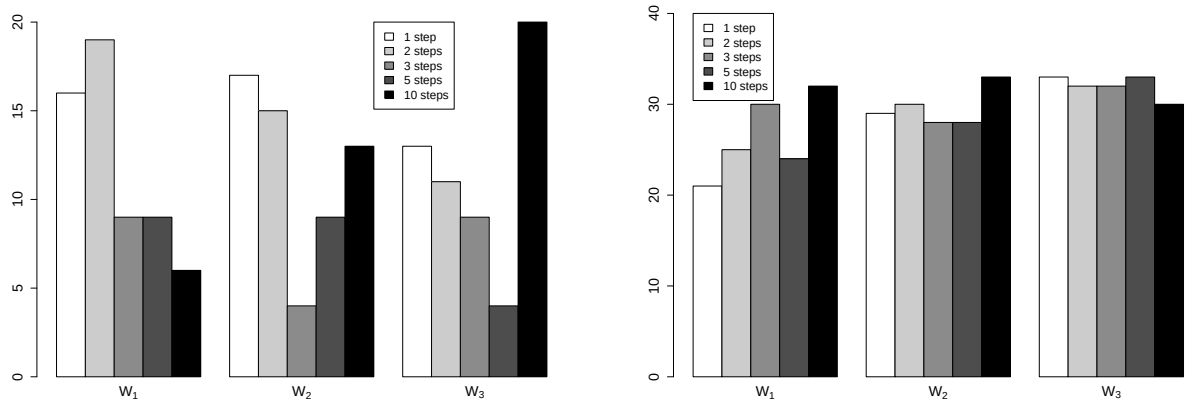


Fig. 3. Precision at fixed recall levels, averaged across the 51 therapeutic DrugBank classes with more than 15 annotated drugs, with $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ pharmacological networks. (a) S_{kNN}; (b) S_{AV}; (c) S_{NN}; (d) *RWR*; (e) *RW* 1 step; (f) *GBA*.

TABLE 2
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. 4. 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. Note that the sum of wins for each network is larger than the number of TCs: this is due to the "ties" of winning methods.

constructed using the chemical similarity between drugs and a TC well characterized from a chemical standpoint (e.g. Cephalosporins) 1-step RW kernels work nicely, but using networks obtained from drug-chemical interactions, if the drugs are involved in the same pathway relevant for a specific TC, it is possible that by introducing more than 2-3 steps we can also achieve better results. To better understand this topic, in the next section, we performed some experiments to analyze the influence of the choice of the number of steps in kernelized score functions for specific TCs.

We performed experiments also with "small" DrugBank TCs, including less than 15 drugs (Table 4 in Supplemental Material). Average AUc and P40R across classes, as expected, register a certain decrement with respect to "large" TCs with more than 15 drugs. Very interestingly, the network integration introduces a more consistent increment in both AUC and P40R: all the methods approximately double the precision passing from $\boldsymbol{W}_1$ to $\boldsymbol{W}_3$, and S_{AV} and S_{kNN} with random walk kernels achieve the best results. A full description and discussion of the experimental results with "small" TCs is available in the Supplemental Material.

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

To get more insights into the significance of the number of steps needed to effectively rank drugs in pharmacological networks, we compared 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 $\boldsymbol{W}_3$ (Tab. 2). Interestingly enough, with most pharmacological networks, there is no a statistically significant difference between 2, 3, 5 and 10 steps random walk kernels according to the Wilcoxon paired signed rank test, at 0.005 significance level (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. 1), we can conclude that also increasing the number of steps, on the average, there is no performance decay in terms of average AUC and P40R.

To gain more insights into the reasons underlying this learning behavior, we counted how many times each k-steps random walk kernel achieved the maximal AUC or P40R (Fig. 4) across TCs. For a specific TC, we say that a method "wins" if its AUC or P40R is the largest among all the considered methods. More precisely, we analyzed the number of "wins" in terms of AUC and P40R among S_{AV} with 1, 2, 3, 5 and 10 steps random walk kernels. 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. 4 (b)), while for the AUC (Fig. 4 (a)) we can observe a quite interesting "peak of wins" for the 10 steps random walk kernel with the full integrated $\boldsymbol{W}_3$ pharmacological space. A possible explanation of these results could consists in the fact that STITCH data involve also drug-chemical interactions: if different chemicals are part of a pathway relevant for the TC under study, and if they are targets of different drugs, it is possible that in the projected pharmacological space also drugs "relatively far" from each other (in terms of the weighted path interconnecting them), could be "pharmacologically connected", since they act on the same pathway relevant for the TC under study.

These results show that, according to the specific characteristics of each therapeutic class, different number of steps should 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.

4.4 Per Class AUC and Precision at a Fixed Recall Results

Fig. 1 in the Supplemental material provides a global view of the AUC results achieved by each drug ranking method for each DrugBank TC. The TCs 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. 1 and 2 in the Supplemental material, for each method we used the parameters listed in Tab. 1. For the correspondences between the TC name abbreviations used in Fig. 1 and 2 and the full DrugBank names, please see Tab. 1 in Supplemental Material.

By moving from $\boldsymbol{W}_1$ (Fig. 1 (a)) to $\boldsymbol{W}_2$ (Fig. 1 (b)) and $\boldsymbol{W}_3$ (Fig. 1 (c), Suppl. material), the heatmap "tones" from yellow to dark red, showing the effectiveness of the ψ*NetPro* approach, independently of the considered drug ranking method. 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 $\boldsymbol{W}_3$. The same general trend can be also observed with the precision at 40% recall (Fig. 2, Supplemental material), even if in this case the results are distributed across a wider range of values. Note 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} 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.

While for some TCs 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. 1 (a), Supplemental material), 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 both Penicillins and Cephalosporins are highly characterized from a chemical standpoint (the average Tanimoto structural similarities in these classes are 0.6743 and 0.6112 respectively versus an average Tanimoto similarity of 0.1748 in the whole drugs reference set) and hence can be effectively predicted by using the similarities between their chemical structures, while for other chemically more heterogeneous TCs, such as "Anti-HIV Agents", drug-target or drug-chemicals relationships play a central role for their characterization.

This is also more evident when we consider the precision (Fig. 2, Supplemental material). Several TCs need the ψ*NetPro* projection and integration to achieve an acceptable precision: for instance "Antiparkinson_Agents" and "Antidyskinetics" substantially increment their P40R values while moving from $\boldsymbol{W}_1$ to 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. 2 (a), third row of the heatmap in the Supplemental material), while with $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ we can obtain a very significant P40R increment.

For most classes S_{kNN} and S_{AV} achieve the best results, and also with respect to the worst ranked TCs we can obtain reasonable results (see Tab. 2 in Supplemental Material). These results show that for several classes we could obtain better results by integrating further informative sources of data projected into homogeneous pharmacological spaces through ψ*NetPro*.

4.5 Comparison of Network Integration Methods

In this section we compare the results obtained using the networks integration methods presented in Section 2.2.4. We considered the ranking tasks involving the 51 DrugBank TCs with more than 15 drugs (Table 1 of Supplemental Material). The performance values have been estimated in terms of AUC averaged across all the TCs (Table 3) and in terms of precision at 40% recall (Table 4), using a stratified 5 folds cross validation repeated 10 times. In the left part of Table 3 and 4, are reported the results obtained with a single network: $structSim$ is the network representing the direct chemical similarities between drugs; $drugTarget$ is the pharmacological network obtained by network projections from DrugBank bipartite drug-target networks, and $drugChem$ the network obtained from the bipartite chemical-chemical networks from the STITCH data base (see Section 2.2.2). The average results across TCs show that the most informative network is $drugTarget$, both in terms of AUC and P40R, while the worst results are obtained with $drugChem$, at least in terms of average P40R. However we warn the reader that this comparison should be considered with caution, since the three networks are composed by different sets of drugs: indeed $structSim$ is the largest one, while $drugTarget$ includes most but not all the nodes/drugs of $structSim$, and $drugChem$ approximately half of the nodes of $structSim$ (see Section 2.2 for more details).

In the right side of Table 3 and 4 are summarized the results obtained with different network integration methods. Independently of the ranking method used

TABLE 3
Comparison of network integration strategies: average AUC across the DrugBank categories.

Ranking methods	Single networks			Network integration methods					
	$structSim$	$drugTarget$	$drugChem$	WA	WAP	UA	PUA	MAX	MIN
S_{AV} 1 step	0.6455	0.7934	0.7657	**0.9307**	0.9291	0.9300	0.9301	0.9286	0.0772
S_{kNN} 2 steps k=31	0.8215	0.8853	0.7862	0.9365	**0.9367**	0.9351	0.9361	0.9346	0.2193
S_{NN} 3 steps	0.8271	0.8847	0.7849	0.9224	**0.9227**	0.9216	0.9219	0.9176	0.2338
$RWR\ \theta = 0.6$	0.7850	0.8930	0.7733	**0.9253**	0.9251	0.9227	0.9241	0.9206	0.3095
RW 1 step	0.6417	0.7682	0.7428	**0.9219**	0.9201	0.9212	0.9209	0.9187	0.0752
GBA	0.5911	0.7547	0.7572	0.9081	0.9048	**0.9083**	0.9074	0.9046	0.0662
RW	0.6567	0.5468	0.5294	0.5654	**0.5833**	0.5633	0.5633	0.5603	0.1378

TABLE 4
Comparison of network integration strategies: average precision at 40% recall across the DrugBank categories.

Ranking methods	Single networks			Network integration methods					
	$structSim$	$drugTarget$	$drugChem$	WA	WAP	UA	PUA	MAX	MIN
S_{AV} 1 step	0.5205	0.5289	0.4531	**0.7031**	0.6876	0.6791	0.6801	0.6546	0.2633
S_{kNN} 2 steps k=31	0.5333	0.5426	0.4668	0.6949	**0.7008**	0.6818	0.6808	0.6675	0.2673
S_{NN} 3 steps	0.3800	0.3727	0.3639	0.4618	**0.4678**	0.4586	0.4589	0.4236	0.2410
$RWR\ \theta = 0.6$	0.5200	0.5467	0.4340	0.6855	**0.6875**	0.6809	0.6727	0.6481	0.2515
RW 1 step	0.5024	0.5142	0.4165	0.6721	0.64361	0.6620	**0.6753**	0.6216	0.2613
GBA	0.3099	0.3433	0.2812	**0.3964**	0.3764	0.3886	0.3869	0.3492	0.2174
RW	0.1882	0.1063	0.0418	0.0429	0.0474	0.0373	0.0372	0.0370	**0.1975**

(except for *RW*), network integration assures a substantial increment of both AUC and P40R. Interestingly enough, this increment is common to all the considered network integration methods (except for *MIN*), and the best results are comparable with those obtained with the progressive integration method (Table 1). The best performing network integration strategy for each ranking method is highlighted in bold. Table 3 and 4 show that the best results are achieved with the weighted integration methods *WA* and *WAP*, even if the difference with respect to the other non-weighted methods is not always statistically significant according to the Wilcoxon rank sum test (at 0.005 significance level). Among the considered network ranking methods, kernelized score function S_{AV} and S_{kNN} achieve the best results, especially with the P40R metric (the difference is statistically significant at 0.005 significance level according to the Wilcoxon rank sum test).

These results show that even if the best results are obtained with the weighted average techniques (*WA* and *WAP*), that take into account the accuracy of the ranking methods on each different network, also the non-weighted integration methods (*UA*, *PUA*, *MAX*) may achieve competitive results, without the overload of computing the weights for each separated network.

4.6 Preliminary Analysis of Top Ranked False Positives

A thorough analysis of the results relative to each TC is out of the scope of this investigation, but 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 integration of different pharmacological networks, 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 [44], 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 it was 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 [45].

5 Conclusions and Developments

The combination of bipartite network projections, weighted integration of different pharmacological spaces

and kernelized score functions with random walk kernels plays a key role to significantly improve the drug ranking results with respect to DrugBank TCs.

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. We outline that we proposed a general algorithmic scheme for drug ranking from which different algorithms can be derived by choosing appropriate distance measures and by designing kernels well-suited to capture the structural and functional similarities between drugs in the underlying pharmacological space. From this standpoint we think that novel research could build on our results to design novel algorithms for drug ranking and repositioning and to construct novel drug networks embedding more information derived from properly chosen heterogeneous bipartite networks.

Indeed in our experiments we integrated three different pharmacological spaces, but the same network projection and integration approach can be applied 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 [47]), or from manually curated pathways databases such as Reactome [48], or from large collections of gene expression signatures as the ones included in the Connectivity Map public repository [5], or also from data obtained through Next Generation Sequencing techniques, one of the most promising biotechnologies for drug discovery and development [49].

According to our experimental results and considerations about the characteristics of the pharmacological spaces, we recommend firstly to apply S_{AV} or S_{kNN} score functions with 2-3 steps random walk kernels, since these method and parameter settings achieve, on the average, the best results. Nevertheless, 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 pharmacological 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 [46] to include and combine random walk kernels with different number of steps may significantly improve the performances of the kernelized score functions.

It is worth noting that even if our proposed methods provide basically a ranking of the drugs with respect to a TC without an explanation of the reasons why an association drug-TC is made, we can derive an interpretation of the results, at least if we apply a relatively simple algorithm from our algorithmic scheme. For instance, if we apply a S_{NN} score with a 1-step RW kernel to a pharmacological network obtained from a bipartite drug-target network, an association of a drug d and a TC T can be explained by the fact that another drug d', known to be associated to T, has a common target with d, and this target is relevant for the therapeutic category under study. Of course the interpretation of an association become difficult if we use more complex score functions, such as a 5-steps RW kernel or a diffusion kernel. In these more general cases, our algorithms explore large parts of the network, thus implicitly considering a multiplicity of functional similarities embedded in the pharmacological network, as well as subtle and "remote" relationships between drugs. In these conditions a clear explanation of the reasons why an association is made is very difficult or not feasible at all.

We would like also to outline that kernelized score ranking methods could be applied to 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 potentially finding initial therapeutic indications for unknown drugs.

We experimented with relatively simple binary network projections, but other approaches based on simple thresholding on the number of bipartite edges, or weighted bipartite projections could improve the robustness of the resulting homogeneous networks in the pharmacological space. We think that the analysis and development of novel network projection algorithms, well-suited to the characteristics of drug ranking and drug repositioning problems, is a promising research line that could be explored in future research work.

ACKNOWLEDGMENTS

We thank the editor and the reviewers for their comments and suggestions. 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] T. Ashburn *et al.*, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews*, vol. 3, no. 8, pp. 28–55, 2004.

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

[3] E. Kotelnikova *et al.*, "Computational approaches for drug repositioning and combination therapy design," *Journal of Bioinformatics and Computational Biology*, vol. 8, pp. 593–606, 2010.

[4] J. Li, X. Zhu, and J. Chen, "Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts," *PLoS Comp. Biol.*, vol. 5, no. 7, 2009.

[5] 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.

[6] F. Iorio, *et al.*, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *PNAS*, vol. 107, no. 33, pp. 14 621–14 626, 2010.

[7] A. Gottlieb, *et al.*, "PREDICT, a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology*, vol. 7, no. 496, 2011.
[8] 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.
[9] M. Keiser, V. Setola, J. Irwin, *et al.*, "Predicting new molecular targets for known drugs," *Nature*, vol. 462, pp. 175–181, 2009.
[10] Y. Yamanishi, *et al.*, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. ISMB 2010, pp. i246–i254, 2010.
[11] 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.
[12] A. Bertoni and G. Valentini, "Discovering multi-level structures in bio-molecular data through the Bernstein inequality," *BMC Bioinformatics*, vol. 9, no. S2, 2008.
[13] T. Hastie, R. Tibshirani, and R. Friedman, *The Elements of Statistical Learning, Second Edition*. New York: Springer, 2009.
[14] B. Chen, D. Wild, and R. Guha, "PubChem as a source of polypharmacology," *J Chem Inf Model*, vol. 49, pp. 2044–2055, 2009.
[15] S. Ekins, J. Mestres, and B. Testa, "PubChem as a source of polypharmacology," *Br J Pharmacol*, vol. 152, pp. 9–20, 2007.
[16] Y. Li and P. Agarwal, "A pathway-based view of human diseases and disease relationships," *PLoS One*, vol. 4, p. e4346, 2009.
[17] M. Campillos, M. Kuhn, and A. Gaviv, "Drug target identification using side-effect similarity," *Science*, vol. 321, pp. 263–266, 2008.
[18] 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.
[19] C. Knox *et al.*, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res.*, vol. 39, no. Jan, pp. D1035–41, 2011.
[20] A. Ma'ayan, "Network integration and graph analysis in mammalian molecular systems biology," *IET Syst. Biol.*, vol. 2, no. 5, pp. 206–221, 2008.
[21] 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.
[22] A. Fraser and E. Marcotte, "A probabilistic view of gene function," *Nature Genetics*, vol. 36, no. 6, pp. 559–564, 2004.
[23] H. Lee *et al.*, "Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug," *BMC Syst Biol.*, vol. 6, no. 1:80, 2012.
[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*, vol. 9, no. 6, pp. 1812–1818, 2012.
[27] M. Re and G. Valentini, "Cancer module genes ranking using kernelized score functions," *BMC Bioinformatics*, vol. 13, no. Suppl 14/S3, 2012.
[28] M. Mayer and P. Hieter, "Protein networks - guilt by association," *Nature Biotechnology*, vol. 18, no. 12, pp. 1242–1243, 2000.
[29] H. Kashima, K. Tsuda, and A. Inokuchi, "Kernels for graphs," in *Kernel Methods in Computational Biology*, B. Scholkopf, K. Tsuda, and J. Vert, Eds. Cambridge, MA: MIT Press, 2004, pp. 155–170.
[30] M. Re and G. Valentini, "Large scale ranking and repositioning of drugs with respect to drugbank therapeutic categories," in *ISBRA 2012*, LNCS, vol. 7292. Springer, 2012, pp. 225–236.
[31] M. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Phys Rev E*, vol. 64, no. 016, p. 131, 2001.
[32] A. Smola and I. Kondor, "Kernel and regularization on graphs," in *Proc. of the Annual Conf. on Computational Learning Theory*, LNCS Springer, 2003, pp. 144–158.
[33] M. Kuhn *et al.*, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Res.*, vol. 36, no. Jan, pp. D684–8, 2008.
[34] L. Gong *et al.*, "PharmGKB: an integrated resource of pharmacogenomic data and knowledge," *Curr. protoc. Bioinformatics*, vol. 14, no. 17, 2008.
[35] A. Davis *et al.*, "The Comparative Toxicogenomics Database: update 2011," *Nucleic Acids Res.*, vol. 39, pp. D1067–D1072, 2011.
[36] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity - a review," *QSAR Comb. Sci.*, vol. 22, no. 9-10, pp. 1006–1026, 2003.
[37] D. Weininger, "Smiles, a chemical language and information system," *J Chem Inf Model*, vol. 28, no. 31, 1988.
[38] M. Kuhn *et al.*, "STITCH 2: an interaction network database for small molecules and proteins," *Nucleic Acids Res.*, vol. 38, no. Jan, pp. D552–6, 2010.
[39] Arabidopsis Interactome Mapping Consortium, "Evidence for network evolution in an Arabidopsis interactome map," *Science*, vol. 333, pp. 601–607, 2011.
[40] L. Lovasz, "Random Walks on Graphs: a Survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, pp. 1–46, 1993.
[41] I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proc ICML 2002*, pp. 315–322.
[42] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proc. World Wide Web 7*. Amsterdam: Elsevier Science Publishers B. V., 1998, pp. 107–117.
[43] 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.
[44] R. Sandyk and H. Fisher, "L-tryptophan supplementation in parkinson's disease." *Int J Neurosci.*, vol. 45, no. (3-4), pp. 215–219, 1989.
[45] R. MacDonald and L. Jeffery, "Benzodiazepines specifically modulate GABA-mediated postsynaptic inhibition in cultured mammalian neurones," *Nature*, vol. 271, pp. 563–564, 1976.
[46] M. Re and G. Valentini, "Ensemble methods: a review," in *Advances in Machine Learning and Data Mining for Astronomy*, Chapman & Hall, 2012, pp. 563–594.
[47] M. Kuhn *et al.*, "A side effect resource to capture phenotypic effects of drugs," *Mol Syst Biol*, vol. 6, no. 343, 2010.
[48] D. Croft *et al.*, "Reactome: A database of reactions, pathways and biological processes," *Nucleic Acids Res.*, vol. 39, no. Jan, pp. D691–D697, 2010.
[49] P. Woollard *et al.*, "The application of next-generation sequencing technologies to drug discovery and development," *Drug Discovery Today*, vol. 16, no. 11-12, pp. 512–519, 2011.

Matteo Re received the "laurea" degree in Biological Science from the University of Milano, and the Ph.D. in Cellular and Molecular Biology from the same university. 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 30 papers published in international peer-reviewed journals, books and conference proceedings.

Giorgio Valentini received the Ph.D. in Computer Science from the University of Genova. He is currently associate professor at DI, Computer Science Department of the University of Milano. 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.