

# Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome

Chen Wang and Lukasz Kurgan

Corresponding author: Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, Richmond, 23284, USA.  
Tel.: +1-804-827-3986; E-mail: lkurgan@vcu.edu

## Abstract

Drug–protein interactions (DPIs) underlie the desired therapeutic actions and the adverse side effects of a significant majority of drugs. Computational prediction of DPIs facilitates research in drug discovery, characterization and repurposing. Similarity-based methods that do not require knowledge of protein structures are particularly suitable for druggable genome-wide predictions of DPIs. We review 35 high-impact similarity-based predictors that were published in the past decade. We group them based on three types of similarities and their combinations that they use. We discuss and compare key aspects of these methods including source databases, internal databases and their predictive models. Using our novel benchmark database, we perform comparative empirical analysis of predictive performance of seven types of representative predictors that utilize each type of similarity individually and all possible combinations of similarities. We assess predictive quality at the database-wide DPI level and we are the first to also include evaluation over individual drugs. Our comprehensive analysis shows that predictors that use more similarity types outperform methods that employ fewer similarities, and that the model combining all three types of similarities secures area under the receiver operating characteristic curve of 0.93. We offer a comprehensive analysis of sensitivity of predictive performance to intrinsic and extrinsic characteristics of the considered predictors. We find that predictive performance is sensitive to low levels of similarities between sequences of the drug targets and several extrinsic properties of the input drug structures, drug profiles and drug targets. The benchmark database and a webserver for the seven predictors are freely available at <http://biomine.cs.vcu.edu/servers/CONNECTOR/>.

**Key words:** drug–protein interactions; computational prediction; drug structure similarity; drug profile similarity; target sequence similarity; drug repurposing

---

Chen Wang is a postdoctoral fellow at the Columbia University. He received a PhD from the Computer Science Department at the Virginia Commonwealth University in 2018. His research interests include applications of machine learning and statistics in bioinformatics, in particular in characterization and prediction of protein structures, functions, and protein–ligand interactions.

Lukasz Kurgan is a Qimonda Endowed Professor of Computer Science at the Virginia Commonwealth University in Richmond, Fellow of AIMBE and Fellow of Kosciuszko Foundation Collegium of Eminent Scientists. His research focuses on high-throughput structural and functional characterization of proteins and small RNAs. More details about his research group are at <http://biomine.cs.vcu.edu/>.

Submitted: 7 May 2018; Received (in revised form): 26 June 2018

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

## Introduction

Therapeutic activity of most of the drugs is determined by their interactions with target molecules, of which about 95% are proteins [1–3]. Knowledge of the associations between drug compounds and their protein targets is essential for a wide range of pharmaceutical and bioinformatics studies. These studies include screening drug candidates that targets specific disease-associated genes/proteins [4–7], drug repurposing (searching for targets associated with diseases that are not yet known to benefit from the existing drugs) [8–12], discovery and characterization of side effects that stem from the interactions with non-therapeutic off-targets [13–17] and elucidation of the druggable human proteome, which is defined as the complement of human proteins that interact with drugs [2, 18–21]. Consequently, many databases of drug–protein interactions (DPIs) have been released [22–69]. Besides offering the annotations of interactions, these databases provide relevant information about drugs and proteins, such as drug structures [34–38, 66–69] and sequences of protein targets [39–44, 52–56]. However, most of these databases are limited to a small portion of the human proteome that includes proteins with known structures [70]. The databases are complemented by computational methods that generate putative drug–protein pairs on a large, up to the whole proteome scale. These predictions can be used to facilitate discovery of new DPIs by constraining the scope of expensive and time consuming *in vitro* and *in vivo* experiments that are employed to discover and validate the putative DPIs [71–73]. They are also utilized to develop databases of pre-computed putative DPIs [74–76] and were used to facilitate discovery of the therapeutic mechanisms and side effects of drugs [17, 77, 78].

Computational prediction of DPIs typically incorporates information about drugs and structures or sequences of target proteins. The computational methods could be categorized into protein structure-based and similarity-based methods [79–81]. The protein structure-based methods utilize the three-dimensional (3D) structure of the target protein(s) and typically perform molecular docking of an often known or otherwise modeled structure of the compound into the protein structure. However, recent works estimate that only about 20% to 30% of the human proteins have 3D structures [82–84], which means that the protein structure-based methods are limited to a relatively small fraction of the human proteome. Moreover, they cannot be used for proteins with a substantial amount of intrinsically disordered regions [21] while recent works show that as many as 44% of human proteins contain at least one long (30 or more consecutive residues) disordered region [85, 86]. Finally, the computational cost of the structure-based predictions is relatively high due to the long runtime of docking [79]. At the same time, a full complement of sequences is available for the human proteome and can be used by the similarity-based methods. The similarity-based methods are developed based on assertions that similar drugs likely target the same proteins and that similar proteins tend to interact with the same drugs [79, 87–89]. This boils down three types of similarities between drug structures, drug profiles and protein target sequences. Various similarity-based methods rely on different ways to represent and quantify one or more of these three similarities. The predictive quality of these approaches depends on how the similarities are quantified, how many are used and how they are combined together. We focus on the similarity-based methods given that they can be applied over the entire human proteome and that they are more computationally tractable when compared to the structure-based methods.

A total of 11 surveys that were written in the past 5 years have focused on the similarity-based methods [79, 90–99]. We also found two other articles that focus on related topics including biological profiles [100] and machine learning tools [101] that are employed in the DPI predictions. The 11 surveys summarize relevant databases, review representative methods and discuss common criteria that are used to evaluate predictive performance. *Supplementary Table S1* assesses scope of these articles with the number of methods they reviewed, coverage of recent predictors, inclusion of summary of databases used by the considered predictors and presence of empirical assessment and analysis of sensitivity of predictive quality. The two most comprehensive surveys cover 28 [98] and 27 predictors [91] while the remaining nine articles discuss fewer than 20 methods. In most cases, over half of the methods that were reviewed was published in the past 5 years. This reflects growing interest in the similarity-based predictors. However, at least eight methods that were published in 2016 and 2017 [102–109] were not included in any of these articles. While most of the surveys provide a brief discussion of the available DPI databases, they are not necessarily the databases that are used by specific predictors to compute the similarity values. Actually only one article provides a more complete discussion that introduces specific databases that are utilized by the 10 predictors that this survey covers [79]. A complete summary of the DPI databases that covers modern methods is still missing, which is why we include this topic. We survey 35 predictors and investigate two types of relevant databases: internal databases that are intrinsic to the predictors and source databases that are used to derive the internal databases. Moreover, only four surveys provide empirical assessment of predictive performance of similarity-based DPI predictors [79, 90, 98, 99]. While they provide quantitative and comparative analysis, they focus on a limited set of approaches. In particular, they only consider certain types and combinations of similarities. They also do not perform sensitivity analysis, i.e. they do not analyze whether and how predictive performance changes with intrinsic and extrinsic characteristics of the considered predictors. To summarize, while the previous surveys offer useful insights, they also have several drawbacks. In particular, they miss several recent similarity-based DPI predictors, provide shallow treatment of databases used by these predictors, offer incomplete empirical analysis and do not include sensitivity analysis.

To this end, we (1) focus on a comprehensive set of 35 similarity-based DPI predictors including the largest set of recent predictors, (2) offer in-depth summary of the internal and source databases that these predictors use, (3) provide complete empirical assessment of predictive quality for representative predictors that use different similarities and (4) include sensitivity analysis. We cover the latest advances. The considered predictors include 22 that were published in the past 5 years and 8 that were published since 2016. We define the three main types of similarities that these 35 methods use to predict compound–protein interactions, describe key architectural details of these tools and comment on their impact. Our empirical evaluation encompasses representative approaches that rely on each of three main types of drug and protein similarities and ensemble predictors that combine two and three types of similarities. Moreover, we compare these ensembles to the methods based on a single similarity. Besides providing the overall predictive performance of the considered predictors, we are the first to consider predictive quality across individual drugs. Finally, our arguably the most comprehensive to date sensitivity analysis investigates whether and how the predictive performance

depends on several intrinsic and extrinsic factors, e.g. the number of targets that are *a priori* known for a given drug.

## Qualitative review of the similarity-based methods for the prediction of DPIs

### Selection of similarity-based methods

We searched the PUBMED repository [110] in April 2018 for relevant articles using the following query: (predict\* [Title/Abstract] AND (drug target interaction [Title/Abstract] OR drug protein interaction [Title/Abstract])). The query generated 170 possibly germane articles. We manually processed these results to select articles that describe similarity-based predictors that were published in reputable journals, i.e. journals with impact factor > 3.5. We collected the impact factors from the latest Journal Citation Reports that was released by Clarivate Analytics (formerly Thomson Reuters) on June 14, 2017 [111]. The resulting 35 similarity-based DPI predictors are included in this review [102–109, 112–138]. This list is longer than the lists covered in the prior surveys that also include methods published in conference proceedings, lower impact factor journals as well as methods that target ligand/chemical-protein interactions besides DPIs.

### Overview and classification of the similarity-based methods

The similarity-based predictors of DPIs are implemented based on the assertions that similar drugs may share the same protein targets and that similar proteins may interact with the same drugs. Therefore, the core aspect that defines these predictions is how to measure the drug–drug and protein–protein similarities. These similarities are typically quantified using the chemical structures of drugs, side-effect profiles of drugs and sequences of their protein targets. In other words, the predictors rely on three main types of similarities: drug structure similarity (DSS), drug profile similarity (DPS) and protein sequence similarity (PSS).

Prediction of DPIs is performed in three steps. First, a user provides inputs in the form of drug structure, drug profile and/or protein sequence(s) of known target(s) for the input drug; whatever is necessary for the selected similarity-based approach. The predictors rely on two components: a predictive model and a database of known DPIs. In the 2nd step, similarities between the input drug structure and/or profile and the structures and profiles of the drugs in the database are computed. If target(s) of the query drug are known or can be collected from the database (the query drug is already included in the database), then similarity between the target of the query drug and all proteins targets in the database is also computed. In the 3rd step, the predictive model combines the similarities to produce a propensity that quantifies likelihood that the query drug interacts with protein targets included in the database.

The three types of similarities can be combined to formulate an ensemble predictor with the underlining goal to improve predictive quality when compared to using a single similarity. Many of the predictors proposed in recent years rely on the ensembles that include two or three similarity types. Consequently, we categorize the 35 considered predictors into three groups: the methods that apply one type of similarity (1S), ensemble models that combine two types of similarities (2S) and ensembles of all three similarities (3S). Table 1 summarizes how these predictors quantify and combine different similarities.

Table 1 includes six 1S predictors. Four methods [104, 108, 127, 135] measure DSS based on Tanimoto coefficient [139] that

quantifies similarity between drug structures that are represented using molecular fingerprints [140, 141]. One predictor applies a statistical test over a set of DSSs based on Tanimoto coefficients between an input drug and a group of drugs in its internal database that are known to bind the input protein target [112]. The other method, DrugMiner, converts protein sequences into numeric feature vectors that represent physicochemical properties and amino acid compositions. Subsequently, it quantifies PSS for these feature vectors using machine learning algorithms and estimates druggability of a given protein based on its PSS to the known drug targets [102].

There are 23 2S models that utilize 2 types of similarities. Besides listing the similarity types we also define how they are combined together to generate predictions. Almost all of the 2S methods (21 out of 23) integrate DSS and PSS. Eleven predictors [113, 116, 119–121, 124, 126, 132–134, 136] quantify DSS using the SIMCOP algorithm [142, 143] while two other predictors [131, 137] utilize Tanimoto coefficient; these 13 methods rely on sequence alignment algorithms such as BLAST [144] and Smith–Waterman [145] to measure PSS. A set of eight other methods represents drug structures and protein sequences using feature vectors that quantify DSS and PSS and they apply machine learning algorithms to generate predictions. These algorithms include kernel functions [105, 106, 115, 118, 123, 138], neural networks [103] and decision trees [109]. DPS is adopted rarely. Only two 2S predictors combine DPS with DSS [114, 125]. They define drug side-effect profile as a vector of binary values that indicate the presence or absence of specific side-effect terms. They quantify DPS based on weighted Tanimoto coefficients that measure similarities between side-effect profiles. The predictive models applied in these methods reveal that DPS could be used to infer DPIs in a similar way to DSS. This approach assumes that drugs that have similar side effects are likely to target the same proteins. The drug profiles that were used in Ref. [114] have motivated the development of the SIDER database that stores drug side-effect information [146, 147]. This resource was later used in other works that we review including another 2S model [125] and three 3S methods [107, 122, 130]. However, SIDER focuses on marketed drugs and so the information about drug side effects it provides is narrower when compared to drug structures that encompass a wider spectrum of drugs and protein sequences that span the entire human proteome. This is likely the reason why DPS is less frequently employed. Moreover, none of the 23 2S methods combines DPS with PSS. There are two ways in which the predictors combine two similarities. Five predictors apply a simple summation and weighted summation of two similarities [114, 124–126, 137]. The other predictors employ a more complex approach that typically involves operations such as maximum, multiplication and geometric mean of multiple similarities.

There are six 3S methods that combine all three similarities. They utilize the same approaches to quantify DSS and PSS as the 1S and 2S models. Three of the six methods compute DPS based on the cosine correlation coefficients between side-effect profiles [117, 122, 128] while two other models use Tanimoto coefficients [107, 130]. The 6th predictor [148] defines drug profiles using the ATC codes, which represent hierarchical classification of drugs and measure semantic similarity between ATC codes [129]. Like for the 2S methods, the 3S models utilize either a simple summation [128, 130] or a more complex approach to combine similarities [107, 117, 122, 129].

The same types of similarities are also utilized in other related predictive efforts. For instance, they are used to predict drug–disease associations [149, 150].

**Table 1.** Comparison of 35 similarity-based DPI predictors that were published in high impact venues. We group these predictors into three categories (separated by horizontal line): predictors using one similarity (1S), predictors combining two types of similarities (2S) and predictors integrating three type of similarities (3S). Methods are sorted chronologically within each group. We summarize approaches that are used to quantify similarities and to compute ensembles of similarities for each predictor. We also summarize their empirical evaluation with information about inclusion of assessment of predictive quality, assessment of benefits of use of ensembles, assessment at per DPI and per drug levels and analysis of sensitivity of predictive performance to intrinsic characteristics (similarity of inputs to the database) and extrinsic characteristics (information about the inputs) of predictors. The check symbol ✓ means a given assessment is included for a given method (row) while a blank cell means it is not included

Type of predictor	Predictor	Year	Similarities and their ensembles <sup>1</sup>				Predictive quality		Benefit of ensemble		Sensitivity analysis	
			DSS	DPS	PSS	Ensemble <sup>2</sup>	per DPI	per drug	per DPI	per drug	Intrinsic	Extrinsic
1S	SEA [112]	2007	SE		N/A		✓		N/A	N/A		
	PRW & NB [127]	2013	TC		N/A		✓		N/A	N/A		
	DrugMiner [102]	2016		ML	N/A		✓		N/A	N/A		
	SDTNBI [135]	2016	TC		N/A		✓		N/A	N/A		
	Peón et al. [104]	2017	TC		N/A		✓		N/A	N/A		
	bSDTNBI [108]	2017	TC		N/A		✓		N/A	N/A		
2S	KRM [113]	2008	SI	AL	C		✓					
	Campillos et al. [114]	2008	TC	TC	S		✓			✓		
	COPICAT [115]	2009	KF		KF	C	✓					
	BLM [116]	2009	SI	AL	C		✓			✓		
	Yabuuchi et al. [118]	2011	KF		KF	C	✓			✓		
	GIP [119]	2011	SI	AL	C		✓					
	NBI [120]	2012	SI	AL	C		✓			✓		
	KBMF2K [121]	2012	SI	AL	C		✓					
	Cao et al. [123]	2012	KF		KF	C	✓					
	BLM-NII [124]	2012	SI		AL	S	✓					
	Cheng et al. [125]	2013	TC	TC		S & C	✓			✓		
	DT-Hybrid [126]	2013	SI		AL	S	✓					
	RWR [131]	2015	TC		AL	C	✓					
	SLP & RLS [132]	2015	SI		AL	C	✓					
	RLS-KF [133]	2016	SI		AL	C	✓					
	NRLMF [134]	2016	SI		AL	C	✓					
	DASPfind [136]	2016	SI		AL	C	✓					
	DrugE-Rank [137]	2016	TC		AL	S	✓			✓		
3S	DBN [103]	2017	NN		NN	C	✓					
	EnsemDT/KRR [138]	2017	KF		KF	C	✓					
	PUDTI [105]	2017	KF		KF	C	✓					
	DVM [106]	2017	KF		KF	C	✓					
	iDTI-ESBoost [109]	2017	ML		ML	C	✓					
	Yamanishi et al. [117]	2010	SI	CO	AL	C	✓			✓		
	PKR [122]	2012	KF	CO	KF	C	✓			✓		
This review	DINIES [128]	2014	SI	CO	AL	S	✓			✓		
	Shi et al. [129]	2015	TC	SS	AL	C	✓					
	Liu et al. [130]	2015	TC	TC	AL	S	✓					
	DTINet [107]	2017	TC	TC	AL	C	✓					
	This review	2018	TC	TC	AL	S	✓	✓	✓	✓	✓	✓

<sup>1</sup>DSS: drug structure similarity; DPS: drug profile similarity; PSS: protein sequence similarity; SE: SEA algorithm that measures DSS [112]; SI: SIMCOMP tool that quantifies DSS [142, 143]; TC: Tanimoto coefficient or other similarity score that measures similarity between binary or numeric vectors [139]; KF: kernel function that measures similarity between feature vectors used by machine learning algorithms; NN: neural networks; CO: correlation; SS: semantic similarity of drug profiles represented by the ATC codes [129]; AL: sequence alignment using PSI-BLAST [144] or Smith–Waterman algorithms [145]; and ML: machine learning algorithm that measures similarity between feature vectors.

<sup>2</sup>The ensembles are established based on either a simple summation (S) or a more complex (C) approach that involves operations such as maximum, multiplication and geometric mean of multiple similarities.

### Assessment of predictive performance and sensitivity of predictive performance to characteristics of predictors

Another important aspect of the considered DPI predictors is the assessment of their predictive performance. The predictions are always provided at an interaction level, which means that the predictive models identify whether a given drug–protein pair interacts. Two early predictors were assessed solely using

the interacting drug–protein pairs (see ‘Scope of evaluation’ column in [Supplementary Table S2](#)) [114, 127]. While such assessment evaluates true positive rate (TPR), which is defined as fraction of interactions that are correctly identified, it lacks the ability to assess specificity, which measures the fraction of correctly predicted non-interactions. The other 33 methods provide assessment for both interacting and non-interacting drug–protein pairs. This type of assessment is more complete

as it measures not only TPR but also the capability to correctly identify the non-interacting pairs that constitute majority of all possible drug–protein pairs. Moreover, the considered 35 predictors report their predictive performance over all drug–protein pairs that span multiple drugs. This means that they did not assess and compare predictive quality for individual drugs (see ‘Predictive quality’ column in *Table 1*). However, drugs differ in their structures, profiles and the number and type of protein partners that they bind to, and these factors likely impact the predictive performance.

Besides quantifying predictive performance, it is also essential to justify the use of ensembles. Since the ensembles require more computations and are more complex to implement than the 1S models, they must offer other benefits to balance these drawbacks. In particular, these benefits typically boil down to higher predictive performance when compared to the 1S methods. Only 6 out of 23 2S methods compare 2S model with 1S to quantify the corresponding improvement in the predictive quality evaluated per interaction (see ‘Benefit of ensemble’ column in *Table 1*) [114, 116, 118, 120, 125, 137]. Additionally, three 3S methods provide empirical evidence that the ensemble of three similarities boosts the predictive quality at interaction level when compared to the 2S ensembles [117, 122, 128]. While this suggests that ensemble models may provide more accurate results than the 1S models, follow-up studies are needed to explore the benefits of ensembles using both per interaction and per drug evaluation. Moreover, none of these methods has comprehensively compared all three 1S, three 2S and one 3S models, which would provide more insights on the use of the similarity-based predictors.

Another interesting aspect is sensitivity of predictive performance to certain characteristics of predictors. For example, in a scenario where the query drug is dissimilar to the drugs in the internal database of a given predictor, DSS and DPS are likely to be ineffective. On the other hand, if the query drug has sufficient levels of similarity to some of the drugs in the internal database then we can use these similar drugs to likely accurately predict the protein targets. Similarity between the user-provided inputs and the contents of the internal database of a given predictor relies on an intrinsic property of this predictor (i.e. its database). On the other hand, predictive performance could be also affected by extrinsic to the predictor and known *a priori* properties of the query drug and targets, such as the drug promiscuity. For instance, it might be easier to identify DPIs for a drug that targets numerous proteins because this increases a chance to find many other similar targets in the internal database. Analysis of sensitivity of performance to these characteristics would help to calibrate the expectations of predictive quality and optimize the design of predictors for specific use cases. However, this type of sensitivity analysis was never addressed in the past studies (see ‘Sensitivity analysis’ column in *Table 1*).

Besides quantifying predictive performance in silico, some of these predictive tools were also assessed experimentally. In that case, the correctness of the putative DPIs that they generate was verified by *in vitro* binding assays and cell assays. This type of validation was done for the earliest predictor [112], the method that first defined DPS [114], and in one recent study [107].

### Impact of the similarity-based methods

One way to measure the impact is to consider the impact factors of venues where these methods were published. We also collect and compare their citation counts. *Supplementary Table S2* lists

the journal impact factors that were collected from the Journal Citation Reports (Clarivate Analytics, 2017) [111]. While by design these methods were published in journals with impact factor  $> 3.5$ , the median impact factor of these journals is 5 and 4 journals have impact factor  $> 10$ . The table also shows total and annual counts of citations that were collected from Google Scholar on April 1, 2018. The annual count is measured as an average citation frequency per calendar year (365 days) calculated over the period from publication date until April 1, 2018. The three pioneering methods that were published in 2007 and 2008 [112–114] have together accumulated 2260 citations. This number is greater than the combined citation count (1966) of the remaining 33 predictors. These three highly cited predictors include the earliest 1S and 2S models that defined the three main types of similarities for the DPI prediction. Moreover, most of the 35 considered methods enjoy high levels of annual citations (median = 10), in many cases exceeding the corresponding journal impact factors (median = 5). This result suggests that the similarity-based DPI predictors elicit strong interest of the scientific community.

### Databases used by the similarity-based methods

One of the key factors that affect the quality of the similarity-based methods is data contents of internal databases, which are a part of these methods. The internal databases contain native DPIs that are derived from anywhere between one and six source databases. These sources include MDDR [22, 23], PDSP Ki [24], BRENDA [25–33], BindingDB [34–38], TTD [39–44], KEGG BRITE [45–51], DrugBank [52–56], GLIDA [57, 58], KEGG DRUG [45–51], SuperTarget [59, 60], Matador [59], STITCH [61–65] and ChEMBL [66–69] (see ‘Source databases’ column in *Supplementary Table S2*). *Supplementary Table S2* shows that most of the internal databases cover between 276 and 7739 drugs. Some of the internal databases also include other drug-like molecules, and correspondingly their sizes are larger and they span between 22 839 and 105 946 molecules. The number of protein targets in the internal databases ranges between 246 and 6662, with the median of 989. These are typically human proteins, except for nine methods that do not explicitly specify organism information of the proteins that they use (see ‘Focus on human proteins’ column in *Supplementary Table S2*). These numbers are comparable to an expected scale of druggable human proteome, which is estimated to include between 1000 and 3000 proteins [2, 19–21]. Another important measure that quantifies coverage of the internal databases is the number of DPIs. This number varies between 1731 and 155 208 with median of 5127. While some databases include a large number of drugs or targets (e.g. databases for SEA, SDTNBI and EnsemDT/KRR predictors), they cover a relatively low number of DPIs per drug. The density of DPIs, which is calculated as the average number of DPIs per drug, varies widely between 1.1 and 11.5 (see ‘DPUs/drug’ column in *Supplementary Table S2*). Databases with low densities might be missing a substantial number of DPIs and consequently the corresponding predictors might provide inaccurate values of similarities. Moreover, higher density of interactions is likely associated with a more complete coverage of interactions for individual drugs. The corresponding internal databases can be used to provide a more reliable estimation of predictive quality at the drug level.

To summarize the qualitative review of the 35 similarity-based DPI predictors, we show that they rely on a variety of approaches to quantify DSS, DPS and PSS. Most of these

predictors utilize ensembles that combine two or three types of similarities, were published in high impact venues and are highly cited. These studies assess predictive performance at the interaction level and some of them have demonstrated that the use of certain ensembles leads to improved predictive performance when compared to the use of a single similarity. However, a comprehensive empirical comparison of all types of similarities and their ensembles is still missing. Although the current works perform empirical assessments over a large number of drug–protein pairs, primarily in the druggable human proteome, the density of DPIs in the databases that were utilized in these assessments is rather low. Moreover, per drug assessment of the predictive quality was not yet performed. Similarly, sensitivity of predictive performance to characteristics of predictors was not yet studied. To this end, we provide comprehensive and in-depth empirical comparative analysis for a representative set of similarity-based predictors. First, we build a novel database that includes drugs characterized by high density of DPIs. Next, we comprehensively evaluate a representative set of predictors that are based on each of the three main types of similarities and the corresponding four possible ensembles. Besides evaluating predictive performance at the interaction level, we are the first to perform the drug level assessment and to offer a comprehensive analysis of sensitivity of predictive quality to the intrinsic and extrinsic characteristics of predictors, i.e. we consider 1S, 2S and 3S models and a relatively large number of characteristics.

## Empirical comparative analysis of similarity-based methods for prediction of DPIs

We empirically evaluate seven representative similarity-based DPI predictors. These seven predictors rely on each of the three individually used similarities (DSS, DPS and PSS) plus four ensembles that combine each pair and all three similarities. We delineate the setup for the assessment by describing the seven predictors and defining measures and a novel high-quality benchmark database of DPIs that we use to assess the predictive performance. Like in the previous studies, we analyze predictive performance at the interaction level. Moreover, we compare all seven representative methods, measure and compare predictive performance at the drug level and provide a comprehensive analysis of sensitivity of predictive performance to characteristics of input drugs and proteins.

### Experimental setup

#### Computation of similarities

The key underpinning concept that defines similarity-based predictors of DPIs is how they quantify and combine the three similarities. We selected one representative approach drawn from previous works to measure each type of similarity. The Tanimoto similarity of chemical structures has been widely applied to quantify DSS [139]. Recent comparative studies have demonstrated that Tanimoto similarity is one of the best DSS measures [141, 151, 152]. 11 selected predictors have used the Tanimoto coefficient to quantify DSS [104, 107, 108, 114, 125, 127, 129–131, 135, 137]. The above reasons motivate our choice of this measure to quantify DSS. Tanimoto coefficient is computed based on molecular fingerprint that converts a chemical structure of a given drug into a bit vector of fixed size. For a given compound, the fingerprint represents presence of certain substructures with a bit value of 1 and their absence with a bit value of 0. Similarity

of two drug structures is represented by the similarity between their corresponding bit vectors A and B:

$$\text{DSS} (A, B) = a_1 b_1 / (a_1 + b_1 - a_1 b_1),$$

where  $a_1 b_1$  is a count of corresponding positions where both vectors A and B have values 1, and  $a_1$  and  $b_1$  are the counts of positions with value 1 in vectors A and B, respectively. Higher value of the Tanimoto similarity means the two molecules are more structurally similar since they have a higher fraction of similar substructures in common. We use the Chemistry Development Kit software [153–156] to generate 1024-bit molecular fingerprints of drug structures and to compute the Tanimoto similarity between drugs.

We quantify DPS based on a drug side-effect profile similarity introduced in one of the pioneering predictors that utilized this type of similarity for the 1st time [114]. A given drug is represented by a profile vector that denotes presence/absence of certain side-effect term with a bit value of 1/0. The profile similarity is computed using the weighted Tanimoto coefficient between the side-effect profiles of a given pair of drugs. Each side-effect term has its own weight. The weight is derived based on an abundance of the corresponding term in a large collection of drugs (all drugs included in the internal database) and a correlation of this term with the other terms. Higher weights are associated with terms that are less frequent (specific to a small number of drugs) and that are dissimilar to other terms (they are infrequently used together with other terms for the same drugs). Consequently, higher value of DPS means that a given pair of drugs shares a high fraction of side-effect terms with high values of weights. This approach to quantify DPS is arguably more accurate than the approaches taken by some other methods that use a simpler definition of weights [130] or unweighted Tanimoto coefficients [125].

We use sequence alignment to estimate PSS. This is the most popular approach that was adopted by 18 predictors [107, 113, 116, 117, 119–121, 124, 126, 128–134, 136, 137]. First, we run BLAST [157, 158] with default parameters to perform sequence alignment between a sequence of a protein target of the input drug provided by user and each protein target in the database of known DPIs. BLAST finds the closest protein from the database based on a distance defined as the fraction of identical and similar residues in the pairwise alignment. Therefore, proteins from the database that have high value of sequence similarity are arguably likely to also interact with the input drug by the virtue of being similar to the known target of this drug.

We create the four ensemble models using a linear combination of multiple types of similarities. This approach is in line with a number of published ensemble methods [114, 124–126, 137]. We consider the 3S model that combines the three types of similarities and three 2S ensembles that combine each pair of the three similarities, which we denote as DSS + DPS, DSS + PSS and DPS + PSS. Some of the past methods combine similarities using more sophisticated approaches, such as machine learning models and drug–protein networks. A typical machine learning-based predictor uses an algorithm such as Support Vector Machine (SVM) to combine multiple kernel functions/matrices derived from two or three similarities [128]. However, SVM requires computationally costly parameterization and this is a black-box method, which means that the model is so complex that it is impossible to understand how predictions are performed. Network-based predictors rely on a linear combination of similarities to quantify weights associated with edges in the network [124]. These predictors require knowledge of drug–drug or drug–target connections (network edges) for an input

drug. Consequently, they cannot be utilized to predict DPIs for drugs that are not already integrated in the network [79]. We opt to use the linear combination of similarities which is computationally efficient, white-box (linear function explicitly defines how similarities are combined), and works with novel drugs. We also compare this linear model with three other diverse and popular machine learning models including Naïve Bayes, C4.5 decision tree and RBF network.

#### **Prediction of DPIs**

To perform the prediction, we screen the query drug against all candidate proteins from the internal database of known DPIs to evaluate how likely they may interact with this drug. Given the three types of similarities, the input information for the query drug may include drug structure, drug profile and/or protein sequences of its known targets. The procedure to perform predictions depends on which inputs are available. If the drug structure is available then for every candidate protein in the database we compute DSS between the query drug and every single drug that is known to bind to that candidate protein (excluding the query structure itself if the candidate protein is already known as a binding partner of this drug). The maximal value among these DSS values is used as the propensity of DPI between the query drug and the candidate protein. A higher propensity indicates that the candidate protein is more likely to bind to the query drug because this protein is known to be a target of a drug that is structurally similar to the query drug. This is consistent with an observation that similar drug structures share the same protein target [79, 87–89]. A similar procedure is used when the drug profile is available. The propensity of DPI for the input drug and a given candidate protein is computed as the maximal DPS between this drug and all drugs that are known to bind to this candidate protein in the database of known DPIs. Such prediction relies on the assertion that drugs with similar side-effect profiles likely target the same protein [114]. If the protein targets of the input drug are known and their sequences are available then we compute pairwise PSS between each candidate protein and all known targets of this drug (excluding the candidate protein itself if this protein is already known to interact with the query drug in the database). The maximal PSS among all known targets is used as the propensity of DPI for the query drug and the candidate protein. The sequence similarity-based prediction is motivated by an observation that similar proteins tend to interact with the same drug [79, 87–89]. If multiple inputs are available then they are combined together. We assess whether such ensembles of two or three similarities would provide improved quality of predictions when compared to using these similarities individually.

#### **Benchmark database of DPIs**

The DPI predictors rely on the internal database that contains native DPIs. We develop a novel benchmark database that we use as the internal database to evaluate these predictors. The benchmark database integrates three resources: Drug2Gene [159], Therapeutic Target Database (TTD) [39–44] and IUPHAR/BPS Guide to Pharmacology database (GtP) [160–165]. Our database is primarily derived from Drug2Gene that integrates 19 source databases. These sources include four databases, PDSP Ki [24], TTD [39–44], DrugBank [52–56] and ChEMBL [66–69], which are used in the 35 considered predictors to derive their internal databases. However, we also include 15 other sources: CGDGP [166], ChEBI [167], CTD [168], HGNC [169], GtP [160–165], Ligand

Expo [170], MICAD [171], NCBI Gene [172], Pathway Commons DB [173], PDBsum [174], PharmGKB [175], Pubchem Bioassay [176], PubChem Compound [177], PubChem Substance [177] and UniProt [178]. Considering that the 35 predictors adopt no more than 6 source databases, Drug2Gene provides us with more complete information about DPIs. Another survey also highlights Drug2Gene as the most comprehensive DPI database [179]. In particular, use of Drug2Gene helps us to acquire high density of DPIs in the benchmark database. However, Drug2Gene (version 5) incorporates older versions of the TTD database (release 2011) and GtP database (release 2012) when compared its other sources. Thus, we combined Drug2Gene with newer versions of TTD (release 2015) and GtP (release 2015). These three resources altogether cover 10 515 compounds, 16 880 proteins and 161 736 interactions. The density of DPIs for this database is 15 DPIs/drug, before removal of duplicates across these sources (see 1st row in *Supplementary Table S3*).

The seven predictors that we assess rely on quantification of the three types of similarities. Thus, drug structures, drug profiles and protein sequences are required for every drug/protein entry in the benchmark database. Drug structures were collected from PubChem [177] by querying their compound identifiers (CIDs) that are available in three sources databases. Drug profiles were obtained from SIDER 2 [146, 147], a database of drug side effects for marketed drugs, by matching PubChem CIDs. Protein sequences were retrieved from UniProt [178] using their UniProt accession numbers (UniProt IDs) that are available in TTD and GtP, and with the Entrez Gene [172] identifiers that are available in Drug2Gene. Consequently, drugs and proteins are respectively indexed by PubChem CIDs and UniProt IDs along with drug structures from PubChem, drug profiles from SIDER and protein sequences from UniProt in the benchmark database. We found redundant drug/protein entries because older versions of TTD and GtP have already been imported into Drug2Gene. We removed the duplicate drugs and proteins by comparing PubChem CID and UniProt ID, respectively. SIDER 2 provides side effects for 996 drugs, which limits the number of drugs in our benchmark database. After collecting information for computing similarities and cleaning the duplicates, the resulting benchmark database contains 965 drugs, 7565 proteins, 93 015 interactions and 96 DPIs/drug (see 2nd row in *Supplementary Table S3*). Like most of the studies in this area [102, 105–109, 113, 114, 116, 117, 119–124, 126–130, 133–137], we focus on the human protein targets. Thus, we eliminated non-human proteins based on their taxonomic identifiers from UniProt. Consequently, 563 drugs, 1537 proteins, 35 352 interactions and 63 DPIs/drug are left in the benchmark database (see 3rd row in *Supplementary Table S3*). Next, to ensure that sufficient amount of information about drug structure and protein sequence similarities is available to perform empirical assessment (particularly at the drug level) and sensitivity analysis, we removed small compounds that have fewer than 10 atoms and drugs that have fewer than 5 known protein targets, respectively. The final version of the benchmark database is composed of 449 drugs, 1469 proteins and 34 456 interactions and includes 76 DPIs/drug. When compared to the internal databases of the considered 35 predictors, this benchmark database is among the top 11% with respect to the number of DPIs (*Supplementary Table S2*). Moreover, an arguably key indicator of a high-quality benchmark DPI database is the density of DPIs. High density of interactions facilitates reliable estimates for per-drug assessment of predictive performance. The benchmark database has the highest density of DPIs at 76 DPIs/drug (*Supplementary Table S2*). This is a direct result of merging the largest number of sources of the DPI data.

The source DPI databases include annotations of native interacting drug–protein pairs, but they lack the annotations of non-interacting pairs. In the past, these databases were used to assess predictors of DPIs in several ways. A few early studies assessed predictive performance using only the native interacting pairs [114, 127]. These assessments are incomplete since they measured only the rate of correctly predicted native DPIs while they did not evaluate the predictive performance for the non-interacting pairs, i.e. they did not evaluate the number of false positives. More recent works have assessed predictions using both interacting and non-interacting drug–protein pairs. Most of these studies assume that the drug–protein pairs that are not annotated as interacting are non-interacting. This assumption may result in mislabeling some of the native interacting pairs as non-interacting; in particular, those that were not yet discovered. This commonly applied approach to annotate the non-interacting pairs relies on an assertion that the number of such mislabeled pairs is much smaller than the number of correctly assumed non-interacting pairs. We are the first to study this assumption empirically. We annotate weak binders among the native drug–protein binding pairs, and we compare similarity between the assumed non-interacting drug–protein pairs and the similarity among the weakly binding pairs. We hypothesize that the assumed non-interacting drug–protein pairs can be used to represent the native non-interacting pairs if the difference between similarity of the non-interacting drug–protein pairs and similarity among the native interacting pairs is significantly higher than the difference between similarity of weak interactors and the similarity of the native interacting pairs.

We map the drug–protein pairs from the benchmark database into the PDSP Ki database that provides binding affinities [24]. A high-binding affinity (affinity constant  $K_i < 10 \mu\text{M}$ ) is typically considered to indicate an interacting drug–protein pair [114, 180]. Consequently, we divide the mapped drug–protein pairs into three subsets: the native interacting pairs with  $K_i < 10 \mu\text{M}$ , weak native interacting pairs with  $K_i > 10 \mu\text{M}$  and the non-interacting pairs that include all pairs that are not annotated as DPIs. We compute DSS, DPS and PSS for the drug–protein pairs in each subset. *Supplementary Figure S1* compares distributions of the values of three similarities for the native interacting pairs, weak native interacting pairs (borderline pairs) and non-interacting pairs. The distributions are represented by the 5<sup>th</sup>, 20<sup>th</sup>, 50<sup>th</sup> (median), 80<sup>th</sup> and 95<sup>th</sup> percentiles. The results are consistent across the three similarities. The similarity values for the non-interacting drug–protein pairs are significantly lower when compared to the similarities for the weak native interactors ( $P\text{-value} < 0.01$ ). They are also significantly lower when compared to the similarities for the native interacting pairs ( $P\text{-value} < 0.01$ ). In other words, *Supplementary Figure S1* reveals that only a small portion of the non-interacting pairs has DSS, DPS and PSS values that are comparable to the DSS, DPS and PSS values of the native DPIs, respectively. We also observe that the distributions for DPS and PSS for the weak native interactors are similar to the distributions for the native interactors ( $P\text{-value} = 0.11$  and  $0.65$ , respectively). While the difference between weak native interactors and native interactors for DSS is significant, the median DSSs are  $0.65$ ,  $0.41$  and  $0.28$  for the native interacting pairs, weak native interacting pairs and non-interacting pairs, respectively. Over 95%/80% of the native interacting pairs have higher DSS values than at least 50%/80% of the non-interacting pairs. In agreement with previous studies, these results suggest that the assumed non-interacting pairs can be used to represent non-interacting pairs since their similarity is significantly lower than the similarity

of the weak native interactors and the native interactors, while the differences between the latter two sets of drug–protein pairs are substantially smaller. We also observe that the number of non-interacting pairs that have relatively high PSS is greater than those that have relatively high DSS and PSS. This implies that predictions using PSS might be less accurate than the predictions with DSS or DPS.

#### Assessment of predictive performance

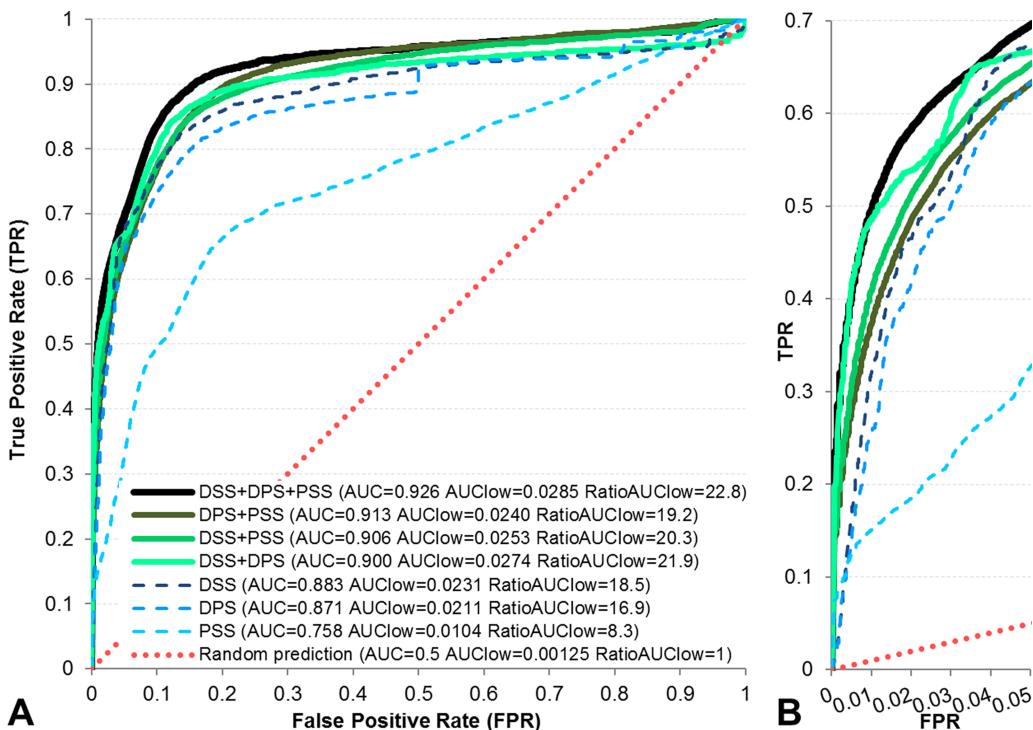
Using the benchmark database, we empirically test predictions for each of the 449 drugs by comparing the putative targets generated by a given method with the known native interactors for this drug. When testing predictions for a given drug, we remove information about its known targets from the database. We analyze the results across all drugs and separately for each drug. Like virtually all considered in this review methods we use the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) to assess quality of DPI predictions [102, 103, 105–109, 112, 113, 115–126, 128–138]. ROC curve is a plot of TPRs against false positive rates (FPRs):

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}),$$

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN}),$$

where the TP (true positive) and the TN (true negative) are the numbers of native interacting pairs and a native non-interacting pairs that are predicted correctly, respectively; the FN (false negative) is the number of native interacting pairs predicted as non-interacting; and the FP (false positive) corresponds to the number of native non-interacting pairs predicted as interacting. The predicted propensities are transformed into binary predictions (positive versus negative) using threshold values, which are next used to compute the TPR and FPR. Drug–protein pairs with propensity above the threshold are assumed to bind (positive), otherwise they are assumed to be non-binding (negative). The thresholds are set to be equal to all unique values of the predicted propensities of DPIs in order to derive the most accurate ROC curve. Higher AUC values denote better predictive performance, i.e. they correspond to predictions that offer higher TPRs (more correct predictions of DPIs) for lower FPRs (fewer incorrect predictions of non-interactions as DPIs).

The fraction of the native interacting drug–protein pairs among all drug–protein pairs in the benchmark database is at about 4%. Thus, for example, if  $\text{FPR} = 8\%$  then the number of predicted putative DPIs is about twice the number of actual native interacting pairs. This corresponds to a substantial over-prediction of putative DPIs. Moreover, the AUC values are determined primarily by the part of the curve where FPR is high, focusing on the over-prediction, i.e. area under the curve for the low FPR values is much smaller than for high FPR values. Thus, motivated by recent works [181–185], we focus the lower part of ROC with FPR less than 4% ( $\text{ROC}_{\text{low}}$ ) where the number of incorrectly predicted DPIs is not larger than the number of native DPIs. Accordingly, we calculate the corresponding area under  $\text{ROC}_{\text{low}}$  curve ( $\text{AUC}_{\text{low}}$ ). Moreover, to ease interpretation of this index, we report  $\text{Ratio}_{\text{AUClow}} = \text{AUC}_{\text{low}}/\text{AUC}_{\text{low-random}}$ , where  $\text{AUC}_{\text{low-random}}$  is  $\text{AUC}_{\text{low}}$  for a random predictor. A random predictor generates TPRs that are equal to FPRs, resulting in a diagonal line in the TPR–FPR space.  $\text{Ratio}_{\text{AUClow}}$  measures to what degree a given predictor is better than a random predictor.  $\text{Ratio}_{\text{AUClow}} > 1$  means the method is better than random.



**Figure 1.** Comparison of predictive performance for the 1S, 2S and 3S DPI predictors measured at the interaction level. Panel A shows ROC curves and panel B shows ROC<sub>low</sub> for which FPRs  $\leq 0.05$ . Diagonal dotted line denotes the ROC curve for a random predictor for which TPR = FPR. Values of AUC, AUC<sub>low</sub> and RatioAUC<sub>low</sub> are shown in the legend for panel A.

### Empirical analysis of predictive performance

We perform evaluation of the predictive quality of the three 1S predictors as well as their 2S and 3S ensembles. We assess predictive performance for each drug by removing it from the benchmark database, predicting its interactions with each proteins target in that database and evaluating these predictions against the removed native interactions of this drug. We aggregate these results over all the drugs and 659 581 drug–protein pairs. A query drug might be dissimilar to the drugs in the benchmark database in terms of its structure and side-effect profile. To simulate a practical scenario, we limit the similarity between the query drug and the drugs in database. Specifically, given a query drug, we not only remove this drug but we also remove the drugs with above median DSS and DPS from the benchmark database when performing the evaluations. Use of the median represents an average case. Note that the protein sequences are easy to collect and all of them are available for every prediction. The complete set of sequences for human proteins can be obtained from UniProt. Thus, the predictions for a given query drug are generated based on the drugs with the below-median DSS and DPS, and the complete set of protein sequences.

### Assessment of predictive performance at the DPI level

Figure 1 shows the ROCs and ROCs<sub>low</sub> that measure the predictive quality at the interaction level over all drug–protein pairs in the benchmark database for the three 1S, three 2S and one 3S predictors. The 3S model has the highest AUC, AUC<sub>low</sub> and RatioAUC<sub>low</sub> values, which are given in the legend for Figure 1. The 1S models register a modest drop in predictive quality when compared to the 2S models. Specifically, the relative difference in AUC (in AUC<sub>low</sub>) between 3S and the best 2S model equals [0.926 –

0.913]/0.913 = 1.4% ( $[0.0285 - 0.0274]/0.0274 = 4.0\%$ ) (Figure 1A). The corresponding relative difference when comparing 3S and 1S models is at least  $[0.926 - 0.883]/0.883 = 4.9\%$  for AUC and  $[0.0285 - 0.0231]/0.0231 = 23.4\%$  for AUC<sub>low</sub>. These results consistently demonstrate that the use of ensembles results in improvement in predictive quality. Moreover, our results show that adoption of the 3S, the three 2S and DSS and DPS models provides high-quality predictions. The only relatively low results for PSS could be perhaps explained by the larger overlap of similarities between the native DPIs and non-DPIs observed in Supplementary Figure S1. Although DSS provides more accurate predictions in terms of AUC when compared to DPS and PSS, the ensemble of DPS + PSS outperforms the two 2S predictors that rely on DSS. Combing DPS and PSS provides the largest improvement over these similarities used individually. Specifically, the relative difference in AUC (in AUC<sub>low</sub>) is 4.8% (13.7%) and 20.4% (130%) when comparing DPS + PSS with DPS and PSS, respectively. The corresponding improvements for the other two 2S models are more modest. DSS + PSS provides a relative increase in AUC (in AUC<sub>low</sub>) when compared to DSS and PSS by 2.6% (9.5%) and 19.5% (143%), respectively. Similarly, the value of AUC (AUC<sub>low</sub>) for DSS + DPS is better by 1.9% (18.6%) and 3.3% (29.9%) than the values for DSS and DPS, respectively. One possible explanation for these differences is that each of the three similarities offers good predictive performance while DPS and PSS are lack any correlation (Spearman correlation coefficient (SCC) = 0.05), DSS and PSS are weakly correlated (SCC = 0.14) and DSS and DPS are modestly correlated (SCC = 0.30). Therefore, PSS and DPS complement each other the most which results in the largest improvement in the quality of predictions.

When FPR is greater than 0.05, the number of native non-DPIs that are incorrectly predicted as DPIs is larger than the total number of native DPIs. In other words, we over-predict

**Table 2.** Significance of differences in the predictive performance of DPI predictors measured at the interaction level. The distribution of AUC,  $AUC_{low}$  and  $Ratio_{AUClow}$  values quantified over 100 different datasets (randomly selected 10% of drugs from the benchmark database) is represented by average  $\pm$  standard deviation. The italic and underlined fonts represent results of tests of significance of difference in AUC and in  $AUC_{low}$ , respectively, for each pair of predictors. We quantify the corresponding P-values by running t-test (for measurement with normal distribution) or Wilcoxon signed-rank test (for non-normal data) between each pair of predictors. We test the normality with the Anderson-Darling test at 0.01 significance

Method	Predictive quality (average $\pm$ standard deviation)			Statistical significance (P-value) of the differences between all pairs of predictors						
	AUC	$AUC_{low}$	$Ratio_{AUClow}$	3S	DPS + PSS	DSS + PSS	DSS + DPS	DSS	DPS	PSS
<b>3S</b>	$0.925 \pm 0.005$	$0.0284 \pm 0.0005$	$22.7 \pm 0.4$		<u>5.4E-26</u>	<u>4.0E-77</u>	<u>2.3E-46</u>	<u>3.7E-59</u>	<u>1.6E-54</u>	<u>8.5E-104</u>
<b>DPS + PSS</b>	$0.912 \pm 0.004$	$0.0240 \pm 0.0005$	$19.2 \pm 0.4$	<u>1.0E-69</u>		<u>3.3E-10</u>	<u>1.9E-12</u>	<u>1.0E-30</u>	<u>5.5E-51</u>	<u>5.8E-105</u>
<b>DSS + PSS</b>	$0.905 \pm 0.005$	$0.0252 \pm 0.0005$	$20.2 \pm 0.4$	<u>4.7E-64</u>	<u>3.9E-14</u>		<u>7.4E-07</u>	<u>4.1E-37</u>	<u>5.8E-34</u>	<u>4.8E-98</u>
<b>DSS + DPS</b>	$0.899 \pm 0.006$	$0.0273 \pm 0.0007$	$21.9 \pm 0.6$	<u>1.3E-08</u>	<u>1.3E-30</u>	<u>1.7E-18</u>		<u>8.2E-45</u>	<u>8.4E-32</u>	<u>5.5E-81</u>
<b>DSS</b>	$0.882 \pm 0.006$	$0.0230 \pm 0.0006$	$18.4 \pm 0.5$	<u>2.0E-56</u>	<u>3.0E-05</u>	<u>3.2E-22</u>	<u>5.2E-58</u>		<u>3.3E-08</u>	<u>5.1E-74</u>
<b>DPS</b>	$0.869 \pm 0.008$	$0.0210 \pm 0.0007$	$16.8 \pm 0.6$	<u>2.9E-69</u>	<u>3.7E-34</u>	<u>7.9E-39</u>	<u>5.7E-78</u>	<u>8.2E-23</u>		<u>1.5E-66</u>
<b>PSS</b>	$0.759 \pm 0.012$	$0.0100 \pm 0.0029$	$8.0 \pm 2.3$	<u>3.2E-92</u>	<u>9.2E-87</u>	<u>2.0E-88</u>	<u>2.1E-72</u>	<u>1.6E-65</u>	<u>1.2E-57</u>	

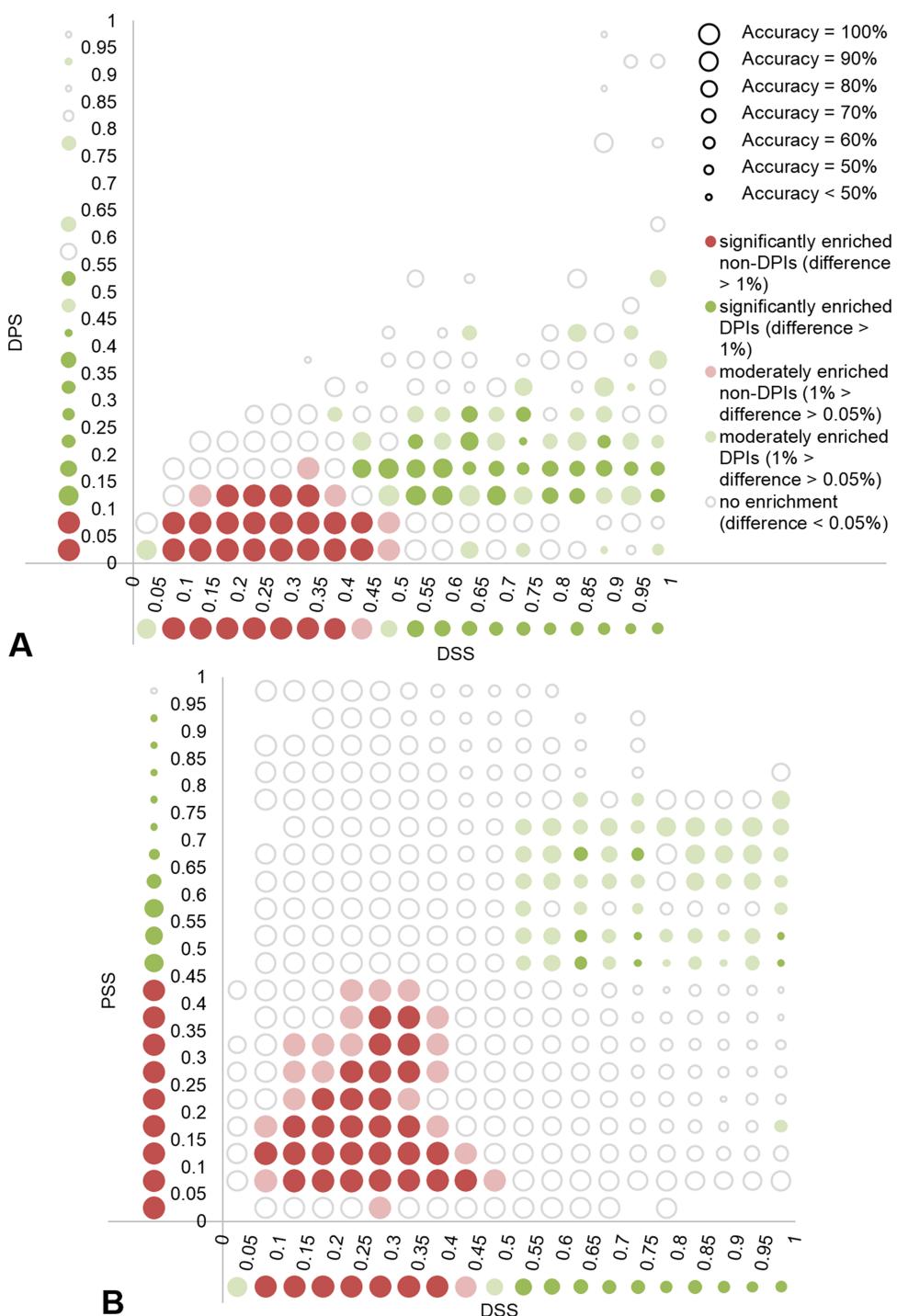
DPIs. Therefore, we utilize  $AUC_{low}$  that considers the ROC curve for  $FPR < 0.05$ . The advantages of using multiple similarities, i.e. 3S versus 2S and 2S versus 1S, are maintained for  $AUC_{low}$  (Figure 1B). The  $Ratio_{AUClow}$  reveals that 3S model is about 23 times more accurate than random predictions while 2S and 1S models are up to 22 and 19 times more accurate when compared to the random predictions. In particular, 3S predictor identifies 70% and the best 2S model finds 67% of native DPIs while they incorrectly recognize 5% of non-DPIs as DPIs ( $FPR = 0.05$ ). Almost all predictors successfully detect over 63% of native DPIs ( $TPR > 0.6$ ) at 5% level of FPR, except PSS which finds 33% of native DPIs. Similar to the evaluations based on the overall AUC, predictor based on DSS is the most accurate 1S model in terms of  $AUC_{low}$ . Therefore, DSS drives the predictive quality of the ensembles. Moreover, PSS has the lowest performance of the three similarities, which is still eight times better than the random predictions. Moreover, PSS is helpful for the ensemble models given its lack of correlation with the other two similarities. The  $AUC_{low}$  for the 2S ensembles that include PSS register relative improvement by 9.5% and 13.7% when contrasted with DSS and DPS, respectively.

We also evaluate statistical significance of the differences in predictive quality between the 3S model, the three 2S models and the three 1S models. To do that, we evaluate the differences over a diverse collection of 100 datasets drawn from the benchmark database. Each dataset includes 10% of randomly selected (without replacement) drugs from the benchmark database. Therefore, overlap between these datasets is minimal resulting in a large and diverse sampling of subsets of the benchmark datasets. We measure the AUC,  $AUC_{low}$  and  $Ratio_{AUClow}$  for each dataset. The corresponding averages and standard deviations quantified over the 100 datasets are given in Table 2. These averages are consistent with the results on the entire benchmark database that we give in Figure 1A. The P-values that measure significance of differences between all pairs of the considered seven methods are below 0.01, which mean that the differences in the predictive quality are statistically significant. This reveals that the benefits brought by the use of multiple similarities and the higher predictive performance of DSS compared to other 1S models are consistent over a diverse sampling of datasets. Overall, we show that 3S is the most accurate and significantly more accurate than the 2S predictors, while the 2S predictors are significantly better than the 1S models. DSS is significantly more accurate when compared to DPS

and PSS. These observations provide useful hints for future users and designers of the DPI predictors.

We study relationship between the predictive performance and similarity values by comparing predictions between the 3S and 1S models to understand how the predictive performance benefits from combining multiple similarities. We divide values of DSS, DPS and PSS into 20 equally sized intervals and utilize these intervals to partition the benchmark database into subsets in which the drug-protein pairs have comparable similarity values. Figure 2 deconstructs the 3S model into two-dimensional plots of two similarities, each divided into 20 intervals for the total of 400 data points. This figure compares the predictive quality of the 3S model (shown inside of the 1st quadrant) with the three 1S models (left most vertical set of points and bottom most horizontal set of points, which are outside of the quadrant). The predictive quality is measured with accuracy defined as a fraction of correctly identified DPIs and non-DPIs to total number of drug-protein pairs in a given data point; we denote that by the bubble size in Figure 2. We could not use AUC because some of these data points may not have either native DPIs or native non-DPIs. We use colors to denote relative enrichment of native DPIs and non-DPIs for each data point where dark red denotes enrichment in non-DPIs and dark green is for enrichment in DPIs. The enrichment is defined as the difference between the fractions of DPIs and non-DPIs, which are computed by dividing the number of native DPIs (non-DPIs) in a given data point by the total number of DPIs (non-DPIs) in the benchmark database. For instance, a light red point for  $DSS \in [0.25, 0.3]$  and  $PSS \in [0, 0.05]$  in Figure 2B shows that a modest majority of drug-protein pairs with these values of the two similarities are non-DPIs (which is denoted by the light red color) and accuracy of the 3S model for this data point is 100% (which is denoted by the large size of the bubble).

Figure 2A shows that the predictive accuracy for 3S model is separated into two distinct regions. The 1st region includes red data points for  $DSS \in [0.05, 0.45]$  and  $DPS \in [0, 0.15]$ . This region is enriched in native non-DPIs (red color) and has high predictive accuracy (large bubble). The 2nd region includes green data points for  $DSS \in [0.45, 1]$  and  $DPS \in [0.1, 0.3]$ . This region is enriched in native DPIs (green color) and has lower predictive accuracy (smaller bubble) when compared to the red region. As expected, the colors reveal that the native non-DPIs are concentrated in the drug-protein pairs with low values of DSS and DPS while the native DPIs are more enriched in the drug-protein pairs



**Figure 2.** Comparison of predictive performance between 3S and 1S models in the function of values of the three similarities. The values of the three similarities are divided into 20 equally sized intervals. The comparison of the 3S versus the three 1S models is divided into two panels that compare the 3S model versus two sets of 1S models each. Panel A shows the comparison of 3S versus DSS and DPS. Panel B shows the comparison of 3S versus DSS and PSS. Accuracy for the 3S model is shown inside the 1st quadrant while 1S models are given as a horizontal line below x-axis and a vertical line to left from the y-axis. Each bubble represents the drug–protein pairs in a given interval. Bubble sizes denote accuracy, which is defined as the fraction of correctly identified DPIs and non-DPIs to total number of drug–protein pairs in that bubble. Color denotes enrichment in native DPIs/non-DPIs, which is defined as the difference between the fractions of native DPIs and non-DPIs. These are computed by dividing the number of DPIs (non-DPIs) in a given bubble by the total number of DPIs (non-DPIs) in the benchmark database. We discretize the enrichment into five levels: dark green (dark red) bubble denotes significant enrichment in DPIs (non-DPIs) if the difference  $\geq 1\%$ ; light green (light red) bubble indicates moderate enrichment in DPIs (non-DPIs) if the difference  $\geq 0.5\%$ ; and white hollow bubble means no enrichment for which the difference  $< 0.5\%$ . White space in the figure denotes bubbles that we removed because they contain fewer than 100 drug–protein pairs; this is to ensure that statistics are robust.

with high similarities. This demonstrates that similarities could be used to discriminate DPIs from non-DPIs. The bubble sizes reveal that it is easier to provide high predictive performance

for non-DPIs because they contain many drug–protein pairs and most of them include drugs that are very dissimilar to drugs that are known to bind the same protein. Such dissimilar drugs are

**Table 3.** Significance of differences in the predictive performance between different implementations of the 3S model measured at the interaction level. The distribution of AUC,  $AUC_{low}$  and  $\text{Ratio}_{AUClow}$  values quantified over 100 different datasets (randomly selected 10% of drugs from the benchmark database) is represented by average  $\pm$  standard deviation. The italic and underlined fonts represent results of tests of significance of difference in AUC and in  $AUC_{low}$ , respectively, for each pair of predictors. We quantify the corresponding P-values by running t-test (for measurement with normal distribution) or Wilcoxon signed-rank test (for non-normal data) between each pair of predictors. We test the normality with the Anderson–Darling test at 0.01 significance

Algorithm used to ensemble similarities	Predictive quality (average $\pm$ standard deviation)			Statistical significance (P-value) of the differences between all pairs of ensemble predictors			
	AUC	$AUC_{low}$	$\text{Ratio}_{AUClow}$	Naïve Bayes	Regression	C4.5	RBF network
Naïve Bayes	$0.925 \pm 0.012$	$0.0274 \pm 0.0016$	$21.9 \pm 1.2$			<u><math>6.99E-22</math></u>	$6.29E-52$
Regression	$0.922 \pm 0.013$	$0.0282 \pm 0.0016$	$22.6 \pm 1.3$	<u><math>7.00E-37</math></u>		$4.01E-48$	$4.67E-50$
C4.5	$0.895 \pm 0.017$	$0.0305 \pm 0.0017$	$24.4 \pm 1.4$	<u><math>2.39E-53</math></u>	<u><math>6.14E-45</math></u>		$2.88E-01$
RBF network	$0.894 \pm 0.013$	$0.0278 \pm 0.0018$	$22.2 \pm 1.4$	<u><math>2.42E-05</math></u>	<u><math>1.26E-04</math></u>	<u><math>5.19E-37</math></u>	

easy to be correctly predicted as the non-interacting pairs. On the other hand, the predictions are less accurate for the green bubbles. This is because they cover drug–protein pairs for which drug structures are similar to the drugs that interact with the same proteins. Moreover, some of the pairs in these bubbles are still non-DPIs and they are difficult to differentiate from the native DPIs. Most importantly, Figure 2A explains why 3S model is better than the corresponding two 1S models that rely on DSS and DPS. DSS values can be used to reliably separate non-DPIs (dark red bubbles in the bottom horizontal line where  $DSS < 0.4$ ) from DPIs (dark green bubbles where  $DSS > 0.5$ ). However, drug–protein pairs for DSS between 0.4 and 0.5 cannot be easily separated. The separation of these drug–protein pairs into DPIs and non-DPIs can be improved by adding the values of DPS. Specifically, drug–protein pairs with  $DSS \in [0.4, 0.5]$  and  $DPS \in [0, 0.1]$  are mostly non-DPIs (red bubbles) while the drug–protein pairs with  $DSS \in [0.4, 0.5]$  and  $DPS > 0.1$  are mostly DPIs (green bubbles). Moreover, the bubble sizes also suggest that the accuracy of the 3S model is better than the DSS model. The bubble for  $DSS \in [0.45, 0.5]$  has moderate accuracy = 78%. The corresponding column for the 3S model at  $DSS \in [0.45, 0.5]$  has much larger bubbles that correspond to higher accuracies. The accuracy for the 3S model is high (between 89% and 98%) when  $DPS < 0.2$  and  $DSS \in [0.45, 0.5]$ . On average, accuracy of the 3S model for  $DSS \in [0.45, 0.5]$  equals 84% when compared to 78% when DSS is employed alone. Similar observations can be made for Figure 2B. Overall, the 3S model improves over the 1S models in the interface between the red and green regions.

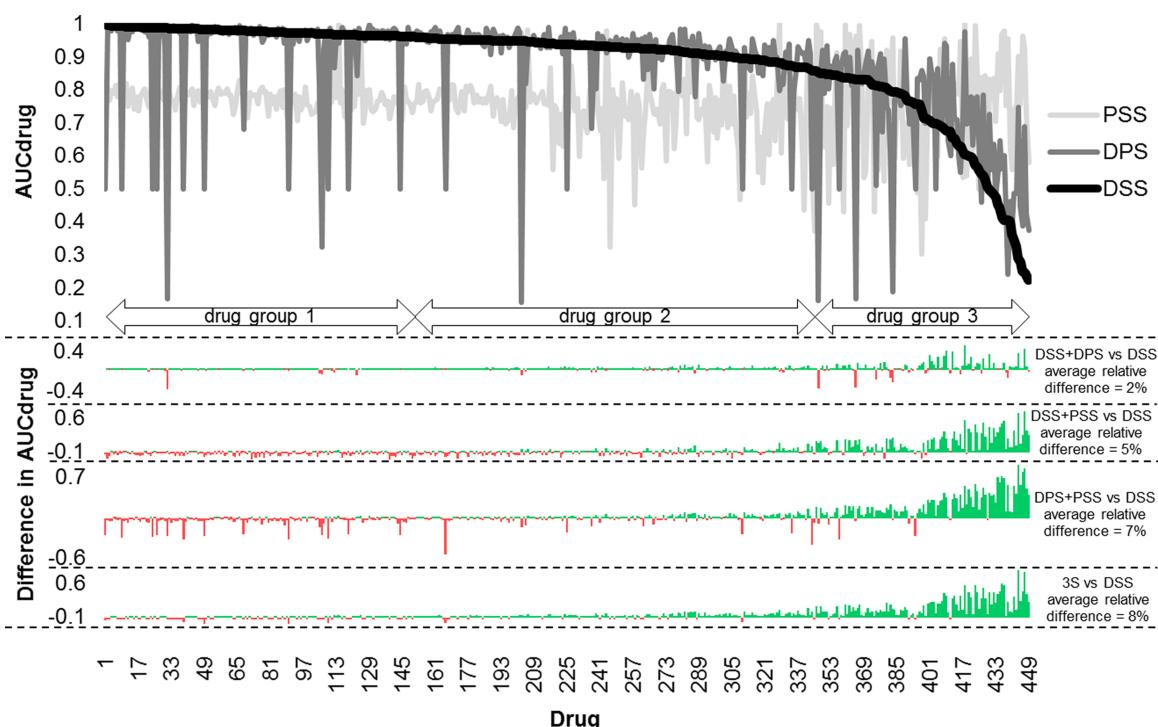
We compare the 3S model that uses linear function (regression) to combine the three similarity types with three other machine learning models: Naïve Bayes, C4.5 decision tree and RBF network. The results are summarized in Table 3. We find that there is no clear winner. The two methods that offer favorable predictive performance that is balanced between AUC and  $AUC_{low}$  measures are Naïve Bayes and regression. While the Naïve Bayes-based ensemble provides a slightly better AUC score, the regression-based ensemble offers a modestly higher  $AUC_{low}$  and  $\text{Ratio}_{AUClow}$ . We argue that given the imbalanced nature of our dataset (native interacting DPIs constitute about 4% of the dataset) the regression is slightly more suitable for these predictions. The other two ensemble models that rely on the decision tree and the RBF network offer substantially lower AUC values. However, we observe that the decision tree-based ensemble outperforms Naïve Bayes and regression-derived ensembles when focusing solely on the  $AUC_{low}$ , as the trade-off for the substantially reduced AUC.

#### Assessment of predictive performance at drug level

All previous studies evaluate predictive performance at the interaction level, which measures accuracy of the DPI predictions over drug–protein pairs that span across multiple drugs. Considering the fact that different drugs may share different levels of similarity to the benchmark database, it would be useful to evaluate these drugs individually. We provide first-of-its-kind assessment of predictive performance for DPI predictions at the drug level. The drug-level assessment provides further insight into complementarity and differences between various 1S, 2S and 3S predictors. We measure the AUC over the DPI predictions for each individual drug ( $AUC_{drug}$ ), which evaluates the quality of associations that are predicted between a given drug and all of the 1469 druggable proteins. We ensure that each drug in the benchmark database has at least five protein targets, which allows us to provide relatively robust estimates of the  $AUC_{drug}$  values.

The upper half of Figure 3 shows the  $AUC_{drug}$  for the three 1S predictors over each of the 449 drugs. The drugs are sorted by the corresponding  $AUC_{drug}$  values of DSS, which is the most accurate 1S model when measured at the interaction level (Figure 1). We divide the drugs into three groups. The 1st group of 150 drugs has  $AUC_{drug}$  for  $DSS > 0.96$ . The  $AUC_{drug}$  for DPS and PSS is virtually always lower than  $AUC_{drug}$  for DSS for this group. The 2nd group of 195 drugs has  $AUC_{drug}$  for DSS between 0.96 and 0.86. The values of  $AUC_{drug}$  for PSS are almost always lower than DSS, while  $AUC_{drug}$  for DPS and DSS are similar for these 195 drugs. The 3rd group of the remaining 104 drugs has  $AUC_{drug}$  values for DSS that are on average lower than DPS and PSS. Although DSS and DPS are much better than PSS for the 1st two drug groups, the  $AUC_{drug}$  for PSS (0.752) for the 3rd group of drugs is on average higher than the  $AUC_{drug}$  for DSS and DPS (0.674 and 0.709). This reveals that although overall PSS provides lower predictive performance, it outperforms the other two measures of similarity for 60/449 = 13% of drugs.

Table 4 provides additional insights into the three drug groups. It summarizes differences in the DSS and DPS values between drugs in the same versus other drug groups. These results reveal that the drugs in the 1st and 2nd groups are similar to each other within the groups, both in terms of their structure ( $DSS = 0.66$ ) and profile ( $DPS = 0.21$  and  $0.20$ ), when compared with the drugs in the 3rd group that are substantially more diverse ( $DSS = 0.56$  and  $DPS = 0.16$ ). The similarities between groups 1 and 2 are also relatively high ( $DSS \geq 0.63$  and  $DPS \geq 0.20$ ) when contrasted with the similarities between group 3 and the other two groups ( $DSS$  between 0.55 and 0.60;  $DPS$  between 0.15 and 0.18). Altogether, these results suggest



**Figure 3.** Comparison of predictive performance for the considered DPI predictors at the drug level. Black and gray lines at the top of the figure show the AUC measured per drug ( $AUC_{drug}$ ) for each of the three 1S models. Drugs are sorted by their AUC produced by the most accurate DSS model. The four sets of bar charts at the bottom of the figure show the difference in drug-level AUC between each ensemble model and DSS. Green bar (red bar) means improvement (decline) in the  $AUC_{drug}$  values.

**Table 4.** Analysis of DSS and DPS values for the three drug groups

Similarity type	Assigned drug group	Average similarity to the nearest drug from group		
		1	2	3
DSS	1	0.66	0.65	0.55
	2	0.63	0.66	0.57
	3	0.57	0.60	0.56
DPS	1	0.21	0.21	0.15
	2	0.20	0.20	0.16
	3	0.16	0.18	0.16

that stronger predictive performance for the drugs in the 1st and 2nd groups likely stems from the high levels of similarities between these drugs, while the drugs in the 3rd group suffer lower predictive performance due to larger diversity of their structures and profiles.

The interaction-level AUC of DSS across the 449 drugs is 0.883 (Figure 1). DSS achieves  $AUC_{drug} > 0.883$  for 324/449 = 72% of drugs. Moreover,  $AUC_{drug}$  for either DPS or PSS  $> AUC_{drug}$  for DSS for 100 out of the 125 drugs for which  $AUC_{drug}$  for DSS  $< 0.883$ . These drugs should be of particular focus for the designers of future DPI predictors since they should be predicted with the help of the DPS and PSS rather than the most accurate at the interaction level DSS. This reflects trade-offs between the three measures of similarity and motivates the development of predictors that use consensus of multiple similarities.

We also compare the drug-level performance for the three 2S and the one 3S predictor with the most accurate 1S predictor, DSS. The lower half of Figure 3 shows the difference in  $AUC_{drug}$

between DSS and each of the four ensemble predictors. Overall, the ensemble models are more accurate than DSS for most of the drugs (green bars in Figure 3). Specifically, DSS is outperformed by DSS + DPS, DSS + PSS, DPS + PSS and 3S model for 370, 222, 244 and 349 drugs, respectively. We measure relative difference in  $AUC_{drug}$ , which is defined as the absolute difference in  $AUC_{drug}$  divided by the  $AUC_{drug}$  for DSS. The average relative differences in  $AUC_{drug}$  over the 449 drugs between DSS and the four ensembles are 2% (when compared to DSS + DPS), 5% (DSS + PSS), 7% (DPS + PSS) and 8% (3S). Thus, as expected based on the interaction-level results, the 3S model is also the most accurate predictor at the drug level. The drug-level results for the 2S models are also consistent with the evaluations at the interaction level. DPS + PSS is the most accurate 2S predictor that secures 7% improvement over the DSS model. In particular, it benefits from PSS for the drugs in the 3rd drug group.  $AUC_{drug}$  for DSS and DPS are correlated for the 3rd drug group ( $SCC = 0.58$ ). On the other hand,  $AUC_{drug}$  for PSS is not correlated with either DSS or DPS ( $SCC = 0.07$  and  $0.05$ ) and it secures relatively high  $AUC_{drug}$  values. This implies that PSS is a strong predictive input that complements the other two similarities, which explains why the ensemble predictors outperform DSS for the 3rd group of drugs.

Our novel drug-level assessment of predictive performance shows that no single similarity could provide accurate DPI predictions across all drugs. DSS is the most accurate 1S model and the primary driver of the ensemble-based predictors for the 1st two groups of drugs. For 23% of drugs (the 3rd drug group) for which DSS does not provide higher predictive quality, the use of the other two similarities substantially increases the predictive performance. The 1S models are strongly dependent on the assumption that similar drugs share the same targets

and similar proteins tend to interact with the same drug. The complementarity between PSS and DSS/DPS allows us to accurately cover a wider range of drugs for which only some of these similarities work individually.

### Sensitivity of predictive performance to characteristics of predictors

The predictive performance can be influenced by two factors: (1) intrinsic characteristics of the similarity-based predictors of DPIs, which are defined based on similarity of the query drug or its known targets to the benchmark database; and (2) extrinsic characteristics that are defined solely based on properties of the query drug and its known targets. For instance, if the query drug is similar to many of the drugs from the benchmark database then we anticipate it might be more likely to accurately find its protein targets when compared to a drug that is dissimilar to the drugs in the database. Also, a query drug that has a large number of known targets might be easier to predict, irrespective of the predictive system used. This is because its targets can be exploited to find more potential new targets from the database when compared to a query drug with no or few known targets. Some prior works have analyzed sensitivity of predictive performance to the intrinsic and extrinsic factors. For instance, two articles that use 2S model (DSS and PSS) for the prediction of DPIs have assessed impact of the number of binding drugs per target (extrinsic characteristic) and the DSS values (intrinsic characteristic) on the predictive quality [186, 187]. We provide a more comprehensive assessment of the sensitivity of predictive quality to a set of practical intrinsic and extrinsic characteristics of predictors. We evaluate 1S, 2S and 3S predictors and we cover a wide range of extrinsic characteristics for each of the three types of similarities.

#### Sensitivity to intrinsic characteristics

Similarity-based DPI predictors rely on finding drugs that are similar to a query drug and proteins that are similar to the known targets of the query drug. The quality of DPI predictions depends on the values of the similarities between query drug/protein and the benchmark database, which in turn are dependent on the size and coverage of the database. In principle, the benchmark database is incomplete as it only covers the currently known drug structures, drug profiles and druggable human sequences. In a typical practical scenario, a query drug might be dissimilar to all the drugs in the benchmark database. This reduces the chance that a DSS- or DPS-based predictor successfully identifies potential DPIs for the query drug, even if the query drug actually shares targets with the drugs in the benchmark database. We simulate such scenarios by excluding drugs (proteins) that share certain levels of similarity to the query drug (target) from the benchmark database when making predictions for the query drug (target). We partition the drugs and proteins from the benchmark databases into clusters using average-linkage hierarchical clustering [188]. As a result, the average similarity of drugs or proteins between any two clusters is lower than a predefined similarity threshold used for the clustering. In other words, the drugs (proteins) from one cluster are dissimilar to drugs (proteins) from another cluster. Subsequently, given a query drug (protein), we exclude the drugs (proteins) that belong to its cluster and use the only remaining drugs (proteins) to compute DSS/DPS/PSS for prediction. Thus, the predictions are based on drugs and proteins that are dissim-

ilar to the inputs. Specifically, we use the 80<sup>th</sup>, 50<sup>th</sup> (median) and 20<sup>th</sup> percentile of similarity values over all the drug–protein pairs in the benchmark database as the clustering thresholds. We also consider a scenario where all drugs and proteins are utilized for the predictions except for the query drug or protein itself. This scenario is equivalent to the clustering using the 100<sup>th</sup> (maximal) percentile of similarity values. Additionally, random predictions that correspond to the 0<sup>th</sup> percentile are also considered. We use these five scenarios to study the sensitivity of the predictive performance to the similarities between the query drug (protein) and the benchmark database. The predictive performance is quantified with the interaction-level AUCs.

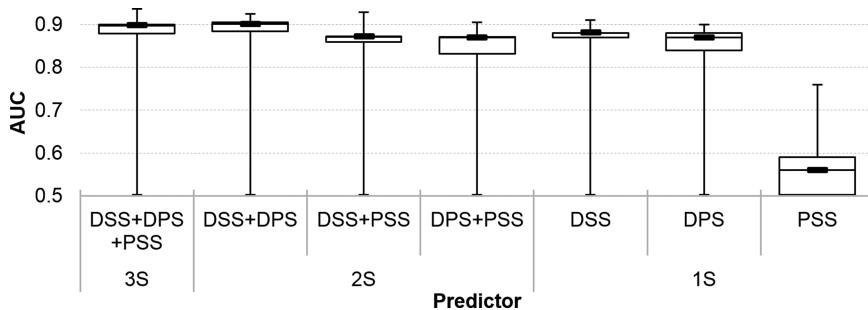
Supplementary Figure S2 shows the ROC curves for the five scenarios and the three 1S, three 2S and one 3S model. Figure 4 summarizes the corresponding AUC values. In general, all but one model are not sensitive to the three intrinsic characteristics. The range of AUC values between 100<sup>th</sup> and 20<sup>th</sup> percentiles of similarities for the 3S, the three 2S, DSS and DPS models is narrow. This means that predictive quality drops by only between 0.04 (DSS + DPS and DSS) and 0.07 (DSS + PSS and DPS + PSS) with the decrease in similarities by 80%. Consequently, the six models offer high predictive quality even when dealing with the input drugs that have low similarity to the benchmark database. The one exception is the PSS model for which the predictive performance is sensitive to the degree of similarity to the targets in the database. Its AUC value substantially drops from 0.76 (100<sup>th</sup> percentile; clustering threshold for PSS = 0.4 which corresponds to 40% pairwise sequence similarity) to 0.56 (50<sup>th</sup> percentile; threshold = 19% pairwise sequence similarity). The results for the 20<sup>th</sup> percentile of PSS (threshold = 10% pairwise sequence similarity) are equivalent to random predictions. This is consistent with previous studies that unveiled that protein sequences with identity <0.25 are likely to have different functions and structures [189, 190]. This is also related to our observations in the Supplementary Figure S1 where the distributions of the PSS values overlap between interacting and non-interacting drug–protein pairs, in particular when compared to DSS and DPS.

This analysis also confirms that use of a larger set of similarities in general results in an improved predictive performance. In particular, for the 100<sup>th</sup> percentile scenario, the AUC of the best 1S model is 0.91, the best 2S model is 0.93 and the 3S model is 0.94. Because the adoption of low PSS similarities results in lower predictive performance, combining it with DSS or DPS into a 2S model does not enhance predictive performance (compare DSS + PSS versus DSS, and DPS + PSS versus DPS plots in Figure 4).

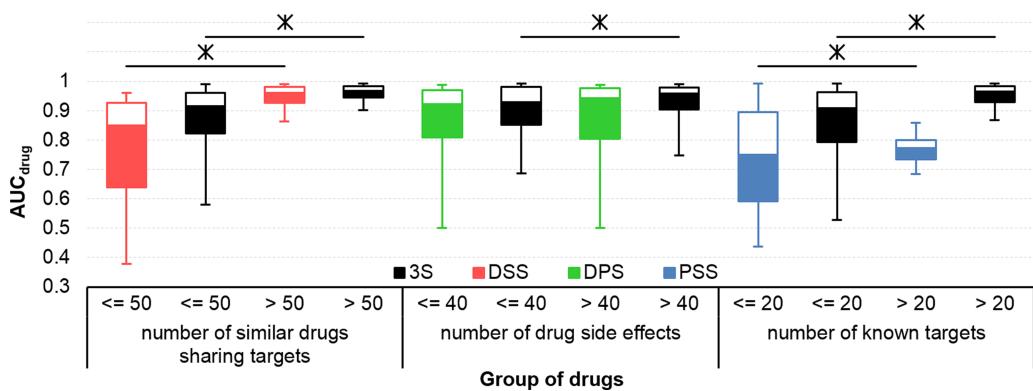
The results of analysis of sensitivity to the intrinsic characteristics of predictors reveal that the predictive performance is not sensitive to the values of DSS and DPS, while it is sensitive to the values of PSS. Moreover, combining PSS with DSS and/or DPS makes the predictors robust to the low values of PSS. We conclude that use of PSS only model is risky when the known targets of query drug are dissimilar to the targets in the benchmark database.

#### Sensitivity to extrinsic characteristics

Predictive performance could also be sensitive to extrinsic characteristics that describe known *a priori* input drug structure, drug profile and protein target. We investigate a wide range of extrinsic characteristics for each of the three types of similarities. They include surface area, charge, shape, weight, size, lipophilicity, number of similar drugs, number of similar drugs sharing targets and similarity to the most similar drugs that are computed from the drug structure. We also considered three



**Figure 4.** Sensitivity of predictive performance to the intrinsic characteristics of predictors. The boxplots denote the AUC values for the scenarios where drugs (proteins) similar to the query drug (protein) are excluded from predictions based on clustering using the 100<sup>th</sup>, 80<sup>th</sup>, 50<sup>th</sup> and 20<sup>th</sup> percentile of the similarities. The bottom bar denotes random predictions that correspond to the 0<sup>th</sup> percentile (AUC = 0.5).



**Figure 5.** Sensitivity of predictive quality to extrinsic characteristics of predictors. The boxplots denote the 5th, 20th, 50th, 80th and 95th percentiles of AUC<sub>drug</sub> values for the 1S and 3S model over the corresponding two groups of drugs separated with help of a given marker that characterizes the input drug/targets. The asterisk indicates that the difference in the AUC<sub>drug</sub> is significant ( $P$ -value  $< 0.05$ ) between drug groups. We assess significance of differences between the corresponding two groups of drugs using the Wilcoxon rank sum test; we use the non-parametric test since these data are not normal according to the Anderson-Darling test at 0.01 significance.

markers of drug profiles: number of side effects, similarity of the profile to the most similar profile and number of similar profiles. Finally, we investigated three markers of the input drug targets: number of known targets, number of known targets that interact with similar drugs and known targets that interact with the most similar drug. We discuss one ‘best’ marker for each type of similarity, which we selected based on the fact that it significantly affects the predictive performance and is easy to quantify. These markers include the number of similar drug structures that share targets with the input drug, the number of side effects and the number of known targets.

We study sensitivity of predictive quality to these extrinsic markers by partitioning the considered drugs into two groups for which we observe large differences in predictive performance. For the best DSS marker, we compare 177 drugs that share targets with <50 similar drugs (which corresponds to the median level of DSS) versus the remaining drugs that share targets with a larger number of similar drugs. Similarly, for the best DPS characteristic, we contrast 55 drugs that have relatively small side-effects profile (<40) versus the remaining drugs that have a larger profile. Finally, for the best PSS marker, we compare 128 drugs with <20 native targets versus the remaining more promiscuous drugs. Figure 5 compares distributions of AUC<sub>drug</sub> values between these groups of drugs when considering the corresponding 1S models and the 3S model.

Figure 5 shows that the AUC<sub>drug</sub> for DSS varies substantially between 0.38 and 0.96 with a median = 0.85 for the drugs that share targets with fewer similar drugs. For the remaining drugs

that share targets with the larger number of similar drugs, the AUC<sub>drug</sub> for DSS is always >0.86 with a median = 0.96. These drugs have lower variation and significantly higher values for AUC<sub>drug</sub> ( $P$ -value  $< 0.05$ ) when compared to the drug that shares targets with fewer similar drugs. Similarly, the 3S model also has significantly higher AUC<sub>drug</sub> values with narrower spread for the drugs with larger number of similar drugs sharing targets. These observations reveal that the predictive quality is significantly higher when working with drugs that share targets with a larger number of similar drugs in the benchmark database. Based on the assertion that similar drugs tend to target the same proteins, that makes it easier to accurately identify targets for such drugs when using DSS. In contrast, the drugs that share targets with a low number of similar drugs are harder to accurately predict.

The AUC<sub>drug</sub> for DPS is lower for the drugs that have a smaller profile (median AUC<sub>drug</sub> = 0.92) than for the drugs that have larger profile (median AUC<sub>drug</sub> = 0.94), but this difference is not statistically significant ( $P$ -value = 0.17). The corresponding difference for the 3S model is slightly larger and statistically significant ( $P$ -value  $< 0.05$ ): median AUC<sub>drug</sub> = 0.93 versus 0.96, respectively. Interestingly, while the difference is not significant when DPS is used individually, use of larger profiles is expected to provide significantly higher predictive performance when DPS is combined with PSS and DSS. Since DPS and PSS are not correlated ( $SCC = 0.05$ ), while DPS and DSS are modestly correlated ( $SCC = 0.30$ ), this suggests that relationship between DPS and DSS is exploited by the ensemble model to further benefit from the increased profile size.

The differences in AUC<sub>drug</sub> between the less versus more promiscuous drugs are significant for the PSS and 3S models ( $P$ -value < 0.05). The median AUC<sub>drug</sub> for the 3S model is 0.91 for the less promiscuous drugs while it equals 0.96 for the more promiscuous drugs. Similarly, for the 1S model that relies on PSS the median AUC<sub>drug</sub> = 0.75 versus 0.77. This trend could be explained by an observation that the drugs that interact with many targets are more likely to have targets that are highly similar to the proteins in the benchmark database. This is supported by the assertion that similar proteins tend to interact with the same drugs.

The results of the analysis of sensitivity to the extrinsic characteristics reveal that the predictive performance is sensitive to several easy to estimate characteristics of the input drug structures, drug profiles and drug targets. As expected, we find that it is more likely to secure higher predictive quality when a more comprehensive information is available for the input drugs, such as a larger number of similar drugs sharing targets, a larger drug profile and a higher number of known targets.

## CONNECTOR webserver

PredicToR of compouNd-proteiN intEraCTiOn based on ensemble of similaRities (CONNECTOR) is a webserver that implements the seven predictors that we evaluate in this review. The webserver and the benchmark database are freely available at <http://biomine.cs.vcu.edu/servers/CONNECTOR/>.

CONNECTOR facilitates prediction for any combination of the three inputs including any individual input (drug structure, drug profile and protein sequence), any pair of inputs and all three inputs. CONNECTOR uses the benchmark dataset as its internal database. The computations are performed on the server side and thus the end user only needs to access the webserver page via any major web browser and provide the inputs. The prediction is performed in three steps: (1) provide any combination of the three inputs including drug structure in the SMILES format, comma separated list of side effects (drug profile) and/or protein sequences in the FASTA format; (2) provide email address where a unique link to the results will be sent once the predictions are ready; and (3) click 'Run CONNECTOR'. Each submitted job is entered into a FIFO queue of jobs on the server. User is notified via the browser window about the current placement in the queue and when the job reaches the front of the queue. The results are displayed in the browser window and the unique link to the page with the results is sent to the user-identified email address. We provide four views of the results:

- All DPIs, which shows all putative drug–protein pairs that include query drug and target(s) and similar compounds and proteins.
- DPIs that include the query drug structure, which cover interactions between the query drug structure and its putative protein targets (if drug structure was provided as one of the inputs).
- DPIs that include the query drug profile, which include interactions between the query drug represented by its profile and its putative protein targets (if drug profile was provided as one of the inputs).
- DPIs that include the query protein sequence, which gives interactions between the query target sequence and the drugs that are predicted to interact with this target (if target sequence was provided as one of the inputs).

Each putative interaction is coupled with the corresponding value of similarity. Drugs are linked to their entries in the Pub-

Chem databases and target sequences are linked to their UniProt entries. We also provide a text-based version of the results for download. We will store the page with the results for at least 3 months.

## Summary and conclusions

Computational prediction of DPIs is critical to facilitate pharmacological research and promote our understanding of therapeutic mechanisms and side effects of drugs. We survey 35 high-impact similarity-based DPI predictors including eight recent methods that were published since 2016. We classify these methods according to the similarities they exploit. We summarize their source databases, provide details about their internal databases and predictive models and perform first-of-its-kind empirical comparative analysis of predictive performance for a representative set of seven methods.

We categorize these methods based on the type and number of similarities that they use. The predictors rely on three types of similarities: DSS, DPS and PSS, which can be used either individually (1S models) or as ensembles (2S and 3S models). Most of these predictors utilize 2S and 3S ensembles, with 20 2S and 5 3S ensembles out of the 35 considered tools. Only two 2S ensembles include DPS, likely because collection of these data and calculation of this similarity are challenging [114, 125]. We find that most DPI predictors are highly cited and were published in high impact venues. We show that the three pioneering methods [112–114] that were published in 2007 and 2008 already received over 2200 citations, which is more than all the subsequent methods combined together. However, majority of the 35 considered predictors also received high citation counts that exceed the impact factors of the journals where they were published. One of the key defining factors of the similarity-based methods is the size and scope of their internal databases. While some internal databases have a large number of drugs or targets, we found that in general they provide a relatively sparse mapping of their interactions. The density of DPIs in these databases varies between 1 and 10.

We have developed a novel benchmark database that includes drugs characterized by high density of interactions and we use it to evaluate a representative set of 1S, 2S and 3S predictors. This database integrates data coming from three resources (19 source databases) and includes 449 drugs, 1469 druggable human proteins and 34 456 native DPIs, resulting in high density of native DPIs per drug that equals 76. We observe that most of the surveyed methods have evaluated predictive performance on both DPIs and non-DPIs, which allowed for computation of TPR, FPR and the resulting AUC values. However, these methods accessed predictive performance only at the interaction level that spans multiple drugs that may differ in structure, profile and the number and type of protein partners that they bind to. We are the first to perform both interaction- and drug-level assessment and to comprehensively and side-by-side compare all 1S, 2S and 3S models.

At the interaction level, we reveal that the use of ensembles results in improved predictive quality. Specifically, the 3S model significantly outperforms the 1S and 2S models, while the 2S models are significantly more accurate than the 1S models. We empirically analyze reason for these improvements and we show that they stem from an improved handling of interactions characterized by modest levels of similarities. Among the three 1S models, DSS and DPS are more accurate than PSS. We find a modest level of correlation between DSS and DPS while correlations between PSS and DSS as well as PSS and DPS are low.

The modest or low levels of correlations further explain why the ensembles benefit from the integration of multiple similarities and why the majority of current predictors are ensembles. We also compare four algorithms that can be used to develop the 3S ensembles. Among these algorithms, the linear regression offers the most balanced and slightly more accurate predictions, while decision trees should be used when focusing on the predictions with low FPRs.

At the drug level, we demonstrate that the predictive quality varies widely between drugs and these differences depend on the type of similarity used. We show that no single 1S model could provide accurate DPI predictions across all the drugs. Moreover, we demonstrate that ensembles improve over the 1S models. The average relative improvements in  $AUC_{\text{drug}}$  over the 449 drugs between the best 1S model, DSS and the four ensembles are 2% (compared to DSS + DPS), 5% (DSS + PSS), 7% (DPS + PSS) and 8% (3S). We divide drugs into three groups: group one includes 150 drugs with high  $AUC_{\text{drug}}$  for DSS, group two includes 195 drugs with moderate  $AUC_{\text{drug}}$  for DSS and group three with 104 drugs for which  $AUC_{\text{drug}}$  for DSS is on average lower than DPS and PSS. We found that the use of 3S ensemble provides particularly large improvements by on average 29% in  $AUC_{\text{drug}}$  over the use of DSS for the 23% of drugs that constitute the 3rd group. Our empirical analysis also reveals that although per-interaction PSS provides the lowest predictive performance, it outperforms the other two 1S models for 13% of drugs when assessed at the drug level. These drugs should be of particular focus for the development of future DPI predictors because more accurate predictions could be obtained with the help of the DPS and PSS rather than with the overall most accurate at the interaction-level DSS that dominates the current methods.

DPI prediction relies on accurate values of similarities between the query drug (its known target) and the drugs (proteins) in the benchmark database. The quality of these values is likely affected by the intrinsic characteristics of predictors, such as the similarity between inputs and the benchmark database. We provide comprehensive sensitivity analysis of predictive quality for each of the three similarities. We find that the predictive quality is not sensitive to the values of DSS and DPS, which means that the predictive performance of the methods that utilize DSS/DPS does not substantially change when the input drug structures or profiles are dissimilar to the structures and profiles in the internal database. In contrast, the predictive quality for PSS drops significantly when the input protein sequences are dissimilar to the sequences in the database. However, such sensitivity to low values of PSS is mitigated when using methods that combine PSS with other similarities. These observations demonstrate that the ensembles are not only accurate at the interaction and drug levels but they are also robust to low values of similarities.

We also investigate sensitivity of predictive quality to extrinsic characteristics of the predictive models. These extrinsic characteristics are intrinsic to the inputs and include information extracted directly from drug structures, drug profiles and drug targets. We find that the 3S model and the corresponding 1S models are more likely to provide accurate predictions when the input drugs share targets with a larger number of similar drugs in the benchmark database and when the input drugs are more promiscuous. The 3S model also benefits from scenarios when the input drugs have larger side-effect profiles. This suggest that a larger amount of available data (number of known targets and/or side effects) leads to calculations of higher quality and higher values of similarities, which in turn results in higher predictive performance.

## Key Points

- Drug structure, drug profile and protein sequence similarity are the main types of similarity measures that are used for DPI prediction.
- A benchmark database with high density of DPIs is developed. This database is particularly suitable for comparative empirical assessment of DPI predictors.
- Ensembles of multiple similarities offer higher predictive quality than models that rely on single or fewer similarities. This result is consistent for evaluations at both interaction and drug level.
- A novel drug-level assessment of predictive performance provides additional insights. It shows that different combination of similarities should be used for different drugs.
- A comprehensive sensitivity analysis of predictive performance to the intrinsic and extrinsic characteristics of DPI predictors reveals sensitivity to lower values PSS and to several extrinsic properties of the input drug structures, profiles and targets.

## Supplementary Data

Supplementary Data are available online <https://academic.oup.com/bib>.

## Funding

This work was supported in part by the Qimonda Endowed Chair position to L. K.

## References

1. Rask-Andersen M, Almén MS, Schiöth HB. Trends in the exploitation of novel drug targets. *Nat Rev Drug Discov* 2011;10:579–90.
2. Rask-Andersen M, Masuram S, Schiøth HB. The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Ann Rev Pharmacol Toxicol* 2014;54:9–26.
3. Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* 2017;16:19–34.
4. Schneider G. Virtual screening: an endless staircase? *Nat Rev Drug Discov* 2010;9:273–6.
5. Núñez S, Venhorst J, Kruse CG. Target–drug interactions: first principles and their application to drug discovery. *Drug Discov Today* 2012;17:10–22.
6. Dalkas GA, Vlachakis D, Tsagkrasoulis D, et al. State-of-the-art technology in modern computer-aided drug design. *Brief Bioinform* 2013;14:745–52.
7. Tseng CY, Tuszyński J. A unified approach to computational drug discovery. *Drug Discov Today* 2015;20:1328–36.
8. Chong CR, Sullivan DJ. New uses for old drugs. *Nature* 2007;448:645–6.
9. Haupt VJ, Schroeder M. Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief Bioinform* 2011;12:312–26.
10. Oprea TI, Mestres J. Drug repurposing: far beyond new targets for old drugs. *AAPS J* 2012;14:759–63.
11. Hu Y, Bajorath J. Compound promiscuity: what can we learn from current data? *Drug Discov Today* 2013;18:644–50.

12. Li J, Zheng S, Chen B, et al. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;17:2–12.
13. Lounkine E, Keiser MJ, Whitebread S, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 2012;486:361–7.
14. Wang J, Li Z-X, Qiu C-X, et al. The relationship between rational drug design and drug side effects. *Brief Bioinform* 2012;13:377–82.
15. Kuhn M, Al Banchaabouchi M, Campillos M, et al. Systematic identification of proteins that elicit drug side effects. *Mol Syst Biol* 2013;9:663.
16. Tarczay Á, Keserú GM. Contributions of molecular properties to drug promiscuity. *J Med Chem* 2013;56:1789–95.
17. Hu G, Wang K, Groenendyk J, et al. Human structural proteome-wide characterization of Cyclosporine A targets. *Bioinformatics* 2014;30:3561–6.
18. Cimermancic P, Weinkam P, Rettenmaier TJ, et al. CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J Mol Biol* 2016;428:709–19.
19. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov* 2002;1:727–30.
20. Russ AP, Lampel S. The druggable genome: an update. *Drug Discov Today* 2005;10:1607–10.
21. Hu G, Wu Z, Wang K, et al. Untapped potential of disordered proteins in current druggable human proteome. *Curr Drug Targets* 2016;17:1198–205.
22. Schuffenhauer A, Zimmermann J, Stoop R, et al. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J Chem Inf Comput Sci* 2002;42:947–55.
23. Southan C, Várkonyi P, Muresan S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J Cheminform* 2009;1:10.
24. Roth BL, Lopez E, Patel S, et al. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* 2000;6:252–62.
25. Schomburg I, Hofmann O, Baensch C, et al. Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene Funct Dis* 2000;1:109–18.
26. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* 2002;30:47–9.
27. Schomburg I, Chang A, Ebeling C, et al. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004;32:D431–3.
28. Barthelmes J, Ebeling C, Chang A, et al. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* 2007;35:D511–4.
29. Chang A, Scheer M, Grote A, et al. BRENDA, AMENDA and FRENDA: the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* 2008;37:D588–92.
30. Scheer M, Grote A, Chang A, et al. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* 2010;39:D670–6.
31. Schomburg I, Chang A, Placzek S, et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res* 2012;41:D764–72.
32. Chang A, Schomburg I, Placzek S, et al. BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res* 2014;43:D439–46.
33. Placzek S, Schomburg I, Chang A, et al. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res* 2017;45:D380–8.
34. Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen* 2001;4:719–25.
35. Chen X, Lin Y, Liu M, et al. The Binding Database: data management and interface design. *Bioinformatics* 2002;18:130–9.
36. Chen X, Lin Y, Gilson MK. The binding database: overview and user's guide. *Biopolymers* 2002;61:127–41.
37. Liu T, Lin Y, Wen X, et al. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2007;35:D198–201.
38. Gilson MK, Liu T, Baitaluk M, et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 2016;44:D1045–53.
39. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res* 2002;30:412–5.
40. Zhu F, Han B, Kumar P, et al. Update of TTD: therapeutic target database. *Nucleic Acids Res* 2010;38:D787–91.
41. Zhu F, Shi Z, Qin C, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;40:D1128–36.
42. Qin C, Zhang C, Zhu F, et al. Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res* 2013;42:D1118–23.
43. Yang H, Qin C, Li YH, et al. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;44:D1069–74.
44. Li YH, Yu CY, Li XX, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 2018;46:D1121–7.
45. Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;34:D354–7.
46. Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2008;36:D480–4.
47. Kanehisa M, Goto S, Furumichi M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2009;38:D355–60.
48. Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2011;40:D109–14.
49. Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2013;42:D199–205.
50. Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;44:D457–62.
51. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353–61.
52. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;34:D668–72.
53. Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledge-base for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;36:D901–6.
54. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res* 2011;39:D1035–41.

55. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;42:D1091–7.
56. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;46:D1074–82.
57. Okuno Y, Yang J, Taneishi K, et al. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res* 2006;34:D673–7.
58. Okuno Y, Tamon A, Yabuuchi H, et al. GLIDA: GPCR—ligand database for chemical genomics drug discovery—database and tools update. *Nucleic Acids Res* 2008;36:D907–12.
59. Günther S, Kuhn M, Dunkel M, et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2008;36:D919–22.
60. Hecker N, Ahmed J, von Eichborn J, et al. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res* 2012;40:D1113–D1117.
61. Kuhn M, von Mering C, Campillos M, et al. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res* 2008;36:D684–D688.
62. Kuhn M, Szklarczyk D, Franceschini A, et al. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res* 2010;38:D552–6.
63. Kuhn M, Szklarczyk D, Franceschini A, et al. STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Res* 2012;40:D876–80.
64. Kuhn M, Szklarczyk D, Pletscher-Frankild S, et al. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Res* 2014;42:D401–7.
65. Szklarczyk D, Santos A, von Mering C, et al. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 2016;44:D380–D384.
66. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;40:D1100–7.
67. Bento AP, Gaulton A, Hersey A, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 2014;42:D1083–90.
68. Davies M, Nowotka M, Papadatos G, et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* 2015;43:W612–20.
69. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res* 2017;45:D945–54.
70. Southan C, Sitzmann M, Muresan S. Comparing the chemical structure and protein content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. *Mol Inform* 2013;32:881–97.
71. Hughes JP, Rees S, Kalindjian SB, et al. Principles of early drug discovery. *Brit J Pharmacol* 2011;162:1239–49.
72. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. *Clin Pharmacol Therapeut* 2016;99:285–97.
73. Pessetto ZY, Chen B, Alturkmani H, et al. In silico and in vitro drug screening identifies new therapeutic approaches for Ewing sarcoma. *Oncotarget* 2017;8:4079–95.
74. Gao Z, Li H, Zhang H, et al. PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics* 2008;9:104.
75. Li L, Bum-Erdene K, Baenziger PH, et al. BioDrugScreen: a computational drug design resource for ranking molecules docked to the human proteome. *Nucleic Acids Res* 2010;38:D765–73.
76. Wang C, Hu G, Wang K, et al. PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. *Bioinformatics* 2016;32:579–86.
77. Xie L, Li J, Xie L, et al. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput Biol* 2009;5:e1000387.
78. Xie L, Evangelidis T, Xie L, et al. Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput Biol* 2011;7:e1002037.
79. Ding H, Takigawa I, Mamitsuka H, et al. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief Bioinform* 2014;15:734–47.
80. Schomburg KT, Rarey M. What is the potential of structure-based target prediction methods? *Future Med Chem* 2014;6:1987–9.
81. Somody JC, MacKinnon SS, Windemuth A. Structural coverage of the proteome for pharmaceutical applications. *Drug Discov Today* 2017.
82. Mizianty MJ, Fan X, Yan J, et al. Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr D Biol Crystallogr* 2014;70:2781–93.
83. Liu T, Altman RB. Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. *J Chem Inf Model* 2015;55:1483–94.
84. Zhang QC, Petrey D, Deng L, et al. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 2012;490:556.
85. Pentony MM, Jones DT. Modularity of intrinsic disorder in the human proteome. *Proteins* 2010;78:212–21.
86. Peng Z, Yan J, Fan X, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci* 2015;72:137–51.
87. Mitchell JB. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. *J Chem Inf Comput Sci* 2001;41:1617–22.
88. Schuffenhauer A, Floersheim P, Acklin P, et al. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci* 2003;43:391–405.
89. Klabunde T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Brit J Pharmacol* 2007;152:5–7.
90. Pahikkala T, Airola A, Pietila S, et al. Toward more realistic drug-target interaction predictions. *Brief Bioinform* 2015;16:325–37.
91. Mousavian Z, Masoudi-Nejad A. Drug-target interaction prediction via chemogenomic space: learning-based methods. *Expert Opin Drug Metab Toxicol* 2014;10:1273–87.
92. Chen X, Yan CC, Zhang X, et al. Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2016;17:696–712.
93. Cichonska A, Rousu J, Aittokallio T. Identification of drug candidates and repurposing opportunities through compound-target interaction networks. *Expert Opin Drug Discov* 2015;10:1333–45.
94. Lavecchia A, Gerchia C. In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov Today* 2016;21:288–98.
95. Hart T, Xie L. Providing data science support for systems pharmacology and its implications to drug discovery. *Expert Opin Drug Discov* 2016;11:241–56.

96. Lotfi Shahreza M, Ghadiri N, Mousavi SR, et al. A review of network-based approaches to drug repositioning. *Brief Bioinform* 2017;bbx017–7.
97. Fang J, Liu C, Wang Q, et al. In silico polypharmacology of natural products. *Brief Bioinform* 2017;bbx045–5.
98. Ezzat A, Wu M, Li X-L, et al. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform* 2018;bbx002–2.
99. Hao M, Bryant SH, Wang Y. Open-source chemogenomic data-driven algorithms for predicting drug–target interactions. *Brief Bioinform* 2018;bbx010–0.
100. Vilar S, Hripcsak G. The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug–drug interactions. *Brief Bioinform* 2017;18:670–81.
101. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 2015;20:318–31.
102. Jamali AA, Ferdousi R, Razzaghi S, et al. DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discov Today* 2016;21:718–24.
103. Wen M, Zhang Z, Niu S, et al. Deep-learning-based drug–target interaction prediction. *J Proteome Res* 2017;16:1401–9.
104. Peón A, Naulaerts S, Ballester PJ. Predicting the reliability of drug–target interaction predictions with maximum coverage of target space. *Sci Rep* 2017;7:3820.
105. Peng L, Zhu W, Liao B, et al. Screening drug–target interactions with positive–unlabeled learning. *Sci Rep* 2017;7:8087.
106. Li Z, Han P, You Z-H, et al. In silico prediction of drug–target interaction networks based on drug chemical structure and protein sequences. *Sci Rep* 2017;7:11174.
107. Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017;8:573.
108. Fang J, Wu Z, Cai C, et al. Quantitative and systems pharmacology. 1. In silico prediction of drug–target interactions of natural products enables new targeted cancer therapy. *J Chem Inf Model* 2017;57:2657–71.
109. Rayhan F, Ahmed S, Shatabda S, et al. iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. *Sci Rep* 2017;7:17731.
110. Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2017;45:D12–7.
111. 2017 Journal Citation Reports®. Clarivate Analytics.
112. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;25:197–206.
113. Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;24:i232–40.
114. Campillos M, Kuhn M, Gavin AC, et al. Drug target identification using side-effect similarity. *Science* 2008;321:263–6.
115. Nagamine N, Shirakawa T, Minato Y, et al. Integrating statistical predictions and experimental verifications for enhancing protein–chemical interaction predictions in virtual screening. *PLoS Comput Biol* 2009;5:e1000397.
116. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009;25:2397–403.
117. Yamanishi Y, Kotera M, Kanehisa M, et al. Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;26:i246–54.
118. Yabuuchi H, Niijima S, Takematsu H, et al. Analysis of multiple compound–protein interactions reveals novel bioactive molecules. *Mol Syst Biol* 2011;7:472.
119. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 2011;27:3036–3043.
120. Cheng F, Liu C, Jiang J, et al. Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012;8:e1002503.
121. Gonen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 2012;28:2304–10.
122. Takarabe M, Kotera M, Nishimura Y, et al. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics* 2012;28:i611–8.
123. Cao D-S, Liu S, Xu Q-S, et al. Large-scale prediction of drug–target interactions using protein sequences and drug topological structures. *Analytica Chimica Acta* 2012;752:1–10.
124. Mei JP, Kwoh CK, Yang P, et al. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 2013;29:238–45.
125. Cheng F, Li W, Wu Z, et al. Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J Chem Inf Model* 2013;53:753–762.
126. Alaimo S, Pulvirenti A, Giugno R, et al. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 2013;29:2004–8.
127. Koutsoukas A, Lowe R, Kalantarmotamedi Y, et al. In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window. *J Chem Inf Model* 2013;53:1957–66.
128. Yamanishi Y, Kotera M, Moriya Y, et al. DINIES: drug–target interaction network inference engine based on supervised analysis. *Nucleic Acids Res* 2014;42:W39–45.
129. Shi J-Y, Yiu S-M, Li Y, et al. Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods* 2015;83:98–104.
130. Liu H, Sun J, Guan J, et al. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 2015;31:i221–9.
131. Seal A, Ahn YY, Wild DJ. Optimizing drug–target interaction prediction based on random walk on heterogeneous networks. *J Cheminform* 2015;7:40.
132. Kuang Q, Xu X, Li R, et al. An eigenvalue transformation technique for predicting drug–target interaction. *Sci Rep* 2015;5:13867.
133. Hao M, Wang Y, Bryant SH. Improved prediction of drug–target interactions using regularized least squares integrating with kernel fusion technique. *Analytica Chimica Acta* 2016;909:41–50.
134. Liu Y, Wu M, Miao C, et al. Neighborhood regularized logistic matrix factorization for drug–target interaction prediction. *PLoS Comput Biol* 2016;12:e1004760.
135. Wu Z, Cheng F, Li J, et al. SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning. *Brief Bioinform* 2016.
136. Ba-Alawi W, Soufan O, Essack M, et al. DASPfind: new efficient method to predict drug–target interactions. *J Cheminform* 2016;8:15.

137. Yuan Q, Gao J, Wu D, et al. DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 2016;32:i18–27.
138. Ezzat A, Wu M, Li X-L, et al. Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods* 2017.
139. Willett P, Barnard JM, Downs GM. Chemical similarity searching. *J Chem Inf Comput Sci* 1998;38:983–96.
140. Bender A, Jenkins JL, Scheiber J, et al. How similar are similarity searching methods?: a principal component analysis of molecular descriptor space. *J Chem Inf Model* 2009;49.
141. Cereto-Massagué A, Ojeda MJ, Valls C, et al. Molecular fingerprint similarity search in virtual screening. *Methods* 2015;71:58–63.
142. Hattori M, Okuno Y, Goto S, et al. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 2003;125:11853–65.
143. Hattori M, Tanaka N, Kanehisa M, et al. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res* 2010;38:W652–6.
144. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–402.
145. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;147:195–7.
146. Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;6:343.
147. Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;44:D1075–9.
148. ATC classification index with DDDs. Oslo, Norway: WHO Collaborating Centre for Drug Statistics Methodology, 2017.
149. Lu L, Yu H. DR2DI: a powerful computational tool for predicting novel drug-disease associations. *J Comput Aided Mol Des* 2018;32:633–42.
150. Gottlieb A, Stein GY, Ruppin E, et al. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;7:496.
151. Todeschini R, Consonni V, Xiang H, et al. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J Chem Inf Model* 2012;52:2884–901.
152. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* 2015;7:20.
153. Steinbeck C, Han Y, Kuhn S, et al. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* 2003;43:493–500.
154. Steinbeck C, Hoppe C, Kuhn S, et al. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des* 2006;12:2111–20.
155. May JW, Steinbeck C. Efficient ring perception for the Chemistry Development Kit. *J Cheminform* 2014;6:3–3.
156. Willighagen EL, Mayfield JW, Alvarsson J, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 2017;9:33.
157. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
158. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;215:403–10.
159. Roider HG, Pavlova N, Kirov I, et al. Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinformatics* 2014;15:68.
160. Harmar AJ, Hills RA, Rosser EM, et al. IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res* 2009;37:D680–5.
161. Sharman JL, Mpamhangwa CP, Spedding M, et al. IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res* 2011;39:D534–8.
162. Sharman JL, Benson HE, Pawson AJ, et al. IUPHAR-DB: updated database content and new features. *Nucleic Acids Res* 2013;41:D1083–8.
163. Pawson AJ, Sharman JL, Benson HE, et al. The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledge-base of drug targets and their ligands. *Nucleic Acids Res* 2014;42:D1098–106.
164. Southan C, Sharman JL, Benson HE, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res* 2016;44:D1054–68.
165. Harding SD, Sharman JL, Faccenda E, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res* 2018;46:D1091–106.
166. Grabowski M, Niedzialkowska E, Zimmerman MD, et al. The impact of structural genomics: the first quindecennial. *J Struct Funct Genomics* 2016;17:1–16.
167. Hastings J, Owen G, Dekker A, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;44:D1214–9.
168. Davis AP, Grondin CJ, Johnson RJ, et al. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* 2017;45:D972–8.
169. Yates B, Braschi B, Gray KA, et al. *Genenames.org*: the HGNC and VGNC resources in 2017. *Nucleic Acids Res* 2017;45:D619–25.
170. Feng Z, Chen L, Maddula H, et al. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* 2004;20:2153–5.
171. Chopra A, Shan L, Eckelman WC, et al. Molecular Imaging and Contrast Agent Database (MICAD): evolution and progress. *Mol Imag Biol* 2012;14:4–13.
172. Brown GR, Hem V, Katz KS, et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res* 2015;43:D36–42.
173. Cerami EG, Gross BE, Demir E, et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011;39:D685–90.
174. de Beer TAP, Berka K, Thornton JM, et al. PDBsum additions. *Nucleic Acids Res* 2014;42:D292–D296.
175. Whirl-Carrillo M, McDonagh EM, Hebert JM, et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Therapeut* 2012;92:414–7.
176. Wang Y, Bryant SH, Cheng T, et al. PubChem BioAssay: 2017 update. *Nucleic Acids Res* 2017;45:D955–63.
177. Kim S, Thiessen PA, Bolton EE, et al. PubChem Substance and Compound databases. *Nucleic Acids Res* 2016;44:D1202–13.

178. Consortium TU. UniProt: the universal protein knowledge-base. *Nucleic Acids Res* 2017;45:D158–69.
179. Glaab E. Building a virtual ligand screening pipeline using free software: a survey. *Brief Bioinform* 2016;17:352–66.
180. Paolini GV, Shapland RHB, van Hoorn WP, et al. Global mapping of pharmacological space. *Nat Biotech* 2006;24:805–815.
181. Meng F, Kurgan L. DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* 2016;32:i341–50.
182. Yan J, Kurgan L. DRNAPred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;45:e84.
183. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform* 2017.
184. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief Bioinform* 2017.
185. Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Brief Bioinform* 2015;16:780–94.
186. Lim H, Gray P, Xie L, et al. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Sci Rep* 2016;6.
187. Lim H, Poleksic A, Yao Y, et al. Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing. *PLoS Comput Biol* 2016;12.
188. Cios KJ, Pedrycz W, Swiniarski RW, et al. *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.
189. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;12:85–94.
190. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective1. *J Mol Biol* 2001;307:1113–43.