Send Orders for Reprints to reprints@benthamscience.ae

Current Medicinal Chemistry, **Year,** *Volume*

Survey of Similarity-based Prediction of Drug-protein Interactions

Chen Wang[a] and Lukasz Kurgan[*a]

[a]*Department of Computer Science, Virginia Commonwealth University, Richmond, VA, 23284, USA*

Abstract: Therapeutic activity of a significant majority of drugs is determined by their interactions with proteins. Databases of drug-protein interactions (DPIs) primarily focus on the therapeutic protein targets while the knowledge of the off-targets is fragmented and partial. One way to bridge this knowledge gap is to employ computational methods to predict protein targets for a given drug molecule, or interacting drugs for given protein targets. We survey a comprehensive set of 35 methods that were published in high-impact venues and that predict DPIs based on similarity between drugs and similarity between protein targets. We analyze the internal databases of known PDIs that these methods utilize to compute similarities, and investigate how they are linked to the 12 publicly available source databases. We discuss contents, impact and relationships between these internal and source databases, and well as the timeline of their releases and publications. The 35 predictors exploit and often combine three types of similarities that consider drug structures, drug profiles, and target sequences. We review the predictive architectures of these methods, their impact, and we explain how their internal DPIs databases are linked to the source databases. We also include a detailed timeline of the development of these predictors and discuss the underlying limitations of the current resources and predictive tools. Finally, we provide several recommendations concerning future development of the related databases and methods.

Keywords: Drug-protein interactions; Drug-protein interaction prediction; Drug repurposing; Drug side-effects; Databases; Drug structure; Protein sequence.

1. INTRODUCTION

Drugs are chemical substances that are used to prevent, treat, and cure diseases. They work through interactions with their biological targets that include proteins, DNA, RNAs, and membrane components such as lipid and carbohydrates [1]. Recent comprehensive analysis has summarized the landscape of drug targets showing that 96% of 893 mechanistic drug targets are mapped to proteins, and these protein targets are responsible for 93% of all identified drug-target interactions [2]. Hence, protein targets play a predominant role in understanding the significant majority of drug-target interactions.

Drug discovery and development requires a significant amount of money and time to address identification of the potential targets, to search for drug leads, to analyze massive quantities of data to select promising leads, to validate leads in a wet-lab, and to perform clinical trials [3]. High-quality identification and validation of drug-protein interactions (DPIs) is a vital prerequisite to investigate new drugs and determine their targets. This is important since drugs may interact not only with the desired therapeutic targets but also with undesired off-targets. The latter may result in adverse events or side-effects, precluding drug development and usage. Chemical screening with cell assays is used to measure drug-protein binding affinity (how tightly a drug lead compound binds to a selected protein target). However, these experiments are limited in scope as they screen against a relatively small panel of protein targets [4, 5]. For example, SafetyScreen44 panel screens against 44 targets recommended by four major pharmaceutical companies including AstraZeneca, GlaxoSmithKline, Novartis, and Pfizer [6]. Novartis uses a panel of 24 targets [7] and Pfizer screens against between 15 and 30 targets [8]. Therefore, there is a clear need to develop high-throughput computational methods for the prediction of DPIs. These methods can screen a given drug against a comprehensive set of thousands of protein targets that make up the human proteome. The computational methods that identify putative DPIs exploit information about both drugs and proteins. Depending on whether the three-dimensional (3D) structure of a protein is used/known, the prediction methods could be grouped into two classes: protein structure-based methods and similarity-based methods that do not rely on protein structures [9-11]. The protein structure-based methods require the 3D structures of the target proteins to develop and run the predictive models [12-17]. However, recent estimates show that only about 20% to 30% of the human proteins have 3D structures, even when including both experimental and predicted structures [18-20]. Thus, the protein-structure based methods are limited to a relatively small portion of the human proteome. However, the entire human proteome that has about ~70,000 protein sequences when including isoforms (source: UniProt reference proteome ID UP000005640) can be covered by the similarity-based predictions of DPIs. The similarity-based methods rely on two assertions: 1) some similar drugs share the same target(s); and 2) some similar targets interact with the same drug(s) [9, 21-23]. This is motivated in part by a

[*]Address correspondence to this author at the Department of Computer Science, Virginia Commonwealth University, Richmond, VA, 23284, USA; E-mail: lkurgan@vcu.edu

XXX-XXX/14 $58.00+.00

quote from James W. Black, winner of the 1988 Nobel Prize in Medicine: "the most fruitful basis for the discovery of a new drug is to start with an old drug" [24]. We focus on this class of methods motivated by an observation that the underlying assertions are valid to a degree that allows for relatively accurate predictions [9, 25] and because they can predict DPIs for complete proteomes.

We found 14 surveys that touched on the topic of the similarity-based predictions of DPIs that were published in the last five years [9, 25-37]. Broadly speaking, these surveys summarize and discuss the two main components of these predictors: internal databases of native DPIs and predictive models that produce putative DPIs. Seven reviews have discussed details of source databases that are used to derive the internal databases [9, 28-30, 32, 34, 35]. These details include a listing of the source databases used, quality and quantity of annotations of interactions, information about binding affinities, characteristics of drugs and target sequences, all of which influence the quality of the internal databases. All 14 articles discuss some aspects of the predictive methods. These typically include details about different types of similarities that are applied to make predictions, about how to quantify these similarities and whether and how they are combined. They also typically scrutinize types of algorithms employed to derive these models, and their inputs, outputs, and availability. While the 14 reviews provide sufficient coverage of databases and predictive models, they are lacking on the inclusion of recent predictors and do not analyze relations between the source databases, internal databases, and predictive models. A defining feature of our survey is the comprehensive coverage of all high-impact predictors that were published in reputable journals [38-75]. We include eight most recent methods [62, 67, 69-73, 75] that have been published since 2016 and which were not included in the other surveys. In total, we provide an in-depth discussion of 35 predictors and all 12 source databases that were utilized to derive the internal databases of these methods. We are also the first to provide a detailed timeline and analysis of how these various source databases, internal databases, and predictive models are related to each other and how they were combined to deliver similarity-based predictions of DPIs.

2. INTRODUCTION TO SIMILARITY-BASED PREDICTION OF DRUG-PROTEIN INTERACTIONS

The similarity-based prediction of DPIs relies on two components: an internal database of known DPIs and a predictive model that produces putative DPIs using that database. The predictive model primarily depends on the definition of similarities between drugs and between drug targets. Typically, the similarity between drugs is computed either between drug structures or between drug profiles, while the similarity between targets is quantified by the similarity between target sequences. The prediction works in three steps:

1) A user provides inputs in the form of drug structure, drug profile and/or its target sequence(s), whatever is available.

2) If the drug structure (profile) is available, then the similarities between the input drug structure (profile) and the structures (profiles) of the drugs from the internal database of known DPIs are computed. If the sequences of the target(s) of the query drug are given as input or can be retrieved from the database, i.e., the query drug is already included in the database, then the similarities between the query target and all targets in the internal database are also computed.

3) The predictive model combines the similarities to produce a propensity which quantifies a likelihood that the query drug interacts with protein targets that are included in the internal database, i.e., the propensity of putative DPIs.

To cover a complete landscape of top-tier similarity-based predictors of DPIs, we collected corresponding articles that were published in high-impact venues. We searched PubMed [76] in April 2018 using the following query: (predict* [Title/Abstract] AND ("drug target interaction" [Title/Abstract] OR "drug protein interaction" [Title/Abstract])). Among the 170 resulting possibly relevant manuscripts, we manually selected reputable articles that introduced similarity-based methods. Specifically, we picked the articles that were published in journals with the impact factor [77] greater than 3.5. The journal impact factors were collected from the 2017 Journal Citation Reports (JCR) that was released by Clarivate Analytics (formerly Thomson Reuters) on June 14, 2017. This version of JCR is based on the citation data in 2016 and reveals the scientific impact of a JCR-indexed journal by quantifying the ratio between the number of 2016 citations to the articles published in this journal in 2014-2015 and the number of articles published in this journal in 2014-2015. Using these criteria, we found 35 predictors that we include in this review [38-75].

The review starts with the discussion of the source databases to build up the background to survey the selected predictors. These databases include native DPIs that are used to implement predictions by the similarity-based predictors. We summarize their timeline, impact, data contents, and overlap. Next, we investigate the timeline, impact, and availability of the 35 selected predictors. Moreover, we compare the contents of their internal databases and review the types of similarities that they utilize. We also detail seven highly cited and publicly available predictors. After that, we provide a timeline that links the chronological record of source databases with the emergence of the 35 predictors. We also analyze how different types of similarities were used and combined over time to develop these predictors.

3. SOURCE DATABASES

Similarity-based predictors of DPIs are composed of an internal database of native DPIs and a predictive model. The architectures of these predictive models are designed and tuned for their corresponding internal databases. The internal databases typically consist of a set of native DPIs that are aggregated and collated from multiple source databases which store curated annotations of DPIs. We investigated the 35 predictors to come up with a list of all source databases that they use to derive the corresponding internal databases. In total, we found 12 publicly accessible source databases. They

include, in chronological order: PDSP Ki [78], BRENDA [79-87], BindingDB [88-92], TTD [93-98], KEGG BRITE [99-105], DrugBank [106-110], GLIDA [111, 112], KEGG DRUG [99-105], SuperTarget [113, 114], Matador [113], STITCH [115-119], and ChEMBL [120-123]. One of the selected predictors, SEA (Similarity Ensemble Approach) [38, 39], utilizes a collection of drugs and associated targets from a commercial MDL Drug Data Report (MDDR) [124, 125]. This database is not available publicly, and thus it is excluded from our analysis. Next, we summarize the contents, timeline, impact, and relationships between these 12 publicly available source databases.

3.1 Timeline and impact

Besides storing the data and providing facilities to conveniently query and access the data, databases must be maintained and regularly updated. They also should be periodically disseminated to inform the users about their contents and the available features. One way to measure the impact of these databases is to tally the citation counts for the scientific articles that introduce these databases and their updated versions.

Table **1** provides a summary of the 12 source databases. It includes information about publications that introduce the original and the updated versions of these databases, as well as the corresponding citation data. The source databases are sorted chronologically according to the date of their first publication. Specifically, we use the date of the early access online publication when it was available. Otherwise, we use the date of the journal issue in which the first publication has appeared. Apart from the date of the first publication, we also list the date when the database was first made available, which typically is before the database was published. We collect these first release dates from the release notes or time stamps recorded on the database websites, if available, and we use the first publication date otherwise. In this chronological order, the four earliest source databases debuted between August 2000 and January 2002. The next seven databases were published several years later, between 2006 and 2007. The last source database was published in 2011. Eleven out of the twelve sources were published within about one year after their first public release. The one exception, the ChEMBL database was first published two years after its initial release. ChEMBL was originally a commercial database called StARlite that was launched before 2005, acquired by EMBL-EBI in 2008, released to the public in 2009, and finally published in 2011 [126-128]. Given that all these databases were originally released by 2007 or built based on an earlier database, the data stored in these 12 source databases are being accumulated for at least ten years.

Most of the source databases have been regularly updated and republished. Besides just the addition of new data, these updates typically include new features and improved user interface. We list the publication date of the latest republished article and the most recent release date for each source database as of April 1, 2018. These two dates together with the first release and the first publication dates provide interesting insights into the progress of the database development. Considering the latest republishing and release dates, GLIDA, SuperTarget, and Matador have not been republished or updated in the last six years. This suggests that they are no longer actively maintained. The PDSP Ki database is actively and frequently updated, but it has never been republished since it was originally published 17 years ago. Meanwhile, the other eight source databases are being updated and republished regularly. Frequent dissemination informs the users about new contents and features and also may help to attract additional users. In general, these frequently updated sources are relatively more mature since they gradually accumulate DPIs, include more recent data, and typically offer a more refined interface and a longer list of features.

Another relevant aspect is to mark when the source databases were de facto used to build the similarity-based predictors of DPIs. Thus, Table **1** shows the date when the earliest predictor has utilized a given source database to derive its internal database of known DPIs. The list of the source databases that were first used to develop the predictors includes PDSP Ki, BRENDA, KEGG BRITE, DrugBank, SuperTarget, and Matador. The earlier adoption of these source databases reflects to a certain degree their popularity and impact. Moreover, these source databases were also used for other purposes including protein structure-based prediction of DPIs [129-131] and development of various cheminformatics and bioinformatics methods and datasets [132-136].

Table **1** summarizes citations for the 12 source databases. The citations are one way to quantify the impact of these resources. The citation counts were collected from Google Scholar (http://scholar.google.com/) on April 1, 2018. We list the total number of citations that include citations to the first publication and all subsequent publications of a given database. Every source database has received at least about two hundred citations. BRENDA, BindingDB, DrugBank, ChEMBL, KEGG BRITE, KEGG DRUG, and STITCH have accumulated over one thousand citations. A notable exception is the KEGG BRITE and KEGG DRUG databases, which are part of the Kyoto Encyclopedia of Genes and Genomes (KEGG) project since 2005. They were published together with the KEGG database and all of its other affiliated databases in 2006 [99]. Currently, KEGG includes 23 individual databases including, for example, KEGG PATHWAY that provides pathway maps of molecular interaction, reaction and relation networks. The citation data for KEGG BRITE and KEGG DRUG includes the citations to the entire KEGG database; these citations cannot be attributed to individual KEGG resources. The citation counts to KEGG BRITE and KEGG DRUG are so high because they reflect the citations to all 23 databases affiliated with KEGG.

An arguably more robust measure to quantify impact are the annual citation counts. The annual counts are defined as the average citation frequency per one calendar year (365 days) computed over the period from the date of the first publication until the date when we acquired the citation data (April 1, 2018). These counts accommodate for the differences in the age of the source databases. PDSP Ki, GLIDA, SuperTarget, and Matador received moderate (<50) numbers of annual citations. These relatively low citation counts could be a result of a lack of effort to update GLIDA,

SuperTarget, and Matador which were last updated in 2011 or earlier. PDSP Ki is frequently updated but it was last published in 2000. On the other hand, BRENDA, BindingDB, TTD, and STITCH have received relatively high (50-150) annual citations. This is likely because these are mature resources that have ten years of history of regular republishing and updates. Noticeably, DrugBank and ChEMBL attract over 300 citations per year. They were also regularly updated and republished. Their success can be attributed to the high-quality of their data contents, a broad range of functional features, and a user-friendly web interface. Next, we review the data contents for the 12 source databases.

3.2 Data contents

Typically, multiple source databases are used to derive an internal database of a given predictor. They are used as a source for information about drugs, their protein targets, and native DPIs. Table **2** summarizes information about the number of relevant drugs or drug-like compounds that are known to interact with protein targets, the number of these targets, and the number of annotated drug- and compound-protein interactions for the 12 source databases. These data were collected from the latest release of each database on April 1, 2018. We captured the numbers from the release notes or statistics page if they were available. Otherwise, we tallied the numbers from the data dumps. BRENDA, a database dedicated to enzyme functions, does not provide specific statistics and data downloads. Thus, we could not calculate the quantities for this database.

Table **2** categorizes the source databases into two types depending on if they are dedicated to drugs or a more generic set of bioactive compounds. The first type of six databases including PDSP Ki, TTD, DrugBank, Matador, KEGG BRITE, and KEGG DRUG focus on approved, under clinical trial, and experimental drugs. The other six source databases include both drugs and other bioactive compounds that typically are small molecules with drug-like properties. Consequently, the first type of databases has fewer compounds, between 8 hundred and 23 thousand, compared to the second group that includes between 23 thousand and over 2 million compounds. Table **2** is sorted by the number of compounds within each of the two categories. Except for BRENDA for which data are not available and GLIDA that is dedicated solely to the G protein-coupled receptors (GPCR), the bioactive compound-centric databases have more DPIs and protein targets than the drug-centric databases. Specifically, the six smaller databases include between 15 and 60 thousand interactions, compared to the group of larger sources that features up to 148 million interactions. The largest compound repository, ChEMBL, stores a comprehensive set of two million bioactive compounds. This number is 90 times higher than the total number of the drugs in the largest repository that focuses exclusively on drugs, TTD. Unsurprisingly, the number of compound-protein interactions in ChEMBL is 440 times larger than the number of DPIs in TTD.

The main focus of these databases is typically on the human protein targets. The druggable human proteome, which is defined as all human proteins that interact with current drugs, is estimated to comprise of between 1000 and 3000 proteins [1, 137-139]. The numbers of protein targets in the source databases typically vary between about 1000 and 12,000, with a median value of 3036. Half of these databases are larger than the druggable proteome because they cover proteins from other organisms. For example, BindingDB contains targets from over 400 organisms. There are two exceptions that include GLIDA and STITCH. The GLIDA database focuses on the GPCRs, and thus, it is limited to the corresponding 410 GPCR proteins. STITCH covers over nine million proteins from 2031 organisms, which include a substantial number of putative and low-quality annotations of targets. We excluded these predicted and low-confidence interactions (using their confidence score < 0.7) when calculating the numbers for Table **2**. This resulted in a set of about 4 million targets and 149 million interactions. The reason why this database is so large is that STITCH includes both direct and indirect DPIs, while data in the other databases include only the direct interactions. The indirect interactions are derived based on signaling pathways where effects of drugs are propagated onto downstream proteins.

In each of the 12 source databases, the number of DPIs is greater than the number of drugs. This is a result of promiscuity of drugs that typically interact with multiple targets. The field of polypharmacology [140-142] and efforts in drug repurposing [143-145] rely on this promiscuity. However, drug promiscuity may also lead to undesired side-effects and unintended toxicities [146-148]. We measure the drug promiscuity in these source databases by calculating the average number of DPIs per drug (see DPIs/drug in Table **2**). The median degrees of drug promiscuity for the 12 source databases is 2.8, which is close to the promiscuity measured using assays that ranges between 2.6 and 3.4 [149]. Ten databases have between 1.3 and 19.8 DPIs per drug. The STITCH database is again an exception. It includes a relatively dense mapping between proteins and drugs due to the inclusion of a considerable number of indirect interactions. The drug promiscuity has inspired the development of the similarity-based predictive models, where the similarity between targets is used to predict DPIs. It also provides an opportunity to use the currently known targets of a given drug to build models that predict other targets of the same drug.

Besides the interactions, these source databases also encompass rich annotations of the structures, functions, and properties of the drugs and targets, together with the corresponding references. For example, DrugBank provides over 200 such annotations. The native DPIs and the additional knowledge are accessible through the web interfaces of these sources. The corresponding URLs for the 12 source databases are listed in Table **2**.

Table 1. Timeline and impact of the source databases of drug-protein interactions. The source databases were used to derive the internal databases of the 35 selected similarity-based predictors. The timeline is a chronological summary of publications and releases for these source databases. The impact measures citation counts for the publications of databases. This table is sorted chronologically according to the date of the first publication. The data of this table was collected on April 1, 2018.

Source database	Date of the first publication[1]	Date of the first release[2]	Date of the latest publication[3]	Date of the latest release[4]	Date of the first predictor[5]	All citations[6]	Annual citations[7]
PDSP Ki [78]	8/1/2000	11/1/1999	N/A	4/1/2018	7/11/2008	231	13
BRENDA [79-87]	10/1/2000	10/1/2000	10/19/2016	1/1/2018	7/1/2008	2557	146
BindingDB [88-92]	12/1/2001	11/1/2000	10/19/2015	4/1/2018	3/4/2016	1398	86
TTD [93-97]	1/1/2002	1/1/2002	11/13/2017	10/4/2017	3/25/2013	961	59
KEGG BRITE [99-105]	1/1/2006	4/1/2005	11/29/2016	4/1/2018	7/1/2008	*13974	*1140
DrugBank [106-109]	1/1/2006	1/1/2006	11/8/2017	4/2/2018	7/1/2008	5822	475
GLIDA [111, 112]	1/1/2006	1/1/2006	11/5/2007	10/10/2010	3/1/2011	194	16
KEGG DRUG [99-105]	1/1/2006	7/1/2005	11/29/2016	3/29/2018	9/3/2012	*13974	*1140
SuperTarget [113, 114]	10/16/2007	10/16/2007	11/8/2011	11/8/2011	7/1/2008	410	39
Matador [113]	10/16/2007	10/16/2007	N/A	10/16/2007	7/11/2008	316	30
STITCH [115-119]	12/15/2007	8/9/2007	11/20/2015	6/30/2016	6/10/2015	1068	104
ChEMBL [120-123]	9/23/2011	10/27/2009	11/28/2016	5/1/2017	7/8/2013	2434	373

[1] The date when a given source database was originally published in a scientific article. It corresponds to the publication date of early access, if available. Otherwise, the date of the journal issue where the first publication appeared is used.
[2] The date when a given database was first made available, which typically is before the database was originally published. We collect these dates from the release notes or time stamps recorded on the database websites, if available, and we use the date of the first publication, otherwise.
[3] The date of the most recent republished article that introduced an updated version of a given database after the database was originally published. Publication dates of additional earlier republished articles are summarized in Fig. 3.
[4] The date of the most recent release of a given database, which typically is after the latest republishing.
[5] The date when a given source database was first to be utilized to derive the internal database of a predictor.
[6] The "All citations" column is the total number of citations that include citations to the first publication and all republished articles for a given source database. The citation counts were collected from Google Scholar.
[7] The "Annual citations" column is the average citation counts per one calendar year (365 days) over the period from the first publication date until April 1, 2018, rounded to the nearest integer.
* KEGG BRITE and KEGG DRUG are a part of the Kyoto Encyclopedia of Genes and Genomes (KEGG) project. They were published together with the KEGG database and all of its affiliated databases. The citation data for these two databases include the citations to the entire KEGG database, and they cannot be distributed to each affiliated database of KEGG. These counts are relatively high because they reflect the citations to all 23 databases affiliated with KEGG by now.

Table 2. Data contents of the source databases of drug-protein interactions. These source databases were used to derive the internal databases of the 35 selected similarity-based predictors. The data of this table correspond to the latest release of each database as of April 2018. The numerical data were captured from the release notes or statistics page if they were available. Otherwise, we counted the numbers from the data dumps of each database.

Type	Database	Abbr.[1]	Drugs[2]	Proteins[3]	DPIs[4]	DPIs/drug[5]	URL
Drug	Matador	MA	801	2,901	15,843	19.8	http://matador.embl.de
	KEGG BRITE	KB	5,045	1,061	14,222	2.8	http://www.genome.jp/kegg/brite.html
	KEGG DRUG	KD	5,045	1,061	14,222	2.8	http://www.genome.jp/kegg/drug/
	DrugBank	DR	10,562	5,020	23,380	2.2	http://www.drugbank.ca
	PDSP Ki	PK	11,569	1,673	63,619	5.5	http://kidbdev.med.unc.edu/databases/kidb.php
	TTD	TT	23,486	3,036	33,467	1.4	http://bidd.nus.edu.sg/BIDD-Databases/TTD/TTD.asp
Bioactive compound	GLIDA	GL	23,214	410	30,410	1.3	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/
	STITCH	ST	*156,686	*3,908,233	*148,826,348	*949.8	http://stitch.embl.de
	SuperTarget	SU	195,770	6,219	332,828	1.7	http://bioinformatics.charite.de/supertarget/
	BindingDB	BI	644,978	7,042	1,439,799	2.2	http://www.bindingdb.org/bind/index.jsp
	ChEMBL	CH	2,101,843	11,538	14,675,320	7.0	http://www.ebi.ac.uk/chembl/
	BRENDA	BR	Unavailable	Unavailable	Unavailable	Unavailable	http://www.brenda-enzymes.org/index.php

[1] Abbreviated name of a given source database. These abbreviations are used to denote the source databases of each predictor in Table **4**.
[2] The number of drugs and drug-like compounds that are known to interact with protein targets in a given source database.
[3] The number of proteins that are targeted by drugs and drug-like compounds in a given source database.
[4] The "DPIs" column is the number of known DPIs that are stored in a given database.
[5] The "DPIs/drug" column is the average number of DPIs per drug, which is calculated by dividing the "DPIs" column to the "Drugs" column for a given database. Unavailable means that these numbers are missing because BRENDA does not provide specific statistics and downloads for the entire collection of DPIs that it store.
* The numerical data for the STITCH database includes both direct and indirect DPIs, while data for other databases include only the direct interactions. The indirect interactions are based on effects of signaling pathways where effects of drugs are propagated onto proteins that interact with other proteins that directly interact with these drugs. The indirect interactions in this database cannot be separated from the direct interactions, and thus the corresponding numbers are higher than expected.

Table 3. Relationships between source databases. Each row lists a source database, and each column specifies an input where the data of a given source database come from. The inputs include data coming from other source databases, literature, and other types of inputs. The "Literature" column denotes that the data are manually curated from scientific articles, patents, and annual reports of pharmaceutical companies. The "Other inputs" column includes predictions, experimental data, and other databases that are not named in the table. "x" indicates that a given source database draws data from a given input.

Source database	Inputs													
	PDSP Ki	BRENDA	BindingDB	TTD	KEGG BRITE	DrugBank	GLIDA	KEGG DRUG	SuperTarget	Matador	STITCH	ChEMBL	Literature	Other inputs
PDSP Ki													x	x
BRENDA													x	x
BindingDB	x											x	x	x
TTD													x	x
KEGG BRITE													x	x
DrugBank				x									x	x
GLIDA	x					x							x	x
KEGG DRUG													x	x
SuperTarget			x	x	x	x		x					x	x
Matador				x	x	x		x					x	x
STITCH	x		x	x	x	x	x	x		x		x	x	x
ChEMBL			x										x	x

3.3 Relationships between source databases

Each of the 12 source databases includes a different collection of DPIs. However, these source databases also overlap with each other. This is because different source databases collect the interactions from some of the same sources, and because some of them also directly import annotations from the other source database. Table **3** summarizes inputs that are used to derive data stored in a given source database. The inputs include data coming from the 12 source databases, directly from literature, and from other resources. The first 12 inputs in the table indicate whether a given source database directly imports DPIs from another source databases. Seven source databases draw DPIs from between one and nine (for STITCH) other source databases. Moreover, some of the inputs are more popular than the others. Six of the nine input databases provide data for at least three source databases. For example, TTD is used as an input to four source databases including DrugBank, SuperTarget, Matador, and STITCH, compared to BRENDA, SuperTarget, and STITCH that are never used as inputs. The direct inclusion of source databases as inputs results in a substantial overlap between databases. For example, Matador shares a substantial overlap with SuperTarget since they both draw data from the same four source databases while SuperTarget also imports data from one more source. Interestingly, ChEMBL and BindingDB exchange data reciprocally. BindingDB obtains data on compound-protein binding affinities from ChEMBL and exports binding data that were extracted from patents to ChEMBL. Noticeably, STITCH includes the interactions taken from nine other source databases, which explains why Table **2** shows that it hosts the largest collection of interactions. Only the BRENDA database neither imports interactions directly from other source databases nor is used as an input.

Moreover, we list two additional types of inputs: literature and other inputs. The literature includes scientific articles, patents, and annual reports of pharmaceutical companies. The other inputs incorporate predictions, experimental data, and other input databases that exclude the 12 source databases. Table **3** shows that all 12 source databases include data coming directly from the literature. This is yet another factor that contributes to the overlap in the contents of the 12 source databases. The data coming from the literature typically includes information about experiments and assays that were used to validate bioactivity and measure affinities of interactions. This information provides context for the interactions. Every source database also acquires data from other inputs that include predictions, other databases such as the Comparative Toxicogenomics Database [150] and PubChem BioAssay [151], and experimental data that were not stored in a dedicated database, such as the data from Refs. [152, 153].

Our analysis reveals that each of the 12 source databases overlaps with at least one other source database. Table **3** qualitatively summarizes this overlap. A quantitative estimate would be extremely challenging. This is because these databases lack a uniform definition of drugs and targets and a consistent way to name and identify compounds and biomolecules. Even if the same literature is used to extract DPIs the resulting data could be different. For example, when the interaction is annotated at the gene level, the same gene could be mapped to different proteins and different types of protein identifiers, depending on which software and databases were used to perform the mapping. The assignments of information from literature to a precisely and uniquely defined set of drugs and protein targets is still an open challenge [2]. The bottom line that that different databases adopt different nomenclatures and identifiers to represent the drugs and proteins, which makes it virtually impossible to quantify the degree of overlap between the 12 databases. A few works have analyzed the overlap of drugs and targets for a small subset of these source databases [41, 154]. A study over a set of 502 approved drugs shows that 49% and 20% of the DPIs in DrugBank were included in Matador and PDSP Ki, respectively [41]. The data that are unique to DrugBank account for only 46% of its DPIs. Moreover, PDSP Ki shares 42% of its known interactions with Matador. As concluded by another investigation, DrugBank had 52% and 21% of drugs in common with ChEMBL and TTD, respectively [154]. The 74% and 55% of the protein targets stored in DrugBank were also included in ChEMBL and TTD, respectively. Moreover, ChEMBL covered 91% and 55% of the drugs and targets that

were housed in TTD, respectively. These numeric analyses support our observation about the relatively large extend of overlap and also reflect the fact that source databases have their unique data.

We show that the 12 source databases were developed using some of the same data sources and some of them even swap the data with each other. These relationships lead to a certain amount of overlap of information that they store. However, each source database also houses its own unique data, and therefore, they should be combined together to collect the most complete set of known DPIs. This is in fact the case for 22 out of the 35 methods that we survey. They use at least two source databases to develop their internal databases. Moreover, 17 methods, including KRM [40], BLM [44], Yamanishi et al. [45], GIP [47], NBI [48], KBMF2K [49], Cao et al. [51], BLM-NII [52], DT-Hybrid [54], DINIES [56], Shi et al. [57], RLS-KF [61], NRLMF [63], DASPfind [65], PUDTI [70], DVM [71], and iDTI-ESBoost [75] have utilized at least four source databases to create their internal DPI databases.

3.4 Other drug-target interaction databases

Besides the 12 source databases, we discuss another 19 databases that have not yet been adopted to develop the selected similarity-based predictors. These databases house the drug-target interactions accompanied by other information, such as details of mechanisms of DPIs [155], unstudied/dark targets [156], interactions at the gene level [157-163], structures of protein targets [164-167], information about protein-protein and drug-drug interactions [168-176], side-effects of drugs [177], and information focused on specific diseases, such as cancer [178-182]. Some databases are constrained to a particular group of drugs [183] or a specific family of proteins [184]. Next, we discuss these 19 databases in greater depth.

DrugCentral (http://drugcentral.org) is a comprehensive knowledgebase that integrates information about drug actions and pharmacological indications, which can be used to elucidate therapeutic mechanisms mediated through DPIs [155]. The Pharos database (http://pharos.nih.gov) incorporates drug action data taken from DrugCentral to define druggable levels of protein targets and define unstudied/dark protein targets. These dark targets are not yet known to be involved in small molecule activities, but they are potentially druggable [156].

Some resources aggregate biological annotations and disease-related knowledge for the druggable genome, which is defined as a collection of genes that encode druggable proteins. These resources include the PharmGKB database (http://www.pharmgkb.org) [157], DGIdb (http://dgidb.genome.wustl.edu) [158, 159], the Drug2Gene database (http://www.drug2gene.com) [160], IUPHAR/BPS GtPdb (http://www.guidetopharmacology.org) [161, 162], and Open Targets (http://www.targetvalidation.org) [163].

The next three databases rely on the 3D structures of protein targets. The PDB database (http://www.rcsb.org) provides access to an extensive collection of 3D structures of protein-drug complexes [164, 165]. The BioLip database (http://zhanglab.ccmb.med.umich.edu/BioLiP) provides access to residue-level annotations of ligand-binding sites, binding affinity data, and biological functions for a comprehensive collection of proteins that have 3D structures [166]. The PDID (http://biomine.cs.vcu.edu/servers/PDID) is a structural human genome-wide repository of putative and native DPIs that are mapped into the 3D structures of protein targets [167]. It currently stores data about over one million interactions for 51 drugs.

A typical database focuses on DPIs. However, some drugs target protein-protein interactions and targets that are relevant to the same disease or condition can be modulated by multiple drugs. Several databases address these aspects. For example, three databases that focus on the druggability of protein-protein interactions include TIMBAL (http://mordred.bioc.cam.ac.uk/timbal) [168, 169], 2P2Idb (http://2p2idb.cnrs-mrs.fr) [170-172], and iPPI-DB (http://www.ippidb.cdithem.fr) [173, 174]. On the other hand, DCDB (http://www.cls.zju.edu.cn/dcdb) is a resource that centers on the therapeutic effects of multi drug combinations [175, 176].

The IntSide database (http://intside.irbbarcelona.org) focuses on drug side-effects. This database includes data about both therapeutic and off-targets, relevant pathways, biological functions, and chemical traits of drugs [177]. This information is particularly useful to explain and understand undesired responses to drug treatments.

The CancerResource (http://data-analysis.charite.de/care) [178, 179] and canSAR (http://cansar.icr.ac.uk) [180-182] databases aim to bridge cancer research with drug discovery. These two resources provide underlying information about drug-target interactions that are relevant to cancer treatment, such as data of gene expression, mutations in cancer-related genes, drug sensitivity in cancer cell lines, and pathways of drug targets.

Finally, some resources are dedicated to a particular collection of drugs or a specific family of protein targets. For example, the WITHDRAWN database (http://cheminfo.charite.de/withdrawn) includes data on targets and pathways of drugs that were recalled from the market due to toxicity or inefficacy [183]. The GLASS database (http://zhanglab.ccmb.med.umich.edu/GLASS) is exclusively focused on the ligand-protein interactions for the G protein-coupled receptors [184]. About 33% of currently used drugs target this family of proteins [2].

The additional information that can be extracted from these databases complements the information about drug-target interactions that can be obtained from the 12 source databases. We suppose that it would be beneficial for the future similarity-based predictors of DPIs to include these databases as sources.

Table 4. Timeline, impact, and availability of the 35 selected similarity-based predictors of drug-protein interactions. This table is sorted chronologically according to the date of publication. The data of this table was collected on April 1, 2018.

Predictor[1]	Abbr.[2]	Date of publication[3]	JIF[4]	Citations[5]	Annual citations[6]	Availability[7]	URL
SEA [38, 39]	SE	2/7/2007	41.7	922	83	WS	http://sea.bkslab.org
KRM [40]	KR	7/1/2008	7.3	446	46	None	N/A
Campillos et al. [41]	CA_1	7/11/2008	37.2	892	92	None	N/A
COPICAT [42, 43]	CO	6/5/2009	4.5	41	5	WS	http://copicat.dna.bio.keio.ac.jp
BLM [44]	BL	7/15/2009	7.3	265	30	SS	http://members.cbio.mines-paristech.fr/~yyamanishi/bipartitelocal/
Yamanishi et al. [45]	YA_1	6/1/2010	7.3	249	32	None	N/A
Yabuuchi et al. [46]	YA_2	3/1/2011	9.8	102	14	None	N/A
GIP [47]	GI	9/4/2011	7.3	195	30	SS	http://cs.ru.nl/~tvanlaarhoven/drugtarget2011/
NBI [48]	NB	5/10/2012	4.5	339	57	None	N/A
KBMF2K [49]	KB	6/23/2012	7.3	125	22	SS	http://github.com/mehmetgonen/kbmf/
PKR [50]	PK	9/3/2012	7.3	71	13	None	N/A
Cao et al. [51]	CA_2	9/24/2012	5.0	33	6	None	N/A
BLM-NII [52]	BN	11/17/2012	7.3	109	20	None	N/A
Cheng et al. [53]	CH	3/25/2013	3.8	49	10	None	N/A
DT-Hybrid [54]	DH	5/29/2013	7.3	73	15	SS & WS	http://alpha.dmi.unict.it/dtweb/index.php
PRW & NB [55]	PR	7/8/2013	3.8	69	15	SS	http://pubs.acs.org/doi/suppl/10.1021/ci300435j
DINIES [56]	DI	5/16/2014	10.2	44	11	WS	http://genome.jp/tools/dinies/
Shi et al. [57]	SH	5/6/2015	3.8	24	8	SS	http://web.hku.hk/~liym1018/projects/drug/drug.html
Liu et al. [58]	LI	6/10/2015	7.3	24	9	None	N/A
RWR [59]	RW	8/19/2015	4.2	11	4	None	N/A
SLP & RLS [60]	SL	9/9/2015	4.3	6	2	SS	http://pan.baidu.com/s/1dDqDLuD
RLS-KF [61]	RL	1/14/2016	5.0	7	3	SS	http://github.com/minghao2016/RLS-KF/
DrugMiner [62]	DM	1/25/2016	6.4	14	6	WS	http://www.drugminer.org
NRLMF [63]	NR	2/12/2016	4.5	34	16	SS	http://github.com/stephenliu0423/PyDTI/
SDTNBI [64]	SD	3/4/2016	5.1	20	10	SS	http://lmmd.ecust.edu.cn/methods/sdtnbi/
DASPfind [65]	DA	3/16/2016	4.2	14	7	WS	http://cbrc.kaust.edu.sa/daspfind/
DrugE-Rank [66]	DR	6/11/2016	7.3	23	13	WS	http://datamining-iip.fudan.edu.cn/service/DrugE-Rank/
DBN [67]	DB	3/6/2017	4.3	10	9	SS	http://github.com/Bjoux2/DeepDTIs_DBN/
EnsemDT/KRR [68]	EN	5/24/2017	3.8	4	5	None	N/A
Peón et al [69]	PE	6/19/2017	4.3	3	4	None	N/A
PUDTI [70]	PU	8/14/2017	4.3	0	0	None	N/A
DVM [71]	DV	9/11/2017	4.3	1	2	None	N/A
DTINet [72]	DT	9/18/2017	12.1	4	7	SS	http://github.com/luoyunan/DTINet/
bSDTNBI [73, 74]	BS	9/28/2017	3.8	2	4	SS	http://lmmd.ecust.edu.cn/methods/bsdtnbi/
iDTI-ESBoost [75]	IE	12/18/2017	4.3	1	4	WS + SS	http://farshidrayhan.pythonanywhere.com/iDTI-ESBoost/

[1] The name of each predictor is either provided in relevant publications or otherwise, named using the last name of its first author.
[2] The two-letter abbreviated name of each predictor. Subscript numbers are added to distinguish duplicate abbreviations. These abbreviations are used in Fig. 3.
[3] The date of early access online publication or the date of journal issue where a given predictor was originally published.
[4] The "JIF" column is the journal impact factor of the journal where a given predictor appeared. The data were collected from the 2017 Journal Citation Reports that was released by Clarivate Analytics (formerly Thomson Reuters) on June 14, 2017.
[5] The "Citations" column is the number of citations to the article(s) of a predictor, which were collected from Google Scholar.
[6] The "Annual citations" column is the average citation counts per one calendar year (365 days) over the period from the publication date until April 1, 2018, rounded to the nearest integer.
[7] The type of publicly availability of implementations, where WS stands for webserver and SS for standalone software including either compiled or source code. None means the implementation is not offered and thus the corresponding URL is not applicable (N/A).

4. SIMILARITY-BASED PREDICTORS

We review the 35 selected high-impact similarity-based predictors of DPIs. We summarize the timeline, impact, and availability of these methods. We discuss their predictive architectures and link their internal databases to the specific source databases that we discussed in section 3. Finally, we provide a more detailed summary for seven highly cited and publicly available tools.

4.1 Timeline, impact, and availability

The 35 similarity-based predictors were developed in the past decade. Table **4** lists these predictors in chronological order by their first publication dates. The first five predictors were published between 2007 and 2009. We note a steady pace of the development of methods between 2010 and 2015. Specifically, eight methods were released between 2010 and 2012, and another eight between 2013 and 2015. The pace has picked up recently and already 14 methods were published between 2016 and the first quarter of 2018. These data reveal an increasing interest in the development of the similarity-based approaches.

The 35 predictors were published in high-impact venues with impact factors > 3.5. Table **4** lists the impact factors for these reputable journals, which are based on the 2017 Journal Citation Reports [77]. The impact factor values for a significant majority of these methods, 33 out of 35, range between 3.8 and 12.1. Two pioneering predictors were published in 2007 and 2008 in journals with the impact factors around 40 [38, 41]. One way to measure the scientific impact of these methods is based on the citation counts for the articles that introduce these methods. Table **4** lists the corresponding total citation counts, which were collected from the Google Scholar on April 1, 2018. The median total of citations over the 35 methods equals 33. Noticeably, the three earliest predictors [38, 40, 41] have accumulated over 400 citations each over the last decade. Their combined number of citations (2260) is greater than the combined count of citations of the remaining 32 methods (1966). These three highly cited predictors have defined and used for the first time the three types of similarities. The SEA method is based on drug structure similarity [38]. The chronologically second method, KRM, was the first to use target protein sequence similarity, which was combined with the drug structure similarity [40]. The third method by Campillos et al. has introduced the drug profile similarity and used it together with the drug structure similarity to make predictions [41]. There are also three other method that secured at least 200 citations. They include BLM [44], the predictor by Yamanishi et al. [45], and NBI [48]. The low citation counts for the recent methods that were published since 2016 should be dismissed because not enough time have yet passed to accumulate citations. We also analyze a more robust number of annual citations. This number is defined as the total number of citations divided by the number of years (365-day periods), measured between the date of publication and April 1, 2018. The median annual citation number equals ten. The three highest cited predictors attract over 40 citations per year. When excluding the recent predictors that were published since 2016, the remaining 21 older methods have received a median of 15 citations per year. Virtually all of these 21 predictors enjoy a level of annual citations that exceeds the corresponding impact factor of the journal where they were published. The above discussion suggests that the similarity-based predictive methods are of significant interest to the scientific community.

Table **4** also summarizes availability of implementations for these predictors. The authors of eight methods [38, 42, 54, 56, 62, 65, 66, 75] have developed webservers that are geared towards less computer savvy users. The webservers are convenient to use because calculations are done on the server side and consequently the end users only need an internet connection and a web browser to process predictions. Fourteen predictors [44, 47, 49, 54, 55, 57, 60, 61, 63, 64, 67, 72, 73, 75] are available as standalone software. In this case, the end users must install and use them on their own hardware. This requires more skill and effort but it also facilitates inclusion of these methods in other computational pipelines. Two methods [54, 75] are available as both standalone software and webserver. The URLs of these 20 publicly available approaches are listed in Table **4**. The other 15 methods are not available publicly.

4.2 Internal databases

Every similarity-based predictor is implemented based an internal database that includes known DPIs and relevant information about these drugs and proteins. The contents of internal databases are derived from the data that were collected from one or more of the 12 source databases that we reviewed in section 3. Table **5** lists the source databases that are used to generate internal databases. Individual predictors utilize between one and six source databases, with a median of three. Specifically, 13 predictors collect data from a single source database, five from two or three sources, 16 from four sources, and one from six sources. The authors of KRM predictor [40] have released their internal database. This database combines DPIs collected from BRENDA (BR), KEGG BRITE (KB), DrugBank (DR), and SuperTarget (SU). This internal database was later reused by another 15 predictors that we review [44, 45, 47-49, 51, 52, 54, 57, 61, 63, 65, 70, 71, 75]. It was also used by a different set of 18 methods which we did not include in our analysis because of the relatively low impact factor of the venues where they were published [185-202]. The frequent reuse of this database explains to some extent why this predictor enjoys high citation counts in Table **4**.

We observe that only up to 6 out of 12 source databases are used to develop an internal database. This is in spite of the fact that each of the 12 source databases includes data that are unique to that source, and that many other source databases are available, including the databases listed in section 3.4. We recommend that future predictors should rely on more comprehensive internal databases that integrate more source databases. However, this would require a significant effort to map and curate data across the sources that utilize different ways to define, name, and identify the drugs and protein targets.

Except for the 15 methods that reuse the internal database of KRM [40], the other predictors employ unique internal databases by combining data coming from different sets of source databases. Some predictors, such as NBI [48], SDTNBI [64], and DASPfind [65], use more than one internal database. In our analysis, we combine the contents of these internal databases for these three methods. Table **5** summarizes the main characteristics of the internal databases including the numbers of drugs, protein targets, DPIs, and average number of DPIs per drug. The medians of these four characteristics over the 35 predictors are 932 drugs, 989 proteins, 5127 DPIs, and 5.1 DPIs per drug. The first three medians correspond to the frequently reused internal database of KRM predictor [40]. The corresponding medians for the latest releases of the 12 source databases in Table **2** are 23124 drugs, 3036 proteins, 33467 DPIs, and 2.8 DPIs per drug. Interestingly, the first three numbers are much higher while the last number is lower when compared to the sizes of the internal databases. This is in spite of the fact that individual predictors combine multiple source databases to derive their internal databases. One of the reasons why internal databases are relatively small is that they focus on particular collections of drugs and proteins. For example, the internal database of COPICAT [42, 43] includes only the 964 FDA-approved drugs, while its source database, DrugBank, also stores five

thousand experimental drugs. The method developed by Liu et al. [58] focuses on *H. sapiens* and *C. elegans*, while its source databases also cover other organisms such as mouse and *E. coli*. Another reason is that the internal databases are not being updated in contrast to the source databases that are frequently updated and grow in size [154]. In other words, some internal databases are based on outdated version(s) of the source database(s). For example, 16 predictors [40, 44, 45, 47-49, 51, 52, 54, 57, 61, 63, 65, 70, 71, 75], including some recent methods that were developed in 2017, utilize the same internal database [40] which has not been updated since it was published in 2008.

Although the internal databases include fewer drugs, proteins, and DPIs than the source databases, their median drug promiscuity at 5.1 DPIs per drug is 80% larger than the median promiscuity of the source databases, which equals 2.8. This increase is due to the aggregation of different DPIs for the same drugs that are coming from different source databases. The higher promiscuity suggests that the information about the interactions in the internal databases is more complete when compared to the individual source databases. This may benefit the similarity-based predictive models. For example, knowledge of a larger number of native targets would likely result in a larger set of candidate protein targets that could be explored to predict novel targets for a given drug. Also, a higher promiscuity increases the chances to identify proteins that are targeted by different drugs.

4.3 Predictive models

The similarity-based prediction methods rely on an underlying premise that similar drugs likely target the same protein(s) and that similar proteins tend to interact with the same drug(s). To identify putative interaction between a given drug and protein, a predictive model typically searches its internal database for drugs that are similar to the given drug and their known targets that are similar to the given target protein. Therefore, the core aspect that defines the architecture of a predictive model is how to measure the drug-drug and the protein-protein similarities. Analysis of the 35 similarity-based predictors reveals that the similarities are typically quantified using the information about the structure of drugs (drug structure similarity [DSS]), therapeutic profiles of these drugs (drug profile similarity [DPS]), and sequences of their protein targets (protein sequence similarity [PSS]). Some predictors employ one type of similarity to infer putative DPIs. Other methods combine multiple types of similarities since this may improve predictive quality when compared to using just a single type of similarity.

Table **5** summarizes the predictive models. It includes information about the similarities used, how the similarities are computed and combined, and how the predictive performance of the selected 35 predictors was evaluated. Except for DrugMiner [62], the other 34 methods utilize DSS to make prediction. Typically, drug structures are represented by binary or numeric vectors, such as molecular fingerprints [203, 204], and then DSS is calculated between these vectors. The first similarity-based predictor, SEA, has applied DSS to infer the probability that a given drug shares target with the drugs which are included in the internal database [38]. Another 13 methods [40, 44, 45, 47-49, 52, 54, 56, 60, 61, 63, 65] compute DSS by applying the SIMCOMP algorithm [205, 206]. This algorithm represents drug structures with graphs in which nodes are atoms and edges are covalent bonds. SIMCOMP measures DSS based on the number of atoms in the common subgraphs between the two graphs that represent drug structures. Eleven algorithms [41, 53, 55, 57-59, 64, 66, 69, 72, 73] compute DSS using the Tanimoto coefficient [207] that quantifies similarity between molecular fingerprints of drugs. Lastly, nine methods [42, 46, 50, 51, 67, 68, 70, 71, 75] utilize machine learning algorithms that use a kernel function or neural networks to measure DSS between feature vectors that represent drug structures.

The second most used similarity is PSS, which is employed by 28 predictors. PSS is measured either directly between protein sequences or between sequence-derived feature vectors that are used to represent the sequences. Sufficiently high sequence similarity suggest that the two proteins may have similar structures [18, 208-211] and functions [212-218]. Hence, if PSS is high then the two proteins could have similar structures and could share similar binding pockets that are targeted by the same drug. The most commonly used way to quantify PSS is sequence alignment, which can be computed using PSI-BLAST [219] or Smith-Waterman algorithms [220]. The sequence alignment-based approach to calculate PSS is adopted by 18 predictors [40, 44, 45, 47-49, 52, 54, 56-61, 63, 65, 66, 72]. The other ten methods use machine learning algorithms, including kernel-based models [42, 43, 46, 50, 51, 68, 70, 71] and neural networks [62, 67]. These models do not use the sequences but instead they quantify PSS between numeric feature vectors that represent the sequences. For example, a sequence could be represented by amino acid composition (20-dimensional vector quantifies faction of each amino acid type in the sequence) and/or a set of physicochemical characteristics of residues in a sequence (e.g., average hydrophobicity computed over all amino acids in the sequence).

DPS is the least often used similarity. It was applied in eight predictive models. These methods typically represent a drug with a profile vector composed of binary values that indicate the presence/absence of a specific side-effect term. In this case, DPS quantifies the degree to which two drugs result in the same or similar adverse effects. Four models [41, 53, 58, 72] utilize the Tanimoto coefficient or its derivations to measure the similarities between the side-effect profiles of drugs. Three other algorithms [45, 50, 56] compute correlations between side-effect profiles to quantify DPS. Alternatively, drugs are represented by the ATC codes that denote hierarchical classification of drugs [221]. This approach is used by two predictors that measure semantic similarity between these codes [53, 57].

Only six methods utilize a single type of similarity. Specifically, five predictors [38, 55, 64, 69, 73] rely on DSS and another predictor [62] on PSS. DPS is never used alone. The other 29 approaches use an ensemble of at least two types of similarities. The "Ensemble" column in Table **5** categorizes these ensemble-based predictors into two groups based on the techniques that they use to combine multiple similarities. The first group of seven methods applies a simple approach to

produce ensembles, typically based on summation of multiple similarities. The second group of 23 methods utilizes a more complex approach which involves operators like maximum and multiplication. Moreover, Cheng et al. proposed two models that rely on both simple and complex approaches [53]. One of these two models utilizes a summation of DSS and DPS and the other model is based on a geometric mean of the same two similarities.

The similarities can be directly used to estimate propensity of DPIs or they can be input into predictive models. Examples of the latter solutions include the use of similarities as elements of a kernel matrix for machine learning algorithms and as weights in an adjacency matrix for network-based approaches. Motivated by a recent survey [34], Table **5** classifies the architectures of the 35 predictive models into three categories: similarity score-based, machine learning-based, and network-based architectures. The most commonly used architectures are based on drug-protein networks and machine learning algorithms, with only three models that directly use similarity scores.

Table **5** also explains how the predictive performance of these methods was assessed. All 35 models were evaluated in a pairwise manner at the DPI level. In other words, they quantify how many drug-protein pairs were correctly predicted. Interestingly, none of these methods has provided assessment per drug. Such assessment would measure how many drug-protein pairs are correctly predicted for a particular drug. The drug-wise assessment would provide insights into ability of specific methods to accurately predict interactions for a specific drug or class of drugs. Availability of the drug-wise assessment would also allow to investigate whether certain characteristics of drugs, such as molecular weight and/or promiscuity, affect predictive performance of the similarity-based methods.

4.4 Timeline of the use of similarities and their ensembles

There are seven potential types of predictive models including three based on one similarity, three based on combinations of two similarities, and one ensemble of three similarities. Fig. **1** provides a timeline of the use of these seven types of predictive models for the 35 considered methods. In general, the ensemble-based models are much more widely utilized than the single similarity-based methods. The most frequently developed ensemble includes DSS and PSS. At least one of these methods was published every year except only for 2007, 2010, and 2014. The second most commonly used type of methods combines the three types of similarities. These methods were shown to outperform ensembles that are based on specific pairs of similarities, such as ensemble of DSS and PSS and ensemble of DPS and PSS [50]. However, a significant majority of current predictors combines only two similarities: DSS and PSS, without DPS. This is likely because the information on side-effects is not available in most of the source databases and this information used to be difficult to collect, especially before databases such as SIDER [222, 223] and MetaADEDB [224] were developed. Also, the mapping of drugs in the side-effect databases into the drugs in the internal databases is non-trivial. Another reason why DPS is not an attractive option is the fact that the side-effect information is limited to the marketed drugs. Consequently, this limits the internal databases to this group of drugs. Unlike the two broadly adopted types of models mentioned above, the other types of models are published less frequently. These models rely on DSS, PSS, and the ensemble of DSS and DPS. Some designs, such as methods that use DPS and the ensemble of DPS and PSS, have not been employed among the 35 considered methods. In summary, the most popular configuration is the ensemble that combines DSS with another type(s) of similarity.

Fig. **2** shows how often an individual type of similarity was used over the last decade. We include their use individually and also as part of ensembles. Each bar in Fig. **2** shows a cumulative number of times these similarities were used up to a given year. The colors inside the bars reflect relative fraction of use of specific types of similarities. The first bar reveals that the first predictor that was published in 2007 was based on DSS [38]. The relative rate of use of DSS has gradually decreased between 2008 and 2010 when compared to the other two types of similarities. After 2010, the relative rate of use of the three types of similarities has steadied; this is reflected by similar proportions of the three colors. The last bar reveals that in total the similarities were used 70 times by the considered set of 35 methods, which correspond to two similarities per method on average. DSS was used 49% of the time, while PSS and DPS were used 40% and 11% of the time, respectively. These fractions are consistent with the observation that the ensemble of DSS and PSS is the most frequently utilized type of predictive model. The relatively infrequent use of DPS is likely due to low drug coverage and difficulty to collect the information on drug side-effects. We believe that with the release and improvements to the databases that provide access to drug profiles, including SIDER [222, 223] and MetaADEDB [224], this type of similarity will play a more prominent role in the development of future predictors.

Table 5. Summary of source databases, internal databases, and predictive models that are utilized in the 35 selected predictors. The data of this table was collected on April 1, 2018.

Predictor[1]	Source databases		Internal database[3]				Predictive model[4]						
	Sources[2]	No. of sources	Drugs	Proteins	DPIs	DPIs/ drug	DSS	DPS	PSS	No. of similarities	Ensemble[5]	Archi-tecture[6]	Assessment[7]
SEA [38, 39]	Unavailable	1	65,241	246	71,094	1.1	SE			1	N/A	SS	Pairwise
KRM [40]	BR KB DR SU	4	932	989	5,127	5.5	SI		AL	2	C	NE	Pairwise
Campillos et al. [41]	PK DR MA	3	502	N/A	4,857	9.7	TC	TC		2	S	SS	Pairwise
COPICAT [42, 43]	DR	1	964	456	1,731	1.8	KF		KF	2	C	ML	Pairwise
BLM [44]	BR KB DR SU	4	932	989	5,127	5.5	SI		AL	2	C	NE	Pairwise
Yamanishi et al. [45]	BR KB DR SU	4	443	989	2,649	6.0	SI	CO	AL	3	C	NE	Pairwise
Yabuuchi et al. [46]	GL	1	866	317	5,207	6.0	KF		KF	2	C	ML	Pairwise
GIP [47]	BR KB DR SU	4	932	989	5,127	5.5	SI		AL	2	C	NE	Pairwise
NBI [48]	BR KB DR SU	4	5,330	4,785	17,610	3.3	SI		AL	2	C	NE	Pairwise
KBMF2K [49]	BR KB DR SU	4	932	989	5,127	5.5	SI		AL	2	C	NE	Pairwise
PKR [50]	KD	1	2,423	436	6,769	2.8	KF	CO	KF	3	C	ML	Pairwise
Cao et al. [51]	BR KB DR SU	4	932	989	5,127	5.5	KF		KF	2	C	ML	Pairwise
BLM-NII [52]	BR KB DR SU	4	932	989	5,127	5.5	SI		AL	2	S	NE	Pairwise
Cheng et al. [53]	TT DR	2	621	893	3,195	5.1	TC	TC		2	S & C	NE	Pairwise
DT-Hybrid [54]	BR KB DR SU	4	5,330	4,773	17,573	3.3	SI		AL	2	S	NE	Pairwise
PRW & NB [55]	CH	1	105,946	894	155,208	1.5	TC			1	N/A	ML	Pairwise
DINIES [56]	PK TT DR KD MA CH	6	678	277	1,804	2.7	SI	CO	AL	3	S	NE	Pairwise
Shi et al. [57]	BR KB DR SU	4	932	989	5,127	5.5	TC	SS	AL	3	C	ML	Pairwise
Liu et al. [58]	DR MA ST	3	2,486	3,356	7,369	3.0	TC	TC	AL	3	S	ML	Pairwise
RWR [59]	DR	1	684	627	2,557	3.7	TC		AL	2	C	NE	Pairwise
SLP & RLS [60]	DR	1	786	809	3,681	4.7	SI		AL	2	C	NE	Pairwise
RLS-KF [61]	BR KB DR SU	4	932	989	5,127	5.5	SI		AL	2	C	NE	Pairwise
DrugMiner [62]	DR	1	1,396	1,224	4,729	3.4			ML	1	N/A	ML	Pairwise
NRLMF [63]	BR KB DR SU	4	932	989	5,127	5.5	SI		AL	2	C	NE	Pairwise
SDTNBI [64]	BI DR CH	3	22,839	1,541	57,726	2.5	TC			1	N/A	NE	Pairwise
DASPfind [65]	BR KB DR SU	4	3,897	6,662	12,919	3.3	SI		AL	2	C	NE	Pairwise
DrugE-Rank [66]	DR	1	1,242	1,324	5,701	4.6	TC		AL	2	S	ML	Pairwise
DBN [67]	DR	1	1,412	1,520	6,262	4.4	NN		NN	2	C	ML	Pairwise
EnsemDT/KRR [68]	DR	1	7,739	4,902	17,483	2.3	KF		KF	2	C	ML	Pairwise
Peón et al [69]	CH	1	745	1,427	8,535	11.5	TC			1	N/A	SS	Pairwise
PUDTI [70]	BR KB DR SU	4	932	989	5,127	5.5	KF		KF	2	C	ML	Pairwise
DVM [71]	BR KB DR SU	4	932	989	5,127	5.5	KF		KF	2	C	ML	Pairwise
DTINet [72]	DR	1	708	1,512	1,923	2.7	TC	TC	AL	3	C	NE	Pairwise
bSDTNBI [73, 74]	BI CH	2	276	453	1,796	6.5	TC			1	N/A	NE	Pairwise
iDTI-ESBoost [75]	BR KB DR SU	4	932	989	5,127	5.5	ML		ML	2	C	ML	Pairwise

[1] The name of each predictor is either provided in relevant publications or otherwise, named using the last name of its first author.

[2] The "Sources" column gives two-letter abbreviated names of the source databases that were used to derive the internal databases. The corresponding full names of the source databases can be found in Table **1**. Unavailable means that the source used by SEA method is not publicly accessible.

[3] Internal database is the set of known DPIs that is used to implement the corresponding predictor. N/A means that the exact number of proteins was not disclosed in Ref. [41] and we could not quantify it.

[4] Predictive model lists the approaches used to measure drug structure similarity (DSS), drug profile similarity (DPS), and protein sequence similarity (PSS). SE: SEA algorithm that measures DSS [38, 39]; SI: SIMCOMP tool that quantifies DSS [205, 206]; TC: Tanimoto coefficient which measures similarity between binary or numeric vectors [207]; KF: kernel function that measures similarity between feature vectors used by machine learning algorithms; NN: neural network; CO: correlation; SS: semantic similarity of drug profiles represented by the ATC codes [57]; AL: sequence alignment using PSI-BLAST [219] or Smith-Waterman algorithms [220]; and ML: machine learning algorithms that use kernel functions, distance metric, and neural networks to quantify similarity between feature vectors. A blank cell indicates that the given type of similarity is not used by the given method.

[5] The "Ensemble" column is the type of approach that is used to combine multiple similarities, including simple (S) and complex (C) approaches. Simple (complex) indicates use of a summation (use of other more complex operators) to combine multiple similarities, and N/A denotes the single similarity-based predictors.

[6] The "Architecture" column summarizes architectures of the 35 considered predictors into three categories: similarity score-based (SS), machine learning-based (ML), and network-based (NE) methods.

[7] The "Assessment" column indicates if a predictor was evaluated per drug-protein pair (pairwise) or per drug (drug-wise).

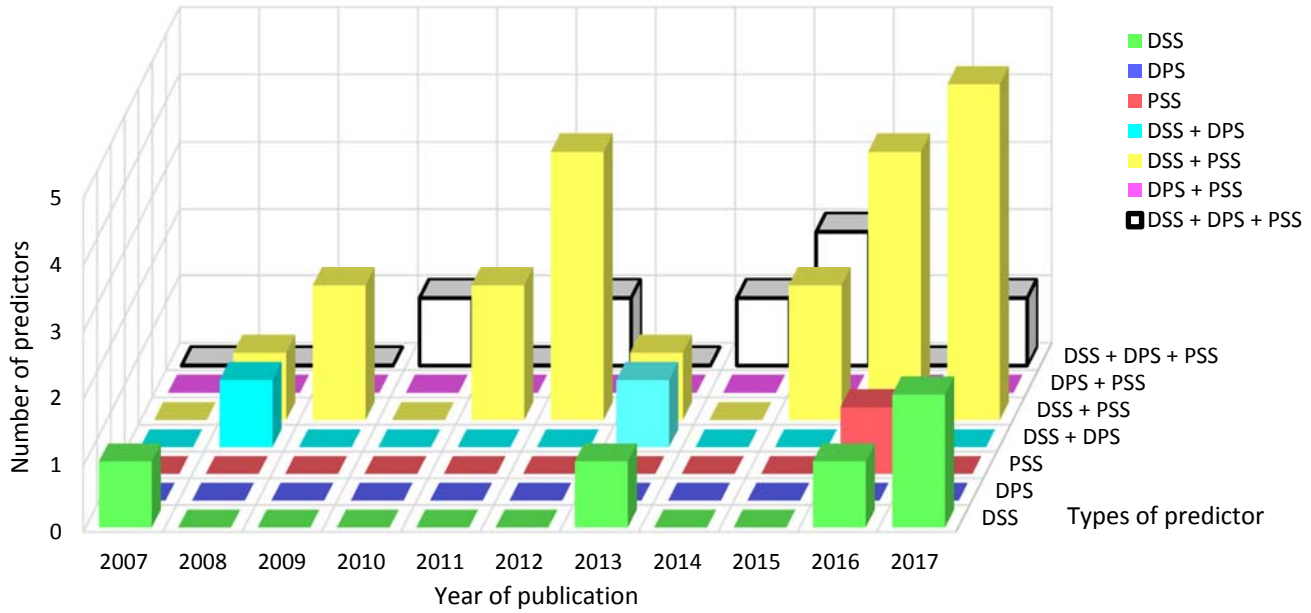

Fig. 1. Timeline of similarity-based predictors of drug-protein interactions. The *x*-axis (width) denotes when the similarity-based predictors were published. The *y*-axis (depth) denotes the type of predictors. The *z*-axis (height) shows the number of a given predictors that were published in a given year. Green, blue, and red cubes represent drug structure (DSS), drug profile (DPS), and protein sequence similarity-based predictors (PSS), respectively. The corresponding ensembles are color-coded according to the mixture of these three base colors shown in the figure legend.

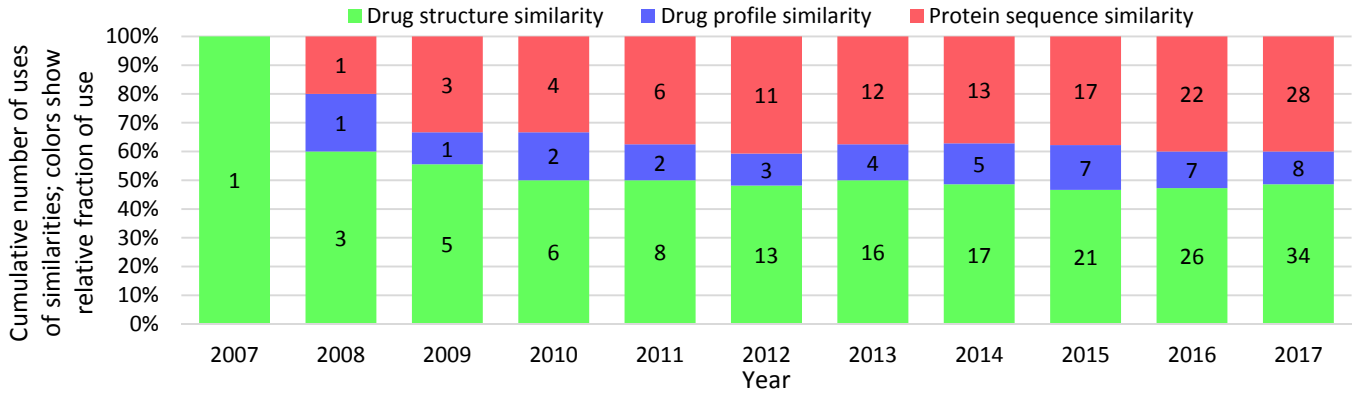

Fig. 2. Use of individual similarities in predictors of drug-protein interactions. The *x*-axis denotes the year. The *y*-axis shows the fractions of cumulated numbers of uses of drug structure (green bars), drug profile (blue), and protein sequence (red) similarities up to a given year. Numbers inside the bars indicate the cumulative count of uses of each similarity. The similarity is counted when it is used individually by a single similarity-based predictor, and in tandem with other similarities by ensemble-based predictors.

4.5 Selected similarity-based predictors

In total, 20 out of 35 methods offer either online or standalone implementations, see Table **4**. This facilitates use of these tools by the end users, compared to the remaining methods that the end users would have to implement. Among these publicly available models, we highlight seven predictors that are relatively influential. This means that each of these tools has received at least a median number of 15 citations per year. This median was calculated across 21 out of the 35 predictors that exclude recent methods which were published since 2016 for which there was not enough time yet to accumulate reliable citation counts. These seven methods are sorted in the descending order by their annual citation numbers. For each selected method, we list the authors, briefly summarize its architecture, describe format of input and output, note the availability of its implementation and internal database, and we comment on its limitations.

SEA

SEA (Similarity Ensemble Approach) method was developed by Shoichet's group and Irwins' group at the University of California, San Francisco in 2007 [38]. This is the earliest similarity-based predictor of DPIs. SEA quantifies propensity of DPIs using a statistical test for the set-wise DSS between a given drug and the group of drugs that are known to bind to a given protein.

Input: Structure of a query drug in the SMILES [225] string format.

Output: Propensity of putative DPIs. The propensities are produced for the input drug and each protein target from the internal database.

Availability: The webserver at http://sea.bkslab.org. The internal database is also accessible for browsing.

Limitations: 1) only predicts for the input proteins that are already included in the internal database; 2) has a small internal database with below median number of proteins (246 proteins); and 3) only uses drug structures for prediction.

BLM

BLM (Bipartite graph inference with Local Models) predictor was proposed by Bleakley and Yamanishi at the Mines ParisTech [44]. This method predicts putative DPIs by applying two SVM models for which kernel matrices quantify DSS and PSS.

Input: Structure of a query drug represented in SMILES [225] or MDL MOL [226] format and amino acid sequence of a query protein.

Output: Propensity and binary annotation of putative DPI for the input drug and protein.

Availability: The source code in MATLAB and the internal database can be downloaded at http://members.cbio.mines-paristech.fr/~yyamanishi/bipartitelocal/.

Limitations: 1) only predicts for the input drugs that are already included in the internal database; 2) only predicts for the input proteins that are already included in the internal database; 3) uses a black box model (the predictive model is not human readable); 4) relies on the internal database that includes annotations of non-interacting drug-protein pairs there were not validated experimentally; and 5) is limited to expert users who can compile and run the source code.

GIP

GIP (regularized least squares with Gaussian Interaction Profile kernels) algorithm was produced by Marchiori's group at the Radboud University [47]. This approach relies on a regularized least squares algorithm that combines DSS and PSS to estimate the confidence of putative interactions.

Input: Structure of a query drug represented in SMILES or MDL MOL format and amino acid sequence of a query protein.

Output: Propensity of putative interaction for the input drug and protein.

Availability: The source code in MATLAB are provided at http://cs.ru.nl/~tvanlaarhoven/drugtarget2011/. The internal database is the same as the one used by BLM.

Limitations: 1) only predicts for the input drugs that are already included in the internal database; 2) only predicts for the input proteins that are already included in the internal database; 3) relies on the internal database that includes annotations of non-interacting drug-protein pairs there were not validated experimentally; and 4) is limited to expert users who can compile and run the source code.

KBMF2K

KBMF2K (Kernelized Bayesian Matrix Factorization with 2 Kernels) method was released by Gönen at the Aalto University [49]. The predictor combines DSS and PSS and applies a Bayesian model to estimate the likelihood of interaction for a given drug and protein.

Input: Structure of a query drug represented in SMILES or MDL MOL format and amino acid sequence of a query protein.

Output: Propensity and binary annotation of putative DPI for the input drug and protein.

Availability: The source code in MATLAB and R is available at http://github.com/mehmetgonen/kbmf/. The implementation uses the internal database of BLM.

Limitations: 1) relies on the internal database that includes annotations of non-interacting drug-protein pairs there were not validated experimentally; and 2) is limited to expert users who can compile and run the source code.

NRLMF

NRLMF (Neighborhood Regularized Logistic Matrix Factorization) algorithm was designed by Liu et al. at the Agency for Science, Technology and Research (A*STAR) in Singapore [63]. The authors use a matrix factorization-based recommendation algorithm that integrates DSS and PSS to predict putative DPIs.

Input: Structure of a query drug represented in SMILES or MDL MOL format and amino acid sequence of a query protein.

Output: Propensity of putative interaction for the input drug and protein.

Availability: The source code in Python is downloadable at http://github.com/stephenliu0423/PyDTI/. The internal database is the same as the one used by the BLM method.

Limitations: 1) relies on the internal database that includes annotations of non-interacting drug-protein pairs there were not validated experimentally; and 2) is limited to expert users who can compile and run the source code.

DT-Hybrid

DT-Hybrid (Domain Tuned-Hybrid network-based inference) model was created by Pulvirenti's group at the University of Catania [54]. This method is a recommendation

algorithm that employs a linear combination of DSS and PSS to score and rank putative drug targets.

Input: Structure of a query drug represented in SMILES or MDL MOL format and amino acid sequence of a query protein.

Output: Propensity of putative interaction for the input drug and protein.

Availability: The webserver, source code in R, and internal database are available at http://alpha.dmi.unict.it/dtweb/.

Limitations: 1) only predicts for the input drugs that are already included in the internal database; 2) only predicts for the input proteins that are already included in the internal database; and 3) relies on the internal database that includes annotations of non-interacting drug-protein pairs there were not validated experimentally.

PRW & NB

PRW & NB (Parzen-Rosenblatt Window and Naive Bayes) predictor was built by Bender's group and Glen's group at the University of Cambridge [55]. This method relies on DSS and applies Bayes' theorem to infer the probability of drug-target interaction.

Input: The extended-connectivity fingerprints (ECFPs) [227] that represent the structure of a given drug.

Output: Propensity of putative interaction for the input drug and each target in the internal database.

Availability: The source code and the internal database are available at http://pubs.acs.org/doi/suppl/10.1021/ci300435j.

Limitations: 1) only predicts for the input proteins that are already included in the internal database; 2) only uses drug structures for prediction; 3) relies on the internal database that includes annotations of non-interacting drug-protein pairs there were not validated experimentally; and 4) is limited to expert users who can compile and run the source code.

5. RELATIONSHIP BETWEEN THE DEVELOPMENT OF SOURCE DATABASES AND PREDICTORS

Fig. 3 is a detailed timeline that links the development of the 12 source databases with the development of the 35 similarity-based DPI predictors. It shows an annual timeline of development for each source database that includes the first publication, every subsequent republished article, every predictor that utilized this source database to derive its internal database, and its latest release. The source databases are sorted chronologically by the time of the first publication.

Fig. 3 extends Table **1** by providing a timeline of all publications of each source database and visually linking the source databases with the predictors. It reveals that Matador, GLIDA, and SuperTarget have not been updated since 2007, 2010, and 2011, respectively. PDSP Ki is being continually updated even though this resource has not been republished since 2000. The lack of recent updates/releases and publications, as it is the case for PDSP Ki, GLIDA, SuperTarget, and Matador, has adversely affected their impact. These four databases secure the lowest annual citation counts among the 12 source databases, see Table **1.** In contrast, BRENDA, TTD, KEGG BRITE, KEGG DRUG, STITCH, and ChEMBL are updated frequently and are biennially republished. Similarly, a major new version of DrugBank is released and published triennially since 2007. Finally, BindingDB was republished recently after about ten years since the previous publication. Biennial publishing is a typical trend for majority (8 out of 12) of the source databases. This is primarily dictated by the policy of the venue where they are published, *Nucleic Acids Research*, which requires a two year period between publications of the same resource. The regular maintenance/updates and periodic dissemination informs current users about additional data and novel functionality and attracts new users, broadening the impact and attracting additional citations.

We put a two-letter abbreviated name of each similarity-based predictor (see full names in Table **4**) in the timeline to mark when each source database was utilized to derive its internal database. Fig. 3 shows that PDSP Ki, BindingDB, TTD, GLIDA, KEGG DRUG, Matador, STITCH, and ChEMBL are adopted by between one and five predictors. The other four source databases, BRENDA, KEGG BRITE, DrugBank, and SuperTarget are used more often. As we discussed in section 4.2, KRM [40] has introduced an internal database that was compiled from these four source databases. Fifteen other predictors reused this internal database. Therefore, at least 16 predictors rely on these four source databases. Moreover, DrugBank is the most popular source database. It was utilized by 29 out of 35 selected predictors. This observation is consistent with the fact that DrugBank is the most cited source database, except for KEGG BRITE and KEGG DRUG that share citation data with the other 21 KEGG-affiliated databases.

The timeline covers predictors that were developed over the last decade. At first, six source databases were employed in the development of the two earliest predictive approaches in 2008. By 2012, the mid-point of the decade, eight different source databases had been adopted by 12 predictors. Since 2013, 11 different source databases were utilized by another 22 predictors. This reveals that more source databases are used to build the internal databases by the newer methods. However, the predictors that use the most sources databases are still limited to no more than six of them. Given the availability of so many more source databases, new methods should exploit a more complete collection of known DPIs by combing data from more sources.

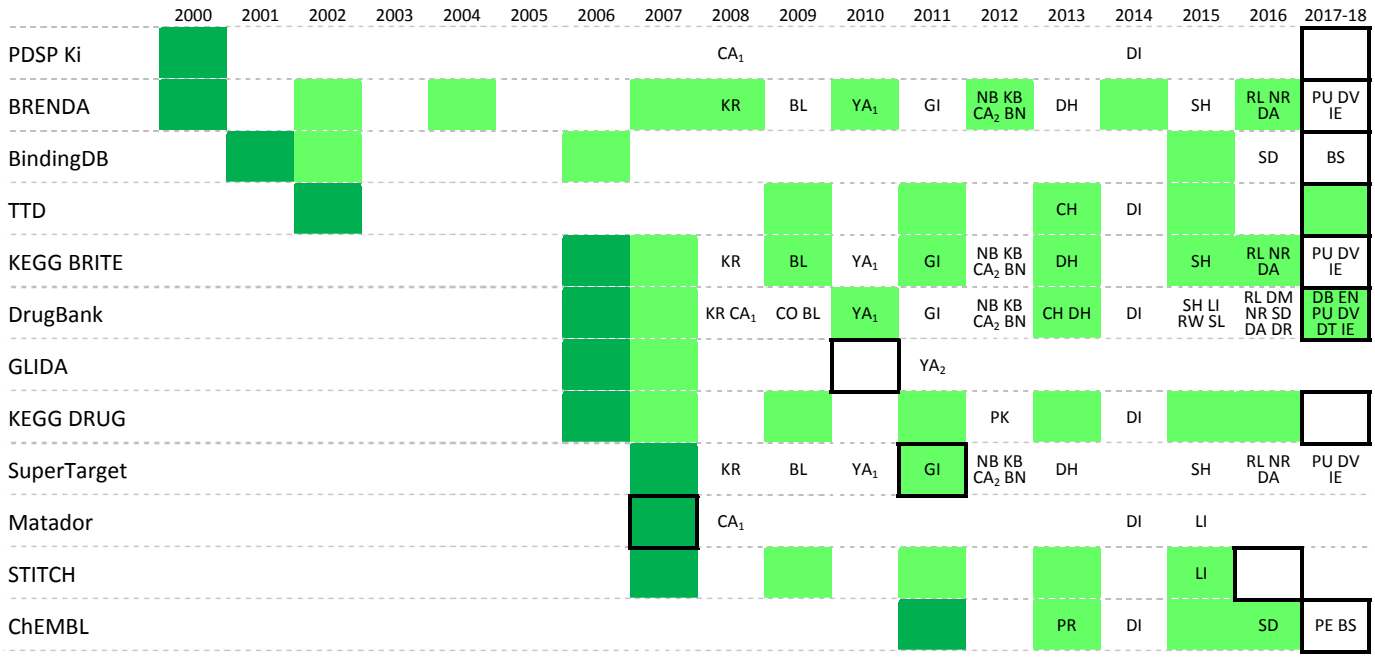

Fig. 3. Timeline of similarity-based predictors of drug-protein interactions and source databases. Each row corresponds to one source database while each column corresponds to one specific year. Dark green indicates the year of the first publication of a given source database. Bright green indicates the year when a given source database was subsequently republished. Black border indicates the year of the latest release of the database. Two-letter abbreviations represent the predictors that utilize a given source database (row) in a specific year (column). The names and references corresponding to these abbreviations are defined in Table **4**.

6. CONCLUSIONS

We have discussed the timeline, impact, availability, contents, and architectures for the 35 high-impact similarity-based predictors of DPIs and the related 12 source databases.

The source databases store curated annotations of DPIs that are used to derive the internal databases of the predictors. Their contents were obtained from relevant literature, experiments, and other data repositories. Most of the 12 source databases also directly import the annotations of DPIs from the other source databases. Consequently, each source database stores its unique data and also a certain amount of data that overlap with the other source databases. Moreover, some source databases focus on the interaction data for the approved and experimental drug compounds, while the others include the protein targets for a more generic collection of bioactive compounds that includes drugs and drug-like molecules. This contributes to the diversity of contents of the source databases, in terms of the number and type of compounds, protein targets, and DPIs. Drugs that are included in the source databases typically target multiple proteins. This drug promiscuity defines the field of polypharmacology and benefits the development of the similarity-based predictors. The source databases have accumulated data for over ten years and have been frequently used by the community. This is reflected by the fact that they were cited at least 190 times each. Most of the databases are continually updated and periodical republished. Our analysis shows that the source databases that are more frequently updated and republished are also more often cited.

The similarity-based predictors have been developed at a steady pace over the past decade. These methods rely on the internal databases that typically include data derived from multiple source databases. We found that recent methods have used a larger number of source databases than the older methods, although this number is still relatively low compared to the number of available source databases. The internal databases generally include lower numbers of drugs, proteins, and interactions but higher degrees of drug promiscuity when contrasted with the corresponding source databases. A higher drug promiscuity allows the similarity-based predictors to screen a more complete set of candidate protein targets and increases likelihood of identifying targets shared by multiple drugs. Given these advantages, the future predictive models should exploit an even more comprehensive set of DPIs that would be collected from a larger number of source databases.

Most of the predictors have received a relatively high annual citation counts when compared to the corresponding impact factors of the journals where they were published. This points to the substantial impact of the similarity-based methods. These methods incorporate predictive models that quantify similarities between the input drugs and proteins and the drugs and their known targets in the internal databases. Our survey reveals that the 35 predictive models that we considered have utilized three types of similarities: drug structure similarity (DSS), drug profile similarity (DPS), and protein sequence similarity (PSS), as well as their ensembles. The three earliest predictors are the most cited. They were the first to use DSS and the ensemble of DSS and PSS/DPS. These pioneering works resulted in the DSS-centric trend for

the development of the similarity-based models. The vast majority of the 35 methods applies DSS or ensemble models that combine DSS with DPS and/or PSS. The ensemble of DSS and PSS is the most commonly utilized type of predictive model. Our analysis of the frequencies of use of similarities has found that DSS and PSS have been utilized about 90% of the time, while DPS is relatively underutilized. The infrequent use of DPS likely results from the incompleteness and difficulty of use of drug profiles. This motivates the need to further develop the drug side-effect profile databases, thereby facilitating a new generation of methods that more heavily rely on DPSs.

Based on the observations made in this survey, we formulate a few observations and recommendations:

1) The source databases should be regularly updated and disseminated. This would improve effectiveness, completeness, and impact of these databases. Periodic publications and peer review will also boost quality of underlying data and features.

2) A larger number of source databases should be integrated to derive the internal databases. The existing predictors have considered no more than six source databases while at least twice as many are available. This includes the drug-target databases discussed in section 3.4 that were never used by the considered top-tier similarity-based predictors. Each source database has unique data that complements contents of the other databases. Thus, combining more source databases would likely further increase the completeness of information about drug promiscuity in the internal databases. This would benefit predictive quality of the similarity-based predictors and would also enable the development of higher-quality benchmark datasets.

3) A high-quality publicly available internal database should be developed. Most of the current predictors rely on their unique internal databases. Both development and maintenance of these databases take a considerable amount of time and effort, which would not be duplicated when a public database would be reused. Instead, authors could focus on building high-quality predictive models. A public database would also make it more consistent and easier to empirically evaluate and compare the predictive performance of different methods. Although the internal database from KRM predictor [40] has been reused several times, at this point this database is outdated and would need to be expanded and improved.

4) Future research should focus on methods that combine multiple similarities. We note that four methods that rely on a single similarity were recently published [62, 64, 69, 73]. The ensemble-based models that combine multiple types of similarities are likely to provide more accurate results than the methods that use one similarity. In particular we advocate further development of drug side-effect profile databases which would drive the development of novel tools that combine all three types of similarities.

5) Most of the current development focuses on protein targets that are structured [139]. However, about 30% of eukaryotic proteins are either fully disordered or have long regions of intrinsic disorder [228, 229]. These disordered proteins are implicated in a wide range of diseases [230-237]. Moreover, certain protein families are enriched in the intrinsic disorder, such as nuclear receptors, kinases, and various enzymes [139, 229, 238-241]. These protein families include important therapeutic drug targets and druggable proteins [1, 2, 137, 138, 242]. This prompts our final recommendation that disorder-specific databases and methods should be developed in the near future.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported in part by the Qimonda Endowment funds to L.K.

REFERENCES

[1] Hopkins, A.L.; Groom, C.R., The druggable genome. *Nat. Rev. Drug Discov.,* **2002**, *1*, (9), 727-730.
[2] Santos, R.; Ursu, O.; Gaulton, A.; Bento, A.P.; Donadi, R.S.; Bologa, C.G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T.I.; Overington, J.P., A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.,* **2017**, *16*, (1), 19-34.
[3] Paul, S.M.; Mytelka, D.S.; Dunwiddie, C.T.; Persinger, C.C.; Munos, B.H.; Lindborg, S.R.; Schacht, A.L., How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.,* **2010**, *9*, (3), 203-214.
[4] Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R.V., Data completeness—the Achilles heel of drug-target networks. *Nat Biotech,* **2008**, *26*, (9), 983-984.
[5] Lavecchia, A.; Giovanni, C.D., Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.,* **2013**, *20*, (23), 2839-2860.
[6] Bowes, J.; Brown, A.J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S., Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discov.,* **2012**, *11*, (12), 909-922.
[7] Urban, L. In *4th Annual Predictive Toxicology Summit*: London, UK, **2012**.
[8] Wang, X.Y.; Greene, N., Comparing Measures of Promiscuity and Exploring Their Relationship to Toxicity. *Molecular Informatics,* **2012**, *31*, (2), 145-159.
[9] Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S., Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief. Bioinform.,* **2014**, *15*, (5), 734-747.
[10] Schomburg, K.T.; Rarey, M., What is the potential of structure-based target prediction methods? *Future Med. Chem.,* **2014**, *6*, (18), 1987-1989.
[11] Somody, J.C.; MacKinnon, S.S.; Windemuth, A., Structural coverage of the proteome for pharmaceutical applications. *Drug Discov. Today,* **2017**.
[12] Xie, L.; Bourne, P.E., A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics,* **2007**, *8 Suppl 4*, S9.

[13] Xie, L.; Xie, L.; Bourne, P.E., A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics,* **2009**, *25*, (12), i305-312.
[14] Hu, G.; Gao, J.; Wang, K.; Mizianty, M.J.; Ruan, J.; Kurgan, L., Finding protein targets for small biologically relevant ligands across fold space using inverse ligand binding predictions. *Structure,* **2012**, *20*, (11), 1815-1822.
[15] Brylinski, M.; Feinstein, W.P., eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput.-Aided Mol. Des.,* **2013**, *27*, (6), 551-567.
[16] Feinstein, W.P.; Brylinski, M., eFindSite: Enhanced Fingerprint-Based Virtual Screening Against Predicted Ligand Binding Sites in Protein Models. *Molecular Informatics,* **2014**, *33*, (2), 135-150.
[17] Litfin, T.; Zhou, Y.; Yang, Y., SPOT-ligand 2: improving structure-based virtual screening by binding-homology search on an expanded structural template library. *Bioinformatics,* **2017**, *33*, (8), 1238-1240.
[18] Mizianty, M.J.; Fan, X.; Yan, J.; Chalmers, E.; Woloschuk, C.; Joachimiak, A.; Kurgan, L., Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr. D Biol. Crystallogr.,* **2014**, *70*, (Pt 11), 2781-2793.
[19] Liu, T.; Altman, R.B., Relating Essential Proteins to Drug Side-Effects Using Canonical Component Analysis: A Structure-Based Approach. *J. Chem. Inf. Model.,* **2015**, *55*, (7), 1483-1494.
[20] Zhang, Q.C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C.A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; Maniatis, T.; Califano, A.; Honig, B., Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature,* **2012**, *490*, (7421), 556-560.
[21] Mitchell, J.B., 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*, (6), 1617-1622.
[22] Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E., Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, (2), 391-405.
[23] Klabunde, T., Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.,* **2007**, *152*, (1), 5-7.
[24] Raju, T.N.K., The Nobel Chronicles. *The Lancet,* **2000**, *355*, (9208), 1022.
[25] Pahikkala, T.; Airola, A.; Pietila, S.; Shakyawar, S.; Szwajda, A.; Tang, J.; Aittokallio, T., Toward more realistic drug-target interaction predictions. *Brief. Bioinform.,* **2015**, *16*, (2), 325-337.
[26] Mousavian, Z.; Masoudi-Nejad, A., Drug-target interaction prediction via chemogenomic space: learning-based methods. *Expert Opin. Drug Metab. Toxicol.,* **2014**, *10*, (9), 1273-1287.
[27] Lavecchia, A., Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today,* **2015**, *20*, (3), 318-331.
[28] Cichonska, A.; Rousu, J.; Aittokallio, T., Identification of drug candidates and repurposing opportunities through compound-target interaction networks. *Expert Opin Drug Discov,* **2015**, *10*, (12), 1333-1345.
[29] Lavecchia, A.; Cerchia, C., In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov. Today,* **2016**, *21*, (2), 288-298.
[30] Glaab, E., Building a virtual ligand screening pipeline using free software: a survey. *Brief. Bioinform.,* **2016**, *17*, (2), 352-366.
[31] Vilar, S.; Hripcsak, G., The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. *Brief. Bioinform.,* **2016**.
[32] Chen, X.; Yan, C.C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y., Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.,* **2016**, *17*, (4), 696-712.
[33] Hart, T.; Xie, L., Providing data science support for systems pharmacology and its implications to drug discovery. *Expert Opin Drug Discov,* **2016**, *11*, (3), 241-256.
[34] Fang, J.; Liu, C.; Wang, Q.; Lin, P.; Cheng, F., In silico polypharmacology of natural products. *Brief. Bioinform.,* **2017**, bbx045-bbx045.
[35] Lotfi Shahreza, M.; Ghadiri, N.; Mousavi, S.R.; Varshosaz, J.; Green, J.R., A review of network-based approaches to drug repositioning. *Brief. Bioinform.,* **2017**, bbx017-bbx017.
[36] Hao, M.; Bryant, S.H.; Wang, Y., Open-source chemogenomic data-driven algorithms for predicting drug–target interactions. *Brief. Bioinform.,* **2018**, bby010-bby010.
[37] Ezzat, A.; Wu, M.; Li, X.-L.; Kwoh, C.-K., Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief. Bioinform.,* **2018**, bby002-bby002.
[38] Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K., Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.,* **2007**, *25*, (2), 197-206.
[39] Keiser, M.J.; Setola, V.; Irwin, J.J.; Laggner, C.; Abbas, A.I.; Hufeisen, S.J.; Jensen, N.H.; Kuijer, M.B.; Matos, R.C.; Tran, T.B.; Whaley, R.; Glennon, R.A.; Hert, J.; Thomas, K.L.; Edwards, D.D.; Shoichet, B.K.; Roth, B.L., Predicting new molecular targets for known drugs. *Nature,* **2009**, *462*, (7270), 175-181.
[40] Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M., Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics,* **2008**, *24*, (13), i232-240.
[41] Campillos, M.; Kuhn, M.; Gavin, A.C.; Jensen, L.J.; Bork, P., Drug target identification using side-effect similarity. *Science,* **2008**, *321*, (5886), 263-266.
[42] Nagamine, N.; Shirakawa, T.; Minato, Y.; Torii, K.; Kobayashi, H.; Imoto, M.; Sakakibara, Y., Integrating statistical predictions and experimental verifications for

enhancing protein-chemical interaction predictions in virtual screening. *PLoS Comput. Biol.,* **2009**, *5*, (6), e1000397.
[43] Sakakibara, Y.; Hachiya, T.; Uchida, M.; Nagamine, N.; Sugawara, Y.; Yokota, M.; Nakamura, M.; Popendorf, K.; Komori, T.; Sato, K., COPICAT: a software system for predicting interactions between proteins and chemical compounds. *Bioinformatics,* **2012**, *28*, (5), 745-746.
[44] Bleakley, K.; Yamanishi, Y., Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics,* **2009**, *25*, (18), 2397-2403.
[45] Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S., Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics,* **2010**, *26*, (12), i246-254.
[46] Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y., Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.,* **2011**, *7*, 472.
[47] van Laarhoven, T.; Nabuurs, S.B.; Marchiori, E., Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics,* **2011**, *27*, (21), 3036-3043.
[48] Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y., Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.,* **2012**, *8*, (5), e1002503.
[49] Gonen, M., Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics,* **2012**, *28*, (18), 2304-2310.
[50] Takarabe, M.; Kotera, M.; Nishimura, Y.; Goto, S.; Yamanishi, Y., Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics,* **2012**, *28*, (18), i611-i618.
[51] Cao, D.-S.; Liu, S.; Xu, Q.-S.; Lu, H.-M.; Huang, J.-H.; Hu, Q.-N.; Liang, Y.-Z., Large-scale prediction of drug–target interactions using protein sequences and drug topological structures. *Anal. Chim. Acta,* **2012**, *752*, 1-10.
[52] Mei, J.P.; Kwoh, C.K.; Yang, P.; Li, X.L.; Zheng, J., Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics,* **2013**, *29*, (2), 238-245.
[53] Cheng, F.; Li, W.; Wu, Z.; Wang, X.; Zhang, C.; Li, J.; Liu, G.; Tang, Y., Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J. Chem. Inf. Model.,* **2013**, *53*, (4), 753-762.
[54] Alaimo, S.; Pulvirenti, A.; Giugno, R.; Ferro, A., Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics,* **2013**, *29*, (16), 2004-2008.
[55] Koutsoukas, A.; Lowe, R.; Kalantarmotamedi, Y.; Mussa, H.Y.; Klaffke, W.; Mitchell, J.B.; Glen, R.C.; Bender, A., In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naive Bayes and Parzen-Rosenblatt window. *J. Chem. Inf. Model.,* **2013**, *53*, (8), 1957-1966.
[56] Yamanishi, Y.; Kotera, M.; Moriya, Y.; Sawada, R.; Kanehisa, M.; Goto, S., DINIES: drug-target interaction network inference engine based on supervised analysis. *Nucleic Acids Res.,* **2014**, *42*, (Web Server issue), W39-45.
[57] Shi, J.-Y.; Yiu, S.-M.; Li, Y.; Leung, H.C.M.; Chin, F.Y.L., Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods,* **2015**, *83*, 98-104.
[58] Liu, H.; Sun, J.; Guan, J.; Zheng, J.; Zhou, S., Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics,* **2015**, *31*, (12), i221-229.
[59] Seal, A.; Ahn, Y.Y.; Wild, D.J., Optimizing drug-target interaction prediction based on random walk on heterogeneous networks. *J. Cheminform.,* **2015**, *7*, 40.
[60] Kuang, Q.; Xu, X.; Li, R.; Dong, Y.; Li, Y.; Huang, Z.; Li, Y.; Li, M., An eigenvalue transformation technique for predicting drug-target interaction. *Sci. Rep.,* **2015**, *5*, 13867.
[61] Hao, M.; Wang, Y.; Bryant, S.H., Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique. *Anal. Chim. Acta,* **2016**, *909*, 41-50.
[62] Jamali, A.A.; Ferdousi, R.; Razzaghi, S.; Li, J.; Safdari, R.; Ebrahimie, E., DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discov. Today,* **2016**, *21*, (5), 718-724.
[63] Liu, Y.; Wu, M.; Miao, C.; Zhao, P.; Li, X.L., Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *PLoS Comput. Biol.,* **2016**, *12*, (2), e1004760.
[64] Wu, Z.; Cheng, F.; Li, J.; Li, W.; Liu, G.; Tang, Y., SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. *Brief. Bioinform.,* **2016**.
[65] Ba-Alawi, W.; Soufan, O.; Essack, M.; Kalnis, P.; Bajic, V.B., DASPfind: new efficient method to predict drug-target interactions. *J. Cheminform.,* **2016**, *8*, 15.
[66] Yuan, Q.; Gao, J.; Wu, D.; Zhang, S.; Mamitsuka, H.; Zhu, S., DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics,* **2016**, *32*, (12), i18-i27.
[67] Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H., Deep-Learning-Based Drug–Target Interaction Prediction. *J. Proteome Res.,* **2017**, *16*, (4), 1401-1409.
[68] Ezzat, A.; Wu, M.; Li, X.-L.; Kwoh, C.-K., Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods,* **2017**.
[69] Peón, A.; Naulaerts, S.; Ballester, P.J., Predicting the Reliability of Drug-target Interaction Predictions with Maximum Coverage of Target Space. *Sci. Rep.,* **2017**, *7*, (1), 3820.
[70] Peng, L.; Zhu, W.; Liao, B.; Duan, Y.; Chen, M.; Chen, Y.; Yang, J., Screening drug-target interactions with positive-unlabeled learning. *Sci. Rep.,* **2017**, *7*, (1), 8087.
[71] Li, Z.; Han, P.; You, Z.-H.; Li, X.; Zhang, Y.; Yu, H.; Nie, R.; Chen, X., In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Sci. Rep.,* **2017**, *7*, (1), 11174.

[72] Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J., A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications,* **2017**, *8*, (1), 573.

[73] Fang, J.; Wu, Z.; Cai, C.; Wang, Q.; Tang, Y.; Cheng, F., 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*, (11), 2657-2671.

[74] Wu, Z.; Lu, W.; Yu, W.; Wang, T.; Li, W.; Liu, G.; Zhang, H.; Pang, X.; Huang, J.; Liu, M.; Cheng, F.; Tang, Y., Quantitative and systems pharmacology 2. In silico polypharmacology of G protein-coupled receptor ligands via network-based approaches. *Pharmacol. Res.,* **2018**, *129*, 400-413.

[75] Rayhan, F.; Ahmed, S.; Shatabda, S.; Farid, D.M.; Mousavian, Z.; Dehzangi, A.; Rahman, M.S., iDTI-ESBoost: Identification of Drug Target Interaction Using Evolutionary and Structural Features with Boosting. *Sci. Rep.,* **2017**, *7*, (1), 17731.

[76] Coordinators, N.R., Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.,* **2017**, *45*, (D1), D12-D17.

[77] *2017 Journal Citation Reports*®. Clarivate Analytics, **2017**.

[78] Roth, B.L.; Lopez, E.; Patel, S.; Kroeze, W.K., The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist,* **2000**, *6*, (4), 252-262.

[79] Schomburg, I.; Hofmann, O.; Baensch, C.; Chang, A.; Schomburg, D., Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene Funct. Dis.,* **2000**, *1*, (3-4), 109-118.

[80] Schomburg, I.; Chang, A.; Schomburg, D., BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.,* **2002**, *30*, (1), 47-49.

[81] Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D., BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.,* **2004**, *32*, (suppl_1), D431-D433.

[82] Barthelmes, J.; Ebeling, C.; Chang, A.; Schomburg, I.; Schomburg, D., BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.,* **2007**, *35*, (suppl_1), D511-D514.

[83] Chang, A.; Scheer, M.; Grote, A.; Schomburg, I.; Schomburg, D., BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res.,* **2008**, *37*, (suppl_1), D588-D592.

[84] Scheer, M.; Grote, A.; Chang, A.; Schomburg, I.; Munaretto, C.; Rother, M.; Söhngen, C.; Stelzer, M.; Thiele, J.; Schomburg, D., BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.,* **2010**, *39*, (suppl_1), D670-D676.

[85] Schomburg, I.; Chang, A.; Placzek, S.; Söhngen, C.; Rother, M.; Lang, M.; Munaretto, C.; Ulas, S.; Stelzer, M.; Grote, A., BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.,* **2012**, *41*, (D1), D764-D772.

[86] Chang, A.; Schomburg, I.; Placzek, S.; Jeske, L.; Ulbrich, M.; Xiao, M.; Sensen, C.W.; Schomburg, D., BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.,* **2014**, *43*, (D1), D439-D446.

[87] Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D., BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.,* **2017**, *45*, (D1), D380-D388.

[88] Chen, X.; Liu, M.; Gilson, M.K., BindingDB: a web-accessible molecular recognition database. *Combinatorial Chem. High Throughput Screening,* **2001**, *4*, (8), 719-725.

[89] Chen, X.; Lin, Y.; Liu, M.; Gilson, M.K., The Binding Database: data management and interface design. *Bioinformatics,* **2002**, *18*, (1), 130-139.

[90] Chen, X.; Lin, Y.; Gilson, M.K., The binding database: overview and user's guide. *Biopolymers,* **2002**, *61*, (2), 127-141.

[91] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K., BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.,* **2007**, *35*, (Database issue), D198-201.

[92] Gilson, M.K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J., BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.,* **2016**, *44*, (D1), D1045-D1053.

[93] Chen, X.; Ji, Z.L.; Chen, Y.Z., TTD: therapeutic target database. *Nucleic Acids Res.,* **2002**, *30*, (1), 412-415.

[94] Zhu, F.; Han, B.; Kumar, P.; Liu, X.; Ma, X.; Wei, X.; Huang, L.; Guo, Y.; Han, L.; Zheng, C.; Chen, Y., Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.,* **2010**, *38*, (Database issue), D787-791.

[95] Zhu, F.; Shi, Z.; Qin, C.; Tao, L.; Liu, X.; Xu, F.; Zhang, L.; Song, Y.; Liu, X.; Zhang, J.; Han, B.; Zhang, P.; Chen, Y., Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.,* **2012**, *40*, (Database issue), D1128-1136.

[96] Qin, C.; Zhang, C.; Zhu, F.; Xu, F.; Chen, S.Y.; Zhang, P.; Li, Y.H.; Yang, S.Y.; Wei, Y.Q.; Tao, L., Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.,* **2013**, *42*, (D1), D1118-D1123.

[97] Yang, H.; Qin, C.; Li, Y.H.; Tao, L.; Zhou, J.; Yu, C.Y.; Xu, F.; Chen, Z.; Zhu, F.; Chen, Y.Z., Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.,* **2016**, *44*, (D1), D1069-D1074.

[98] Li, Y.H.; Yu, C.Y.; Li, X.X.; Zhang, P.; Tang, J.; Yang, Q.; Fu, T.; Zhang, X.; Cui, X.; Tu, G.; Zhang, Y.; Li, S.; Yang, F.; Sun, Q.; Qin, C.; Zeng, X.; Chen, Z.; Chen, Y.Z.; Zhu, F., Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.,* **2018**, *46*, (D1), D1121-D1127.

[99] Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.;

Hirakawa, M., From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.,* **2006**, *34*, (suppl_1), D354-D357.
[100] Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y., KEGG for linking genomes to life and the environment. *Nucleic Acids Res.,* **2008**, *36*, (suppl_1), D480-D484.
[101] Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M., KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.,* **2009**, *38*, (suppl_1), D355-D360.
[102] Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M., KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.,* **2011**, *40*, (D1), D109-D114.
[103] Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M., Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.,* **2013**, *42*, (D1), D199-D205.
[104] Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M., KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.,* **2016**, *44*, (D1), D457-D462.
[105] Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K., KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.,* **2017**, *45*, (D1), D353-D361.
[106] Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J., DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.,* **2006**, *34*, (suppl_1), D668-D672.
[107] Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M., DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.,* **2008**, *36*, (suppl_1), D901-D906.
[108] Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.C.; Wishart, D.S., DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.,* **2011**, *39*, (suppl_1), D1035-D1041.
[109] Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z.T.; Han, B.; Zhou, Y.; Wishart, D.S., DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.,* **2014**, *42*, (Database issue), D1091-1097.
[110] Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M., DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.,* **2018**, *46*, (D1), D1074-D1082.
[111] Okuno, Y.; Yang, J.; Taneishi, K.; Yabuuchi, H.; Tsujimoto, G., GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.,* **2006**, *34*, (suppl_1), D673-D677.
[112] Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C., GLIDA: GPCR—ligand database for chemical genomics drug discovery—database and tools update. *Nucleic Acids Res.,* **2008**, *36*, (suppl_1), D907-D912.
[113] Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J.; Schneider, R.; Skoblo, R.; Russell, R.B.; Bourne, P.E.; Bork, P.; Preissner, R., SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.,* **2008**, *36*, (suppl_1), D919-D922.
[114] Hecker, N.; Ahmed, J.; von Eichborn, J.; Dunkel, M.; Macha, K.; Eckert, A.; Gilson, M.K.; Bourne, P.E.; Preissner, R., SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.,* **2012**, *40*, (Database issue), D1113-1117.
[115] Kuhn, M.; von Mering, C.; Campillos, M.; Jensen, L.J.; Bork, P., STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.,* **2008**, *36*, (suppl_1), D684-D688.
[116] Kuhn, M.; Szklarczyk, D.; Franceschini, A.; Campillos, M.; von Mering, C.; Jensen, L.J.; Beyer, A.; Bork, P., STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.,* **2010**, *38*, (Database issue), D552-556.
[117] Kuhn, M.; Szklarczyk, D.; Franceschini, A.; von Mering, C.; Jensen, L.J.; Bork, P., STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Res.,* **2012**, *40*, (D1), D876-D880.
[118] Kuhn, M.; Szklarczyk, D.; Pletscher-Frankild, S.; Blicher, T.H.; von Mering, C.; Jensen, L.J.; Bork, P., STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.,* **2014**, *42*, (Database issue), D401-407.
[119] Szklarczyk, D.; Santos, A.; von Mering, C.; Jensen, L.J.; Bork, P.; Kuhn, M., STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.,* **2016**, *44*, (D1), D380-D384.
[120] Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J.P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.,* **2012**, *40*, (Database issue), D1100-1107.
[121] Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Kruger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J.P., The ChEMBL bioactivity database: an update. *Nucleic Acids Res.,* **2014**, *42*, (Database issue), D1083-1090.
[122] Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J.P., ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.,* **2015**, *43*, (W1), W612-W620.
[123] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis,

L.J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M.P.; Overington, J.P.; Papadatos, G.; Smit, I.; Leach, A.R., The ChEMBL database in 2017. *Nucleic Acids Res.,* **2017**, *45*, (D1), D945-D954.

[124] Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.-J.; Lecchini, S.; Jacoby, E., An Ontology for Pharmaceutical Ligands and Its Application for in Silico Screening and Library Design. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, (4), 947-955.

[125] Southan, C.; Várkonyi, P.; Muresan, S., Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminform.,* **2009**, *1*, (1), 10.

[126] Euskirchen, G., Integrative approaches in molecular medicine. *Pharmacogenomics,* **2004**, *5*, (4), 357-360.

[127] Warr, W.A., ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput.-Aided Mol. Des.,* **2009**, *23*, (4), 195-198.

[128] Bender, A., Databases: Compound bioactivities go public. *Nat. Chem. Biol.,* **2010**, *6*, (5), 309-309.

[129] Zhou, H.; Gao, M.; Skolnick, J., Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci. Rep.,* **2015**, *5*, 11090.

[130] Chartier, M.; Morency, L.-P.; Zylber, M.I.; Najmanovich, R.J., Large-scale detection of drug off-targets: hypotheses for drug repurposing and understanding side-effects. *BMC Pharmacology and Toxicology,* **2017**, *18*, (1), 18.

[131] Brylinski, M., Aromatic interactions at the ligand–protein interface: Implications for the development of docking scoring functions. *Chem. Biol. Drug Des.,* **2017**, 1-11.

[132] Tatonetti, N.P.; Ye, P.P.; Daneshjou, R.; Altman, R.B., Data-Driven Prediction of Drug Effects and Interactions. *Sci. Transl. Med.,* **2012**, *4*, (125), 125ra131-125ra131.

[133] Schomburg, K.T.; Rarey, M., Benchmark Data Sets for Structure-Based Computational Target Prediction. *J. Chem. Inf. Model.,* **2014**, *54*, (8), 2261-2274.

[134] Wishart, D.; Arndt, D.; Pon, A.; Sajed, T.; Guo, A.C.; Djoumbou, Y.; Knox, C.; Wilson, M.; Liang, Y.; Grant, J.; Liu, Y.; Goldansaz, S.A.; Rappaport, S.M., T3DB: the toxic exposome database. *Nucleic Acids Res.,* **2015**, *43*, (D1), D928-D934.

[135] Legehar, A.; Xhaard, H.; Ghemtio, L., IDAAPM: integrated database of ADMET and adverse effects of predictive modeling based on FDA approved drug data. *J. Cheminform.,* **2016**, *8*, (1), 33.

[136] Shameer, K.; Glicksberg, B.S.; Hodos, R.; Johnson, K.W.; Badgeley, M.A.; Readhead, B.; Tomlinson, M.S.; O'Connor, T.; Miotto, R.; Kidd, B.A.; Chen, R.; Ma'ayan, A.; Dudley, J.T., Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. *Brief. Bioinform.,* **2017**, bbw136-bbw136.

[137] Russ, A.P.; Lampel, S., The druggable genome: an update. *Drug Discov. Today,* **2005**, *10*, (23–24), 1607-1610.

[138] Rask-Andersen, M.; Masuram, S.; Schioth, H.B., The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu. Rev. Pharmacol. Toxicol.,* **2014**, *54*, 9-26.

[139] Hu, G.; Wu, Z.; Wang, K.; Uversky, V.N.; Kurgan, L., Untapped Potential of Disordered Proteins in Current Druggable Human Proteome. *Curr. Drug Targets,* **2016**, *17*, (10), 1198-1205.

[140] Paolini, G.V.; Shapland, R.H.B.; van Hoorn, W.P.; Mason, J.S.; Hopkins, A.L., Global mapping of pharmacological space. *Nat Biotech,* **2006**, *24*, (7), 805-815.

[141] Hopkins, A.L., Drug discovery: Predicting promiscuity. *Nature,* **2009**, *462*, (7270), 167-168.

[142] Anighoro, A.; Bajorath, J.; Rastelli, G., Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.,* **2014**, *57*, (19), 7874-7887.

[143] Chong, C.R.; Sullivan, D.J., New uses for old drugs. *Nature,* **2007**, *448*, (7154), 645-646.

[144] Haupt, V.J.; Schroeder, M., Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief. Bioinform.,* **2011**, *12*, (4), 312-326.

[145] Hu, Y.; Bajorath, J., Compound promiscuity: what can we learn from current data? *Drug Discov. Today,* **2013**, *18*, (13-14), 644-650.

[146] Lounkine, E.; Keiser, M.J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J.L.; Lavan, P.; Weber, E.; Doak, A.K.; Cote, S.; Shoichet, B.K.; Urban, L., Large-scale prediction and testing of drug activity on side-effect targets. *Nature,* **2012**, *486*, (7403), 361-367.

[147] Tarcsay, Á.; Keserű, G.M., Contributions of Molecular Properties to Drug Promiscuity. *J. Med. Chem.,* **2013**, *56*, (5), 1789-1795.

[148] Hu, G.; Wang, K.; Groenendyk, J.; Barakat, K.; Mizianty, M.J.; Ruan, J.; Michalak, M.; Kurgan, L., Human structural proteome-wide characterization of Cyclosporine A targets. *Bioinformatics,* **2014**, *30*, (24), 3561-3566.

[149] Jasial, S.; Hu, Y.; Bajorath, J., Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS One,* **2016**, *11*, (4), e0153873.

[150] Davis, A.P.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; King, B.L.; McMorran, R.; Wiegers, J.; Wiegers, T.C.; Mattingly, C.J., The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.,* **2017**, *45*, (D1), D972-D978.

[151] Wang, Y.; Bryant, S.H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B.A.; Thiessen, P.A.; He, S.; Zhang, J., PubChem BioAssay: 2017 update. *Nucleic Acids Res.,* **2017**, *45*, (D1), D955-D963.

[152] Anastassiadis, T.; Deacon, S.W.; Devarajan, K.; Ma, H.; Peterson, J.R., Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat Biotech,* **2011**, *29*, (11), 1039-1045.

[153] Davis, M.I.; Hunt, J.P.; Herrgard, S.; Ciceri, P.; Wodicka, L.M.; Pallares, G.; Hocker, M.; Treiber, D.K.; Zarrinkar, P.P., Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotech,* **2011**, *29*, (11), 1046-1051.

[154] 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*, (11-12), 881-897.

[155] Ursu, O.; Holmes, J.; Knockel, J.; Bologa, C.G.; Yang, J.J.; Mathias, S.L.; Nelson, S.J.; Oprea, T.I., DrugCentral: online drug compendium. *Nucleic Acids Res.,* **2017**, *45*, (D1), D932-D939.

[156] Nguyen, D.-T.; Mathias, S.; Bologa, C.; Brunak, S.; Fernandez, N.; Gaulton, A.; Hersey, A.; Holmes, J.; Jensen, L.J.; Karlsson, A.; Liu, G.; Ma'ayan, A.; Mandava, G.; Mani, S.; Mehta, S.; Overington, J.; Patel, J.; Rouillard, A.D.; Schürer, S.; Sheils, T.; Simeonov, A.; Sklar, L.A.; Southall, N.; Ursu, O.; Vidovic, D.; Waller, A.; Yang, J.; Jadhav, A.; Oprea, T.I.; Guha, R., Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.,* **2017**, *45*, (D1), D995-D1002.

[157] Whirl-Carrillo, M.; McDonagh, E.M.; Hebert, J.M.; Gong, L.; Sangkuhl, K.; Thorn, C.F.; Altman, R.B.; Klein, T.E., Pharmacogenomics Knowledge for Personalized Medicine. *Clin. Pharmacol. Ther.,* **2012**, *92*, (4), 414-417.

[158] Griffith, M.; Griffith, O.L.; Coffman, A.C.; Weible, J.V.; McMichael, J.F.; Spies, N.C.; Koval, J.; Das, I.; Callaway, M.B.; Eldred, J.M.; Miller, C.A.; Subramanian, J.; Govindan, R.; Kumar, R.D.; Bose, R.; Ding, L.; Walker, J.R.; Larson, D.E.; Dooling, D.J.; Smith, S.M.; Ley, T.J.; Mardis, E.R.; Wilson, R.K., DGIdb: mining the druggable genome. *Nat Meth,* **2013**, *10*, (12), 1209-1210.

[159] Wagner, A.H.; Coffman, A.C.; Ainscough, B.J.; Spies, N.C.; Skidmore, Z.L.; Campbell, K.M.; Krysiak, K.; Pan, D.; McMichael, J.F.; Eldred, J.M.; Walker, J.R.; Wilson, R.K.; Mardis, E.R.; Griffith, M.; Griffith, O.L., DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res.,* **2016**, *44*, (D1), D1036-D1044.

[160] Roider, H.G.; Pavlova, N.; Kirov, I.; Slavov, S.; Slavov, T.; Uzunov, Z.; Weiss, B., Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinformatics,* **2014**, *15*, 68.

[161] Pawson, A.J.; Sharman, J.L.; Benson, H.E.; Faccenda, E.; Alexander, S.P.H.; Buneman, O.P.; Davenport, A.P.; McGrath, J.C.; Peters, J.A.; Southan, C.; Spedding, M.; Yu, W.; Harmar, A.J., The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res.,* **2014**, *42*, (D1), D1098-D1106.

[162] Southan, C.; Sharman, J.L.; Benson, H.E.; Faccenda, E.; Pawson, A.J.; Alexander, Stephen P.H.; Buneman, O.P.; Davenport, A.P.; McGrath, J.C.; Peters, J.A.; Spedding, M.; Catterall, W.A.; Fabbro, D.; Davies, J.A., The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.,* **2016**, *44*, (D1), D1054-D1068.

[163] Koscielny, G.; An, P.; Carvalho-Silva, D.; Cham, J.A.; Fumis, L.; Gasparyan, R.; Hasan, S.; Karamanis, N.; Maguire, M.; Papa, E.; Pierleoni, A.; Pignatelli, M.; Platt, T.; Rowland, F.; Wankar, P.; Bento, A.P.; Burdett, T.; Fabregat, A.; Forbes, S.; Gaulton, A.; Gonzalez, C.Y.; Hermjakob, H.; Hersey, A.; Jupe, S.; Kafkas, Ş.; Keays, M.; Leroy, C.; Lopez, F.-J.; Magarinos, M.P.; Malone, J.; McEntyre, J.; Munoz-Pomer Fuentes, A.; O'Donovan, C.; Papatheodorou, I.; Parkinson, H.; Palka, B.; Paschall, J.; Petryszak, R.; Pratanwanich, N.; Sarntivijal, S.; Saunders, G.; Sidiropoulos, K.; Smith, T.; Sondka, Z.; Stegle, O.; Tang, Y.A.; Turner, E.; Vaughan, B.; Vrousgou, O.; Watkins, X.; Martin, M.-J.; Sanseau, P.; Vamathevan, J.; Birney, E.; Barrett, J.; Dunham, I., Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.,* **2017**, *45*, (D1), D985-D994.

[164] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E., The Protein Data Bank. *Nucleic Acids Res.,* **2000**, *28*, (1), 235-242.

[165] Rose, P.W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A.R.; Christie, C.H.; Costanzo, L.D.; Duarte, J.M.; Dutta, S.; Feng, Z.; Green, R.K.; Goodsell, D.S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A.S.; Shao, C.; Tao, Y.-P.; Valasatava, Y.; Voigt, M.; Westbrook, J.D.; Woo, J.; Yang, H.; Young, J.Y.; Zardecki, C.; Berman, H.M.; Burley, S.K., The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.,* **2017**, *45*, (D1), D271-D281.

[166] Yang, J.; Roy, A.; Zhang, Y., BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.,* **2013**, *41*, (D1), D1096-D1103.

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

[168] Higueruelo, A.P.; Schreyer, A.; Bickerton, G.R.J.; Pitt, W.R.; Groom, C.R.; Blundell, T.L., Atomic Interactions and Profile of Small Molecules Disrupting Protein–Protein Interfaces: the TIMBAL Database. *Chem. Biol. Drug Des.,* **2009**, *74*, (5), 457-467.

[169] Higueruelo, A.P.; Jubb, H.; Blundell, T.L., TIMBAL v2: update of a database holding small molecules modulating protein–protein interactions. *Database,* **2013**, *2013*, bat039-bat039.

[170] Bourgeas, R.; Basse, M.-J.; Morelli, X.; Roche, P., Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database. *PLoS One,* **2010**, *5*, (3), e9598.

[171] Basse, M.J.; Betzi, S.; Bourgeas, R.; Bouzidi, S.; Chetrit, B.; Hamon, V.; Morelli, X.; Roche, P., 2P2Idb: a structural database dedicated to orthosteric modulation of protein–protein interactions. *Nucleic Acids Res.,* **2013**, *41*, (D1), D824-D827.

[172] Basse, M.-J.; Betzi, S.; Morelli, X.; Roche, P., 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein–protein interactions. *Database,* **2016**, *2016*, baw007-baw007.

[173] Labbé, C.M.; Laconde, G.; Kuenemann, M.A.; Villoutreix, B.O.; Sperandio, O., iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. *Drug Discov. Today,* **2013**, *18*, (19), 958-968.

[174] Labbé, C.M.; Kuenemann, M.A.; Zarzycka, B.; Vriend, G.; Nicolaes, G.A.F.; Lagorce, D.; Miteva, M.A.;

Villoutreix, B.O.; Sperandio, O., iPPI-DB: an online database of modulators of protein–protein interactions. *Nucleic Acids Res.,* **2016**, *44*, (D1), D542-D547.
[175] Liu, Y.; Hu, B.; Fu, C.; Chen, X., DCDB: Drug combination database. *Bioinformatics,* **2010**, *26*, (4), 587-588.
[176] Liu, Y.; Wei, Q.; Yu, G.; Gai, W.; Li, Y.; Chen, X., DCDB 2.0: a major update of the drug combination database. *Database,* **2014**, *2014*, bau124-bau124.
[177] Juan-Blanco, T.; Duran-Frigola, M.; Aloy, P., IntSide: a web server for the chemical and biological examination of drug side effects. *Bioinformatics,* **2015**, *31*, (4), 612-613.
[178] Ahmed, J.; Meinel, T.; Dunkel, M.; Murgueitio, M.S.; Adams, R.; Blasse, C.; Eckert, A.; Preissner, S.; Preissner, R., CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res.,* **2011**, *39*, (suppl_1), D960-D967.
[179] Gohlke, B.-O.; Nickel, J.; Otto, R.; Dunkel, M.; Preissner, R., CancerResource—updated database of cancer-relevant proteins, mutations and interacting drugs. *Nucleic Acids Res.,* **2016**, *44*, (D1), D932-D937.
[180] Halling-Brown, M.D.; Bulusu, K.C.; Patel, M.; Tym, J.E.; Al-Lazikani, B., canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res.,* **2012**, *40*, (D1), D947-D956.
[181] Bulusu, K.C.; Tym, J.E.; Coker, E.A.; Schierz, A.C.; Al-Lazikani, B., canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.,* **2014**, *42*, (D1), D1040-D1047.
[182] Tym, J.E.; Mitsopoulos, C.; Coker, E.A.; Razaz, P.; Schierz, A.C.; Antolin, A.A.; Al-Lazikani, B., canSAR: an updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.,* **2016**, *44*, (D1), D938-D943.
[183] Siramshetty, V.B.; Nickel, J.; Omieczynski, C.; Gohlke, B.-O.; Drwal, M.N.; Preissner, R., WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic Acids Res.,* **2016**, *44*, (D1), D1080-D1086.
[184] Chan, W.K.B.; Zhang, H.; Yang, J.; Brender, J.R.; Hur, J.; Özgür, A.; Zhang, Y., GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics,* **2015**, *31*, (18), 3035-3042.
[185] He, Z.; Zhang, J.; Shi, X.-H.; Hu, L.-L.; Kong, X.; Cai, Y.-D.; Chou, K.-C., Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features. *PLoS One,* **2010**, *5*, (3), e9603.
[186] Xia, Z.; Wu, L.-Y.; Zhou, X.; Wong, S.T.C., Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.,* **2010**, *4*, (2), S6.
[187] Yu, W.; Jiang, Z.; Wang, J.; Tao, R., Using feature selection technique for drug-target interaction networks prediction. *Curr. Med. Chem.,* **2011**, *18*, (36), 5687-5693.
[188] Chen, X.; Liu, M.-X.; Yan, G.-Y., Drug-target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems,* **2012**, *8*, (7), 1970-1978.
[189] Chen, H.; Zhang, Z., A Semi-Supervised Method for Drug-Target Interaction Prediction with Consistency in Networks. *PLoS One,* **2013**, *8*, (5), e62975.
[190] van Laarhoven, T.; Marchiori, E., Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. *PLoS One,* **2013**, *8*, (6), e66952.
[191] Yu, W.; Yan, Y.; Liu, Q.; Wang, J.; Jiang, Z., Predicting drug–target interaction networks of human diseases based on multiple feature information. *Pharmacogenomics,* **2013**, *14*, (14), 1701-1707.
[192] Cao, D.-S.; Zhang, L.-X.; Tan, G.-S.; Xiang, Z.; Zeng, W.-B.; Xu, Q.-S.; Chen, A.F., Computational Prediction of Drug□Target Interactions Using Chemical, Biological, and Network Features. *Molecular Informatics,* **2014**, *33*, (10), 669-681.
[193] Huang, Y.-A.; You, Z.-H.; Chen, X., A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences. *Curr. Protein Peptide Sci.,* **2016**.
[194] Nascimento, A.C.A.; Prudêncio, R.B.C.; Costa, I.G., A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics,* **2016**, *17*, (1), 46.
[195] Shi, J.-Y.; Li, J.-X.; Lu, H.-M., Predicting existing targets for new drugs base on strategies for missing interactions. *BMC Bioinformatics,* **2016**, *17*, (8), 282.
[196] Wang, L.; You, Z.-H.; Chen, X.; Yan, X.; Liu, G.; Zhang, W., RFDT: A Rotation Forest-based Predictor for Predicting Drug-Target Interactions using Drug Structure and Protein Sequence Information. *Curr. Protein Peptide Sci.,* **2016**.
[197] Yan, X.-Y.; Zhang, S.-W.; Zhang, S.-Y., Prediction of drug-target interaction by label propagation with mutual interaction information derived from heterogeneous network. *Molecular BioSystems,* **2016**, *12*, (2), 520-531.
[198] Buza, K.; Peška, L., Drug–target interaction prediction with Bipartite Local Models and hubness-aware regression. *Neurocomputing,* **2017**, *260*, 284-293.
[199] Keum, J.; Nam, H., SELF-BLM: Prediction of drug-target interactions via self-training SVM. *PLoS One,* **2017**, *12*, (2), e0171839.
[200] Meng, F.-R.; You, Z.-H.; Chen, X.; Zhou, Y.; An, J.-Y., Prediction of Drug–Target Interaction Networks from the Integration of Protein Sequences and Drug Chemical Structures. *Molecules,* **2017**, *22*, (7), 1119.
[201] Shen, C.; Ding, Y.; Tang, J.; Xu, X.; Guo, F., An Ameliorated Prediction of Drug–Target Interactions Based on Multi-Scale Discrete Wavelet Transform and Network Features. *International Journal of Molecular Sciences,* **2017**, *18*, (8), 1781.
[202] Zhang, J.; Zhu, M.; Chen, P.; Wang, B., DrugRPE: Random projection ensemble approach to drug-target interaction prediction. *Neurocomputing,* **2017**, *228*, (Supplement C), 256-262.
[203] Bender, A.; Jenkins, J.L.; Scheiber, J.; Sukuru, S.C.K.; Glick, M.; Davies, J.W., How similar are similarity searching methods?: a principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.,* **2009**, *49*.

[204] Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G., Molecular fingerprint similarity search in virtual screening. *Methods,* **2015**, *71*, 58-63.
[205] Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M., 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*, (39), 11853-11865.
[206] Hattori, M.; Tanaka, N.; Kanehisa, M.; Goto, S., SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.,* **2010**, *38*, (suppl_2), W652-W656.
[207] Willett, P.; Barnard, J.M.; Downs, G.M., Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.,* **1998**, *38*, (6), 983-996.
[208] Wood, T.C.; Pearson, W.R., Evolution of protein sequences and structures11Edited by J. M. Thornton. *J. Mol. Biol.,* **1999**, *291*, (4), 977-995.
[209] Baker, D.; Sali, A., Protein Structure Prediction and Structural Genomics. *Science,* **2001**, *294*, (5540), 93-96.
[210] Liu, J.; Rost, B., Target space for structural genomics revisited. *Bioinformatics,* **2002**, *18*, (7), 922-933.
[211] Ginalski, K., Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.,* **2006**, *16*, (2), 172-177.
[212] Aravind, L.; Koonin, E.V., Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches11Edited by J. M. Thornton. *J. Mol. Biol.,* **1999**, *287*, (5), 1023-1040.
[213] Wilson, C.A.; Kreychman, J.; Gerstein, M., Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores1. *J. Mol. Biol.,* **2000**, *297*, (1), 233-249.
[214] Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K.O.; Ofran, Y., Automatic prediction of protein function. *Cellular and Molecular Life Sciences CMLS,* **2003**, *60*, (12), 2637-2650.
[215] Lee, D.; Redfern, O.; Orengo, C., Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.,* **2007**, *8*, (12), 995-1005.
[216] Sangar, V.; Blankenberg, D.J.; Altman, N.; Lesk, A.M., Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics,* **2007**, *8*, (1), 294.
[217] Addou, S.; Rentzsch, R.; Lee, D.; Orengo, C.A., Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer. *J. Mol. Biol.,* **2009**, *387*, (2), 416-430.
[218] Clark, W.T.; Radivojac, P., Analysis of protein function and its prediction from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics,* **2011**, *79*, (7), 2086-2096.
[219] Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **1997**, *25*, (17), 3389-3402.
[220] Smith, T.F.; Waterman, M.S., Identification of common molecular subsequences. *J. Mol. Biol.,* **1981**, *147*, (1), 195-197.
[221] *ATC classification index with DDDs*. WHO Collaborating Centre for Drug Statistics Methodology: Oslo, Norway, **2017**.
[222] Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L.J.; Bork, P., A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.,* **2010**, *6*, 343.
[223] Kuhn, M.; Letunic, I.; Jensen, L.J.; Bork, P., The SIDER database of drugs and side effects. *Nucleic Acids Res.,* **2016**, *44*, (D1), D1075-1079.
[224] Cheng, F.; Li, W.; Wang, X.; Zhou, Y.; Wu, Z.; Shen, J.; Tang, Y., Adverse Drug Events: Database Construction and in Silico Prediction. *J. Chem. Inf. Model.,* **2013**, *53*, (4), 744-752.
[225] Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.,* **1988**, *28*, (1), 31-36.
[226] Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J., Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.,* **1992**, *32*, (3), 244-255.
[227] Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.,* **2010**, *50*, (5), 742-754.
[228] Peng, Z.; Mizianty, M.J.; Kurgan, L., Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins: Structure, Function, and Bioinformatics,* **2014**, *82*, (1), 145-158.
[229] Peng, Z.; Yan, J.; Fan, X.; Mizianty, M.J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V.N.; Kurgan, L., Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.,* **2015**, *72*, (1), 137-151.
[230] Mark, W.-Y.; Liao, J.C.C.; Lu, Y.; Ayed, A.; Laister, R.; Szymczyna, B.; Chakrabartty, A.; Arrowsmith, C.H., Characterization of Segments from the Central Region of BRCA1: An Intrinsically Disordered Scaffold for Multiple Protein–Protein and Protein–DNA Interactions? *J. Mol. Biol.,* **2005**, *345*, (2), 275-287.
[231] Cheng, Y.; LeGall, T.; Oldfield, C.J.; Dunker, A.K.; Uversky, V.N., Abundance of Intrinsic Disorder in Protein Associated with Cardiovascular Disease. *Biochemistry,* **2006**, *45*, (35), 10448-10460.
[232] Uversky, V.N.; Oldfield, C.J.; Dunker, A.K., Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annual Review of Biophysics,* **2008**, *37*, (1), 215-246.
[233] Uros, M.; Christopher, J.O.; Dunker, A.K.; Zoran, O.; Vladimir, N.U., Unfoldomics of Human Genetic Diseases: Illustrative Examples of Ordered and Intrinsically Disordered Members of the Human Diseasome. *Protein Peptide Lett.,* **2009**, *16*, (12), 1533-1547.
[234] Uversky, V.N.; Oldfield, C.J.; Midic, U.; Xie, H.; Xue, B.; Vucetic, S.; Iakoucheva, L.M.; Obradovic, Z.; Dunker, A.K., Unfoldomics of human diseases: linking

protein intrinsic disorder with diseases. *BMC Genomics,* **2009**, *10*, (1), S7.

[235] Rajagopalan, K.; Mooney, S.M.; Parekh, N.; Getzenberg, R.H.; Kulkarni, P., A majority of the cancer/testis antigens are intrinsically disordered proteins. *J. Cell. Biochem.,* **2011**, *112*, (11), 3256-3267.

[236] Casu, F.; Duggan, Brendan M.; Hennig, M., The Arginine-Rich RNA-Binding Motif of HIV-1 Rev Is Intrinsically Disordered and Folds upon RRE Binding. *Biophys. J.,* **2013**, *105*, (4), 1004-1017.

[237] Uversky, V.N.; Davé, V.; Iakoucheva, L.M.; Malaney, P.; Metallo, S.J.; Pathak, R.R.; Joerger, A.C., Pathological Unfoldomics of Uncontrolled Chaos: Intrinsically Disordered Proteins and Human Diseases. *Chem. Rev.,* **2014**, *114*, (13), 6844-6879.

[238] Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T., Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.,* **2004**, *337*, (3), 635-645.

[239] Kathiriya, J.J.; Pathak, R.R.; Clayman, E.; Xue, B.; Uversky, V.N.; Dave, V., Presence and utility of intrinsically disordered regions in kinases. *Molecular BioSystems,* **2014**, *10*, (11), 2876-2888.

[240] Wang, C.; Uversky, V.N.; Kurgan, L., Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics,* **2016**, *16*, (10), 1486-1498.

[241] DeForte, S.; Uversky, V.N., Not an exception to the rule: the functional significance of intrinsically disordered protein regions in enzymes. *Molecular BioSystems,* **2017**, *13*, (3), 463-469.

[242] Imming, P.; Sinning, C.; Meyer, A., Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.,* **2006**, *5*, (10), 821-834.