Drug targets prediction using chemical similarity

Diego Galeano
Department of Computer Science
Centre for Systems and Synthetic Biology
Royal Holloway, University of London
Egham Hill, Egham, UK
Email: Diego.Galeano.2014@live.rhul.ac.uk

Alberto Paccanaro
Department of Computer Science
Centre for Systems and Synthetic Biology
Royal Holloway, University of London
Egham Hill, Egham, UK
Email: alberto@cs.rhul.ac.uk

***Abstract*—The growing productivity gap between investment in drug research and development (R&D) and the number of new medicines approved by the US Food and Drug Administration (FDA) in the past decade is concerning. This productivity problem raises the need for innovative approaches for drug-target prediction and a deeper understanding of the interplay between drugs and their target proteins. Chemogenomics is the interdisciplinary field which aims to predict gene/protein/ligand relationships. The predictions are based on the assumption that chemically similar compounds should share common targets. Here, we exploit our understanding of the network-based representation of the protein-protein interaction (PPI network) to introduce a distance between drug-targets and could verify whether it correlates with their chemical similarity. We build a fully connected graph composed of US Food and Drug Administration (FDA) - approved drugs using the Tanimoto 2D similarity based on fingerprints from the SMILES representation of the chemical structure. Our analysis of 1165 FDA-approved drugs indicates that the chemical similarity of drugs predicts closeness of their targets in the human interactome.**

I. Introduction

The scientific discoveries in the past decade have been substantial in increasing our knowledge about the molecular basis of diseases, the network-based structure of diseases [1] and the mode of action of a small molecule[1] in a disease pathway to alleviate the symptoms. Despite the modern advances and huge investment in modern technologies such as molecular biology methods, high-throughput screening, structure-based drug design, combinatorial and parallel chemistry, and the sequencing of the human genome, the number of drug approvals by the US Food and Drug Administration (FDA) registered in a 10-year period from 1999 to 2008 had experienced little effect of these advances [2].

New medicines were discovered in different ways before the genomics era in the 1990s. Fundamentally, because the molecular mechanisms of diseases were mostly unknown. In fact, common drugs such as aspirin, were found based on a serendipitous way through the derivation of the pharmacologically active natural substance from a plant extract. Eder et al. [3] define these approaches as *system-based*, because they consist on a hypothesis-agnostic assay that monitors phenotypic changes *in vitro* or *in vivo*. The fact that the modification of the proteins activity has a fundamental role in the development of a disease has produced an important shift in the drug discovery approach. The past decade, drug discovery was driven mainly by a *target-centric* approach. New technologies were developed for identification of target proteins that are involved in the disease of interest (e.g., RNA interference) and compounds that are likely to interact with these targets (e.g., high-throughput screening[2] and virtual screening[3]).

The dawn of the target-centric approach comes with its drawbacks. The increased rise in late-stage attrition rates in clinical trials in the last decade is concerning. It also reveals that the 'one gene, one drug, one disease' paradigm is ineffective [6]. For instance, Yıldırım et al. [7] compiled data from DrugBank, a public database for FDA drugs that includes approved and experimental drugs, and showed that most drugs target a single protein. Nevertheless, many effective drugs (such as those for cancer and schizophrenia) were shown to act via modulation of multiple proteins rather than a single protein and although this principle is known (as shown in Fig. 1) the rationale of the current paradigm in drug design remains for practicality.

The inherent complexity of the problem is to find a relationship between a ligand[4] and its target proteins. Chemogenomics is an interdisciplinary field that studies the ability of isolated molecular targets to interact with chemical compounds [4]. It attempts to build a relevant set of compounds libraries before the high-throughput screening takes place. By integrating and compiling data from drugs (structural, physical, and biological properties) and structural classification of all known proteins, it aims to predict relationships between drugs and proteins to optimize the screening and thus contribute to a more rational development. The predictions are based on the assumption that chemically similar compounds should share common targets and targets sharing similar ligands should share similarities in their binding sites [8].

Our goal here is to broaden our understanding of the relation between chemical similarity of drugs and their protein targets

[1]Small molecule refers to organic compounds with low molecular weight that may help to regulate a biological process. Most drugs are small molecules.

[2]High-throughput screening is a large-scale, trial-and-error evaluation of compounds in a parallel target-based or cell-based assay [4].

[3]Virtual screening uses computer-based methods to discover new ligands based on their structure [5].

[4]The ligand is a substance that binds to the target protein to change its biological response. It is another term to refer to a drug.

978-1-5090-1633-4/16/$31.00

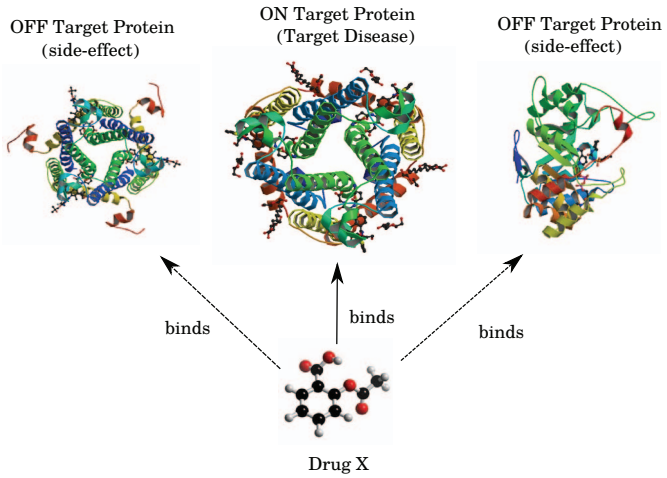


Fig. 1. Paradigm of polypharmacology of the drugs. The small molecule (drug) is a chemical structure that binds to the protein of interest yielding to the expected clinical effect. However, the small molecule can also bind to other proteins, called *off-targets*, producing side-effects. These off-proteins may be involved in other biological processes or phenotypes.

in the context of networks. We combined information about drugs and their targets to assess network-based relationships between chemical similarity of drugs and closeness of their targets on the human interactome. Our hypothesis is based on the similarity property principle introduced by Johnson and Maggiora [9] that establishes that similar chemical structures are deemed to lead to similar biological outcomes [10].

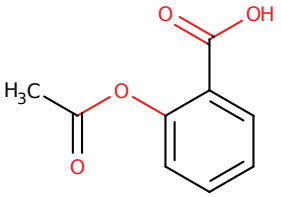
II. Materials and Methods

In this section we will introduce the methods and databases which we have used so far. First, we will explain the binary representation for chemical structure and the Tanimoto similarity measure to compute chemical similarity. Then, the databases used for retrieving data about drugs, their targets, and the interaction between protein targets.

A. Chemical structure as SMILES

The Simplified Molecular-Input Line-Entry System (SMILES) is a specification in line notation for describing the structure of chemical formulas using ASCII strings. SMILES was developed by the Environmental Research Laboratory-Duluth QSAR Research program to facilitate the storage, retrieval, and modelling of chemical structures and chemical information. This notation provides a flexible and unambiguous method for specifying the topological structure of molecules [11]. For example, Fig. 2 shows the SMILES representation of the well-known acetylsalicylic acid (aspirine).

Ideally, the representation of each chemical structure should be unique and function as a *fingerprint*. The aim is to capture the patterns that makes the molecule unique, and different or similar to others. Usually, the molecule fingerprint is generated using information from the molecule itself such as the atoms, bonds, neighbours and so forth. This representation has been widely used for the representation of chemical structure [12].



SMILES REPRESENTATION
CC(=O)OC1=CC=CC=C1C(O)=O

Fig. 2. SMILES representation of the acetylsalicylic acid (Aspirine). The string is an unambiguous representation of the molecule using ASCII code.

B. Tanimoto similarity

There has been extensive debate about which similarity measure is better to reflect the properties or activity values of chemical compounds based on their 2D chemical structure [13]. The Tanimoto similarity measure has proven to be one of the most effective [14]. This similarity is compute based on binary hash fingerprints obtained from the SMILES representation of each drug. Mathematically, it can be defined for two vectors of bits, H_A, H_B, each corresponding to the fingerprints of Drugs A and B, respectively, as follows:

$$T(H_A, H_B) = \frac{\sum_i H_A \cap H_B}{\sum_i H_A + \sum_i H_B - \sum_i H_A \cap H_B} \quad (1)$$

where
$\sum_i H_A$ is the number of bits "on" in H_A;
$\sum_i H_B$ is the number of bits "on" in H_B;
$\sum_i H_A \cap H_B$ is the numbers of bits "on" in both H_A and H_B;

For instance, an intuition of the structure-based similarity is shown in Fig. 3. Drug A and Drug B are chemically equal in the area highlighted in green. For each drug a 2048 bits hash has been computed (binary fingerprint) from the SMILES representation. Using these fingerprints, we compute $T(D_A, D_B)$ for each pair of drug.

C. Protein-protein interaction network

The protein is a biomolecule that performs a function within a living organism. It is a fundamental part of the machinery of the cell. Normally, proteins interact with other proteins to form complexes or catalyzed reactions such as the enzymes. There have been experimental efforts to build a network in which nodes are human proteins and a link between two proteins is related to the likelihood of interaction. The interaction between proteins is an important resource because it may provide insights into protein function and for understanding the principles of cellular functional organization [15].

Biogrid is a repository with data compiled through comprehensive curation efforts. The database contains information about known interactions among proteins [16]. For human, there are 9476 proteins. In the PPI network, every link has a

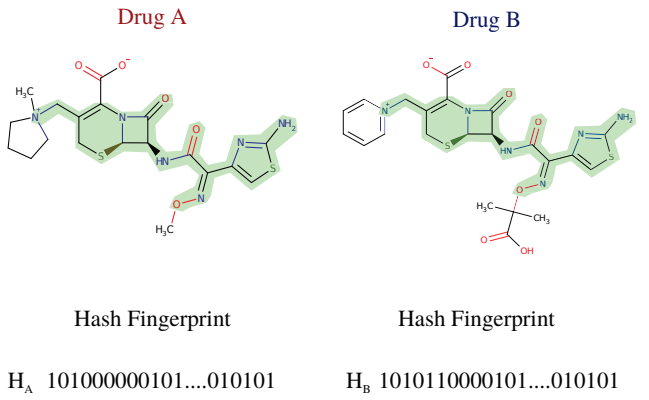


Fig. 3. Example of similarity between two drugs. Each drug is represented with a hash fingerprint (obtained from the SMILES). The area highlighted in green is where the chemical structure is the same. For instance, the Tanimoto similarity for these two drugs is 88%.

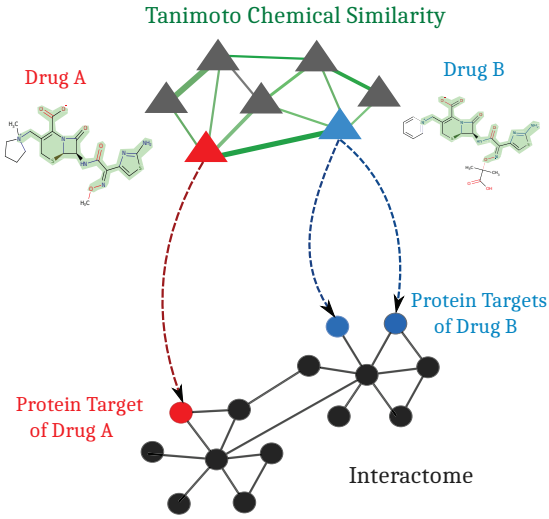


Fig. 4. Mapping of the drugs-targets from the ChemSIM network to the PPI network. **ChemSIM** is a fully connected graph, where each node represents a FDA-approved drug and the weight of the edge is given by the Tanimoto chemical similarity. In the **PPI** network, each node represents a known human protein, and an edge represents interaction between the proteins (binary). The arrows from ChemSIM to the interactome indicate the mapping.

binary value (connected or not) that implies that two connected proteins are likely to interact. We build a matrix that contains all the shortest paths between every pair of human proteins. We use this matrix later to find distances between pairs of drugs-targets.

D. Drugbank database

The Drugbank database [17] is a unique bioinformatics and cheminformatics resource that combines detailed drug data (i.e. chemical, pharmacological and pharmaceutical) with comprehensive drug target information (i.e. chemical structure, SMILES representation, targets). The version 4.3 from the 27th of March 2016 contains 8203 drug entries, including 1991 FDA-approved small molecule drugs.

III. The Addressed Question

In this paper we focus on the relationship between chemical similarity of drugs and the distance of their protein targets on the interactome. We related drugs by the similarity in their chemical structure. The more similar structure, the more likely they produce similar biological outcomes [9]. In addition, proteins exert their functions by interacting with each other. Interacting proteins are more likely to be involve or share common biological functions than non-interacting ones [18]. Therefore, our hypothesis is whether chemical similarity of drugs predicts closeness of their target proteins on the human interactome.

In order to answer our question, we built two networks (Fig. 4): (i) *Chemical similarity network* (ChemSIM), where each node represents an FDA-approved drug and the weight of the edges is the Tanimoto chemical similarity. ChemSIM is a weighted fully connected graph; (ii) *Interactome*, which contains each known protein-protein interaction. The dotted arrows represent the mapping of each drug target on the interactome in order to compute shortest path distances (minimum, average and median) between each pair of drug targets. We are interested in finding relationships between these two networks.

IV. Results

We have found that 1164 of the FDA approved drugs have SMILES representations of their chemical structure and at least one human target on the interactome. In order to get the Tanimoto 2D chemical similarity from the SMILES representation of the drugs, we use the RDKit, an open-source cheminformatics software. The library was integrated to iPython 2.7 to perform the calculations. The distribution of the chemical similarity obtained in this fashion is shown in Fig. 5. The distribution has a mean of $\mu = 0.36$, and a standard deviation of $\sigma = 0.087$. This means that most of the drugs have less than 40% chemical similarity.

One of the properties of the PPI network is that there are multiple paths between proteins. Hence, in order to compute distances between proteins on the interactome, we calculated the shortest path between every pair of proteins. Let us denote the shortest path matrix as $\mathbf{D_{sp}}$. In addition, drugs can have more than one target on the interactome [7]. Hence, let us denote the set of target proteins of Drug A as P_A, and any element of the set as $a_i \in P_A$. We propose three different metrics for measuring the distance (shortest path[5]) between targets of Drug A and Drug B: minimum (d_{min}), average (d_{avg}) and median (d_{med}) shortest path. Mathematically, the different distances between protein targets of Drug A and Drug B can be defined as follows:

$$d_{min} = \min\left(\mathbf{D_{sp}}(a_i, b_j)\right) \qquad \forall a_i \in P_A, b_j \in P_B \qquad (2)$$

[5]The shortest path between two proteins A and B in the PPI is the minimum number of steps required to reach protein B from protein A.

$$d_{avg} = \frac{1}{\mid P_A \mid \times \mid P_B \mid} \sum_{i}^{|P_A|} \sum_{j}^{|P_B|} (\mathbf{D_{sp}}(a_i, b_j)) \quad (3)$$

$$d_{med} = median\,(\mathbf{D_{sp}}(a_i, b_j)) \qquad \forall a_i \in P_A, b_j \in P_B \quad (4)$$

In Fig. 6 we show the distribution of values for the minimum ($\mu = 2.74, \sigma = 1.09$), average ($\mu = 3.59, \sigma = 0.83$) and median ($\mu = 3.64, \sigma = 0.87$) distance. The mean of the minimum distance is lower because it only considers the closest target protein between the two sets. Notably, the distributions of the average and median shortest path are similar.

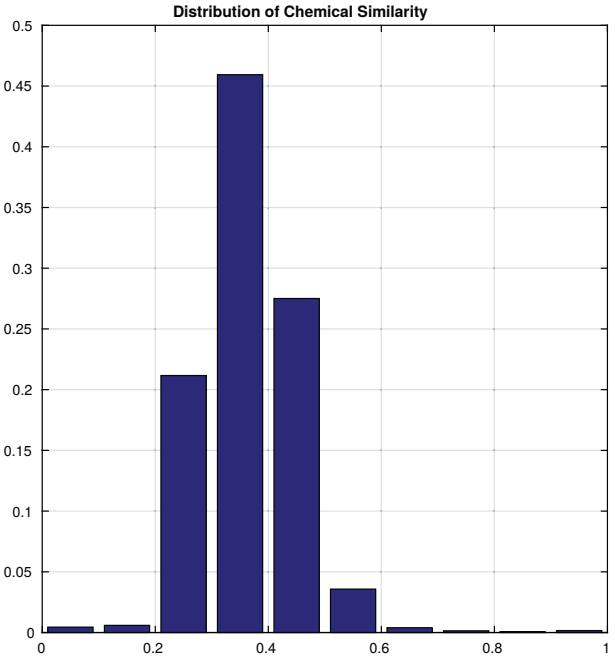


Fig. 5. Distribution of Tanimoto chemical similarity. For 1164 FDA-approved drugs retrieved from Drugbank, most of the drugs have less than 40% similarity in their chemical structure.

A. Predicting the closeness of target proteins

We are interested in finding whether there is any relationship between similar chemical structures of drugs and the closeness of their protein targets. One way of connecting these two networks is through prediction: can chemical similarity predict closeness of protein targets on the interactome? In prediction problems is common to have SCORES and TRUE LABELS (positive and negatives). The performance of the prediction is then computed based on how well the score predicts the true labels. This process is usually done by computing systematically the True Positive Rate (TPR), which corresponds to the proportion of positive data that are correctly considered as positive, and False Positive Rate (FPR), which corresponds to the proportion of negative points that are mistakenly considered as positive. To combine TPF and FPR into a single metric, the score vector is thresholded in many different points and used as a predicted class. By changing the threshold in the score vector is possible to plot the Receiver Operating characteristic (ROC) Curve, which is a single metric that comprises the performance of the prediction. The area under the ROC curve (AUC) is another important metric used to measure the performance [19].

In our problem, the Tanimoto chemical similarity can be used as SCORES for predicting TRUE LABELS (targets are close, targets are not close). Since the method requires defining true labels, which are essentially binary, is necessary to threshold the distance matrices. Let us denote any of distance matrices (minimum, average or median) as $\mathbf{D}$, and an element of the matrix as D_{ij}. Notice that the matrix $\mathbf{D}$ is an $N \times N$ matrix, where N is the number of drugs, and each value D_{ij} represents the minimum, average or median distance (equations 2, 3, 4). The binarized version of the distance matrix is defined as follows:

$$BD_{ij}^{\mathcal{T}} = \begin{cases} 1, & D_{ij} \leq \mathcal{T} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $\mathcal{T} = 0, 1, 2, 3, 4$ is the threshold used to binarize the distance matrices. For instance, for the minimum shortest path distance, $\mathcal{T} = 0$ means that the drugs share at least one target on the interactome. For values $\mathcal{T} > 0$, our notion of closeness is relaxed. Notice that for a given value of $\mathcal{T}$, we consider also as a positive label all distances that are below the given threshold.

The performance of the measure is evaluated by computing the AUC for all the different thresholds $\mathcal{T}$. In Fig. 7, we show a comparative result. We can see that the average and median distances outperform the minimum distance with more than 25%, yielding an AUC of 85%. The fact that the average and median distances perform better than the minimum, can be because they contain more information about the targets. The minimum distance only takes into account the closest target, ignoring the polypharmacology of the drugs. In addition, for all the cases, the AUC decreases when the threshold increases. This phenomenon is expected since our notion of closeness is relaxed. From a biological point of view, the two first cases ($\mathcal{T} = 0$ and 1) are more interesting because the chemical similarity can be used to predict whether two structurally similar drugs have interactive targets on the interactome.

B. Visualization of Chemical Similarity Space

The fact that chemical similarity predicts the closeness of their targets on the interactome can be observed more intuitively by obtaining a 3D graphical representation of FDA-approved drugs. One popular technique for embedding high dimensional data into 3D space is t-SNE [20]. Fig. 8 shows the embedding of drugs into 3D space. In the figure, each point corresponds to a drug and the distance between two drugs relates to the Tanimoto chemical similarity measure. Most of the drugs are coloured in gray, but three pairs (red, green and blue) have been chosen as an example. For every pair of chosen drugs, we have checked the amount of targets

shared on the interactome, as summarize in Table I. In the table, the values of chemical similarity are above the mean. These particular examples, show that the higher the similarity, the more targets they share, but this is not the general case.

TABLE I
EXAMPLE WITH THREE PAIRS OF DRUGS.

DrugBank ID	ChemSIM	Shared Targets PPI
DB04575 DB04574	0.58	1
DB00367 DB00294	0.91	3
DB00318 DB00295	0.99	7

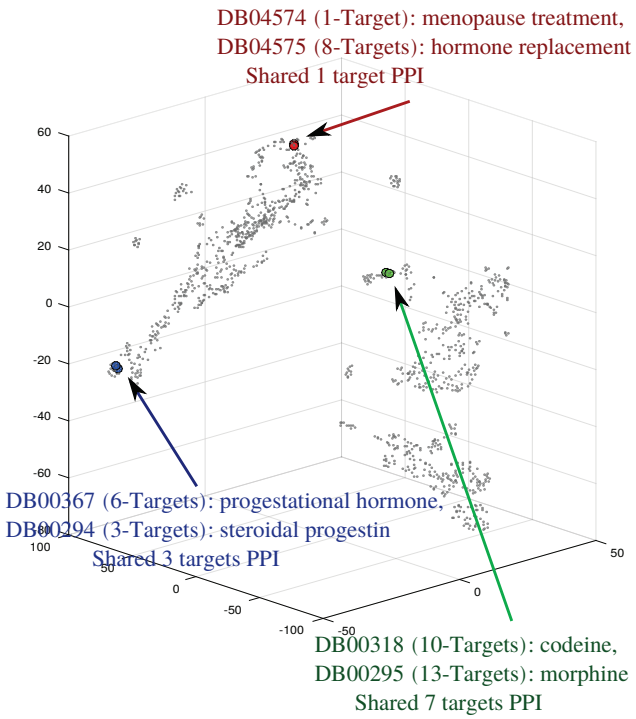


Fig. 8. Embedding of drugs chemical similarity in 3D space using t-SNE. Each point represents a FDA-approved drug. Drugs are coloured in gray. We have highlighted three pairs of chemically similar drugs (red, green, blue). For each pair, we have checked how many targets each of them have, and how many targets they share (if any). The chemical similarity measure can be used as an indication of closeness between drug targets on the interactome.

V. DISCUSSION

Diseases are abnormal conditions that affects the phenotypes of an organism. Typically, defects in multiple genes (and hence proteins) are the causes of diseases. These proteins are called disease proteins because the alteration of multiple proteins can produce changes in the phenotype. In order to cure diseases, pharmaceutical companies design molecules that can target disease-associated proteins. The still long and expensive process of drug discovery requires a deeper understanding of the relation between the chemical structure of the drugs and their target proteins. An understanding of the molecular interaction between a drug and its targets can help to contribute to a more rational and effective design process [2, 7, 21]. Here, we investigated the relation between chemical similarity and the distance, at interaction level, between their protein targets. The systematic analysis in a large set of FDA-approved drugs show that chemical similarity implies closeness of protein targets on the interactome.

In order to understand the relationship between the chemical similarity of the drugs and the interaction distance of their targets, we computed several distances between drug targets on the interactome. To relate chemical similarity with closeness of their targets on the interactome we addressed as a prediction problem, where chemical similarity is used as a score to predict closeness of their targets. The experiments show that chemical similarity performs better in prediction of average and median distances (AUC $\sim 0.85, \mathcal{T} = 0$) than minimum distance (AUC $\sim 0.60, \mathcal{T} = 0$). The average shortest path distance performs slightly better than the median distance for values of the threshold $\mathcal{T} > 0$. Fig. 9 shows a comparison of chemical similarity distribution for all the pairs of drugs, and for those with average shortest path equal to zero. The two distributions are different (Student's t-test $P < 10^{-4}$). We observed that 67% of the pairs of drugs with average shortest path zero have scores in the 95th percentile, indicating that higher similarity value are correlated with shared targets on the interactome.

These findings suggest that chemically similar drugs tend to target the same protein complex, because protein complexes are dense regions containing many connections in PPI networks [22]. The multi-therapeutic category nature of chemically similar drugs can be explained by these findings, given that the protein complex can act in different disease pathways. Furthermore, drugs can have different efficacies depending on the disease pathway in which the protein is acting [23].

Chemical similarity provides a powerful information about the relation between drugs. It has been widely used for drug target prediction and drug repurposing [24, 25] sometimes integrated with other evidence such as side-effect similarity [25]. It would be interesting to use the knowledge presented here to enrich machine learning techniques for drug-target prediction or drug re-purposing.

ACKNOWLEDGMENT

This work was supported by the Programa Nacional de Becas de Postgrado en el Exterior Don Carlos Antonio López (BECAL) from the Republic of Paraguay and in part by the CONACYT Paraguay Grant INVG01-112 (14-INV-088).

The authors would like to thank Juan Cáceres for sharing the code and Jessica Gliozzo for the many useful discussions on the mechanism of drugs.

REFERENCES

[1] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[2] D. C. Swinney and J. Anthony, "How were new medicines discovered?" *Nature reviews Drug discovery*, vol. 10, no. 7, pp. 507–519, 2011.

[3] J. Eder, R. Sedrani, and C. Wiesmann, "The discovery of first-in-class drugs: origins and evolution," *Nature Reviews Drug Discovery*, vol. 13, no. 8, pp. 577–587, 2014.

[4] M. Bredel and E. Jacoby, "Chemogenomics: an emerging strategy for rapid target and drug discovery," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 262–275, 2004.

[5] B. K. Shoichet, "Virtual screening of chemical libraries," *Nature*, vol. 432, no. 7019, pp. 862–865, 2004.

[6] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature chemical biology*, vol. 4, no. 11, pp. 682–690, 2008.

[7] M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drugtarget network," *Nature biotechnology*, vol. 25, no. 10, pp. 1119–1126, 2007.

[8] K. Wilson and J. Walker, *Principles and techniques of biochemistry and molecular biology*. Cambridge university press, 2010.

[9] M. A. Johnson and G. M. Maggiora, *Concepts and applications of molecular similarity*. Wiley, 1990.

[10] S. Croset, "Drug repositioning and indication discovery using description logics," Ph.D. dissertation, University of Cambridge, 2014.

[11] E. Anderson, G. D. Veith, and D. Weininger, *SMILES, a line notation and computerized interpreter for chemical structures*. US Environmental Protection Agency, Environmental Research Laboratory, 1987.

[12] T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, "Structure-based virtual screening for drug discovery: a problem-centric review," *The AAPS journal*, vol. 14, no. 1, pp. 133–141, 2012.

[13] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity–a review," *QSAR & Combinatorial Science*, vol. 22, no. 9-10, pp. 1006–1026, 2003.

[14] X. Chen and C. H. Reynolds, "Performance of similarity measures in 2d fragment-based similarity searching: comparison of structural descriptors and similarity coefficients," *Journal of chemical information and computer sciences*, vol. 42, no. 6, pp. 1407–1414, 2002.

[15] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen *et al.*, "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[16] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[17] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D668–D672, 2006.

[18] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, no. 1, p. 88, 2007.

[19] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

[20] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[21] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[22] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature methods*, vol. 9, no. 5, pp. 471–472, 2012.

[23] S. Galandrin, G. Oligny-Longpré, and M. Bouvier, "The evasive nature of drug efficacy: implications for drug discovery," *Trends in pharmacological sciences*, vol. 28, no. 8, pp. 423–430, 2007.

[24] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran *et al.*, "Predicting new molecular targets for known drugs," *Nature*, vol. 462, no. 7270, pp. 175–181, 2009.

[25] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, 2008.

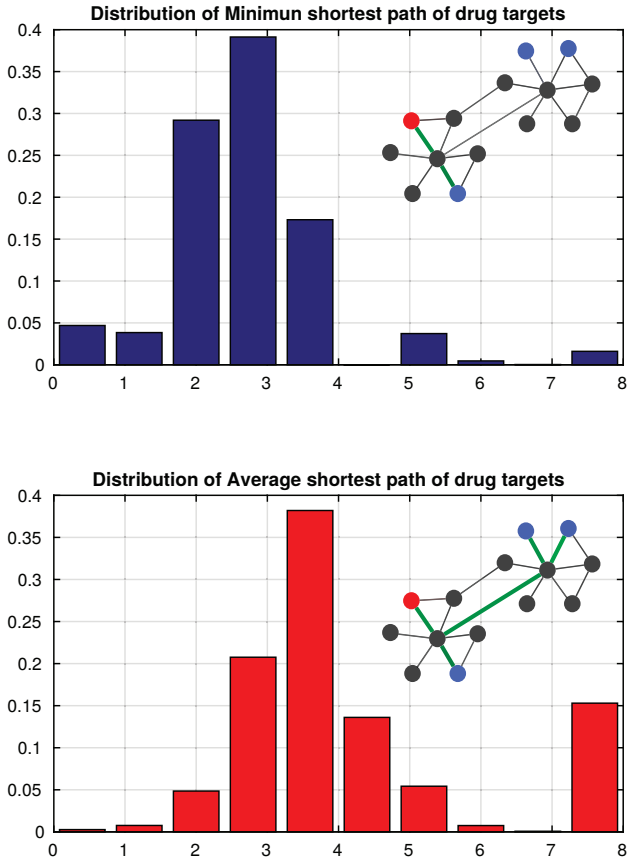


Fig. 6. Distribution of minimum, average and median shortest paths between drug targets on the human interactome. **Top** The minimum shortest path only accounts for the closer targets (eq. 2). The illustration shows that there is only two steps between the closer targets; **Middle** The average shortest path sum up all the possibles shortest paths between targets and divided by the product of the cardinality (eq. 3). The illustration shows that all the targets are considered for the computation of this distance; **Bottom** median shortest path compute the median of the vector of shortest path between all the targets (eq. 4).

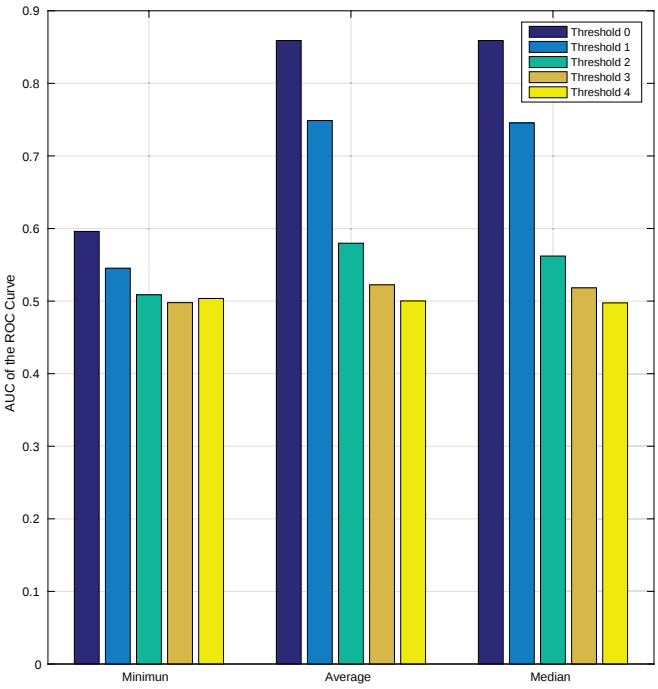


Fig. 7. Performance comparison for the shortest path distance metrics. We have used the chemical similarity to predict binary relationships between drug targets distances on the interactome using different thresholds $\mathcal{T}$. Tanimoto chemical similarity performs better at predicting **average** and **median** distances than **minimum** distance.

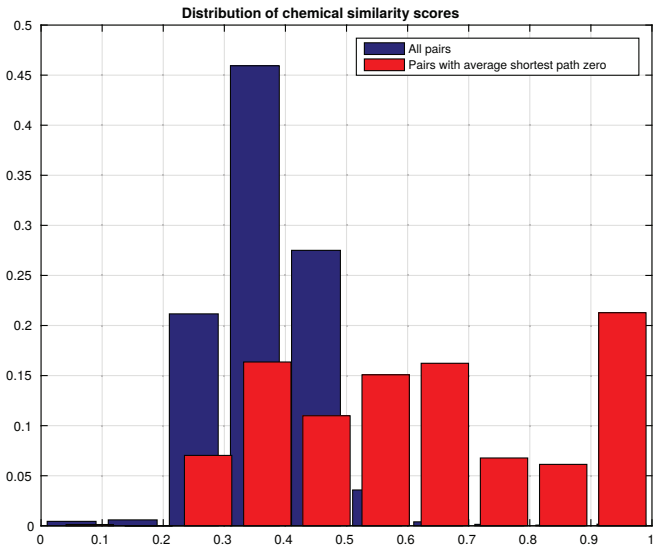


Fig. 9. Comparison of chemical similarity distributions. Distribution of Tanimoto chemical similarity for all pairs of drugs (blue bars) vs distribution of Tanimoto chemical similarity for drug pairs with average shortest path zero (red bars) on the interactome. 67% of the pairs of drugs with average shortest path distance zero, have scores in the 95th percentile.