Large scale ranking and repositioning of drugs with respect to DrugBank therapeutic categories

Matteo Re and Giorgio Valentini

DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, 20135 Milano, Italia.
{re,valentini}@dsi.unimi.it

Abstract. The ranking and prediction of novel therapeutic categories for existing drugs (drug repositioning) is a challenging computational problem involving the analysis of complex chemical and biological networks. In this context we propose a novel semi-supervised learning problem: ranking drugs in integrated bio-chemical networks according to specific DrugBank therapeutic categories. To deal with this challenging problem, we designed a general framework based on bipartite network projections by which homogeneous pharmacological networks can be combined and integrated from heterogeneous and complementary sources of chemical, biomolecular and clinical information. Moreover, we propose a novel method based on kernelized score functions for fast and effective drug ranking in the integrated pharmacological space. Results with 51 therapeutic DrugBank categories involving about 1300 FDA approved drugs show the effectiveness of the proposed approach.

1 Introduction

Drug development is a costly and failure-prone process [1]. In recent years a novel pharmacological research paradigm known as drug repositioning is emerging because of its ability to reduce development costs and to shorten paths to approval [2], which typically takes 10-15 years and upwards $1 billion [3], while revenues due to repurposed drugs can exceeds billions of dollars [4].

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 [5, 6, 7, 8]. Large-scale applications, involving a relatively large number of drugs and diseases, count only a few examples [9, 10, 11, 12].

Different computational tasks related to the the drug repositioning problem have been proposed, ranging from clustering drugs either considering their pharmacophore descriptors [5] or Connectivity Map-based networks [10], to prediction of drug-target interactions [13, 14], or drug-disease associations [15, 11].

In this context, 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 (www.e-therapeutics.ca) and STAT!Ref (AHFS) (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]. For each considered DrugBank therapeutic category we provide a ranking of drugs, since this 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.

To this end, we propose a novel and very fast semi-supervised network method based on kernelized score functions for ranking drugs according to their likelihood to belong to a given therapeutic category. Moreover, we propose a general framework based on bipartite networks projections for the construction of homogeneous pharmacological spaces. The nature of these network-structured projected spaces allows the application of prediction algorithms to homogeneous pharmacological spaces and improves the integration of different sources of chemical, biomolecular and clinical sources of information.

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.

2 Methods

We propose *ψNetPro*, Pharmacological Spaces Integration based on Networks Projections, a general approach to construct and integrate different pharmacological similarity spaces capturing different pharmacological characteristics of drugs, and a novel method for ranking drugs in the integrated pharmacological networks to discover new therapeutic indications for known drugs. In Section 2.1 we introduce the bipartite network projection method to construct homogeneous pharmacological spaces from inhomogeneous spaces represented though bipartite networks. In Section 2.2 we show how to construct and integrate different pharmacological spaces using different sources of chemical, biomolecular and pharmacological data, and finally in Section 2.3 we present our novel approach to rank drugs in pharmacological networks through kernel-based score functions.

2.1 Bipartite networks projection and integration

Bipartite (or two-mode) networks are graphs composed by two types of vertices in which edges are established only between vertices belonging to different sets (Fig. 1 a). Bipartite networks can be transformed into one-mode networks (composed by a single type of nodes) by selecting one of the sets of nodes and linking two nodes from that set if the intersection of their neighborhoods in the two-mode network is not empty (Fig. 1 b).

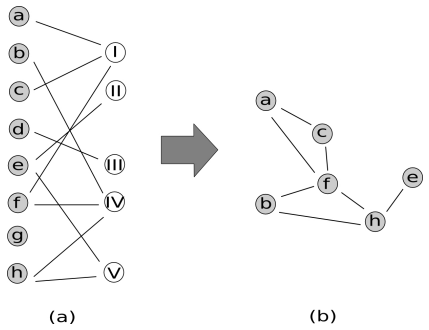


Fig. 1. Bipartite network projection: the two-mode network (a) is projected onto a one-mode network (b). Singleton nodes (i.e. d and g) are removed from the projected network.

More precisely, given a bipartite graph $G =< V, E >$, with two distinct sets of nodes $V_a, V_b \subset V$, $V_a \cup V_b = V, V_a \cap V_b = \emptyset$ and edges $(u, v) \in E \Rightarrow u \in V_a \wedge v \in V_b$, we may induce a projected graph $G_p =< V_p, E_p >$, with $V_p \subseteq V_a$, such that:

$$(u', u'') \in E_p \iff \exists v \in V_b \text{ s.t. } (u', v) \in E \wedge (u'', v) \in E \tag{1}$$

This operation is commonly referred to as "binary mode projection" and is suitable for the induction of a similarity space between vertices $v \in V_a$ (Fig. 1). The binary mode projection produces one-mode networks containing binary edges, but more complex projection schemes can assign edge weights according to the degree of nodes and the edge weights in the bipartite two-mode network. 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 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. These spaces may be integrated using appropriate network integration methods and proper normalization techniques. For instance, we adopted the normalized graph Laplacian $\boldsymbol{L}$ [18] to make comparable the pharmacological networks $G =< V, E >$ represented through the corresponding symmetric adjacency matrices $\boldsymbol{W}$:

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

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:

$$\bar{w}_{ij} = \frac{1}{|D(i,j)|} \sum_{d \in D(i,j)} w_{ij}^d, \qquad D(i,j) = \{d | v_i \in V^d \wedge v_j \in V^d\} \tag{3}$$

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.

2.2 Construction of pharmacological networks

We constructed three pharmacological similarity networks reflecting the pairwise chemical structure similarity between drugs ($\Phi_{chemsim}$), the similarity between drugs derived from common protein targets ($\Phi_{drugtarget}$) and the pairwise similarity from chemical-chemical interactions ($\Phi_{chemint}$) between the considered drugs and other chemicals involved in their pharmacological activity.

Chemical and pharmacological data bases. Data for the computation of $\Phi_{chemsim}$ and $\Phi_{drugtarget}$ have been obtained from DrugBank [16], while data for $\Phi_{chemint}$ have been extracted from the STITCH database [19]. 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 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 [20, 21].

Constructing pharmacological spaces from different sources of data. For $\Phi_{chemsim}$ we directly computed 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.

The simplest similarity space, $\Phi_{chemsim}$, is based on chemical structure similarities and was obtained by computing the Tanimoto similarity scores between each pair of drugs in the reference set [22]. The scores were obtained by comparing the simplified molecular input line entry specification (SMILES) annotations contained in DrugBank entries [23]. The obtained adjacency matrix was then converted to a binary matrix by thresholding the similarity scores according to the procedure reported in [13].

The second considered similarity space, $\Phi_{drugtarget}$, 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 ($\Phi_{chemint}$) has been constructed by processing the chemical-chemical interactions stored in the STITCH 2.0 database [24]. 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). The adjacency matrix was converted to a binary matrix by thresholding the interaction scores to 0.7 in order to ensure a high confidence in the selected STITCH chemical interactions. The thresholding led to a final coverage of 50% of the drugs in our reference set.

Progressive integration of pharmacological networks. We progressively integrated the computed pharmacological networks in order to add different and complementary sources of information and to maintain a high-coverage of drugs for large-scale drug repositioning. To this end we considered at first the $\Phi_{chemsim}$ 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 $\Phi_{drugtarget}$ and $\Phi_{chemint}$. 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. Despite the three networks having the same number of nodes/drugs (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 $\boldsymbol{W}_3$, thus resulting in a high-coverage and a large-scale setting of the drug repositioning problem.

2.3 Ranking methods for drug therapeutic category prediction

By using the adjacency matrices $\boldsymbol{W}$ corresponding to the graphs $G =< V, E >$ obtained by bipartite network projection and integration (Sect. 2.1), we dispose of networks in the pharmacological space well-suited for ranking the drugs $v \in V$ according to their likelihood to belong to a specific therapeutic category C. To this aim we can exploit the pharmacological similarities between pairs of drugs $v_i, v_j \in V$, represented by the weights $w_{ij} > 0$ of the edges $(i, j) \in E$, the overall

topology of the integrated pharmacological spaces, and a subset of drugs $V_C \subset V$ belonging to a priori known therapeutic category C.

In our experiments we compared results obtained with drug ranking algorithms based on random walks on graphs with our novel proposed method that can be interpreted as a kernelized extension of the classical random walks.

Random walks. Random walk (*RW*) algorithms [25] rank 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. If $\boldsymbol{p}_t$ represents the probability vector of finding a "random walker" at step t in the nodes $v \in V$, then the probability at step $t+1$ is:

$$\boldsymbol{p}_{t+1} = \boldsymbol{Q}^T \boldsymbol{p}_t \tag{4}$$

The initial probability of belonging to 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 $v \in V \setminus V_C$, and the update (4) is iterated until convergence. 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 could try to apply the random walk with restart (*RWR*) algorithm: at each step the random walker can move to one of its neighbours or can restart from its initial condition with probability θ:

$$\boldsymbol{p}_{t+1} = (1-\theta)\boldsymbol{Q}^T \boldsymbol{p}_t + \theta \boldsymbol{p}_o \tag{5}$$

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.

Score functions based on kernelized random walks. In this section we propose a novel similarity-based method that on the one hand embeds in a kernel function the random walk strategy and on the other hand uses this kernel within a properly defined kernelized similarity score functions to rank drugs according to the topology of the pharmacological 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:

$$D_{NN}(v, V_C) = \min_{x \in V_C} \left|\left| \phi(v) - \phi(x) \right|\right|^2 \tag{6}$$

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* [18] 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 random walk kernel function $K(\cdot,\cdot)$ is obtained from the adjacency matrix $\boldsymbol{W}$ of the pharmacological network:

$$\boldsymbol{K} = (a-1)\boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}} \tag{7}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{D}$ is a diagonal matrix with elements $d_{ii} = \sum_j w_{ij}$ and a is a value larger than 1.

By developing the square (6) we can derive the following similarity measure:

$$Sim_{NN}(v, V_C) = -\min_{x \in V_C} [K(v,v) - 2K(v,x) + K(x,x)] \tag{8}$$

By assuming an equal auto-similarity $K(x,x)$ for all $x \in V$, we can simplify (8), thus achieving the *nearest neighbours score* S_{NN}:

$$S_{NN}(v, V_C) = -\min_{x \in V_C} -2K(v,x) = 2\max_{x \in V_C} K(v,x) \tag{9}$$

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}:

$$S_{kNN}(v, V_C) = 2\sum_{x \in I_k(v)} K(v,x) \tag{10}$$

where $I_k(v) = \{x \in V_C | x \text{ is ranked among the first } k \text{ in } V_C \text{ according to } K(v,x)\}$. 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:

$$S_{AV}(v, V_C) = \frac{2}{|V_C|}\sum_{x \in V_C} K(v,x) \tag{11}$$

It is worth noting that the S_{AV} score resembles the one proposed by Borgwardt and others in the context of gene function prediction from synthetic lethality networks: from this standpoint our approach can be viewed as an extension of the algorithm proposed in [26].

3 Experiments

3.1 Experimental setup

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.

Table 1. Average AUC results across therapeutic classes of the compared ranking methods using different pharmacological networks $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$.

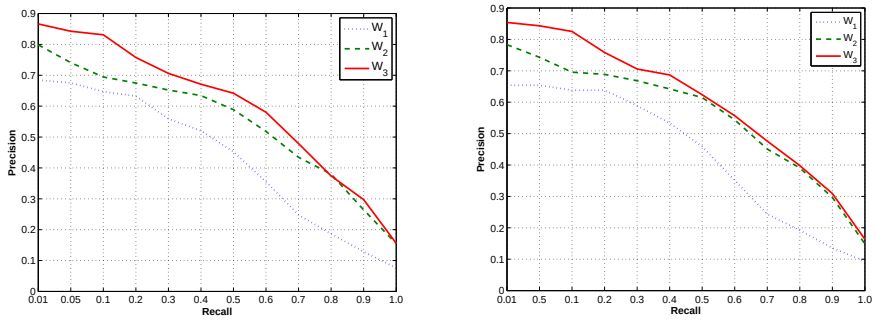
	RW	RWR	S_{AV}	S_{NN}	S_{kNN}
$\boldsymbol{W}_1$	0.6846	0.8037	0.8262	0.8074	0.8277
$\boldsymbol{W}_2$	0.5780	0.9171	0.9232	0.9066	0.9230
$\boldsymbol{W}_3$	0.5334	0.9258	0.9312	0.9129	0.9299

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) by extracting all the drug category annotations excluding the chemical categories (categories reflecting the chemical nature of the considered compounds). We finally 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. 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 averaged across all the considered therapeutic classes.

3.2 Results

Table 1 shows the average AUC across therapeutic classes. We can observe that both *RWR* and kernelized score function methods achieve good results (for several classes the AUC is 1 or very close to 1 when the most informative network $\boldsymbol{W}_3$ is used – data not shown), while the classical *RW* substantially fails in these ranking tasks, since it explores too remote relationships between drugs, thus introducing noise in prediction results. More interestingly, independently of the considered methods (apart from *RW*), the average AUC 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 obtained, e.g. by S_{kNN}, when we pass from 2 to 3 integrated pharmacological spaces is actually statistically significant according to the Wilcoxon ranks sum test (*p-value*< 0.01). These results are also confirmed by the precision at different recall levels outcomes (Fig. 2): we can observe an increment in performance whenever we move from $\boldsymbol{W}_1$ to $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$, no matter the method we apply. For lack of space we reported only *RWR* and S_{kNN} results, but with the other methods (except *RW* that performs poorly also with this metric) we can observe similar trends.

Comparing AUC results between the different methods, according to the Wilcoxon rank sum test, there is no statistically significant difference between S_{AV} and S_{kNN}, while both S_{AV} and S_{kNN} achieve significantly better results than *RWR* and S_{NN} (at 0.005 significance level), independently of the considered pharmacological space. Considering precision at fixed recall levels, S_{NN} performs



(a) (b)

Fig. 2. Precisions at fixed recall levels, with $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ pharmacological networks. (a) Random Walks with Restart (RWR); (b) S_{kNN}. Results are averaged across the 51 therapeutic DrugBank classes.

Table 2. Computational time requirements of the compared methods with the $\boldsymbol{W}_3$ network, using an Intel i7-860 2.80 GHz processor.

	RW	RWR	S_{AV}	S_{NN}	S_{kNN}
time (sec.)	13840	645	5	5	12

significantly worse than the other methods. S_{AV} and S_{kNN} achieve always better or equal results than RWR with both $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ pharmacological networks, while with the most informative $\boldsymbol{W}_3$ network no significant difference between methods can be registered at any recall level (0.05 significance level, Wilcoxon ranks sum test).

Table 2 reports the empirical computational complexity of the different methods for the completion of the entire experimental scheme (5-folds CV repeated 10 times for each of the 51 therapeutic categories). Results show that kernelized score methods are significantly faster than RW and RWR methods.

Cross-validated average results across classes show that our proposed method is 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 [27], 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 [28, 29].

4 Conclusions

Results show that in the context of the drug repositioning problem the construction and integration of informative pharmacological spaces is at least relevant as the design and the choice of proper label ranking algorithms. Indeed the best precision at a given recall results are obtained with the integrated and most-informative pharmacological network $\boldsymbol{W}_3$, independently of the method used (Fig. 2). With the simplest and least-informative pharmacological space $\boldsymbol{W}_1$, based on direct chemical similarities between drugs, S_{AV} and S_{kNN} significantly outperform the other methods, and this is true also with the $\boldsymbol{W}_2$ network. This means that the process of integration of multiple pharmacological spaces by projection of drug-target and drug-chemicals bipartite networks plays a crucial role to improve the information content of the original simple direct chemical similarity space between drugs. Interestingly enough, important increment in performances are also obtained in the ranking of drugs belonging to difficult-to-predict therapeutic classes such as the "Antiparkinson agents" ($\boldsymbol{W}_1$ AUC : 0.7486, $\boldsymbol{W}_2$ AUC : 0.8930, $\boldsymbol{W}_3$ AUC : 0.9316, results obtained using S_{kNN} with $k = 19$). Results averaged across classes show that our proposed approach is able to correctly rank known drugs with respect to their known therapeutic categories. Moreover 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 emphasize that kernelized score ranking methods could be applied to significantly larger drug networks, due to their low computational complexity and scalability (Table 2). Indeed in our experiments we considered about a thousand of FDA-approved drugs, but 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 from annotated side-effects (as the one stored in public databases such as SIDER [30]), or from manually curated pathways databases such as Reactome [31], or from large collections of gene expression signatures as the ones included in the Connectivity Map public repository [9], or also from

data obtained through Next Generation Sequencing techniques, one of the most promising biotechnologies for drug discovery and development [32].

Even if using simple binary projections we obtained high performances in term of AUC, to better exploit the fine-grained information stored in the aforementioned databases, in the future work we plan to experiment with real-valued network projections, to take into account the weights eventually associated to the edges of the bipartite network.

Acknowledgments

We thank 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] DiMasi, J., et al.: New drug development in the United States from 1963 to 1999. Clinical pharmacology and therapeutics **69**(5) (2001) 186–196

[2] Ashburn, T., et al.: Drug repositioning: identifying and developing new uses for existing drugs. Nature reviews **3**(8) (2004) 28–55

[3] DiMasi, J., et al.: The price of innovation: new estimates of drug development costs. Journal of health economics **22**(2) (2003) 151–185

[4] Anand, G.: How drug's rebirth as treatment for cancer fueled price rises: once-demonized thalidomide boosts celgene's sales; patients see costs soar. The Wall Street Jurnal **15**(A1) (2004)

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

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

[7] Kotelnikova, E., Yuryev, A., Mazo, I., Daraselia, N.: Computational approaches for drug repositioning and combination therapy design. Journal of Bioinformatics and Computational Biology **8** (2010) 593–606

[8] Li, J., Zhu, X., Chen, J.: Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. PLoS Computational Biology **5**(e1000450) (2009)

[9] Lamb, J., et al.: The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. Science **313**(5795) (2006) 1929–1935

[10] Iorio, F., Bosotti, R., Scacheri, E., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., di Bernardo, D.: Discovery of drug mode of action and drug repositioning from transcriptional responses. PNAS **107**(33) (2010) 14621–14626

[11] Gottlieb, A., Stein, G., Ruppin, E., Sharan, R.: PREDICT, a method for inferring novel drug indications with application to personalized medicine. Molecular Systems Biology **7**(496) (2011)

[12] Sirota, M., et al.: Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci. Transl. Med. **96**(3) (2011) 96–77

[13] Keiser, M., Setola, V., Irwin, J., et al.: Predicting new molecular targets for known drugs. Nature **462** (2009) 175–181
[14] Yamanishi, Y., Kotera, M., Kaneisha, M., Goto, S.: Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics **26**(ISMB 2010) (2010) i246–i254
[15] Chiang, A., Butte, A.: Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin. Pharmacol. Ther. **86** (2009) 507–510
[16] Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A., Wishart, D.: DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. **39**(Jan) (2011) D1035–41
[17] Dudley, J., Desphonde, T., Butte, A.: Exploiting drug-disease relationships for computational drug repositioning. Briefings in Bioinformatics **12**(4) (2011) 303–311
[18] Smola, A., Kondor, I.: Kernel and regularization on graphs. In Scholkopf, B., Warmuth, M., eds.: Proc. of the Annual Conf. on Computational Learning Theory. Lecture Notes in Computer Science, Springer (2003) 144–158
[19] Kuhn, M., von Mering, C., Campillos, M., Jensen, L., P., B.: STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res. **36**(Jan) (2008) D684–8
[20] Gong, L., et al.: PharmGKB: an integrated resource of pharmacogenomic data and knowledge. Curr. protoc. Bioinformatics **14**(17) (2008)
[21] Davis, A., et al.: The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Res. **39** (2011) D1067–D1072
[22] Nikolova, N., Jaworska, J.: Approaches to measure chemical similarity - a review. QSAR Comb. Sci. **22**(9-10) (2003) 1006–1026
[23] Weininger, D.: Smiles, a chemical language and information system. Journal of Chemical Information and Modeling **28**(31) (1988)
[24] Kuhn, M., Szklarczyk, D., Franceschini, A., Campillos, M., von Mering, C., Jensen, L., Beyer, A., Bork, P.: STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Res. **38**(Jan) (2010) D552–6
[25] Lovasz, L.: Random Walks on Graphs: a Survey. Combinatorics, Paul Erdos is Eighty **2** (1993) 1–46
[26] Lippert, G., Ghahramani, Z., Borgwardt, K.: Gene function prediction from synthetic lethality networks via ranking on demand. Bioinformatics **26**(7) (2010) 912–918
[27] Sandyk, R., Fisher, H.: L-tryptophan supplementation in parkinson's disease. Int J Neurosci. **45**((3-4)) (1989) 215–219
[28] MacDonald, R., Jeffery, L.: Benzodiazepines specifically modulate GABA-mediated postsynaptic inhibition in cultured mammalian neurones. Nature **271** (1976) 563–564
[29] Hanson, S., Czajkowski, C.: Structural mechanisms underlying benzodiazepine modulation of the $GABA_A$ receptor. The Journal of Neuroscience **28**(13) (2008) 3490–3499
[30] Kuhn, M., Campillos, M., Letunic, I., Jensen, L., P., B.: A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol **6**(343) (2010)
[31] Croft, D., O'Kelly, G., Wu, G., Haw, R.and Gillespie, M., Matthews, L., Caudy, M., Garapati, P., et al.: Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Res. **39**(Jan) (2010) D691–D697
[32] Woollard, P., Mehta, N., Vamathevan, J., Van Horn, S., Bonde, B., Dow, D.: The application of next-generation sequencing technologies to drug discovery and development. Drug Discovery Today **16**(11-12) (2011) 512–519