

PDID: Database of Experimental and Putative Drug Targets in Human Proteome

Chen Wang^{*}, Michal Brylinski^{†,‡}, Lukasz Kurgan^{*}

^{*}Department of Computer Science, Virginia Commonwealth University, Richmond, VA, United States

[†]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, United States

[‡]Center for Computation & Technology, Louisiana State University, Baton Rouge, LA, United States

1 INTRODUCTION

Drugs produce biological effects via interactions with target molecules that include proteins, DNA, RNAs, and membrane components (Yang et al., 2016). While a comprehensive mapping of the drug targets remains an open challenge (Santos et al., 2017), numerous studies have shown that well over 90% of the marketed drug targets are proteins (Hopkins & Groom, 2002; Rask-Andersen, Almén, & Schiöth, 2011; Rask-Andersen, Masuram, & Schiöth, 2014; Santos et al., 2017). The knowledge of drug-protein interactions (DPIs) is essential for a diverse set of applications, including screening drug candidates to target specific proteins (Schneider, 2010; Tseng & Tuszynski, 2015), repurposing of drugs (Chong & Sullivan, 2007; Haupt & Schroeder, 2011; Li et al., 2016), and identifying side effects related to interactions with off-targets (Hu et al., 2014; Lounkine et al., 2012; Wang, Li, Qiu, Wang, & Cui, 2012). This information is also essential to elucidate the druggable human proteome/genome, defined as the complement of human proteins interacting with drugs (Cimermancic et al., 2016; Hopkins & Groom, 2002; Hu, Wu, Wang, Uversky, & Kurgan, 2016; Rask-Andersen et al., 2014; Russ & Lampel, 2005).

The studies mentioned are assisted by a number of databases currently offering access to large collections of DPIs (Chen et al., 2016; Glaab, 2016). The two arguably most popular databases of experimentally determined DPIs include DrugBank (Wishart et al., 2006, 2018)

and Therapeutic Target Database (TTD) (Chen, Ji, & Chen, 2002; Li et al., 2018). DrugBank includes biochemical and pharmacological data for over 10,000 drugs including 3254 FDA-approved compounds, 5124 experimental compounds and their 5020 protein targets. TTD covers over 23,000 drugs, including close to 15,000 experimental drugs and 2360 protein targets, and links this information to about 900 diseases. Other similar resources include PDSP Ki (Roth, Lopez, Patel, & Kroeze, 2000), KEGG DRUG (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017; Kanehisa, Goto, Furumichi, Tanabe, & Hirakawa, 2009), Matador (Günther et al., 2008), and PROMISCUOUS (von Eichborn et al., 2011). Table 1 lists their release dates, numbers of drugs and targets they cover, addresses of their websites, and key statistics comprising the total number of DPIs and an average number of DPIs per drug.

TABLE 1 Summary of the Databases of Drug-Protein Interactions (DPIs)

Type	Database	Release Date	Number of Drugs	Number of Targets	Key Statistics	URL
Experimental annotations	PDSP Ki	11/01/1999	11,569	1673	63,619 DPIs 5.5 DPIs per drug	https://pdsp.unc.edu/databases/kidb.php
	TTD	01/01/2002	23,486	3036	33,467 DPIs 1.4 DPIs per drug	http://bidd.nus.edu.sg/BIDD-Databases/TTD/
	KEGG DRUG	07/01/2005	5045	1061	14,222 DPIs 2.8 DPIs per drug	http://www.genome.jp/kegg/drug/
	DrugBank	01/01/2006	10,562	5020	23,380 DPIs 2.2 DPIs per drug	http://www.drugbank.ca
	Matador	10/16/2007	801	2901	15,843 DPIs 19.8 DPIs per drug	http://matador.embl.de
	PROMISCUOUS	11/10/2010	5000	6500	21,500 DPIs 4.3 DPIs per drug	Unavailable (no longer supported)
Putative annotations	BioDrugScreen	11/18/2009	1592	1589	3,066,192 predictions	Unavailable (no longer supported)
	PDID	10/01/2014	51	3746	1,088,789 predictions 16,800 DPIs and 100 DPIs per drug	http://biomine.cs.vcu.edu/servers/PDID/

The databases are divided into two groups that offer access to experimental vs. putative data. They are sorted chronologically by the date of their release. We collected these dates from the release notes or time stamps recorded on the database websites, if available, and we use the date of the first publication otherwise.

These resources have already been available for about a decade or more, are constantly upgraded and updated, and include data on thousands of interactions. There are also numerous databases that expand beyond the drug molecules to cover interactions with many more small, drug-like ligands. These databases include BRENDA (Placzek et al., 2017; Schomburg, Hofmann, Baensch, Chang, & Schomburg, 2000), BindingDB (Chen, Liu, & Gilson, 2001; Gilson et al., 2016), GLIDA (Okuno et al., 2008; Okuno, Yang, Taneishi, Yabuuchi, & Tsujimoto, 2006), SuperTarget (Günther et al., 2008; Hecker et al., 2012), STITCH (Kuhn, von Mering, Campillos, Jensen, & Bork, 2008; Szklarczyk et al., 2016), and ChEMBL (Gaulton et al., 2012, 2017). The largest two of these databases are ChEMBL, which covers over 14 million interactions between over 2.1 million chemical compounds and about 11 thousand protein targets, and STITCH, which contains 1.6 billion interactions between about 0.5 million chemicals and 9.6 million proteins coming from over 2000 species. One caveat of the latter repository is that it stores many low-quality, putative, and indirectly inferred annotations. Another group of relevant databases is dedicated to drugs that target protein-protein interactions. These include TIMBAL (Higueruelo, Jubb, & Blundell, 2013; Higueruelo et al., 2009), 2P2Idb (Basse, Betzi, Morelli, & Roche, 2016; Bourgeas, Basse, Morelli, & Roche, 2010), and iPPI-DB (Labbé et al., 2016; Labbé, Laconde, Kuenemann, Villoutreix, & Sperandio, 2013).

The databases mentioned primarily focus on the already-known drug targets. However, drugs typically interact not only with the therapeutic targets that they were designed for but also with often unidentified and numerous off-targets. Comprehensive identification and cataloguing of the off-targets are crucial to fully understand how drugs work. Although interactions with off-targets may result in adverse events or side effects, they may also present an opportunity to repurpose drugs for diseases that they were not originally intended for (Peters, 2013). Studies have shown that drugs interact with on average 6.3 proteins (Hu & Bajorath, 2013; Mestres, Gregori-Puigjane, Valverde, & Sole, 2008) and this number is likely much higher given that the current information on DPIs is likely highly incomplete (Mestres et al., 2008; Peters, 2013). Experimental screening of drugs is limited to a relatively small panel of protein targets (Lavecchia & Giovanni, 2013; Mestres et al., 2008). Example panels used by pharmaceutical companies include anywhere between 15 and 48 proteins (Bendels et al., 2013; Bowes et al., 2012; Urban, 2012; Wang & Greene, 2012). The high levels of drug promiscuity and the relatively low coverage of the experimental screening motivate the development of high-throughput computational methods that predict DPIs on the druggable-genome and even whole-genome scale (Chen et al., 2016; Ding, Takigawa, Mamitsuka, & Zhu, 2014; Ezzat, Wu, Li, & Kwoh, 2018; Hao, Bryant, & Wang, 2018; Lavecchia & Cerchia, 2016). The availability of these computational tools has also spurred the development of databases providing access to putative DPIs. The release of these databases is primarily motivated by the often high computational cost of running these predictions and the ability to query the predictions across multiple drugs and thousands of protein targets. Table 1 summarizes details on the two databases that directly focus on the putative DPIs: BioDrugScreen (Li et al., 2010) and PDID (Protein-Drug Interaction Database) (Wang et al., 2016). BioDrugScreen is based on results of the docking of about 1600 small drug-like molecules against 1589 human protein targets that were collected from DrugBank and HCPIN (Huang et al., 2008) databases. The molecular docking was performed for nearly 2000 surface pockets in these proteins, producing about 3 million putative compound-protein complexes. However, this resource is no longer available. PDID is based on predictions that rely on structural similarity between a

large database of structures of drug-protein complexes and a query protein and its binding sites for a given query drug. These predictions complement the contents of the BioDrugScreen database. PDID also provides access to the annotations of experimental DPIs that are linked to their sources: PDB (Berman et al., 2000; Rose et al., 2017), DrugBank, and BindingDB. Importantly, PDID offers molecular level details for DPIs, in terms of coordinates of the location of the drugs in the structures of their predicted and experimental protein targets. We also acknowledge the Dr. PIAS database of putative druggable protein-protein interactions (Sugaya & Furuya, 2011; Sugaya, Kanai, & Furuya, 2012). These interactions were predicted with the help of machine-learning algorithms and they cover over 83 thousand protein-protein interactions in human, mouse, and rat. However, they lack associations with specific drugs.

This chapter focuses on the only currently available resource that comprehensively covers the experimental and putative DPIs, the PDID database. We describe the algorithms that are used to make predictions that are stored in PDID, briefly comment on the predictive quality of these algorithms, summarize and analyze content of a current version of PDID, and discuss how to access and use this resource.

2 DEVELOPMENT AND OUTLINE OF THE PDID DATABASE

The fundamental principle behind the predictors that are used to implement PDID is to transfer binding sites from known drug-protein complexes to a protein that is known to interact with the input drug and that is structurally similar to these known drug-complexed proteins. There are two ways to measure similarity between the input protein structure and the known drug-protein complexes. The first quantifies similarity of the corresponding protein folds. The corresponding methods include *eFindSite* (Brylinski & Feinstein, 2013; Feinstein & Brylinski, 2014) and its predecessor FINDSITE (Michal Brylinski & Skolnick, 2008; Skolnick & Brylinski, 2009). The second way exploits similarity of binding sites, with example algorithms that include SMAP (Xie & Bourne, 2007, 2008; Xie, Xie, & Bourne, 2009) and IsoMIF (Chartier, Adriansen, & Najmanovich, 2016; Chartier & Najmanovich, 2015). We also developed the ILbind (inverse ligand binding) method that combines these two types of approaches to improve predictive performance (Hu et al., 2012).

The PDID database provides convenient access to query and retrieve results generated by three predictors: *eFindSite*, SMAP, and ILbind. They represent each of the two types of approaches and their ensemble. These putative annotations are combined with experimental data collected from the DrugBank and BindingDB resources. Fig. 1 overviews the approach used to populate the PDID database with data. The following subsections explain details of the three predictors, which are shown in Fig. 1 using the blue boxes (named boxes in the middle of the figure).

2.1 *eFindSite*

eFindSite makes predictions using template protein(s) that have structure(s) in complex with a given input drug. The template proteins come from a database of templates, typically defined as a set of nonredundant high-quality structures of drug-protein complexes curated

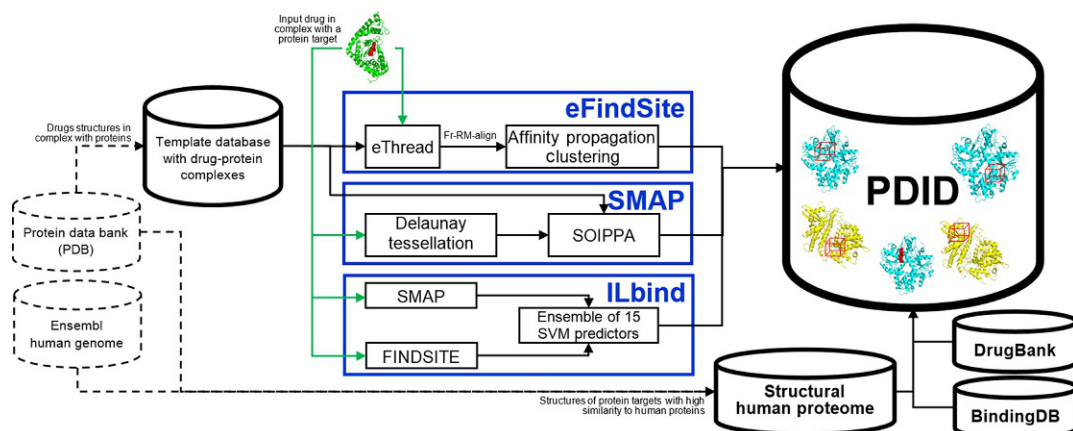


FIG. 1 Flowchart of the approach used to populate the PDID database with data. *Blue boxes* (named boxes in the middle of the figure) represent the three predictors that were used to generate putative DPIs. *Cylinders* shown with solid lines represent databases that are used directly to derive the PDID resources. The *dashed lines* denote sources that were used indirectly to develop the directly used databases. The *green lines* (lines coming from the drug-protein complex) indicate how the drugs covered in PDID were utilized to generate the putative data. The *red cubes* (cubes shown inside the PDID database) signify the fact that PDID provides approximate location of the DPI sites.

from the PDB database; *eFindSite* also accepts putative structures modeled computationally with TASSER (Zhang, Arakaki, & Skolnick, 2005; Zhang & Skolnick, 2004a, 2004b), MODELLER (Martí-Renom et al., 2000; Šali & Blundell, 1993), or PROSPECTOR3 (Skolnick, Kihara, & Zhang, 2004). The *eFindSite*'s predictive protocol includes four steps:

1. A meta-threading algorithm, *eThread* (Brylinski & Feinstein, 2012; Brylinski & Lingam, 2012), is used to find similar template proteins for the protein that is in complex with the input drug; *eThread* applies a Naïve Bayes classifier to build a consensus threading alignment from 10 individual threading algorithms. This algorithm recognizes template proteins that likely have similar structural folds when compared to the input protein based on sequence alignment and predicted secondary structure, given the sequence of the input protein. The identified template proteins are supposed to be structurally similar to the input protein, regardless of whether or not they have high sequence similarity to the input protein.
2. Those template proteins complexed with the input drug are selected from the template set that is obtained by the threading alignment. Then the template set is expanded by including homologous proteins of the current templates. Consequently, we collect a set of template proteins that interact with the input drug, have known three-dimensional structures of drug-protein complex, are likely structurally similar to the input protein, and are possibly remotely homologous to the input protein.
3. A clustering algorithm, Affinity Propagation (Frey & Dueck, 2007), is utilized to group the structures of template drug-protein complexes based on structural similarities between templates computed with fr-TM-align (Pandit & Skolnick, 2008). The clustered template structures are superimposed into the structure of input protein.

4. Finally, the resulting superimposed locations of drug from each clustered template constitute the predicted position of the input drug, which in turn can be used to annotate putative binding sites in the input protein structures. These predicted binding sites are ranked by the number of templates from the corresponding cluster where each binding site (cluster center) comes from.

2.2 SMAP

SMAP works by generating potential binding pockets in the input protein structure and then finding whether these pockets are similar to the known binding pockets in the template drug-protein complexes (Xie et al., 2009; Xie & Bourne, 2007, 2008). It employs a geometric representation of protein structure to characterize binding sites (Xie & Bourne, 2007) and a sequence profile alignment to compare binding sites (Xie & Bourne, 2008). The predictive protocol of SMAP algorithm includes the following five steps:

1. SMAP reduces the representation of an input protein structure by using only the coordinates of alpha carbon (C_α) atoms, which are the first carbon atoms attached to the carboxyl group of an amino acid. The C_α atoms are represented as vertices in a graph. A convex hull algorithm, Delaunay tessellation, is applied to partition the C_α atoms into tetrahedra (triangular pyramids) that are defined by the graph edges (Xie & Bourne, 2007).
2. The Delaunay tessellation is constrained by removing those tetrahedra including edges (atomic distances) longer than 30 Å, because such distance indicates an open binding pocket on the molecular surface, rather than an enclosed sphere. The outside layer of the remained convex hull defines an environmental boundary that surrounds the input protein and its binding pockets. Next, the tetrahedra larger than 7.5 Å are removed. This cut-off length is related to the average radius between two amino acids that are in contact with each other. The remaining tetrahedra on the outside of the structure form a protein boundary. The removed tetrahedra, which are the tetrahedra located between the protein boundary and the environmental boundary, make up the possible positions where drugs could be located.
3. The distance and orientation of each C_α atom to the protein boundary and environmental boundary are used to compute a geometric potential with specific formulas listed in (Xie & Bourne, 2007) for each C_α atom. The geometric potential quantifies the positions of a C_α atom and its neighboring atoms relative to the environmental boundary, and the relative positions between this C_α atom and its neighbor C_α atoms.
4. The possible positions of drugs obtained in step (2) are clustered based on their overlap in circumscribed spheres of the corresponding tetrahedron. The cluster centers represent the predicted potential positions of the drugs that could bind to the given protein. If a C_α atom is located within 10 Å from the predicted potential positions and the edge between these two atoms are not cut by other circumscribed spheres, then this C_α atom (representing a drug-binding amino acid) is predicted as a part of a binding site. This way the amino acids that make up specific binding pockets are defined.
5. The predicted potential positions of drugs generated in step (4) represent a candidate drug-binding position but without specifying for which specific drug. SMAP uses a sequence order independent profile-profile alignment (SOIPPA) method to align the candidate

drug-binding sites (the corresponding amino acids) in the input protein to the known binding sites in the template proteins that are in complex with the input drug in the PDB database (Xie & Bourne, 2008). Next, a candidate binding site is mapped to a known binding site of the specific input drug if the SOIPPA alignment shows that these two sites are similar. The SOIPPA algorithm is designed to compare and align two subgraphs that are extracted from the geometric representations of the input protein and the template protein. The computation of the alignment uses the geometric potential scores computed in step (3). The binding sites from the template protein are aligned and superimposed as the candidate sites in the input protein. An alignment score is computed based on the position specific score matrix (Altschul et al., 1997) to measure the similarity of these binding sites.

2.3 ILbind

ILbind is an ensemble predictor that uses a machine-learning algorithm to predict drug-binding sites for a specific input drug (Hu et al., 2012). This meta-approach exploits the fact that FINDSITE and SMAP use complementary approaches to provide predictions for a wide range of drugs and nutraceuticals. ILbind uses selected outputs generated by FINDSITE and SMAP as its inputs. First, a dataset of ~ 150 drugs was clustered into three structurally similar groups. These clusters were represented by three drugs that correspondingly have diverse structures: *N*-acetyl-D-glucosamine (NAG), adenosine-5'-diphosphate (ADP), and palmitic acid (PLM) (Hu et al., 2012). Structures of five randomly chosen complexes of proteins with each of these three drugs were used to design ILbind, resulting in a total of 15 configurations. Correspondingly, 15 support vector machine (SVM) models were custom designed using a training dataset. Outputs generated by FINDSITE and SMAP were ranked by their average area under the curve (AUC) values on the training set. Next, a wrapper-based best-first search was applied to select the best subset of FINDSITE and SMAP outputs for each configuration using cross-validation tests on the training dataset. Given an input drug and the structure of an input protein, ILbind works in two steps:

1. Compute predictions with the 15 SVM models using the selected outputs of FINDSITE and SMAP as the inputs.
2. Use a consensus (average) of the 15 SVM predictions as the predicted propensity for binding to the input drug.

Since ILbind does not predict the putative position of the center of the input drug, these positions are taken directly from the outputs of FINDSITE and SMAP.

A web server and a standalone version of the ILbind method are available at <http://biomine.cs.vcu.edu/servers/ILbind/>. This is particularly useful for the users who would like to collect predictions for compounds that extend beyond the 51 drugs that are currently covered by PDID.

2.4 Predictive Quality of ILbind, eFindSite, and SMAP

The predictive quality of outputs generated by ILbind, eFindSite, and SMAP was recently evaluated empirically based on a dataset that covers interactions between 25 representative

drugs and a representative collection of human proteins (Wang et al., 2016). The test drugs represent 25 clusters of chemically similar drug structures resulting in a broad sampling of the drug structure space that is included in PDB. The evaluation follows a protocol from (Hu et al., 2014). In short, the native DPIs for the 25 drugs that were collected from PDB, BindingDB, and DrugBank are compared against the predictions from the three methods in the structural human proteome (collection of all structures of human proteins in PDB). AUCs values for *eFindSite*, SMAP, and ILbind that are averaged over the 25 drugs are 0.63, 0.74, and 0.76, respectively. These are reasonably good results given that the AUC values range between 0.5 and 1. As expected, ILbind outperforms the other two methods, but this advantage is not universal. The empirical tests reveal that *eFindSite* provides the highest AUC for 5 drugs, SMAP for 6 drugs, and ILbind for the remaining 14 drugs. About 40% of the native drug targets are predicted within the top 4% of predictions from ILbind and SMAP and among the top 14% of predictions from *eFindSite*. Moreover, the results are better for medium-sized drugs (molecular weight between 200 and 400 Da) when compared to either small (<200 Da) or large drugs (>400 Da) drugs. More specifically, the best overall method ILbind secures AUCs for the small, medium, and large drugs that equal 0.70, 0.86, and 0.59, respectively. To sum up, the empirical results show that the three algorithms offer practical levels of predictive performance. More details concerning the evaluations of these three tools can be found in Hu et al. (2012, 2014) and Wang et al. (2016).

3 CONTENT OF THE PDID DATABASE

The current release 1.1 of the PDID database covers 51 drugs, 3746 protein targets that are represented by 9652 structures, 1,088,789 predicted DPIs, and 730 experimental annotations of DPIs. Several other key statistics are summarized in Table 2. PDID is linked to the relevant drug and protein databases. In particular, PDID links drugs and native drug target annotations to the corresponding PDB, DrugBank, and BindingDB entries, while the protein targets

TABLE 2 Key Statistics of the Current Release of the PDID Database

Number of Drugs		51
Number of proteins (protein structures)		3746 (9652)
Number of predictions of interactions		1,088,789
Number of experimentally determined protein targets		730
ILbind predictions	Number of putative targets	5172
	Median number of putative targets per drug	31
SMAP predictions	Number of putative targets	7184
	Median number of putative targets per drug	23
<i>eFindSite</i> predictions	Number of putative targets	4444
	Median number of putative targets per drug	30

are linked to their UniProt and PDB records. Importantly, PDID also provides approximate location of the drug molecules relative to the structures of their protein targets for each of the 16,800 putative interactions and many of the 730 experimentally determined interactions.

Fig. 2 shows a breakdown of the number of protein targets across the drugs that are covered in PDID. The gray bars denote the number of experimental annotations that are included in PDID and that are linked to PDB, DrugBank, and BindingDB. Erlotinib features the largest number of experimentally annotated targets at 216, while the median number of targets per drug equals 8. The colored bars represent the numbers of putative targets generated by ILbind (orange/second from the bottom), SMAP (green/third from the bottom) and *e*FindSite (blue/fourth from the bottom). The largest number of predictions was generated by SMAP (7184), followed by ILbind (5172) and *e*FindSite (4444). The corresponding median number of targets per drug generated by SMAP, ILbind, and *e*FindSite is 23, 31, and 30, respectively. Moreover, the number at the base of the bars shows the total number of experimental and computationally annotated unique targets, in contrast to the bars where some of the targets are annotated multiple times by different tools. More specifically, the total number of experimental and putative annotations of drug-target interactions across the 51 drugs equals 17,530 while the number of unique drug-target interactions is 13,791. This means that a substantial number of interactions are annotated by two or more methods. We caution the reader that some of the putative interactions in PDID could be false positives and thus the previously mentioned statistics should be taken as a ceiling for the actual number of interactions.

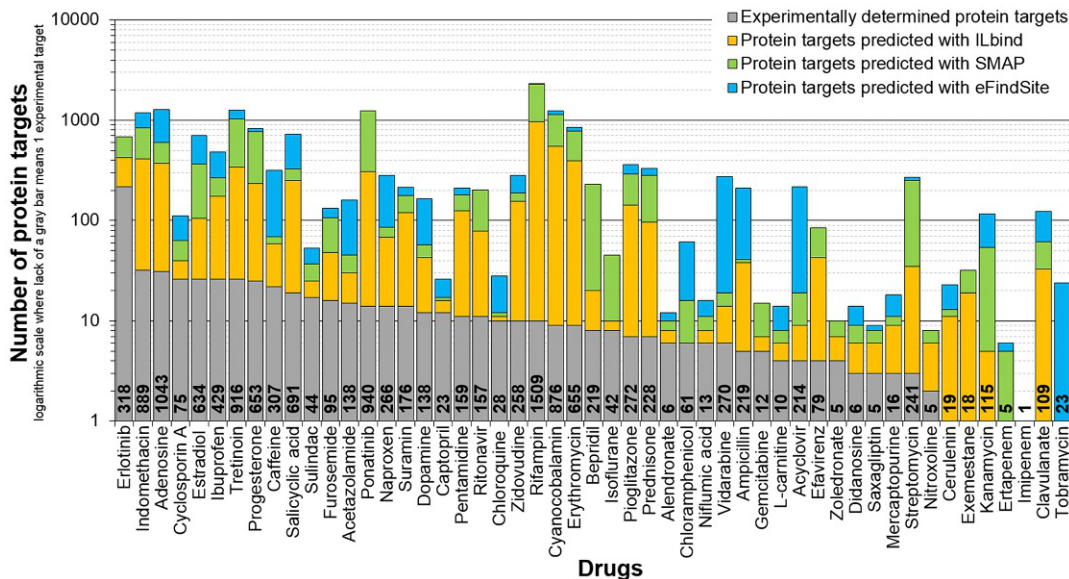


FIG. 2 Number of experimental and putative protein targets for the 51 drugs included in PDID. The numbers at the base of the plot show the total number of unique protein targets across all experimental and putative annotations. The bars represent the number of targets generated by each methodology (experiment, ILbind, SMAP, and *e*FindSite) where multiple methods can annotate the same target. Drugs are sorted by the number of experimentally determined targets. Each protein target could be represented by multiple structures that are stored in PDID.

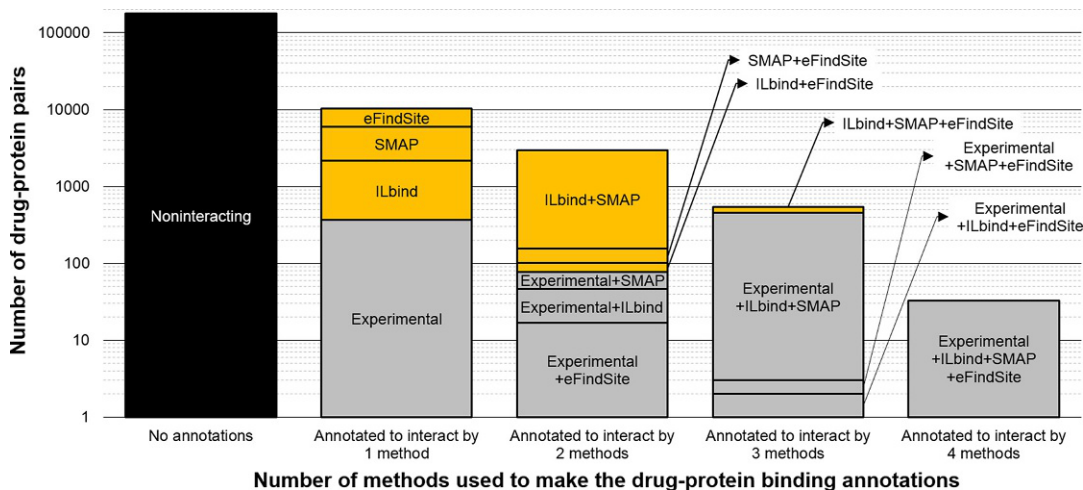


FIG. 3 Breakdown of annotations for all possible drug-protein pairs in the PDID database. Each pair is categorized by the number and types of methods used to annotate it as a potential drug-protein interaction. *Black shading* denotes pairs that are annotated as noninteracting. *Gray* is for drug-protein pairs that have experimental annotation of interaction, which could be also accompanied by a computational prediction. *Orange* (darker gray at the top of the three columns in the middle) represents pairs that feature putative computational annotation of interactions.

Fig. 3 further explores the issue of the multiple annotation of interactions for the same drug-protein pairs. We categorize each of the 191,046 possible drug-protein pairs by the number and types of methods used to annotate it as a potential drug-protein interaction. About 7% of these pairs are predicted and/or known to interact. **Fig. 3** reveals that the majority of drug-protein pairs are annotated as interacting by a single method (10,278 out of 13,791) and 96% of these 10,278 annotations are putative (gray bars vs. orange/darker gray bars at the top of the three columns in the middle in **Fig. 3**). There are 2934 pairs that are predicted by two methods with 97% of them relying solely on the computational predictions. However, the proportion of the putative interactions substantially decreases for those drug-protein pairs annotated by three methods. Only 17% of the drug-protein pairs predicted by three methods are putative. Finally, PDID includes 33 drug-protein pairs that are annotated experimentally to interact and are also predicted as such by the three predictors. This suggests that targets annotated by multiple predictors are more likely to be accurately predicted.

4 USE OF THE PDID DATABASE

The end users need only an internet connection and a modern browser to use the PDID resource. The resource is part of a larger computational platform located at <http://biomine.cs.vcu.edu/> that includes a variety of popular tools such as PPCpred (Mizianty & Kurgan, 2011), fDETECT (Meng, Wang, & Kurgan, 2018; Mizianty, Fan, et al., 2014), MFDp (Mizianty et al., 2010; Mizianty, Uversky, & Kurgan, 2014), MoRFpred (Disfani et al., 2012; Yan, Dunker, Uversky, & Kurgan, 2016), hybridNAP (Zhang, Ma, & Kurgan, 2017), DRNAPred (Yan & Kurgan, 2017), DisoRDPbind (Peng & Kurgan, 2015; Peng, Wang,

Uversky, & Kurgan, 2017), and DFLpred (Meng & Kurgan, 2016). Upon landing on the main PDID page at <http://biomine.cs.vcu.edu/servers/PDID/> users are presented with the page shown in Fig. 4. Starting at the top of the entry page, the “About” link leads to a webpage that includes information about data sources and methods that are used to derive data that are stored in PDID. The “Help and Tutorial” link provides access to a tutorial page that explains layout and details of all webpages that are available in PDID, including the main page and the query result pages. The latter pages are organized around queries for a specific drug and for a specific protein target that can be found either by the PDB identifier or by the sequence. The “Search the Database for Drug/Protein/Sequence” field offers the three corresponding options:

1. The “Search by drug name” allows selection of the desired drug with the help of a pull-down menu. After clicking “Search,” a new window with the detailed information about experimental and putative interactions of the selected drug will be opened. This page includes links to the protein structures and sequences and results from the three predictors. An example is shown in Fig. 5.
2. The “search by PDB identifier of target structure” option is for users who would like to search for a specific protein target. The identifier is in the PDB format that includes four characters followed by one character that denotes the chain identifier (e.g., 12CA_A). An example page that is generated by such a query is given in Fig. 6.
3. The “search by protein sequence” option is directed toward users who would like to identify their protein of interest using the sequence. PDID uses BLAST (McGinnis & Madden, 2004) to align the query sequence to the sequences of all protein targets included in PDID to find the closest match. Correspondingly, the user is asked to enter the query protein sequence in the FASTA format and select the E-value threshold for BLAST alignment using a pull-down menu. The search returns the drug binding details for the most similar protein in PDID, which is required to have better than the threshold similarity to the query protein. The resulting page is similar to Fig. 6, with the addition of a pairwise alignment shown at the top.

The “Materials” section at the bottom of the main PDID webpage provides access to the list of proteins and drugs that are included in PDID, and a complete database file in the MySQL database format. The option at the very bottom of Fig. 4 can be used to download a text-based file with coordinates of the predicted locations of drugs; these coordinates are relative to the structure of the corresponding target proteins. Finally, the “?” symbol indicates availability of a help page that explains particular details of the interface. These links are also available for the query result pages.



Fig. 5 gives the PDID’s webpage that summarizes results for a selected query drug. This page includes links to relevant pages in PDB, DrugBank, and BindingDB, and tabulated detailed information concerning the known and putative interactions (or lack of them) between the selected drug and each of the human proteins included in PDID. Proteins in this table are sorted by default in descending order using the likelihood that they interact with the selected drug quantified with the ILbind score (the most accurate predictor included in PDID). The table includes the following columns:

1. The “PDB ID” column gives identifiers that are linked to the corresponding record in PDB.
2. The “Protein name {synonym(s)}” column is linked to the page that describes results per protein target (Fig. 6).

PDID: Protein-Drug Interaction Database in the structural human proteome

[About](#) | [Help and Tutorial](#) | [Release Notes](#) | [Statistics](#) | [Database](#) | [References](#) | [Materials](#) | [Disclaimer](#) | [Biomine](#)

Search the Database for Drug/Protein/Sequence

HINT: The  symbols indicate availability of (additional) explanations. Clicking  opens a new window with help and hints related to the selected section/task.

Search by drug name

Search by PDB identifier of protein structure

Search by PDB identifier of protein structure

A four-character protein identifier and a one-character chain identifier are both required, and concatenated with underscore as protein_chain. Example: 12CA_A

Search by protein sequence

Enter one protein sequence in FASTA format.

Maximal E-value to filter BLAST output:

References

Upon the usage the users are requested to use the following citation:

- Wang C, Hu G, Wang K, Brylinski M, Xie L, Kurgan LA, 2016. PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. *Bioinformatics*, 32(4): 579-586.

Materials

- Structural human proteome - PDB identifiers of the human and human-like proteins with known 3D structures
- List of drugs - Names of drugs that are available in the database
- Database - The users could download the database file and make local query using MySQL
- Putative targets with coordinates of the predicted centers of ligands


FIG. 4 The main page of the PDID database.

Protein Targets for Salicylic Acid (SAL)

More information about Salicylic Acid could be found from the following sources.

- **PDB:** <http://www.rcsb.org/pdb/ligand/ligandsummary.do?hetid=SAL>
- **DrugBank:** <http://www.drugbank.ca/drugs/DB00936>
- **BindingDB:** <http://www.bindingdb.org/bind/chemsearch/marvin/MolStructure.jsp?monomerid=26193>

The table below lists all proteins from the structural human proteome that are sorted by the ILbind binding propensity.

- The table includes annotations of known binding events from PDB, DrugBank, and BindingDB, and binding propensities predicted by ILbind, SMAP, and eFindSite.
- The table sorts all proteins by their binding propensities predicted by ILbind, but they can be resorted by the predictions from SMAP or eFindSite by clicking on the name of predictor in the rightmost column of header row.
- Click protein name to get all protein-drug interactions corresponding to the selected protein. 


























PDB ID	Protein name (synonym(s))	Organism	Sequence file (FASTA)	Structure file (PDB)	Type of binding annotation	Source	Sequence similarity to known target [%]	Predicted binding propensity		
								ILbind 	SMAP 	eFindSite 
2E1Q_A	XANTHINE DEHYDROGENASE/OXIDASE (XANTHINE DEHYDROGENASE; XD; XANTHINE OXIDASE; XO; XANTHINE OXIDOREDUCTASE)	Homo sapiens			in complex	PDB	100	0.91	104.39	0.41
2E3T_A	XANTHINE DEHYDROGENASE/OXIDASE	Rattus norvegicus			in complex	PDB	99.5	0.91	103.78	0.39
1WYG_A	XANTHINE DEHYDROGENASE/OXIDASE	Rattus norvegicus			in complex	PDB	100	0.91	103.52	0.41
3AN1_A	XANTHINE DEHYDROGENASE/OXIDASE (XANTHINE DEHYDROGENASE; XD; XANTHINE OXIDASE; XO; XANTHINE OXIDOREDUCTASE)	Rattus norvegicus			in complex	PDB	99.6	0.91	97.71	0.4
2VDB_A	SERUM ALBUMIN (HUMAN SERUM ALBUMIN)	Homo sapiens			known to bind	DrugBank	100	0.85	70.72	
1TR2_A	VINCULIN ISOFORM 1	Homo sapiens			predicted to bind			0.85	62.74	
3A6P_A	EXPORTIN-5 (EXPS; RAN-BINDING PROTEIN 21)	Homo sapiens			predicted to bind			0.84	59.72	
3GB8_A	EXPORTIN-1 (EXP1; CHROMOSOME REGION MAINTENANCE 1 PROTEIN HOMOLOG)	Homo sapiens			predicted to bind			0.84	56.97	
3AR4_A	SARCOPLASMIC/ENDOPLASMIC RETICULUM CALCIUM ATPASE 1 (SERCA1; SR CA(2+)-ATPASE 1; CALCIUM PUMP 1; CALCIUM- TRANSPORTING ATPASE SARCOPLASMIC RETICULUM TYPE; FAST TWITCH SKELETAL MUSCLE ISOFORM; ENDOPLASMIC RETICULUM CLASS 1/2 CA(2+) ATPASE)	Oryctolagus cuniculus			predicted to bind			0.84	62.67	
1H7X_A	DIHYDROPYRIMIDINE DEHYDROGENASE (DIHYDROURACIL DEHYDROGENASE; DIHYDROTHYMINE DEHYDROGENASE)	Sus scrofa			predicted to bind			0.84	57.25	
3SQJ_A	SERUM ALBUMIN	Homo sapiens			known to bind	DrugBank	100	0.83	65.69	

FIG. 5 Results of a query for a specific drug, salicylic acid.

RESULTS FOR SERUM ALBUMIN [?](#)

The structure of SERUM ALBUMIN is identified as **3SQJ** (chain A) in PDB.

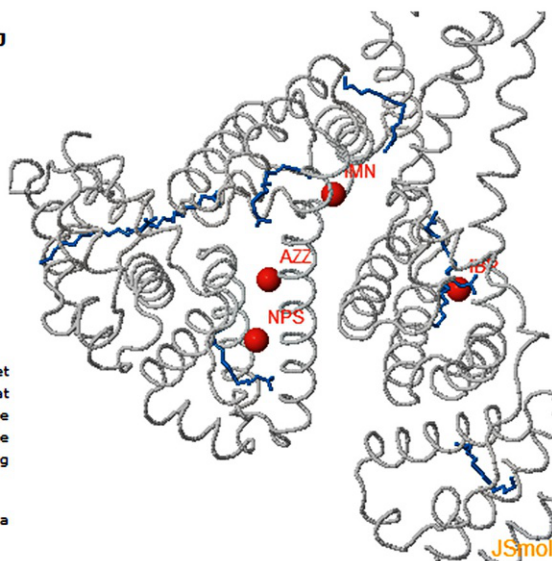
3SQJ shares 100% similarity with **P02768** in UniProt.

The protein sequence file could be downloaded [HERE](#).

The protein structure file could be downloaded [HERE](#).

The panel on the right shows structure of the protein target (using gray trace of the backbone), position of ligands that are in complex with this protein (using blue sticks; they are included in the structure file), and the predicted center of the drug molecule (using red balls) as summarized in the 'Binding to drugs' table.

The structure is editable, it can be rotated and zoomed with a mouse.

Binding to drugs [?](#)

Drug ID	Drug Name	Annotated as Known Target		Predicted as Target			Predicted Coordinates of the Center of the Drug Molecule [X;Y;Z]	Binding Summary	
		Type of interaction	Source database	ILbind binding propensity	SNAP raw score	eFindSite confidence score		Known target?	# methods predicting the protein as target
NPS	NAPROXEN	In Complex	PDB	0.91	161.17	0.36	-9.807;-0.255;82.507	Yes	2
AZZ	ZIDOVUDINE	In Complex	PDB	0.91	202.35	0.34	-1.940;0.699;80.779	Yes	2
IMN	INDOMETHACIN	In Complex	PDB	0.9	241.54	0.37	9.163;-0.880;72.132	Yes	2
IBP	IBUPROFEN	In Complex	PDB	0.9	232.68	0.37	-3.414;0.257;55.829	Yes	2

FIG. 6 Results of a query for a specific protein target, human serum albumin (PDB identifier: 3SQJ_A). Structure of the protein target is shown using a gray trace of the backbone. The positions of ligands complexed with this protein are shown as *blue (dark gray) sticks* (these ligands are included in the PDB structure file). The predicted centers of the drug molecule are visualized with *balls*, and the corresponding positions are given in the "Binding to drugs" table. Visualization of the structure uses the JSmol plugin from <http://sourceforge.net/projects/jsmol/>. The structure can be manipulated (rotated, zoomed, redrawn) using a mouse.

- The "Sequence file (FASTA)" and "Structure file (PDB)" columns provide links to the files with the protein sequence (using the FASTA format) and protein structure (using the Protein Data Bank format).
- The "Type of binding annotation" column provides information on whether a given protein is:
 - Known to bind the selected drug (denoted as "in complex" or "known to bind"). The former means that the protein-drug complex was solved structurally and is available in PDB (the adjacent column provides a link to this structure), while the latter means that

this DPI was deposited into the DrugBank or BindingDB (the adjacent column provides a link to the corresponding record in the DrugBank and/or BindingDB).

- Predicted to bind the selected drug (denoted as “*predicted to bind*”). This is based on the prediction from the ILbind method, i.e., this method has to provide a sufficiently high score (>0.75) that is shown in the last column. The ILbind scores are used to make these annotations since this predictor was found to outperform the other two methods, SMAP and *eFindSite*.
 - Predicted not to bind the selected drug (denoted as “*no interaction*”). This means that this protein is not known and was not predicted to bind the selected drug. Note that some of these protein targets may have high prediction scores (shown in the last column and color-coded in green) from the SMAP and/or *eFindSite* methods, which indicates that it is possible that this interaction occurs.
5. The “*Source*” column provides links to the information from the Protein Data Bank (PDB), DrugBank, and BindingDB for proteins that are known to bind the selected drug.
 6. The “*Sequence similarity to known target [%]*” column gives the sequence similarity measured with BLAST between the protein identified in the first column by the PDB identifier and the protein identified in the BindingDB, DrugBank, or PDB databases, which were used to annotate the experimental protein-drug interactions. A higher value of similarity denotes a more accurate match.
 7. The “*Predicted binding propensity*” column gives scores generated by ILbind, SMAP, and *eFindSite*. Among several scores that ILbind and SMAP generate, we provide one score (binding propensity for ILbind and raw score for SMAP) that was empirically shown to provide the best predictive performance (Hu et al., 2012). The *eFindSite* tool generates only one propensity score. We note that the results can be sorted by each of the three scores by using links located at the top of this column.

Fig. 6 is an example of the PDID webpage that provides information concerning the known and putative DPIs for a selected protein target. This page includes links to the corresponding protein sequence and structure files, visualization of the protein structure in complex with drugs that are known and predicted to bind the selected protein, and tabulated detailed information concerning the known and putative DPIs. The drugs in the table are sorted by the likelihood that they interact with the selected proteins, starting with the drugs that are known to interact and following with the drugs predicted to interact, from higher to lower likelihood of interaction. Only the drugs that are known to interact or are predicted by at least one method to interact are shown. The table shown at the bottom of Fig. 6 has the following key columns:

1. The two “*Annotated as Known Target*” columns identify drugs that are known to interact with the selected target protein. Two types of interactions are possible: “*In Complex*,” which corresponds to the fact that the protein-drug complex was solved structurally and it is available in the PDB database (the adjacent column provides a link to this structure), and “*Known to Bind*,” meaning that this protein-drug interaction was deposited into the DrugBank and/or BindingDB (the adjacent column provides a link to the corresponding record in the DrugBank or BindingDB).
2. The three “*Predicted as Target*” columns give the selected scores generated by ILbind, SMAP, and *eFindSite*.

3. The “*Predicted Coordinates*” column provides coordinates (in the coordinates system based on the structure file that is linked at the top of the page) of the predicted positions of the centers of the drug molecules. These coordinates are shown using balls in the structure shown in the top of the page.
4. The two “*Binding Summary*” columns show whether a given drug is known to interact with the selected protein target, and how many prediction methods (out of three) predict a given interaction.

Overall, the navigation of the PDID database is fairly straightforward and intuitive. Users can query this resource in three ways and can easily navigate between the corresponding three types of screens with results. The results are color-coded to ease the interpretation and they incorporate both the experimental and putative annotations. Users can also effortlessly link to the structures and sequences of protein targets and the source records of the drugs.

5 SUMMARY

The key challenge in pharmacology has shifted from the study of single molecules to the exhaustive exploration of biologically relevant molecular interactions at the level of complete proteomes. Adopting a systems-level approach can help comprehend the underlying principles of the cellular networks of interacting molecules, with practical applications in the discovery of new biopharmaceuticals and repurposing of current drugs. Nonetheless, a relatively low coverage of DPIs by the experimental data necessitates augmenting the existing databases with putative interactions annotated by across-proteome computational modeling. To meet this demand, we recently developed the PDID, a new resource comprising over one million PDIs confidently inferred based on the structural similarity between query proteins and a large database of drug-protein complexes. PDID combines interaction data generated by three state-of-the-art predictors: *eFindSite*, *SMAP*, and *ILbind*. A unique feature of this resource is that it contains the molecular level details for DPIs, such as the location of binding sites in target structures. Encouragingly, the analysis of putative interactions included in the PDID indicates that the majority of them are likely to be accurately predicted. The database was designed to be user-friendly and easily accessible via any modern web browser. Online materials include a full documentation and tutorials explaining how to query the PDID with either a drug or a target protein and interpret the results. The PDID has a wide range of applications, including the identification of novel targets for pharmacotherapy, the development of safer drugs with reduced side effects, the repurposing of existing therapeutics to treat new diseases, and the design of polypharmacological agents simultaneously targeting multiple proteins in complex disorders.

Acknowledgment

This research was funded in part by the Robert J. Mattauch Endowment funds to L.K.

References

- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402. <https://dx.doi.org/10.1093/nar/25.17.3389>.
- Basse, M. -J., Betzi, S., Morelli, X., & Roche, P. (2016). 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. *Database*, 2016, baw007. <https://dx.doi.org/10.1093/database/baw007>.
- Bendels, S., Fasching, B., Fischer, H., Gerebtzoff, G., Guba, W., Hert, J., Kansy, M., Migeon, J., Peters, J., et al. (2013). Safety screening in early drug discovery: an improved assay profile. In: *Paper presented at the Gordon Research Conference on Computer Aided Drug Design, Mount Snow, VT, USA*.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235–242.
- Bourgeois, R., Basse, M. -J., Morelli, X., & Roche, P. (2010). Atomic analysis of protein-protein interfaces with known inhibitors: the 2P2I Database. *PLoS One*, 5(3), e9598. <https://dx.doi.org/10.1371/journal.pone.0009598>.
- Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., & Whitebread, S. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature Reviews. Drug Discovery*, 11(12), 909–922.
- Brylinski, M., & Feinstein, W. (2012). Setting up a meta-threading pipeline for high-throughput structural bioinformatics: eThread software distribution, walkthrough and resource profiling. *Journal of Computer Science and Systems Biology*, 6(1), 001–010. <https://dx.doi.org/10.4172/jcsb.1000094>.
- Brylinski, M., & Feinstein, W. P. (2013). eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *Journal of Computer-Aided Molecular Design*, 27(6), 551–567. <https://dx.doi.org/10.1007/s10822-013-9663-5>.
- Brylinski, M., & Lingam, D. (2012). eThread: a highly optimized machine learning-based approach to meta-threading and the modeling of protein tertiary structures. *PLoS One*, 7(11)e50200. <https://dx.doi.org/10.1371/journal.pone.0050200>.
- Brylinski, M., & Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences*, 105(1), 129–134. <https://dx.doi.org/10.1073/pnas.0707684105>.
- Chartier, M., Adriansen, E., & Najmanovich, R. (2016). IsoMIF Finder: online detection of binding site molecular interaction field similarities. *Bioinformatics*, 32(4), 621–623. <https://dx.doi.org/10.1093/bioinformatics/btv616>.
- Chartier, M., & Najmanovich, R. (2015). Detection of binding site molecular interaction field similarities. *Journal of Chemical Information and Modeling*, 55(8), 1600–1615. <https://dx.doi.org/10.1021/acs.jcim.5b00333>.
- Chen, X., Ji, Z. L., & Chen, Y. Z. (2002). TTD: therapeutic target database. *Nucleic Acids Research*, 30(1), 412–415.
- Chen, X., Liu, M., & Gilson, M. K. (2001). BindingDB: a web-accessible molecular recognition database. *Combinatorial Chemistry & High Throughput Screening*, 4(8), 719–725.
- Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., & Zhang, Y. (2016). Drug-target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics*, 17(4), 696–712. <https://dx.doi.org/10.1093/bib/bbv066>.
- Chong, C. R., & Sullivan, D. J. (2007). New uses for old drugs. *Nature*, 448(7154), 645–646. http://www.nature.com/nature/journal/v448/n7154/supinfo/448645a_S1.html.
- Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., ... Sali, A. (2016). CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *Journal of Molecular Biology*, 428(4), 709–719. <https://dx.doi.org/10.1016/j.jmb.2016.01.029>.
- Ding, H., Takigawa, I., Mamitsuka, H., & Zhu, S. (2014). Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in Bioinformatics*, 15(5), 734–747. <https://dx.doi.org/10.1093/bib/bbt056>.
- Disfani, F. M., Hsu, W. L., Mizianty, M. J., Oldfield, C. J., Xue, B., Dunker, A. K., ... Kurgan, L. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, 28(12), i75–i83. <https://dx.doi.org/10.1093/bioinformatics/bts209>.
- Ezzat, A., Wu, M., Li, X. -L., & Kwok, C. -K. (2018). Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics*, bby002. <https://dx.doi.org/10.1093/bib/bby002>.

- Feinstein, W. P., & Brylinski, M. (2014). eFindSite: enhanced fingerprint-based virtual screening against predicted ligand binding sites in protein models. *Molecular Informatics*, 33(2), 135–150.
- Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814), 972–976. <https://dx.doi.org/10.1126/science.1136800>.
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., ... Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(Database issue), D1100–D1107. <https://dx.doi.org/10.1093/nar/gkr777>.
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., ... Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1), D945–D954. <https://dx.doi.org/10.1093/nar/gkw1074>.
- Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1), D1045–D1053.
- Glaab, E. (2016). Building a virtual ligand screening pipeline using free software: a survey. *Briefings in Bioinformatics*, 17(2), 352–366. <https://dx.doi.org/10.1093/bib/bbv037>.
- Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., ... Preissner, R. (2008). SuperTarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Research*, 36(suppl_1), D919–D922. <https://dx.doi.org/10.1093/nar/gkm862>.
- Hao, M., Bryant, S. H., & Wang, Y. (2018). Open-source chemogenomic data-driven algorithms for predicting drug-target interactions. *Briefings in Bioinformatics*, bby010. <https://dx.doi.org/10.1093/bib/bby010>.
- Haupt, V. J., & Schroeder, M. (2011). Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings in Bioinformatics*, 12(4), 312–326. <https://dx.doi.org/10.1093/bib/bbr011>.
- Hecker, N., Ahmed, J., von Eichborn, J., Dunkel, M., Macha, K., Eckert, A., ... Preissner, R. (2012). SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Research*, 40(Database issue), D1113–D1117. <https://dx.doi.org/10.1093/nar/gkr912>.
- Higuero, A. P., Jubb, H., & Blundell, T. L. (2013). TIMBAL v2: update of a database holding small molecules modulating protein-protein interactions. *Database*, 2013, bat039. <https://dx.doi.org/10.1093/database/bat039>.
- Higuero, A. P., Schreyer, A., Bickerton, G. R. J., Pitt, W. R., Groom, C. R., & Blundell, T. L. (2009). Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL Database. *Chemical Biology & Drug Design*, 74(5), 457–467. <https://dx.doi.org/10.1111/j.1747-0285.2009.00889.x>.
- Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nature Reviews. Drug Discovery*, 1(9), 727–730. http://www.nature.com/nrd/journal/v1/n9/supinfo/nrd892_S1.html.
- Hu, G., Gao, J., Wang, K., Mizianty, M. J., Ruan, J., & Kurgan, L. (2012). Finding protein targets for small biologically relevant ligands across fold space using inverse ligand binding predictions. *Structure*, 20(11), 1815–1822. <https://dx.doi.org/10.1016/j.str.2012.09.011>.
- Hu, G., Wang, K., Groenendyk, J., Barakat, K., Mizianty, M. J., Ruan, J., ... Kurgan, L. (2014). Human structural proteome-wide characterization of Cyclosporine A targets. *Bioinformatics*, 30(24), 3561–3566. <https://dx.doi.org/10.1093/bioinformatics/btu581>.
- Hu, G., Wu, Z., Wang, K., Uversky, V. N., & Kurgan, L. (2016). Untapped potential of disordered proteins in current druggable human proteome. *Current Drug Targets*, 17(10), 1198–1205.
- Hu, Y., & Bajorath, J. (2013). Compound promiscuity: what can we learn from current data? *Drug Discovery Today*, 18(13–14), 644–650. <https://dx.doi.org/10.1016/j.drudis.2013.03.002>.
- Huang, Y. J., Hang, D., Lu, L. J., Tong, L., Gerstein, M. B., & Montelione, G. T. (2008). Targeting the human cancer pathway protein interaction network by structural genomics. *Molecular & Cellular Proteomics*, 7(10), 2048–2060. <https://dx.doi.org/10.1074/mcp.M700550-MCP200>.
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1), D353–D361.
- Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2009). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Suppl. 1), D355–D360.
- Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2008). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research*, 36(Suppl. 1), D684–D688. <https://dx.doi.org/10.1093/nar/gkm795>.
- Labbé, C. M., Kuenemann, M. A., Zarzycka, B., Vriend, G., Nicolaes, G. A. F., Lagorce, D., ... Sperandio, O. (2016). iPPI-DB: an online database of modulators of protein-protein interactions. *Nucleic Acids Research*, 44(D1), D542–D547. <https://dx.doi.org/10.1093/nar/gkv982>.

- Labbé, C. M., Laconde, G., Kuenemann, M. A., Villoutreix, B. O., & Sperandio, O. (2013). iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein-protein interactions. *Drug Discovery Today*, 18(19), 958–968. <https://doi.org/10.1016/j.drudis.2013.05.003>.
- Lavecchia, A., & Cerchia, C. (2016). In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discovery Today*, 21(2), 288–298. <https://dx.doi.org/10.1016/j.drudis.2015.12.007>.
- Lavecchia, A., & Giovanni, C. D. (2013). Virtual screening strategies in drug discovery: a critical review. *Current Medicinal Chemistry*, 20(23), 2839–2860. <https://doi.org/10.2174/09298673113209990001>.
- Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., & Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics*, 17(1), 2–12. <https://dx.doi.org/10.1093/bib/bbv020>.
- Li, L., Bum-Erdene, K., Baenziger, P. H., Rosen, J. J., Hemmert, J. R., Nellis, J. A., ... Meroueh, S. O. (2010). BioDrugScreen: a computational drug design resource for ranking molecules docked to the human proteome. *Nucleic Acids Research*, 38(Database issue), D765–D773. <https://dx.doi.org/10.1093/nar/gkp852>.
- Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., ... Zhu, F. (2018). Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Research*, 46(D1), D1121–D1127. <https://dx.doi.org/10.1093/nar/gkx1076>.
- Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., ... Urban, L. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403), 361–367. <http://www.nature.com/nature/journal/v486/n7403/abs/nature11159.html#supplementary-information>.
- Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., & Šali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(1), 291–325.
- McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32, W20–W25. <https://dx.doi.org/10.1093/nar/gkh435>.
- Meng, F., & Kurgan, L. (2016). DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*, 32(12), i341–i350. <https://dx.doi.org/10.1093/bioinformatics/btw280>.
- Meng, F., Wang, C., & Kurgan, L. (2018). fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. *BMC Bioinformatics*, 18(1), 580. <https://dx.doi.org/10.1186/s12859-017-1995-z>.
- Mestres, J., Gregori-Puigjané, E., Valverde, S., & Sole, R. V. (2008). Data completeness—the Achilles heel of drug-target networks. *Nature Biotechnology*, 26(9), 983–984.
- Mizianty, M. J., Fan, X., Yan, J., Chalmers, E., Woloschuk, C., Joachimiak, A., & Kurgan, L. (2014). Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallographica. Section D, Biological Crystallography*, 70(Pt 11), 2781–2793. <https://dx.doi.org/10.1107/S1399004714019427>.
- Mizianty, M. J., & Kurgan, L. (2011). Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics*, 27(13), i24–i33. <https://dx.doi.org/10.1093/bioinformatics/btr229>.
- Mizianty, M. J., Stach, W., Chen, K., Kedariseti, K. D., Disfani, F. M., & Kurgan, L. (2010). Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, 26(18), i489–i496. <https://dx.doi.org/10.1093/bioinformatics/btq373>.
- Mizianty, M. J., Uversky, V., & Kurgan, L. (2014). Prediction of intrinsic disorder in proteins using MFDp2. *Methods in Molecular Biology*, 1137, 147–162. https://dx.doi.org/10.1007/978-1-4939-0366-5_11.
- Okuno, Y., Tamon, A., Yabuuchi, H., Nijijima, S., Minowa, Y., Tonomura, K., ... Feng, C. (2008). GLIDA: GPCR—ligand database for chemical genomics drug discovery—database and tools update. *Nucleic Acids Research*, 36(suppl_1), D907–D912. <https://dx.doi.org/10.1093/nar/gkm948>.
- Okuno, Y., Yang, J., Taneishi, K., Yabuuchi, H., & Tsujimoto, G. (2006). GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Research*, 34(Suppl. 1), D673–D677. <https://dx.doi.org/10.1093/nar/gkj028>.
- Pandit, S. B., & Skolnick, J. (2008). Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics*, 9(1), 531. <https://dx.doi.org/10.1186/1471-2105-9-531>.
- Peng, Z., & Kurgan, L. (2015). High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Research*, 43(18)e121. <https://dx.doi.org/10.1093/nar/gkv585>.
- Peng, Z., Wang, C., Uversky, V. N., & Kurgan, L. (2017). Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods in Molecular Biology*, 1484, 187–203. https://dx.doi.org/10.1007/978-1-4939-6406-2_14.
- Peters, J. U. (2013). Polypharmacology—foe or friend? *Journal of Medicinal Chemistry*, 56(22), 8955–8971. <https://dx.doi.org/10.1021/jm400856t>.

- Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., & Schomburg, D. (2017). BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Research*, 45(D1), D380–D388.
- Rask-Andersen, M., Almén, M. S., & Schiöth, H. B. (2011). Trends in the exploitation of novel drug targets. *Nature Reviews. Drug Discovery*, 10(8), 579–590. Retrieved from :http://www.nature.com/nrd/journal/v10/n8/suppinfo/nrd3478_S1.html.
- Rask-Andersen, M., Masuram, S., & Schiöth, H. B. (2014). The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual Review of Pharmacology and Toxicology*, 54, 9–26. <https://dx.doi.org/10.1146/annurev-pharmtox-011613-135943>.
- Rose, P. W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., ... Burley, S. K. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, 45(D1), D271–D281. <https://dx.doi.org/10.1093/nar/gkw1000>.
- Roth, B. L., Lopez, E., Patel, S., & Kroeze, W. K. (2000). The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *The Neuroscientist*, 6(4), 252–262.
- Russ, A. P., & Lampel, S. (2005). The druggable genome: an update. *Drug Discovery Today*, 10(23–24), 1607–1610. [https://doi.org/10.1016/S1359-6446\(05\)03666-4](https://doi.org/10.1016/S1359-6446(05)03666-4).
- Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3), 779–815. <https://doi.org/10.1006/jmbi.1993.1626>.
- Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., ... Overington, J. P. (2017). A comprehensive map of molecular drug targets. *Nature Reviews. Drug Discovery*, 16(1), 19–34. <https://dx.doi.org/10.1038/nrd.2016.230>. Retrieved from: <http://www.nature.com/nrd/journal/v16/n1/abs/nrd.2016.230.html#supplementary-information>.
- Schneider, G. (2010). Virtual screening: an endless staircase? *Nature Reviews. Drug Discovery*, 9(4), 273–276. <https://dx.doi.org/10.1038/nrd3139>.
- Schomburg, I., Hofmann, O., Baensch, C., Chang, A., & Schomburg, D. (2000). Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene Function & Disease*, 1(3–4), 109–118.
- Skolnick, J., & Brylinski, M. (2009). FINDSITE: a combined evolution/structure-based approach to protein function prediction. *Briefings in Bioinformatics*, 10(4), 378–391. <https://dx.doi.org/10.1093/bib/bbp017>.
- Skolnick, J., Kihara, D., & Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins: Structure, Function, and Bioinformatics*, 56(3), 502–518. <https://dx.doi.org/10.1002/prot.20106>.
- Sugaya, N., & Furuya, T. (2011). Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions. *BMC Bioinformatics*, 12, 50. <https://dx.doi.org/10.1186/1471-2105-12-50>.
- Sugaya, N., Kanai, S., & Furuya, T. (2012). Dr. PIAS 2.0: an update of a database of predicted druggable protein-protein interactions. *Database: The Journal of Biological Databases and Curation*, 2012, bas034. <https://dx.doi.org/10.1093/database/bas034>.
- Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(D1), D380–D384. <https://dx.doi.org/10.1093/nar/gkv1277>.
- Tseng, C. Y., & Tuszynski, J. (2015). A unified approach to computational drug discovery. *Drug Discovery Today*, 10(11), 1328–1336. <https://dx.doi.org/10.1016/j.drudis.2015.07.004>.
- Urban, L. (2012). Translational value of early target-based safety assessment and associated risk mitigation. In *Paper presented at the 4th Annual Predictive Toxicology Summit, London, UK*.
- von Eichborn, J., Murgueitio, M. S., Dunkel, M., Koerner, S., Bourne, P. E., & Preissner, R. (2011). PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Research*, 39(Database issue), D1060–D1066. <https://dx.doi.org/10.1093/nar/gkq1037>.
- Wang, C., Hu, G., Wang, K., Brylinski, M., Xie, L., & Kurgan, L. (2016). PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. *Bioinformatics*, 32(4), 579–586. <https://dx.doi.org/10.1093/bioinformatics/btv597>.
- Wang, J., Li, Z. -X., Qiu, C. -X., Wang, D., & Cui, Q. -H. (2012). The relationship between rational drug design and drug side effects. *Briefings in Bioinformatics*, 13(3), 377–382. <https://dx.doi.org/10.1093/bib/bbr061>.
- Wang, X. Y., & Greene, N. (2012). Comparing measures of promiscuity and exploring their relationship to toxicity. *Molecular Informatics*, 31(2), 145–159. <https://dx.doi.org/10.1002/minf.201100148>.

- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1), D1074–D1082. <https://dx.doi.org/10.1093/nar/gkx1037>.
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., ... Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(Suppl. 1), D668–D672. <https://dx.doi.org/10.1093/nar/gkj067>.
- Xie, L., & Bourne, P. E. (2007). A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics*, 8(Suppl. 4), S9. <https://dx.doi.org/10.1186/1471-2105-8-S4-S9>.
- Xie, L., & Bourne, P. E. (2008). Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proceedings of the National Academy of Sciences of the United States of America*, 105(14), 5441–5446. <https://dx.doi.org/10.1073/pnas.0704422105>.
- Xie, L., Xie, L., & Bourne, P. E. (2009). A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, 25(12), i305–i312. <https://dx.doi.org/10.1093/bioinformatics/btp220>.
- Yan, J., Dunker, A. K., Uversky, V. N., & Kurgan, L. (2016). Molecular recognition features (MoRFs) in three domains of life. *Molecular BioSystems*, 12(3), 697–710. <https://dx.doi.org/10.1039/c5mb00640f>.
- Yan, J., & Kurgan, L. (2017). DRNAPred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Research*, 45(10)e84. <https://dx.doi.org/10.1093/nar/gkx059>.
- Yang, H., Qin, C., Li, Y. H., Tao, L., Zhou, J., Yu, C. Y., ... Chen, Y. Z. (2016). Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research*, 44(D1), D1069–D1074. <https://dx.doi.org/10.1093/nar/gkv1230>.
- Zhang, J., Ma, Z., & Kurgan, L. (2017). Comprehensive review and empirical analysis of hallmarks of DNA-RNA- and protein-binding residues in protein chains. *Briefings in Bioinformatics*. <https://dx.doi.org/10.1093/bib/bbx168>.
- Zhang, Y., Arakaki, A. K., & Skolnick, J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Structure, Function, and Bioinformatics*, 61(S7), 91–98. <https://dx.doi.org/10.1002/prot.20724>.
- Zhang, Y., & Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20), 7594–7599. <https://dx.doi.org/10.1073/pnas.0305695101>.
- Zhang, Y., & Skolnick, J. (2004b). Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical Journal*, 87(4), 2647–2655. <https://dx.doi.org/10.1529/biophysj.104.045385>.