bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

# Investigating side effect modules in the interactome and their use in drug adverse effect discovery

Emre Guney*

*Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB), c/ Baldiri Reixac 10-12, 08028, Barcelona, Spain
E-mail:emre.guney@irbbarcelona.org

One of the biggest challenges in drug development is increasing costs of bringing new drugs to the market. Many candidate drugs fail during phase II and III trials due to unexpected side effects and experimental methods remain cost ineffective for large scale discovery of adverse effects. Alternatively, computational methods are used to characterize drug side effects, but they often rely on training predictors based on drug and side effect similarity. Moreover, these methods are typically tailored to the underlying data set and provide little mechanistic insights on the predicted associations. In this study, we investigate the role of network topology in explaining observed side effects of drugs. We find that drug targets are closer in the interactome to the proteins inducing the known side effects of the drug compared to the proteins associated with the rest of the side effects. We show that the interactome based proximity can be used to identify side effects and we highlight a use case in which interactome-based side effect prediction can give insights on drug side effects observed in the clinic.

## Introduction

Drug safety is one of the major driving factors beneath the attrition of drugs, contributing to more than 20% of the clinical trial failures and thus increasing costs associated with drug development [1, 2]. Undesired side effects of drugs are also among the leading causes of mortality in Western countries [3], prompting a clear need for better understanding of drug side effects.

*In silico* association of drugs to side effects offers a cost-effective alternative toward characterizing drug side effects. The first examples of such methods using chemical structure similarity to cluster drugs and predict their potential side effects date back to more than a decade ago [4, 5]. Nonetheless, the large-scale prediction of drug side effects gained immense attention after the availability of SIDER, a resource containing side effect information mined from drug labels [6]. Using the data available in SIDER, several studies sought building prediction models incorporating various similarity measures between drugs and side

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

effects [7, 8, 9]. For instance, Atias and Sharan combined diffusion in side effect similarity network and canonical correlation analysis using chemical fingerprint similarity of drugs to link drugs to side effects [7]. Similarly, Duran-Frigola and Aloy turned to building decision tree based classifiers using various features such as chemical structure, small fragments, drug targets, functional associations, and pathway annotations over-represented among known drug-side effect associations [8]. They used these classifiers to explain the side effect profiles of drugs observed in SIDER and found that the side effects can be predicted reliably for a relatively small portion of the data set. Huang *et al.*, on the other hand, included the interactors of drug targets in the protein-protein interaction (PPI) network in addition to the structural properties of the drug to train a support vector machine and suggested that integrating PPI improved the prediction accuracy substantially [9].

The topology of the human interactome encodes biologically relevant information that can be used to discover novel drug-disease [16, 17, 18], and drug-side effect [19, 20] relationships. Although, some side effects can be explained by the proteins the drug is intended to target, many side effects likely to originate from the interactions of the drug with off-targets or the interactions between these proteins [21]. To understand the role of protein interactions in drug induced arrhythmias, Berger and colleagues identified the neighborhood of disease associated genes for long-QT syndrome in the PPI network and used this neighborhood to predict drugs likely to have risks for QT-interval prolongation [19]. They calculated a random-walk based score from each protein in the PPI network to known disease genes involved in long-QT syndrome, corresponding to the reachability of the proteins from the known disease genes. They then used this score to define a long-QT syndrome specific interactome neighborhood and to rank the drugs based on the targets falling in this neighborhood. Moreover, Brouwers *et al.,* investigated whether the side effect similarity between drugs could be explained by the closeness of the drug targets in a functional PPI network [20]. They observed that only a minor fraction (6%) of drugs whose targets were direct neighbors in the network shared similar side effects, emphasizing the need for taking the global topology of the network into account.

In this study, we aim to investigate whether the global topology of the human interactome can characterize drug side effects. We first define side effect modules as the drug targets elucidating the side effects and check the network-based distances between side effect modules and drug targets. We show that drug targets are closer to the proteins associated with the known side effects of the drug in the network compared to the proteins associated with the rest of the side effects. We then use interactome based closeness to systematically identify side effects of the Federal Drug Administration (FDA) approved drugs in the DrugBank database. Finally, we demonstrate how the interactome based closeness can be used to predict side effects of tamoxifen that are not listed in SIDER.

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

# Materials and methods

## Data sets

The drugs used in our analysis were retrieved from DrugBank v4.3 database [22]. For all FDA approved drugs, we extracted drug-protein interactions including drug target, enzyme, transporter and carrier interactions (hereafter we simply refer all these proteins as drug targets). Uniprot ids from DrugBank were mapped to ENTREZ gene ids using Uniprot id mapping file (retrieved on October 2015). The SMILES strings of drugs were also downloaded from DrugBank.

We obtained drug side effect information from SIDER v4 [23], a resource containing side effects extracted from drug labels via text mining and mapped the drug ids to DrugBank ids using the PubChem mapping provided in DrugBank. We represented the side effects with their preferred terms reported in SIDER. To avoid including drugs whose side effects are not well characterized, we only considered drugs with at least five side effects in SIDER.

For validation purposes, in addition to SIDER, we used OFFSIDES [24], cataloging clinically significant drug side effects from FDA adverse event reporting system. We parsed the OFFSIDES flat file and mapped the drug ids to DrugBank ids using the PubChem mapping provided in DrugBank as we did for SIDER. Only the side effects with observed medical effect were included in the analysis.

We used the human interactome curated in a recent study [25], containing physical interactions between proteins from TRANSFAC[26], IntAct[27], MINT[28], BioGRID[29], HPRD[30], KEGG[31], BIGG[32], CORUM[33], PhosphoSitePlus[34], as well as several large scale studies [35, 36, 37]. The coverage and confidence of this integrated interaction network has been showed to be superior to interaction networks coming from yeast-two-hybrid or functional association data sets [25, 18]. Following the methodology in these studies, the largest connected component of the network, containing 141,150 interactions between 13,329 proteins, was used in the analysis.

## Defining side effect modules

To identify drug targets that contribute to the side effects, we followed the procedure presented in Kuhn *et al.* [38]. For each side effect and drug target we counted the number of drugs with and without the side effect for which the drug target was a known target versus the number of drugs with and without the side effect for which the target was not a known target. We used Fisher's exact test to calculate the two sided P-value of the observed occurrence of the target with the side effect as follows: The P-values were then corrected for multiple hypothesis testing using Benjamini and Hochberg's method. We selected the targets that were below 20% false discovery rate to describe the side effect module. In our analysis, we considered the side effects modules that had at least five targets in the interactome. We note that although the proposed

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

approach is applicable to side effects defined by any number of proteins, we use the side effects with at least five proteins to ensure that the side effects in the analysis can be fairly explained by a group of proteins. We provide the side effect module information and the Jupyter Notebook to replicate the analysis in this study at github.com/emreg00/proxide.

## Characterizing closeness between drug targets and side effect modules

Given a network $G(V, E)$, we defined the following topological measures to quantify the network based closeness between targets of a drug, $T$, and proteins in a side effect module, $S$.

(i) *Shortest:* The average pairwise shortest path length between each drug target and side effect module protein.

$$d_{\text{Shortest}}(T, S) = \frac{1}{\|T\| * \|S\|} \sum_{s \in S} \sum_{t \in T} d(t, s)$$

where $d(t, s)$ is the shortest path length between nodes $t$ (a drug target) and $s$ (a side effect protein) in the network. To convert the average shortest path length above to a side effect specific z-score for each drug, we normalized $d_{\text{Shortest}}(T, S)$ using the mean ($\mu_{d_{\text{Shortest}}(T,S)}$) and standard deviation ($\sigma_{d_{\text{Shortest}}(T,S)}$) of $d_{\text{Shortest}}(T_i, S)$ values calculated for all the drugs $\{T_1, T_2, ..., T_n\}$ in the data set. Accordingly, the closeness between drug $T$ and side effect $S$ was given by

$$z_{\text{Shortest}}(T, S) = \frac{d_{\text{Shortest}}(T, S) - \mu_{d_{\text{Shortest}}(T,S)}}{\sigma_{d_{\text{Shortest}}(T,S)}}$$

We used Dijkstra's shortest path algorithm implemented in Python networkx package to calculate the pairwise shortest path length between pairs of proteins in the interactome.

(ii) *Closest:* The average shortest path length to the closest protein in the side effect module from the drug targets, given by

$$d_{\text{Closest}}(T, S) = \frac{1}{\|T\|} \sum_{t \in T} \min_{s \in S} d(t, s)$$

We normalized these values using the mean and standard deviation of the values calculated for all the drugs as it was done above, yielding

$$z_{\text{Closest}}(T, S) = \frac{d_{\text{Closest}}(T, S) - \mu_{d_{\text{Closest}}(T,S)}}{\sigma_{d_{\text{Closest}}(T,S)}}$$

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(iii) *PageRank:* The average PageRank score of the drug targets when the proteins in the side effect module were used to weight the influence of the nodes in the network. We assigned higher priors to the proteins in the side effect module, 1, compared to the rest of the nodes that were assigned 0.01 and calculated the probability that a random walker in the network would end up in a certain node based on the following formula:

$$PR_{i+1}(u) = (1-d) * PR_0(u) + d \sum_{v \in Neighbors(u)} \frac{PR_i(v)}{degree(v)}$$

where $u$ was the current node in consideration, $v$ was a node connected to $u$, $PR_i(u)$ was the PageRank score at iteration $i$ and $d$ is *damping factor* that was set to 0.15. The algorithm was repeated till convergence. The drug - side effect closeness was then defined using

$$z_{\text{PageRank}}(T,S) = \frac{1}{\|T\|} \sum_{t \in T} PR^{normalized}_{i_{convergence}}(t)$$

where $PR^{normalized}_{i_{convergence}}(t)$ is the PageRank score of the target $t$ normalized using the mean and standard deviation of the PageRank scores of all nodes for the given side effect. We used PageRank with priors implementation in GUILD package [15].

(iv) *NetScore:* The average NetScore score of the drug targets when the proteins in the side effect module were used as the source of information passed among the nodes. NetScore scored all the nodes in the network by iteratively propagating the score of the proteins in the side effect module to the neighboring nodes through shortest paths [15]. Unlike conventional shortest path based algorithms, NetScore weighted by the number of alternative shortest paths in between two nodes. We used the implementation of NetScore in GUILD software package [15] and initialized the proteins with a score of 1 if they belong to the side effect module and 0.01, otherwise. We limited the number of repetitions the program used to 3 with each iterations in each of them. The drug - side effect closeness was then defined as

$$z_{\text{NetScore}}(T,S) = \frac{1}{\|T\|} \sum_{t \in T} NS^{normalized}(t)$$

where $NS^{normalized}(t)$ is the NetScore score of the target $t$ normalized using the mean and standard deviation of the NetScore scores of all nodes for the given side effect.

(v) *Proximity:* The significance of the observed average shortest path length to the closest protein in the side effect module from the drug targets. Interactome based proximity [18] first quantified the average shortest path

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

length between the closest protein in the side effect module and the drug targets ($d_{\text{Closest}}(T, S)$ above) and then calculated a z-score corresponding significance of these distances using

$$z_{\text{Proximity}}(T, S) = \frac{d_{\text{Closest}}(T, S) - \mu_{d_{\text{Closest}}(T,S)}}{\sigma_{d_{\text{Closest}}(T,S)}}$$

where $\mu_{d_{\text{Closest}}(T,S)}$ and $\sigma_{d_{\text{Closest}}(T,S)}$ are the mean and the standard deviation of the background distribution of expected minimum shortest path distances between two randomly selected groups of proteins (with the same size and degrees of the original protein sets). The background distance distribution was generated using 1,000 randomly selected protein groups matching drug targets and side effect proteins.

### Drug side effect prediction using network-based closeness

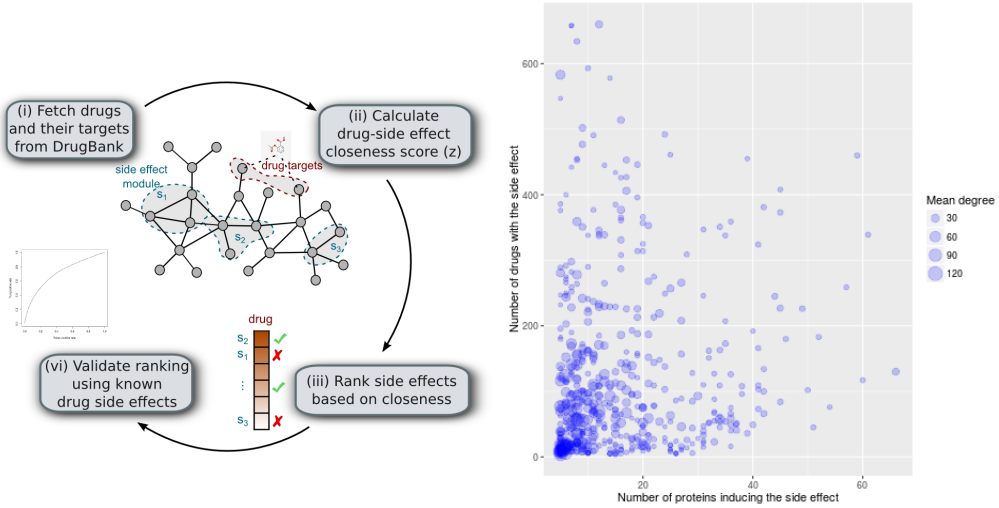
To investigate the whether network-based closeness predict side effects, for each known and unknown drug and side effect pair, we recorded the five topology based closeness scores ($z_{Shortest}$, $z_{Closest}$, $z_{PageRank}$, $z_{NetScore}$, $z_{Proximity}$). We then verified whether these topology based scores could discriminate known drug - side effect pairs from the rest by calculating the number of correctly and incorrectly predicted known and unknown drug - side effect pairs at various score cutoffs and checking the area under ROC curve (AUROC) and area under precision-recall curve (AUPRC). The known drug-side effect associations in SIDER and OFFSIDES databases were used as the gold standard positive instances and the remaining associations were assumed to be negative instances. We employed Python scikit-learn package to calculate AUROC and AUPRC values and R for the statistical tests.

## Results

### Side effect modules in the interactome

The available experimental information on the drug targets contributing to the side effects of drugs is often limited to a handful of drug targets [39, 40], hindering a large scale analysis of drug targets inducing the side effects. Alternatively, over-representation analysis of drug targets and side effects can characterize the targets eliciting side effects [38]. Therefore, we define the side effect modules as the groups of drug targets significantly associated with the side effects using the drug target information in DrugBank [22] and SIDER database [23]. Using 1,530 FDA approved drugs and their targets in DrugBank, we identify 1,177 drug target groups associated with the side effects. To confirm that the proteins defining the side effect modules are biologically relevant, we check the overlap between the side effect targets by Lounkine *et al.* [40]. The side effect modules cover at least one protein associated with the side effect for 164 of 241 side effects that are also in the Lounkine *et al.* study. Furthermore, 130 out 265

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.




(a) (b)

Figure 1: **Side effect modules in the interactome and their use in drug adverse effect characterization. (a)** Schematic overview of the interactome based analysis of drug side effect modules. For each of 817 drugs and 537 side effects, we calculate network based closeness between the drug targets and the proteins inducing the side effect and validate the predictions using known drug-side effect associations. **(b)** Each point represents a side effect consisting of proteins identified to be significantly associated to the side effect. The x-axis is the number of proteins in the side effect module and the y-axis is the number of drugs that shows the side effect. The size of the points scales with the median degree of the proteins in the side effect module.

of the proteins in the identified side effect modules appear among 224 proteins given in the Lounkine data set, covering more than half of the experimentally verified side effect targets.

To understand the interactome based relationship between drug targets and side effect modules, we focus on 537 side effect modules that have at least 5 proteins in the interactome and 817 drugs both known to exert any of these side effects and having at least one target in the interactome. We seek whether topological characteristics of these two groups of nodes, drug targets and side effect module proteins, can explain observed side effects of drugs (Figure 1a). We first turn our attention to the side effect module proteins and ask if the number of proteins in the module or their degree can provide insights on the side effects drugs show. The average module size is $\langle n_{\text{module}} \rangle = 15.8$ among 537 side effects and the largest module, the one of gynaecomastia (enlargement of a man's breasts), contains 66 proteins. Interestingly, the average degree of all the proteins contributing to a side effect is higher than the average degree of the remaining proteins in the interactome ($\langle k_{\text{side effect}} \rangle = 26.5$ vs $\langle k_{\text{non side effect}} \rangle = 21.1$). If

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the proteins within each side effect module are considered independently, however, the average degree of the proteins in the side effect modules is around the average degree of the interactome ($\langle k_{\text{module}} \rangle = 20.8$ vs $\langle k \rangle = 21.2$), with peliosis hepatis, an uncommon vascular condition in liver, being the side effect with the highest average degree ($\langle k_{\text{peliosis hepatis}} \rangle = 123.6$).

To investigate whether the size and the average degree of the identified side effect modules are higher for the "popular" side effects –the side effects that occur frequently in SIDER–, we look at the number of drugs the side effect is observed and the number and mean degree of the proteins in the side effect module (Figure 1b). The significant but low correlation between the number of drugs showing the side effects and the module size (Spearman's rank correlation $\rho = 0.16$, $P = 1.8 \times 10^{-4}$) suggests that the size of the module is not strongly associated to the occurrence of the side effects. On the other hand, the degree of the proteins within the side effect modules is not correlated with the number of drugs the side effect is observed (Spearman's rank correlation $\rho = 0.03$, $P = 0.55$).

## Network based closeness of drugs and side effects

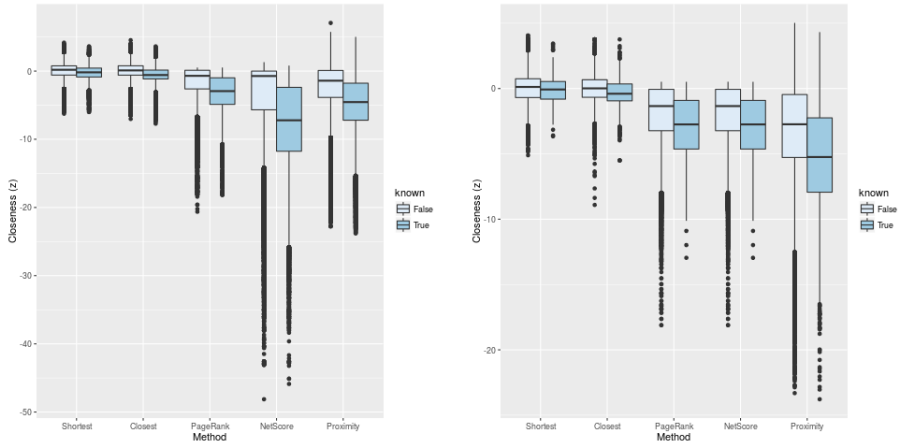
Next, for each drug and side effect pair in our analysis ($817 \times 537$ pairs), we calculate the network based closeness of the drug's targets to the side effect module in the interactome using five topological measures (see Methods). We then investigate how well the calculated closeness scores discriminate the observed drug side effects using the known drug side effect associations in SIDER and OFFSIDES databases (Table 1).

Table 1: Number of drugs, side effects and known drug - side effect associations included in the analysis according to SIDER and OFFSIDES databases.

| | **SIDER** | **OFFSIDES** |
|---|---|---|
| Number of drugs | 817 | 269 |
| Number of side effects | 537 | 118 |
| Number of known drug-side effect associations | 64,885 | 2,060 |
| Percentage of known associations | 14.8% | 6.5% |

We find that the drugs tend to be closer to the proteins inducing the side effects known to be associated with them compared to the proteins in the rest of the side effect modules (Figure 2). The difference in the closeness values of known and unknown drug - side effect pairs is significant using both SIDER and OFFSIDES side effect associations (one-sided Mann–Whitney U test $P \ll 0.05$). We observe that NetScore, the method that takes alternative shortest path between drug targets and side effect module proteins and Proximity, the method that compares observed shortest path length between drug targets and the closest side effect module protein to the distances between randomly selected nodes in the network yield a wider range of closeness scores than the remaining methods.

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.




(a) (b)

Figure 2: **Network based closeness of known and uknown drug - side effect pairs.** The closeness between drug targets and side effects calculated using five topological measures (Closest, Shortest, PageRank, NetScore and Proximity) for each of 817 drugs and 537 side effects. Known drug - side effect associations are taken from **(a)** SIDER and **(b)** OFFSIDES.

We then turn to predicting drug side effects using the network neighborhood information of the side effect modules and quantify the closeness between drug targets and side effect modules in the interactome. We use the drug-side effect associations in SIDER and OFFSIDES as the gold standard data to calculate the precision, recall, false positive rate at various closeness score cutoffs and check the area under the ROC curve (AUROC), the area under the precision-recall curve (AUPRC) and the percentage of the drugs for which the highest scoring prediction is a known side effect (Table 2). We see that, overall, the best performing methods are NetScore and Proximity, showing higher prediction accuracy on both SIDER and OFFSIDES data sets compared to the rest of the methods.

Table 2: AUROC, AUPRC and percentage of correctly predicted highest ranked drug - side effect pair for various network based closeness methods using SIDER and OFFSIDES associations.

| | **AUROC (%)** | | **AUPRC (%)** | | **Correct at top (%)** | |
|---|---|---|---|---|---|---|
| | SIDER | OFFSIDES | SIDER | OFFSIDES | SIDER | OFFSIDES |
| Shortest | 59.8 | 53.9 | 17.8 | 7.1 | 15.9 | 8.2 |
| Closest | 67.9 | 57.7 | 27.6 | 8.5 | 79.6 | 28.6 |
| PageRank | 69.0 | 59.6 | 27.0 | 8.6 | 55.8 | 13.0 |
| NetScore | 71.7 | 61.9 | 28.8 | 9.6 | 52.1 | 14.5 |
| Proximity | 71.1 | 63.6 | 32.8 | 11.4 | 56.7 | 11.5 |

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Despite using only the network topology the AUROCs for NetScore and Proximity scores on SIDER drug-side effect associations are 71.7% and 71.1%, respectively, suggesting that closeness of drugs to side effect modules is predictive of the drug's adverse effects. We also examine the area under precision-recall curve (AUPRC) and find that NetScore and Proximity achieve AUPRC values of 28.8% and 32.8%, respectively. Furthermore, for 52.1% and 56.7% of the drugs used in the analysis, the highest scoring side effect identified by NetScore and Proximity is reported in SIDER, showing that drug-side effect module closeness can provide insights on the side effects of drugs. On the other hand, when the drug-side effect associations in OFFSIDES database is used, the AUROC drops to 61.9% and 63.6% for NetScore and Proximity, still substantially higher than that would be expected from a classifier producing random predictions (50%). Moreover, only for around 10% of the drugs, the highest scoring side effect is in OFFSIDES, an observation we attribute to the lower coverage of known side effects in OFFSIDES database (6.5%) compared to the SIDER (14.8%, Table 1). Accordingly, due to the higher coverage of drugs and side effects, and better prediction accuracy, in the rest of the text, we use SIDER drug - side effect associations as the gold standard.

## Assessing the effect of the data incompleteness

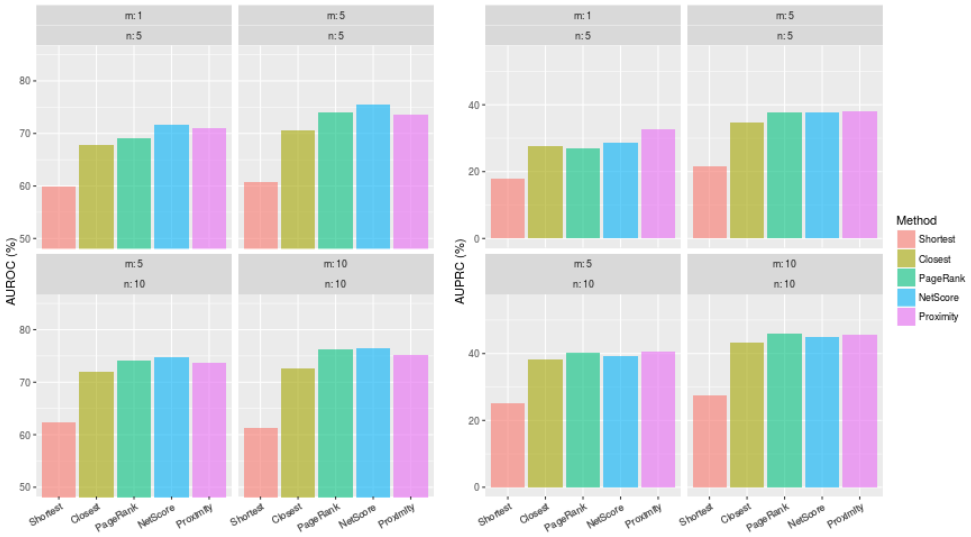
The current knowledge on drug-target interactions represent only a partial view of the possibly many proteins involved in drug's action [41]. To account for the potential implications of incompleteness of the drug target data, we analyze the prediction performance of each method on various subsets of drugs and side-effects categorized with respect to the number of drug targets ($m$) and side effect proteins ($n$). Figure 3 shows the AUROC and AUPRC values ($i$) on the original data set containing 817 drugs with at least one target and 537 side effect modules of at least five proteins ($m \geq 1$, $n \geq 5$) and when we repeat the analysis using ($ii$) 428 drugs and 537 side effects with at least five targets and proteins ($m \geq 5$, $n \geq 5$), ($iii$) 428 drugs with at least five targets and 322 side effect modules with at least ten proteins ($m \geq 5$, $n \geq 10$), and finally, ($iv$) 176 drugs and 322 side effects with at least ten proteins ($m \geq 10$, $n \geq 10$).

We find that, as the drugs and side effects associated with more proteins are used, the closeness based predictions improve. Nonetheless, the improvement mainly stems from the higher number of drug targets, as the change in the accuracy is modest when the number of proteins in the side effect modules increases. On the other hand, the AUROC and AUPRC values increase 3-6% when the drugs with more number of targets are used.

## Case study: Top ranking side effects of Tamoxifen

To highlight how interactome based closeness of drug targets can help identifying side effects, we use Proximity, the method that show high overall accuracy according to various performance measures (Table 2). Using only the target information of a given drug, Proximity calculates a network topology

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.




(a) (b)

Figure 3: **The effect of data incompleteness on prediction performance.** The area under **(a)** the ROC curve (AUROC) and **(b)** the precision-recall curve (AUPRC) values when a subset of the drugs and side effects are excluded from the analysis. In each panel, the drugs having less than $m$ targets in the network and the side effect modules that have less than $n$ proteins in the network are excluded from the analysis.

based significance of the closeness of the drug to all side effects, allowing us to rank the likelihood of all side effects for any drug with drug target information. Notably, among the drugs in our data set for which the top ranking side effect is not reported in SIDER, we see tamoxifen, an estrogen receptor modulator used for the treatment of breast cancer. Although eight out of ten highest scoring side effects are reported in SIDER, two side effects with very strong association scores, "muscular weakness" and "neuropathy peripheral" are not listed in SIDER (Table 3). We find out that the muscle weakness is indeed a known side effect according to the drug information in Medlineplus (nlm.nih.gov/medlineplus/druginfo/meds/a682414.html). Furthermore, while not indicated in neither SIDER nor Medlineplus, the peripheral neuropathy appears to be a clinically relevant condition reported by several patients in message boards (community.breastcancer.org/forum/78/topics/780591, medhelp.org/posts/Breast-Cancer/tamoxifen-and-neuropathy/show/261680, medhelp.org/posts/Breast-Cancer/Can-longer-term-tamoxifen-cause-peripheral-neuropathy/show/1384498).

The Proximity score of Tamoxifine to the 14 proteins associated to peripheral neuropathy is $z = -12.1$, suggesting that the drug targets are highly proximal to the side effect proteins in the interactome as a group. This is largely due to seven enzymes (*CYP1A2, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP3A4,*

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 3: Top 10 side effects predicted for tamoxifen using Proximity.

| Rank | Side effect | in SIDER | Proximity score (z) |
|---|---|---|---|
| 1 | muscular weakness | 0 | -12.9 |
| 2 | musculoskeletal discomfort | 1 | -12.3 |
| 3 | alopecia | 1 | -12.1 |
| 4 | neuropathy peripheral | 0 | -12.1 |
| 5 | drug interaction | 1 | -11.7 |
| 6 | hepatitis | 1 | -11.7 |
| 7 | diarrhoea | 1 | -11.7 |
| 8 | myalgia | 1 | -11.6 |
| 9 | injury | 1 | -11.5 |
| 10 | discomfort | 1 | -11.3 |

*CYP3A7*) and two transporters (*ABCB1, ABCC2*) tamoxifen is known to bind are in the side effect module. Furthermore, protein encoded by *KIT* gene in the side effect module, is known to be inhibited via phosphorylation by Protein kinase C protein family, a family of proteins targeted by tamoxifen, contributing to the observed proximity to the peripheral neuropathy.

# Discussion

Most existing approaches rely on existing drug side effect associations to predict drug side effects, hindering both the interpretability of predicted associations and the ability to discover novel side effects. In contrast, in this study, we investigate the network based closeness of drug targets to the proteins likely to induce the side effects to explain the observed drug adverse effects. We use the interactome based closeness to predict side effects associated with a drug, providing a mechanistic explanation of the predicted association.

We start with defining the proteins inducing side effects and show that the proteins used to define side effect modules have fair coverage of known side effect inducing proteins. We find that though the proteins likely to induce the side effects show a slight tendency to have higher degrees in the interactome, the effect of degree is not prominent when the side effect modules are considered individually. We also find that the size and the average degree of the identified side effect modules are not higher for the side effects that occur frequently in SIDER. Taken together these findings suggest that the number or degrees of the proteins in the modules are not a good descriptor of observed side effects.

The AUROC values for drug adverse effect prediction reported in the literature range between 60-90% depending on the validation scheme, data sets, predictive models and features (see [42] for a recent review). In particular, compared to the predictor combining canonical correlation analysis using chemical similarity and network diffusion on side effect similarity network by Atias and Sharan, Proximity identifies a known side effect as the top scoring side effect for 56.7% of the drugs in contrast to 34.7% of the drugs reported in the original study [7]. In another study, Huang *et al.* reported an AUROC value (70%)

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

similar to that of using Proximity for the support vector machine based predictor that combined various features including PPI network neighborhood of drug targets and drug structural properties [9].

One drawback of network based methods is that they require that at least a drug target known to interact with a protein in the interactome. Furthermore, they can only be applied to side effects for which a set of proteins inducing the side effect can be identified. Yet, we show that interactome based closeness can systematically detect side effects of 817 FDA approved drugs in DrugBank without relying on the known drug-disease associations. Moreover, network based closeness offers an important advantage over similarity based methods by providing interactome-based insights on the likelihood of a drug to induce a given side effect.

## Acknowledgements

The author is grateful to Dr. Patrick Aloy for providing computational resources for this study and the members of the lab for fruitful discussions. EG is supported by EU-cofunded Beatriu de Pinós incoming fellowship from the Agency for Management of University and Research Grants (AGAUR) of Government of Catalunya.

## References

[1] Allison, M.: Reinventing clinical trials. Nature Biotechnology **30**(1), 41–49 (2012). doi:10.1038/nbt.2083

[2] Hay, M., Thomas, D.W., Craighead, J.L., Economides, C., Rosenthal, J.: Clinical development success rates for investigational drugs. Nature Biotechnology **32**(1), 40–51 (2014). doi:10.1038/nbt.2786. Accessed 2016-03-21

[3] Wu, T.-Y., Jen, M.-H., Bottle, A., Molokhia, M., Aylin, P., Bell, D., Majeed, A.: Ten-year trends in hospital admissions for adverse drug reactions in England 1999-2009. Journal of the Royal Society of Medicine **103**(6), 239–250 (2010). doi:10.1258/jrsm.2010.100113

[4] Krejsa, C.M., Horvath, D., Rogalski, S.L., Penzotti, J.E., Mao, B., Barbosa, F., Migeon, J.C.: Predicting ADME properties and side effects: the BioPrint approach. Current Opinion in Drug Discovery & Development **6**(4), 470–480 (2003)

[5] Fliri, A.F., Loging, W.T., Thadeio, P.F., Volkmann, R.A.: Analysis of drug-induced effect patterns to link structure and side effects of medicines. Nature Chemical Biology **1**(7), 389–397 (2005). doi:10.1038/nchembio747. Accessed 2016-03-16

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[6] Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P.: A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology **6**, 343 (2010). doi:10.1038/msb.2009.98

[7] Atias, N., Sharan, R.: An algorithmic framework for predicting side effects of drugs. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology **18**(3), 207–218 (2011). doi:10.1089/cmb.2010.0255

[8] Duran-Frigola, M., Aloy, P.: Analysis of chemical and biological features yields mechanistic insights into drug side effects. Chemistry & Biology **20**(4), 594–603 (2013). doi:10.1016/j.chembiol.2013.03.017

[9] Huang, L.-C., Wu, X., Chen, J.Y.: Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. PROTEOMICS **13**(2), 313–324 (2013). doi:10.1002/pmic.201200337. Accessed 2016-03-16

[10] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics **21**(Suppl 1), 302–310 (2005). doi:10.1093/bioinformatics/bti1054. Accessed 2012-06-04

[11] Chua, H.N., Sung, W.-K., Wong, L.: Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein–Protein Interactions. Bioinformatics **22**(13), 1623–1630 (2006). doi:10.1093/bioinformatics/btl145. Accessed 2012-06-04

[12] Wu, X., Jiang, R., Zhang, M.Q., Li, S.: Network-based global inference of human disease genes. Molecular Systems Biology **4**, 189 (2008). doi:10.1038/msb.2008.27. Accessed 2016-03-18

[13] Kohler, S., Bauer, S., Horn, D., Robinson, P.N.: Walking the Interactome for Prioritization of Candidate Disease Genes. The American Journal of Human Genetics **82**(4), 949–958 (2008). doi:10.1016/j.ajhg.2008.02.013. Accessed 2012-02-22

[14] Navlakha, S., Kingsford, C.: The Power of Protein Interaction Networks for Associating Genes with Diseases. Bioinformatics **26**(8), 1057–1063 (2010). doi:10.1093/bioinformatics/btq076. Accessed 2012-06-04

[15] Guney, E., Oliva, B.: Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE **7**(9), 43557 (2012). doi:10.1371/journal.pone.0043557. Accessed 2013-08-16

[16] Zhao, S., Li, S.: A co-module approach for elucidating drug-disease associations and revealing their molecular basis. Bioinformatics **28**(7), 955–961 (2012). doi:10.1093/bioinformatics/bts057. Accessed 2014-04-02

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[17] Guney, E., Garcia-Garcia, J., Oliva, B.: GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. Bioinformatics (Oxford, England) **30**(12), 1789–1790 (2014). doi:10.1093/bioinformatics/btu092

[18] Guney, E., Menche, J., Vidal, M., Barabási, A.-L.: Network-based in silico drug efficacy screening. Nature Communications **7**, 10331 (2016). doi:10.1038/ncomms10331. Accessed 2016-03-24

[19] Berger, S.I., Ma'ayan, A., Iyengar, R.: Systems pharmacology of arrhythmias. Science Signaling **3**(118), 30 (2010). doi:10.1126/scisignal.2000723

[20] Brouwers, L., Iskar, M., Zeller, G., van Noort, V., Bork, P.: Network neighbors of drug targets contribute to drug side-effect similarity. PloS One **6**(7), 22187 (2011). doi:10.1371/journal.pone.0022187

[21] Berger, S.I., Iyengar, R.: Role of systems pharmacology in understanding drug adverse events. Wiley Interdisciplinary Reviews: Systems Biology and Medicine **3**(2), 129–135 (2011). doi:10.1002/wsbm.114. Accessed 2016-03-24

[22] Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y., Wishart, D.S.: DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Research **42**(Database issue), 1091–1097 (2014). doi:10.1093/nar/gkt1068

[23] Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The SIDER database of drugs and side effects. Nucleic Acids Research **44**(D1), 1075–1079 (2016). doi:10.1093/nar/gkv1075

[24] Tatonetti, N.P., Ye, P.P., Daneshjou, R., Altman, R.B.: Data-driven prediction of drug effects and interactions. Science Translational Medicine **4**(125), 125–31 (2012). doi:10.1126/scitranslmed.3003377

[25] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., Barabási, A.-L.: Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science (New York, N.Y.) **347**(6224), 1257601 (2015). doi:10.1126/science.1257601

[26] Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., Wingender, E.: TRANSFAC: transcriptional regulation, from patterns to profiles. Nucl. Acids Res. **31**(1), 374–378 (2003). doi:10.1093/nar/gkg108. Accessed 2014-09-26

[27] Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J.,

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., Eijk, K.v., Hermjakob, H.: The IntAct molecular interaction database in 2010. Nucleic Acids Research **38**(suppl 1), 525–531 (2010). doi:10.1093/nar/gkp878. Accessed 2014-09-26

[28] Ceol, A., Aryamontri, A.C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., Cesareni, G.: MINT, the molecular interaction database: 2009 update. Nucl. Acids Res. **38**, 983 (2009). doi:10.1093/nar/gkp983. Accessed 2014-09-26

[29] Stark, C., Breitkreutz, B.-J., Chatr-aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., Tyers, M.: The BioGRID Interaction Database: 2011 update. Nucleic Acids Research **39**(Database), 698–704 (2010). doi:10.1093/nar/gkq1116. Accessed 2012-02-23

[30] Prasad, T.S., Kandasamy, K., Pandey, A.: Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol **577**, 67–79 (2009)

[31] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. Nucl. Acids Res. **34**(suppl 1), 354–357 (2006). doi:10.1093/nar/gkj102. Accessed 2014-09-26

[32] Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., Palsson, B.Ø.: Global reconstruction of the human metabolic network based on genomic and bibliomic data. PNAS **104**(6), 1777–1782 (2007). doi:10.1073/pnas.0610772104. Accessed 2014-09-26

[33] Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., Mewes, H.W.: CORUM: the comprehensive resource of mammalian protein complexes. Nucl. Acids Res. **36**(suppl 1), 646–650 (2008). doi:10.1093/nar/gkm936. Accessed 2014-09-26

[34] Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., Sullivan, M.: PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Research **40**(D1), 261–270 (2012). doi:10.1093/nar/gkr1122. Accessed 2014-09-26

[35] Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski,

bioRxiv preprint doi: https://doi.org/10.1101/089730; this version posted November 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M.: Towards a proteome-scale map of the human protein-protein interaction network. Nature **437**(7062), 1173–1178 (2005)

[36] Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M.A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J.M., Cevik, S., Simon, C., de Smet, A.-S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R.R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M.E., Roth, F.P., Hill, D.E., Tavernier, J., Wanker, E.E., Barabási, A.-L., Vidal, M.: An empirical framework for binary interactome mapping. Nature Methods **6**(1), 83–90 (2009). doi:10.1038/nmeth.1280. Accessed 2014-09-26

[37] Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., Assmus, H.E., Andrade-Navarro, M.A., Wanker, E.E.: A directed protein interaction network for investigating intracellular signal transduction. Science Signaling **4**(189), 8–8 (2011). doi:10.1126/scisignal.2001699. Accessed 2014-09-26

[38] Kuhn, M., Banchaabouchi, M.A., Campillos, M., Jensen, L.J., Gross, C., Gavin, A.-C., Bork, P.: Systematic identification of proteins that elicit drug side effects. Molecular Systems Biology **9**(1), 663 (2013). doi:10.1038/msb.2013.10. Accessed 2015-01-26

[39] Ji, Z.L., Han, L.Y., Yap, C.W., Sun, L.Z., Chen, X., Chen, Y.Z.: Drug Adverse Reaction Target Database (DART) : proteins related to adverse drug reactions. Drug Safety **26**(10), 685–690 (2003)

[40] Lounkine, E., Keiser, M.J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J.L., Lavan, P., Weber, E., Doak, A.K., Côté, S., Shoichet, B.K., Urban, L.: Large-scale prediction and testing of drug activity on side-effect targets. Nature **486**(7403), 361–367 (2012). doi:10.1038/nature11159. Accessed 2016-03-16

[41] Mestres, J., Gregori-Puigjané, E., Valverde, S., Solé, R.V.: Data completeness—the Achilles heel of drug-target networks. Nat Biotech **26**(9), 983–984 (2008). doi:10.1038/nbt0908-983. Accessed 2014-04-02

[42] Ivanov, S.M., Lagunin, A.A., Poroikov, V.V.: In silico assessment of adverse drug reactions and associated mechanisms. Drug Discovery Today **21**(1), 58–71 (2016). doi:10.1016/j.drudis.2015.07.018