Article details

Sydow D., Burggraaff L., Szengel A., Vlijmen H.W.T. van, IJzerman A.P., Westen G.J.P. van & Volkamer A. (2019), Advances and Challenges in Computational Target Prediction, Journal of Chemical Information and Modeling 59(5): 1728-1742.
Doi: 10.1021/acs.jcim.8b00832

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Cite This: J. Chem. Inf. Model. 2019, 59, 1728−1742

Review

pubs.acs.org/jcim

Advances and Challenges in Computational Target Prediction

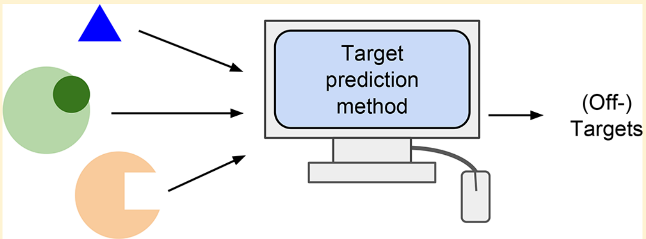
Dominique Sydow,[†,∥] Lindsey Burggraaff,[‡,∥] Angelika Szengel,[†] Herman W. T. van Vlijmen,[§,‡] Adriaan P. IJzerman,[‡] Gerard J. P. van Westen,[*,‡] and Andrea Volkamer[*,†]

[†] In silico Toxicology, Institute of Physiology, Charité − Universitätsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany

[‡] Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, P.O. Box 9502, 2300 RA, Leiden, The Netherlands

[§] Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30, B-2340 Beerse, Belgium

ABSTRACT: Target deconvolution is a vital initial step in preclinical drug development to determine research focus and strategy. In this respect, computational target prediction is used to identify the most probable targets of an orphan ligand or the most similar targets to a protein under investigation. Applications range from the fundamental analysis of the mode-of-action over polypharmacology or adverse effect predictions to drug repositioning. Here, we provide a review on published ligand- and target-based as well as hybrid approaches for computational target prediction, together with current limitations and future directions.



INTRODUCTION

Target prediction is a key aspect in early preclinical drug development, pivotal to determine the clinical application and to initiate drug development campaigns. For instance, orphan compounds may be known from phenotypic screening, showing changes in cell or organism phenotypes upon compound exposure, without the underlying molecular mechanism being known.[1] Targets for orphan compounds can be experimentally identified with techniques based on chemical proteomics such as affinity chromatography and activity-based protein profiling (ABPP), enabling compound testing against the proteome of cell lysates or even intact cells and organisms.[2−4]

Since these experiments are time and cost extensive, computational alternatives to rapidly predict the primary targets have gained momentum and are commonly known as *in silico target prediction*, target identification, or target fishing.[5] Herein, a general distinction can be made between *ligand-based* methods, centered around small molecules, and *structure-based* methods, implementing information from protein structures.[6] Pivotal to most of these approaches is the chemical similarity principle stating that "similar molecules have a similar biological effect" and conversely that "similar proteins bind similar ligands".[7]

One of the main applications of computational target prediction is to elucidate the *mode-of-action* of a compound by identifying its potential target. However, the traditional magic bullet paradigm, wherein a ligand has a high potency and selectivity toward a single target, has shifted to the understanding that a ligand affects multiple targets simultaneously.[8,9] In this context, target prediction methods can be used to explore desired *polypharmacological effects* of ligands to cover disease pathways.[10] Similarly, it can help to spot selectivity or toxicity problems during compound optimization which can potentially lead to unwanted *adverse* or *side effects*.[11] Moreover, approved drugs, and hence clinically tested ligands, can be repurposed for different indications if they are also found to interact with a protein target that is part of another disease mechanism.[12−14] This process is called *drug repositioning* or *drug repurposing*. Whereas the aforementioned applications focus on predicting targets, computational target prediction methods can also be applied to select ligands that have the highest potential to be relevant *chemical probes* used for ABPP to characterize the biological function of a poorly understood target.[15−17]

Designed for computational biologists, medicinal chemists, and neighboring disciplines, this review aims to outline the general principle and potential of computational target prediction together with the underlying methods and their application. The article starts with ligand-based modeling, followed by hybrid approaches (using both ligand and protein data), as well as structure- and interaction-based methods (Figure 1). Finally, potential pitfalls of the different approaches are covered, and a future perspective is given.

LIGAND-BASED TARGET PREDICTION

Central to ligand-based methods is that they rely on the chemical structure of ligands and associated bioactivity of similar ligands. Ligand-based methods are often used to predict the bioactivity of novel compounds for a specific target (Figure 2). However, ligand-based methods can also be applied to predict activities for a range of targets. Generally, this can be

Special Issue: Women in Computational Chemistry

Received: November 21, 2018
Published: February 28, 2019

DOI: 10.1021/acs.jcim.8b00832

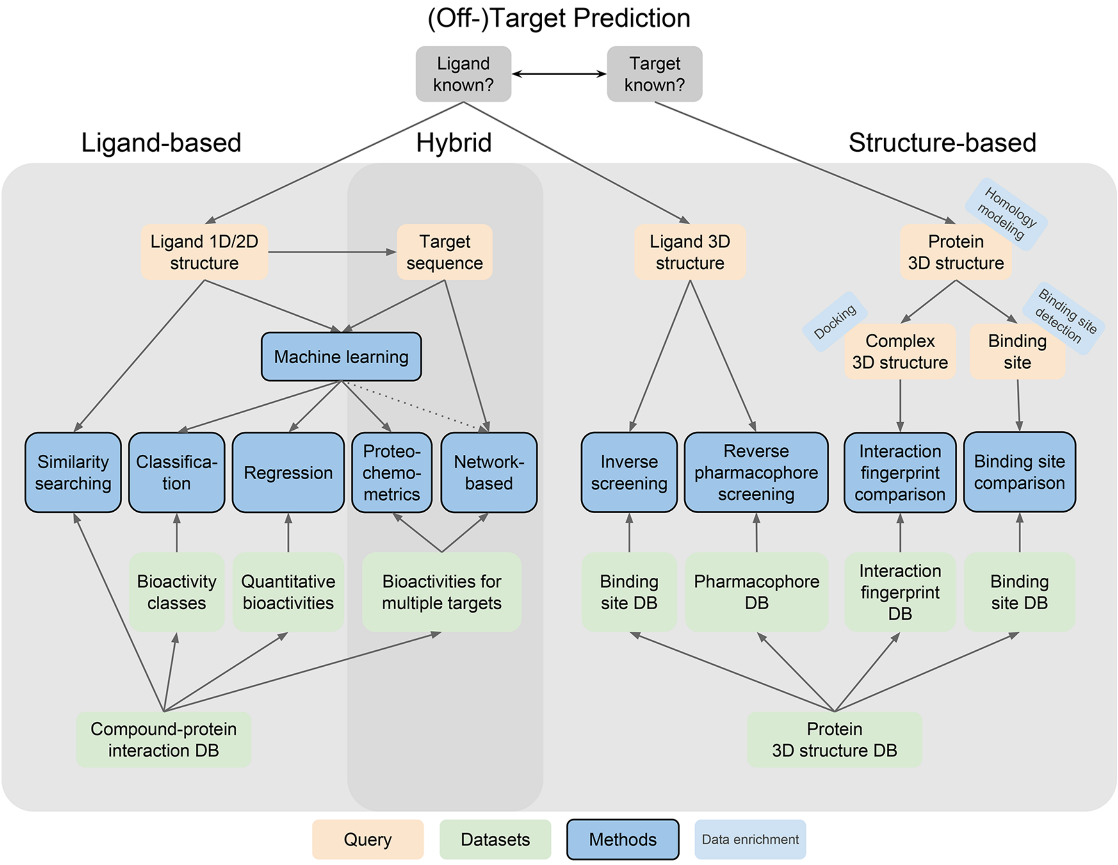


Figure 1. Overview of ligand- and structure-based as well as hybrid methods for target prediction (blue) with optional data enrichment strategies (light blue), using database (DB) or training data input (green), separated by applicability depending on available query data (orange). Necessary and potential connections are displayed with solid and dotted arrows, respectively.

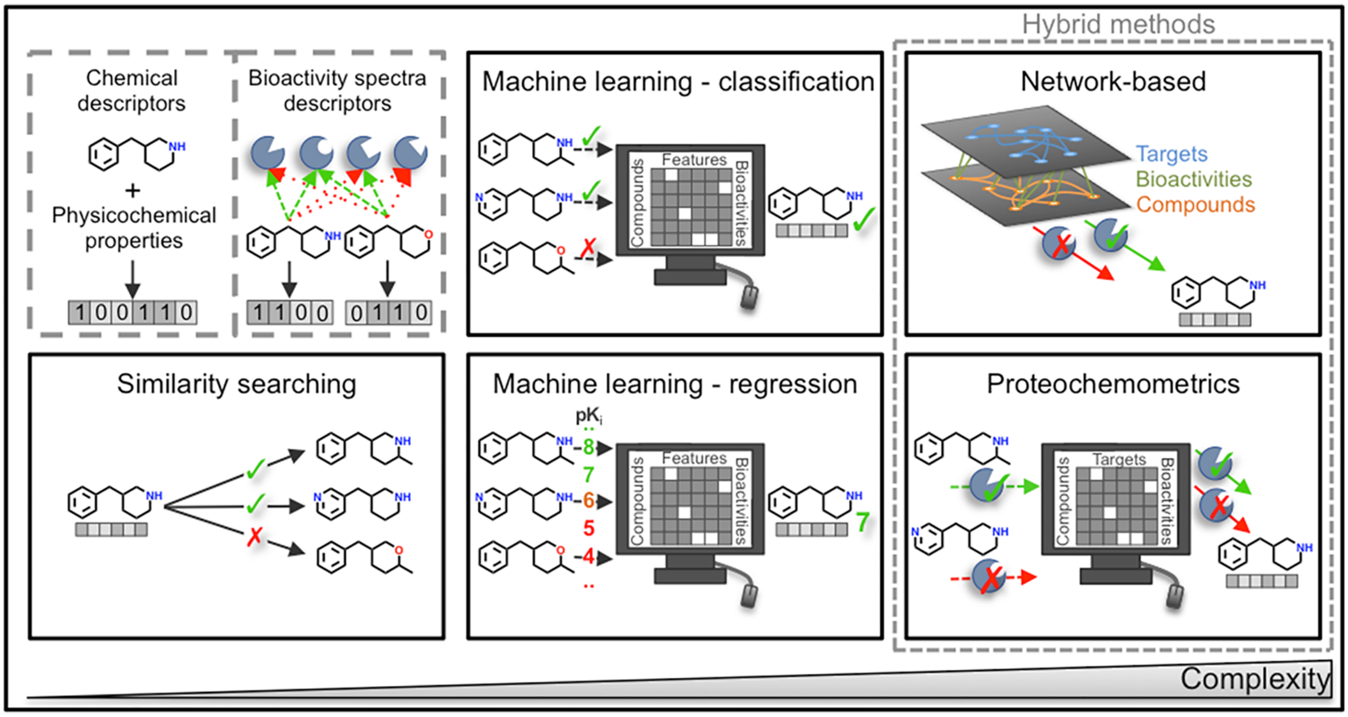


Figure 2. Ligand-based methods for target prediction. Descriptors in ligand-based methods are shown in the dashed-lined boxes on the left. Methods increase in complexity from left to right.

Table 1. Ligand-Based and Hybrid Methods in Target Prediction[a]

	Data in model training				
Name	Compound	Interaction	Training set requirements	Target ranking	Target prediction tools
Ligand-based models					
Similarity searching	Chemical structure	–	–	Targets classified based on similarity threshold of compounds	SwissTarget-Prediction,[35] SuperPred,[36] SEA,[40] OCEAN,[45] ROCS,[72] FTrees[73]
Similarity searching	Bioactivities	–	–	Targets classified based on similarity threshold of bioactivity spectra	BASS,[38] BioSEA[46]
Machine learning: Classification	Chemical structure	Activity class	Balanced (in)active classes	Targets classified based on activity class	PIDGIN[74]
Machine learning: Regression	Chemical structure	Bioactivity	Normally/equally distributed bioactivities	Targets ranked based on bioactivity	–
Hybrid models (ligand- and structure-based)					
Proteochemometrics	Chemical structure	Activity class or bioactivity	Balanced (in)active classes or normally/equally distributed bioactivities	Targets classified or ranked based on bioactivity	ChEMBL models[58,65]
Network-based models	Chemical structure and similarity	Activity class or bioactivity	Sufficient number of connections/ bioactivities	Targets classified or ranked based on bioactivity	DINIES,[68] drugCIPHER[69]

[a] The table gives information on what data is used and how targets are inferred from the model output.

accomplished by ranking targets based on predicted compound activity: the target for which the highest activity is predicted is expected to be the most likely target of that query compound.

Typically, the ChEMBL database[18] occasionally in combination with PubChem,[19] e.g., in the case of the ExCape database,[20] is used as a public source for chemical structures. These databases hold experimentally validated bioactivity data for many compounds tested on a wide range of proteins.

In the following, some general compound descriptors for ligand-based methods are outlined; for specific details, the reader is referred to the review by Rognan.[21] Subsequently, a description of ligand-based methods ordered by increasing complexity coupled to prediction confidence is given (Table 1). The latter is expected to be higher for the more complex methods.

Compound Descriptors. Compounds in ligand-based models are typically described using their 2D chemical structures. Depending on the data source, an intermediate step can be the conversion from a 1D sequential textual format (e.g., SMILES[22]) to a 2D structure, from which more complex binary vectors such as molecular fingerprints are usually obtained.[23] Different fingerprints are available to describe chemical structures, e.g., atom-pair fingerprints, topological-torsion fingerprints, or circular fingerprints, where atom environments are included (e.g., ECFP).[24] Optionally, the 3D shape of compounds is taken into account and translated into similar molecular fingerprints. However, this requires additional information on the 3D conformation of the compounds.[25,26] The use of different chemical fingerprints can impact model performance and was explored by Bender et al.[27] Additionally, physicochemical properties, topological information, and pharmacophore features of compounds can be added as descriptors in a similar way. As a result, each compound is described by an array of numbers forming the compound descriptors. Resemblance between arrays is higher when compounds are more similar to each other.

A more complex representation of compounds, compared to chemical descriptors, are bioactivity spectra descriptors. A spectrum in its simplest form is a binary bitstring representation where each bit represents a protein. Proteins for which a given compound shows activity are marked with a "1" as opposed to those for which this is not the case (marked with "0"). Bioactivity spectra rely on compounds being tested on a range of proteins, instead of compounds being tested on only one or a few targets. Considering compound promiscuity, it is expected that compounds display activity on a number of proteins.[28] Based on the bioactivity spectra, compounds that are not chemically similar but do exert a similar phenotype/ bioactivity might be recognized (so-called activity cliffs[29]). Likewise, this bioactivity profile can form an array of numbers that can be implemented as descriptors for similarity searching or machine learning, where activities can be treated as a bioactivity fingerprint. Recently, the biological annotation of compounds has been extended to include gene expression profiles[30,31] and high content cellular images,[32] providing additional, high-dimensional descriptors that can be added to a bioactivity fingerprint in a straightforward way.

Similarity Searching. The simplest and fastest method for target prediction is based on molecular similarity and is often referred to as similarity search or nearest neighbor search.[33] Using a similarity coefficient of choice (e.g., Tanimoto) and any type of compound descriptors (e.g., ECFP), the similarity between a pair of molecules can be quickly generated. For example, finding the most similar 100 compounds for a given query compound in a PubChem-sized library (~96 million compounds) takes a few seconds using chemfp tools developed by Dalke.[34]

The simplest implementation for target prediction based on similarity is to rank the data set compounds based on their similarity toward the query compound and assume that the biologically tested target of the most similar compounds is also the most likely target of the query compound. Webserver tools that enable the use of this method are, e.g., SwissTargetPrediction[35] and SuperPred.[36] These tools suggest protein targets based on molecular similarity of the query compound to compounds with known bioactivity toward these targets. It should be noted however that these approaches cannot provide a direct quantification of the biological activity of the query compound on the top-ranked targets.

While similarity search is classically performed by comparing chemical descriptors, activity spectra descriptors can also be used (if enough bioactivity data is available). Early work by Kauvar et al.[37] characterized molecular similarity by an affinity fingerprint based on experimental screenings of molecules against a reference panel of selected proteins. Also in BASS[38] (bioactivity profile similarity search), the similarity search is performed based on bioactivity spectra of chemical structures. Here, when the query has experimentally validated activities on a number of targets, additional targets can be predicted based on its bioactivity spectrum. Alternatively, gene expression profiles can be used to predict bioactivities of compounds for targets.[30,39] Both bioactivity spectra and gene expression profiles do not compare the molecular structure of compounds. Therefore, these methods are suited to identify different chemical structures for similar targets.

In contrast to a classical similarity search, **similarity ensemble** methods are applied to identify targets based on a group of known compounds for that target rather than a single compound. The compounds are first grouped based on interactions (e.g., bioactivity) with the same target(s). The similarity between different compound groups is subsequently calculated, and when defined as being similar, the targets that are known to interact with one compound group are identified as targets for the other compound group(s). The added benefit is that this allows the calculation of statistical measures that can score the relevance of a given retrieved target. When ensemble approaches are applied to identify targets for a query compound, the similarity is measured between this compound and the different compound groups. The targets belonging to the most similar groups are then identified as targets for the query compound. The SEA[40] method utilizes the similarity ensemble concept to group proteins based on ligand topology. Within SEA, the retrieved value is then compared to an expected random value (similar to the way this is implemented in BLAST[41,42]), and subsequently, an "E-value" is returned.[43] This E-value represents the extreme value and indicates the quality of the result. The (similarity) score of the selected samples is compared to what is expected when two random samples are taken into account. E-values closer to zero indicate that it is more unlikely that random samples would have equal similarity as the selected samples. The SEA method has been applied by Lounkine et al.[44] in a target prediction challenge. Here, side effects of 656 compounds were predicted based on compound interactions with 73 off-targets. The results were partially validated by data from hold-out databases or experimentally validated *in vitro*. Remarkably, off-targets were identified that had very low sequence similarity with the on-target (e.g., off-target serotonin transporter 5-HTT and on-target histamine H_1 receptor for antihistamine diphenhydramine), indicating that such a ligand-based approach can predict targets without the need of molecular biology information on protein targets. OCEAN[45] is a similar technique, though using different thresholds to determine compound similarities. Finally, BioSEA[46] also applies the same methodology; however, instead of comparing compound similarities based on chemical structure, bioactivity profiles are compared to create ensembles of compounds.

Machine Learning. Similarity search methods consider all features in the compound descriptors as equal. However, statistical methods can weigh the relevance of individual descriptors by connecting them to biological activity of the compounds and are often better suited to extrapolate to new compounds. Machine learning methods require a training phase, which is performed on known active and inactive compounds. Herein, a statistical model is fitted to the data to quantify how chemical descriptors relate to activity. Contrary to the similarity searching example above, this approach returns predicted compound−protein activities rather than a number of compound structures that are similar for a query compound. When applied to a single protein target for a congeneric chemical series, these methods are named *quantitative structure−activity relationship (QSAR)* models.[47] Given a query compound, QSARs can predict its expected activity based on the compound descriptors. In target prediction, however, more than one protein is considered.

Machine learning can both be used for classification (e.g., is the expected affinity higher than a threshold that was defined *a priori* as active?) or for regression (e.g., what is the predicted K_i value for a compound−protein interaction?). Typically, algorithms such as Random Forest,[48] Support Vector Machines,[49] and Naïve Bayes[50] are applied. However, with more data becoming available and to become more independent of the chosen descriptor, recent work is moving toward deep learning, a method able to directly derive features from molecular structures.[51,52]

An example where machine learning was applied in target prediction is the identification of novel inhibitors for the enzyme mycobacterial dihydrofolate reductase.[53] Here, targets were predicted for a set of query compounds using Naïve Bayesian models. The predicted compound−target interactions were validated *in vitro*, which indicates the value of such target prediction methods.

Classification. The most frequently used method in ligand-based target prediction is arguably classification.[1,54] Classification requires the setting of an activity threshold for measured interactions to separate the classes. This interaction can be measured binding affinity (e.g., pK_i) but can also be efficacy or other experimental measurements (e.g., pEC_{50}) or even a combination of multiple measurement types (e.g., pChEMBL value).[55] For classification models, a difference can be made between several approaches:

Single Model Multi-Class (SMMC). In this approach, one model is used that predicts the most probable target for a given compound, and target classes are mutually exclusive, in other words a compound cannot be active on more than one target.[56] Given known ligand promiscuity, the SMMC method provides an inaccurate representation of the behavior of ligands and could even be considered to be at odds with the similarity principle.

Ensemble Model Multi-Label (EMML). With EMML, also referred to as ensemble model multi-class, one model is used per protein, and compounds receive a prediction from each model.[1,57] Thus, the sum of protein models where the compound was predicted active on represents the set of potential target proteins. To build the model per protein, all compounds with an activity for the respective protein above a certain threshold are deemed the active class, and all other compounds are typically pooled in the inactive class. For the EMML approach, pooling constitutes a source of error. It might very well be that although a given compound has not been tested on the protein under consideration, it is indeed active yet pooling defines it to be inactive. Thus, potential targets for the query compound may be missed.

Single Model Multi-Label (SMML). Here, one model is used to predict all potential targets for a given molecule, and compounds can belong to multiple target classes (or labels).[56] The active class for a given protein is defined equally as is described for EMML, but all other compounds are not explicitly pooled in an inactive class, merely the ones that were tested to be inactive are considered. A caveat can be that there are none or too few known inactive compounds for good model fitting.

When a query compound is run through a classification model, the output gives the activity class per target (e.g., active/inactive, depending on the previously described approaches and on the predetermined activity threshold). However, regression can directly predict the affinity of a compound.

Pitfalls Defining an "Active" Class. Typically, the activity threshold in classification models is set at 10 μM (i.e., an affinity better than 10 μM defines active interactions, corresponding to a pK_i of 5). This parameter carries a significant influence on effectiveness and applicability of target prediction methods. In principle, for classification, a balanced set of active and inactive compounds is desired. When the activity threshold is set at 10 μM, this gives a skewed distribution of actives and inactives. Recently, target prediction was performed using an affinity value of ~316 nM (corresponding to 6.5 on a logarithmic scale) as the threshold; this leads to a better distribution of active and inactive classes when using ChEMBL data.[58] An added benefit is that this threshold also provides a more relevant prediction of biological activity. Given that the biological error of assays is on average around ~0.5 log units for mixed pK_i values, a model using a cutoff of pK_i = 6.5 could at worst correspond to an experimental activity of a pK_i = 6.0. When a cutoff of pK_i = 5.0 (10 μM) is used, this error would be at worst pK_i = 4.5 for predicted actives.[57,58] However, the optimal activity threshold for balanced classification sets is dependent on the databases from which compounds and bioactivities are extracted (e.g., ExCape[20] contains more compounds with lower bioactivities than ChEMBL). Furthermore, the targets that are considered can be biased toward reported (in)actives (often in relation to the amount of studies focused on the target, see the Discussion and Future Directions section).

When a reasonable number of inactive compounds is available, but significantly less than the number of active compounds, some workarounds can be applied to train representative models. For instance, active compounds can be divided into smaller subsets in order to train separate models for each subset of actives with the same set of inactives (e.g., random undersampling) and, finally, recombined by ensembling. Ensembling is a technique to combine predictions from multiple models into one prediction that has shown to increase performance.[58,59] The downside of any ensembling method is the unavoidable increase in computational time required as predictions for multiple methods are needed.

Another workaround (which also requires increased computational time) is to construct multiple ligand-based target prediction models at different thresholds (e.g., 10 μM, 1 μM, 100 nM, 10 nM, and 1 nM). However, doing so decreases the available data points for the higher activity thresholds as fewer compounds are known that meet the threshold, and hence, this has a negative effect on the chemical applicability domain. In these cases, regression might allow the use of more data.

Regression. Contrary to classification, regression methods are able to directly train on the strength of a given ligand–protein interaction avoiding the need for a preset threshold. Trained on experimental data, regression models can make quantitative predictions (e.g., K_i values) for compounds based on the chemical structure. These predictions can be directly translated to the interaction (e.g., affinity as a K_i value). Thus, when regression is applied to multiple proteins (using an ensemble of models), the targets can quantitatively be ranked based on predicted compound–protein activity. In addition to predicting activity, the differences in interaction strength for different proteins can be evaluated. Using regression models, the output of a query ligand can constitute a list with ranked targets based on quantitative bioactivity predictions. The output, therefore, does not only define "active" or "inactive" targets but also the activity strength that is reflected by the predicted bioactivity values.

HYBRID METHODS FOR TARGET PREDICTION

Similarity searching and machine learning methods—which are classically built on ligand information—can also be applied in more complex systems where protein information is added. Although the underlying mechanism of the methods is the same (e.g., machine learning), the implementation can be different, in turn leading to other application possibilities. This results in alternate methods to model and analyze the data.

Proteochemometrics. With proteochemometrics (PCM), both compound and protein information are combined by addition of an explicit protein descriptor.[60] The most common approach is to add protein information based on knowledge derived from the protein sequence. Sequences are translated into descriptive scores (e.g., Z-scales[61]), reflecting the properties of the amino acid residues of the proteins.[62] Additionally, when structural protein information is available, this may be used to increase descriptor quality as information on binding site location can be included, making the model more accurate compared to using full sequences.[63] PCM can be applied to expand single target models to multiple targets: based on sequence similarity between proteins, data from one protein can be extrapolated to a related one.[64] Another application is increasing the amount of available data (compared to single target models) in order to increase model performance.[63] Several PCM models for target prediction based on ChEMBL data have been reported.[58,65] Such models predict the activities of a query compound for each of the incorporated targets. When these models are based on regression, the most likely target for a query compound can be derived based on the highest predicted activity for that target compared to other targets. Additionally, a quantitative activity score is given per target; therefore, it can be assessed if activity of the query compound for the highest ranked target(s) is sufficient. Noteworthy, as the combination of compound and protein descriptors defines each compound protein pair as a unique pair, even binary class PCM models behave as SMML models. A compound tested to be inactive on protein A can be distinguished from the same compound tested on protein B by the algorithm based on the protein descriptor.

Network-Based Methods. Protein–protein or protein–ligand interactions can be described as a large network similar to a social network. Here, nodes can be proteins, compounds, or both, with the edges being interactions, similarities, or phenotypic effects. These connections can also be weighted based on the strength of interaction (e.g., pK_i). Using chemical structures and similarities between connections, targets can be identified for query compounds.[51] This has led to the publication of several works that use network analysis tools to predict protein pharmacology.[66,67] Additionally, network-based target prediction tools such as DINIES[68] and drugCIPHER[69] are made available as open source tools to detect ligand–target interactions for query molecules. The concept of network-based models is often based on similarities between chemical structures but can also include similarities between proteins. More simplistic models implement only one similarity (e.g., protein similarity), whereas more complex models can encompass similarities between protein, chemical structures, and interactions, simultaneously. Such a heterologous network was constituted using three different networks by Chen et al.[70] Here, a protein similarity network (based on sequence similarity) was connected to a compound similarity network by using a ligand−protein interaction network.[71] Therefore, in this network, protein and compound similarities can simultaneously be addressed, which is not possible with only similarity searching as described in the section regarding this topic. Targets for a given query compound can be inferred from the network based on activities (or connections) of similar ligands and their corresponding targets.

■ STRUCTURE-BASED TARGET PREDICTION

Methods for structure-based target prediction identify the most likely targets for a query ligand or the most similar targets for a query target, using 3D structural, i.e., steric and physicochemical, information (Figure 3). The former group of approaches focuses on docking a *query ligand* either to a set of targets (*inverse screening*) or to a set of pharmacophores inferred from ligand−target complexes (*reverse pharmacophore screening*), see Table 2. The latter group of methods compares a *query target*, either to a set of targets (*binding site comparison*) or to a set of interactions inferred from ligand−target complexes (*interaction fingerprint comparison*),[5] see Table 3.

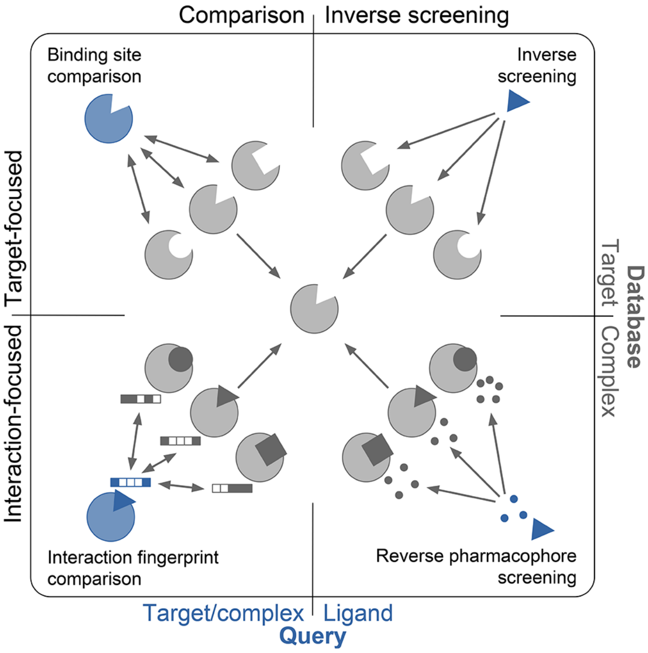


Figure 3. Structure-based target prediction: conceptual representation of the four main approaches, i.e. binding site comparison, inverse screening, reverse pharmacophore screening, and interaction fingerprint comparison.

Typically, the Protein Data Bank (PDB)[75] is used as a public source for protein structures, currently holding more than 140,000 protein structures (accessed in November 2018). Since the binding site is the key to protein function, most methods are proceeded by a binding site annotation step: with a ligand present, binding sites are extracted by a defined ligand−target residue distance cutoff, and without a co-crystallized ligand, binding site detection methods can be invoked.[76] A widely used resource for such annotated binding sites is the scPDB[77] database, containing more than 16,000 ligand-bound binding sites from the PDB and covering about 4700 proteins with 6300 ligands.

Methods for structure-based target prediction are all composed of three main steps, which are described in detail in the individual method paragraphs: (i) binding site encoding, (ii) target screening or comparison, and (iii) target ranking. First, binding sites or ligand−target interactions are encoded using different descriptor techniques and stored in a target database. Second, depending on the method, either a query ligand is screened against the target database, using different docking engines, or a query binding site is compared with the target database, using different similarity measures. Finally, targets are ranked based on a suitable scoring approach.

Inverse Screening. Classically, molecular docking is used to predict both the binding mode and the approximate binding free energy of a set of ligands against one target of interest. In inverse docking, also known as inverse screening or panel docking, this strategy is reversed, and one query ligand is docked to a set of target proteins in order to predict its most likely targets. Most docking tools are theoretically applicable for inverse screening, yet need adaption with respect to inter-target instead of conventional inter-ligand ranking (Table 2).[78,79]

(i) Binding Site Encoding. Since the query compound is screened against each target in the data set, the targets need to be preprocessed accordingly. Target databases for methods using conventional docking engines simply contain structure files for binding sites (e.g., TarFisDock[80] and idTarget[81]) or for whole proteins (INVDOCK[82]), preprocessed as required for the respective docking tool. In contrast, iRAISE[83] prepares for an efficient comparison by encoding binding sites with triangle descriptors, which contain pharmacophoric and shape information and are stored as bitmap database, a specialized index for high-dimensional features.

(ii) Target screening. Most inverse screening methods use conventional docking engines, such as DOCK (TarFisDock), MEDock (idTarget), Glide (VTS[84]), or AutoDock Vina (VinaMPI[85] and IFPTarget[86]), in order to estimated the fit of the query compound against each protein in the target database. High computational costs are addressed by either parallel screening (VinaMPI and IFPTarget) or by search space reduction. The latter can be realized by aborting the search at the first pose reaching a threshold score based on interaction energies from reference ligand−protein complexes (INVDOCK) or by testing one target representative per precalculated target cluster (based on sequence identity) before screening the entire cluster (idTarget). Usually, energy-based functions, such as interaction or binding free energy functions, are used to score the resulting docking poses. In iRAISE, the query ligand is described with triangles, in the same manner as the binding sites before, and is efficiently matched based on bitmap indices, followed by respective superimposition of the ligand and binding site triangles. Finally, iRAISE docking poses are scored using a more extensive approach in the form of a scoring cascade, including a clash test, an interaction energy score, a reference score cutoff (based on the co-crystallized reference ligand), and a ligand and pocket coverage score.

Table 2. Structure-Based Target Prediction: Selected Methods for Inverse Screening and Reverse Pharmacophore Screening

		Target screening			
Name	Encoding	Docking engine	Scoring function	Target ranking	Av.[a]
Inverse screening					
INVDOCK[82]	Sphere-coated surface	DOCK derivative	Interaction energy	–	2
TarFisDock[80]	Sphere-coated surface	DOCK 4.0	Interaction energy	–	2
idTarget[81]	Energetic grid map	MEDock	Binding free energy (AutoDock4 score)	Z-score based on binding free energies of reference complexes	1
VTS[84]	Energetic grid map	Glide	Binding free energy (Glide Gscore)	Gscore comparison to Boltzmann-weighted average of reference Gscores	2
VinaMPI[85]	Energetic grid map	AutoDock Vina	Binding free energy (Vina score)	–	1
iRAISE[83]	Bitmap of binned triangles (3 pharmacophore features and cavity shape)	Index-based bitmap comparison	Scoring cascade: clash test, interaction energy and reference cutoff, ligand and pocket coverage	Gaussian-weighted score based on scores for reference complexes	1
Reverse pharmacophore screening					
PharmMapper[90]	Hash table of binned triangles (5 pharmacophore features)	Geometric hashing	Fit score (based on matching feature types and positions)	Z-score based on fit score distribution of reference complexes	1

[a] Av. = availability: web server, software, or code is (1) free for academic use and/or available upon request or (2) not (yet) available or unclear.

(iii) Target Ranking. Targets are ranked either directly based on the interaction energies of the best docking pose(s) per target (INVDOCK, TarFisDock, and VinaMPI) or based on separate functions tailor-made for inter-target ranking. In the latter approach, each target in the database is profiled beforehand either with a set of ligands using docking (iRAISE and VTS) or with one co-crystallized ligand (idTarget and IFPTarget). These reference profiles are then used to normalize the scores of docking poses of a query ligand and potential targets.

Inverse screening methods have been widely used for target prediction.[78,79] For example, Scafuri et al.[87] applied idTarget to predict potential targets of apple polyphenols, known for their chemo-preventive effect against colorectal cancer. In a bioinformatics-driven function analysis, the gene expression levels for the predicted targets were shown to be significantly altered in colorectal cancer cells, indirectly linking the investigated apple polyphenols to the predicted targets.

Reverse Pharmacophore Screening. Similar to inverse screening, reverse pharmacophore screening consecutively fits a query ligand in the form of a ligand-based pharmacophore into a precalculated panel of pharmacophore models, derived from protein−ligand complexes. A pharmacophore is defined as an ensemble of physicochemical and steric features that are necessary for the recognition of a ligand by a target, triggering or blocking a biological response.[88] Structure-based approaches derive such pharmacophores from a target complex, whereas ligand-based pharmacophores consider solely ligand properties. Several studies have conducted reverse pharmacophore screening for polypharmacology, using available standard software packages that allow for rapid pharmacophore model building and evaluation.[89] However, to the knowledge of the authors, the only available automated workflow for pharmacophore-based target prediction is PharmMapper.[90]

In PharmMapper, the interactions of selected ligand−target complexes are encoded as pharmacophore feature triplets, stored in a hash table, and deposited in a target database (i). For target screening (ii), ligand-based pharmacophores are generated for multiple conformations of the query ligand. Each conformer pharmacophore is described in form of triplets and aligned onto each pharmacophore triplet in the target database, using triangle hashing. Subsequently, targets are scored based on the overlap of feature types and positions between the ligand and target pharmacophores. Finally, each target score is normalized by a reference score for target ranking (iii). The reference score per target reflects the score distribution of matching all ligand pharmacophores extracted from the original protein−ligand complex structures in the database against the target pharmacophore.

Reverse pharmacophore screening was often applied to search for targets of compounds in Chinese traditional medicine (CTM).[79] For example, Liu et al.[91] used PharmMapper to predict the glucocorticoid receptor, p38 mitogen-activated protein kinase, and dihydroorotate dehydrogenase as potential targets of berberine, a compound used in CTM to treat cancers including melanoma. Experimental tests confirmed the predicted targets to be potentially involved in the anti-melanoma effect of berberine.

Binding Site Comparison. Target comparison is based on the assumption that similar proteins—or more precisely binding sites—bind similar ligands. Various binding site comparison methods have been developed, pursuing different strategies to encode binding sites, as well as to measure and score their similarities[92,93] (Table 3).

(i) Binding Site Encoding. The structural complexity of binding sites is reduced to labeled representatives, whose spatial arrangement is encoded and stored in a database, to be compared with a query binding site encoded accordingly. Binding site *representatives* can be per-residue points (e.g., CavBase[94] or (Med-)SuMo[95,96]), binding site surfaces (e.g., ProBis[97]), or binding site volumes (e.g., Volsite/Shaper[98]), with *labels* mostly containing pharmacophoric information. The *spatial arrangement* of these representatives is often encoded as graphs (e.g., CavBase) and triangles/quadruplets. The latter are binned by their edge lengths and vertex labels and stored as fingerprints (e.g., FuzCav[99] and FLAP[100]), hash tables (SiteEngine[101]), or bitmaps (TrixP[102]), whereas (Med-)SuMo[95,96] uses a graph of adjacent triangles. Alternate methods describe binding sites as distance distributions between aforementioned per-residue points (e.g., RAPMAD[103]), or with volume functions (Volsite/Shaper).

(ii) Binding Site Similarity Measure. Common strategies for measuring binding site similarities can be divided into alignment-based (often slower) and alignment-free methods (mostly faster), as well as accelerated alignment-based methods. The latter combine the speed of alignment-free

Table 3. Structure-Based Target Prediction: Selected Methods for Binding Site and Interaction Fingerprint Comparison[a]

	Encoding					
Name	Representatives	Label	Pattern	Comparison	Scoring	Av
Binding Site Comparison						
Alignment-Based Methods						
SiteBase[104]	^{1}Atoms	5 atom types	^{5}On-the-fly triangles	Geometric matching	Matching atoms	2
(Med-) SuMo[95,96]	^{1}Per-residue: pseudocenters (PCs)	Chemical groups	^{5}Triangles as graph of adjacent triangles	Geometric matching, stepwise connection of adjacent matches	Size of connected matches	1,3
SiteEngine[101]	^{1}Per-residue: PCs & surface patches	5 pharmacophoric features	^{5}Triangles in hash table	Geometric hashing	Matching surface patches & PCs	1
CavBase[94]	^{1}Per-residue: PCs & surface patches	5 pharmacophoric features	^{4}Graph	Clique detection	Matching surface patches & PCs	3
eF-site[113]	^{2}Triangulated surface	Electrostatics & surface curvature	^{4}Graph	Clique detection		1
ProBis[97]	^{2}Triangulated surface	5 pharmacophoric features	^{4}Subgraphs	Clique detection (per subgraph)	Matching surface patches & residues	1
PoLiMorph[114]	^{3}Grid-based cavity volume	5 physicochemical features*	^{4}Fuzzy graph/ self-organizing map	Error-tolerant graph matching	Matching vertices	2
Alignment-Free Methods						
Pocket-Match[115]	^{1}Per-residue: C_{α}, C_{β} & centroid = A	5 amino acids groups = B	^{6}90 distance histograms for all $A-B$ combinations	Corresponding histogram comparison	Average matching distance bins	2
RAPMAD[103]	^{1}Per-residue: PCs p_i incl. 2 references p_1 & p_2	7 pharmacophoric features: 7 PC subsets s_i	^{6}14 distance histograms: p_1-p_i and p_2-p_i per s_i	Corresponding histogram comparison	Jensen-Shannon divergence	2
FuzCav[99]	^{1}Per-residue: C_{α}	6 pharmacophoric features	^{5}Triangles as 4833 int fingerprint	Fingerprint comparison	Matching non-zeros	1
KRIPO[116]	^{1}Defined points relative to residue	5 pharmacophoric features	^{5}Triangles as fuzzy fingerprint(s)	Fingerprint comparison	Modified Tanimoto index	1
Pocket-FEATURE[117]	^{1}Per-residue: center with 6 shells = microenvironment (ME)	80 physicochemical features* per shell	^{7}480 int fingerprints per ME	Fingerprint comparison per ME pair if same amino acid	Sum of bestscoring (Tanimoto index) ME pairs	2
Accelerated Alignment-Based Methods						
BSAlign[105]	^{1}Per-residue: PCs	5 physicochemical features*	^{4}Reduced (red.) graph	Clique detection, red. product graph	Matching residues & RMDS	1
SiteAlign[106]	^{2}80-fold triangulated sphere at binding site center	8 topological and chemical descriptors mapped to triangles	^{5}640 int fingerprint	Fingerprint comparison	Average of normalized triangle differences	1
TrixP[102]	^{1}Pharmacophoric feature points	3 pharmacophoric features	^{5}Triangles in bitmap	Index-based bitmap comparison	Matching triangles	2
FLAP[100]	^{2}Clustered GRID-MIFs	5 pharmacophoric features	^{5}Quadruplets as 11 int fingerprints	Fingerprint comparison	Matching quadruplets	3
BioGPS[118]	^{2}Clustered GRID-MIFs	3 pharmacophoric features	^{5}Quadruplets as 11 int fingerprints	Fingerprint comparison	MIF volume overlap per feature	2
Volsite/ Shaper[98]	^{3}Grid-based cavity volume	7 pharmacophoric features	^{7}Volume as smooth Gaussian function	Volume overlap	Matching pharmacophoric features	1
Interaction Fingerprint Comparison						
SIFt[110]	Interacting residues	7 pharmacophoric features	Per-residue: 7 bit vector, concatenated in fixed order	Fingerprint comparison	Tanimoto index	2
TIFP[111]	Pseudoatoms between interacting ligand−target pairs	7 pharmacophoric features	Triplets as 210 int fingerprint	Fingerprint comparison	Tanimoto index	2
SPLIF[112]	Interacting fragments	Atom and bond types	ECFP2 fingerprint	RMSD for all matching fingerprint bits	Matching ligand and protein atoms	2
LIFt[109]	Atom-by-atom ligand−target interactions	10 pharmacophoric features	Interaction fingerprint	Fingerprint comparison	Tanimoto index	2
IFPTarget[86]	Interacting residues	8 pharmacophoric features	Label × residue matrix	Matrix comparison between query and reference complexes	Modified Tanimoto index & energy-based score for query complex	2

[a] Binding sites are encoded based on ^{1}per-residue points, binding site ^{2}surfaces, and ^{3}volume, and are represented as ^{4}graph, ^{5}triangles/quadruplets (e.g. binned into fingerprints/hash tables/bitmaps), ^{6}distance distributions of atom pairs, and ^{7}volume functions. RMSD = root-mean-square deviation; MIFs = molecular interaction fields; Av. = availability: web server, software, or code is (1) free for academic use and/or available upon request, (2) not (yet) available or unclear, or (3) commercially available; *Including pharmacophoric and additional features, e.g. buriedness.

methods with the visual interpretability of alignment-based methods. *Alignment-based methods* calculate and perform the best possible structural superimposition of two binding sites based on their encoded features, using geometric matching and hashing of two triangle sets (e.g., SiteBase[104] and SiteEngine, respectively) or most commonly clique detection between two graphs (e.g., CavBase). The latter approach searches the maximum complete subgraph (clique) in a product graph, which is built from a target and query graph with matching vertices and edges. Many *alignment-free methods* operate on the comparison of fingerprints (e.g., FuzCav) or of distance histograms (e.g., RAPMAD). *Accelerated alignment-based methods* use efficient data structures for rapid comparison, with subsequent binding site alignments for scoring and visual interpretation. Those methods include strategies to reduce graph complexity before clique detection (BSAlign[105]), to compare binding site volumes using smooth Gaussian functions (Volsite/Shaper), and to store binned 3-point pharmacophores in bitmap indices (TrixP). Moreover, properties of a binding site can be projected to a triangulated sphere positioned at its center, stored as fingerprint to be iteratively compared, and aligned to another binding site fingerprint (SiteAlign[106]).

(iii) Binding Site Similarity Ranking. Alignment-based methods score the similarity of binding sites based on the mutual overlap and/or root-mean square deviation (RMSD) of their associated encoded features. In contrast, alignment-free methods mainly calculate fingerprint similarity based on the number of matching fingerprints, if multiple fingerprints exist per binding site (e.g., FLAP), or based on the Tanimoto coefficient, if only one fingerprint per binding site (e.g., FuzCav) is calculated.

An exemplary application of binding site comparison is a study on cross-reactivity using SiteAlign by De Franchi et al.[107] Virtual screening of Pim-1 kinase against ATP-binding sites showed high similarity to synapsin I, a protein regulating neurotransmitter release in the synapse, suggesting a cross-reaction of protein kinase inhibitors with synapsin I. Biochemical validation revealed nanomolar affinities for pan-kinase inhibitor staurosporine and selective Pim-1 kinase inhibitor quercetagetin for synapsin I. These findings were proposed as possible explanations for the observed down-regulation of neutrotransmitter release by some protein kinase inhibitors.

Interaction Fingerprint Comparison. Interaction fingerprints (IFPs), or protein−ligand fingerprints, are vectors that encode information on interacting ligand and target moieties, such as hydrogen bond, hydrophobic, charge, aromatic, and metal-binding interactions. IFPs are often used in combination with screening methods in order to rescore docking poses.[108] Only a few IFP-based pipelines have been published for target prediction so far. Note that they require a ligand placement step for IFP calculation. Thus, for IFP encoding (i), the query ligand has to be docked against the target structure(s). Generally, IFP methods either map detected interactions to ligand atoms (e.g., LIFt[109]), to target binding site residues (e.g., SIFt[110] and IFPTarget[86]), or define a ligand- and target-independent fixed length fingerprint (e.g., TIFP[111] and SPLIF[112]). Similar to the alignment-free fingerprint-based binding site comparison, the comparison of two IFPs is usually based on the Tanimoto coefficient (ii), and targets are rank-ordered accordingly (iii). In the following, two tools are introduced: In the first approach, interactions are mapped on the ligand; thus, ligand IFPs are compared. In the second, information is mapped on the target residues, and subsequently, target IFPs are compared.

Cao and Wang[109] propose a pipeline for off-target prediction exemplified on a tubulin agent with kinase-cross activity. The tubulin agent complex structure is the starting point to generate the ligand-based interaction fingerprint (LIFt) for the query compound. Next, the query ligand is docked to a panel of kinase structures. The best-scoring pose per ligand−kinase complex is encoded as LIFt, documenting interactions per ligand atom. Finally, these predicted panel LIFts are compared (Tanimoto coefficient) to the known reference LIFt and ranked accordingly.

In contrast, IFPTarget by Li et al.[86] first sets up a target database, where the co-crystallized ligand is used to define the reference target IFP, documenting per-residue interactions. Next, the query ligand is docked to the same panel of targets, and the top-scoring pose for each target is used to generate the docked target IFP. Subsequently, reference and docked target IFPs are compared and ranked by a final score that integrates aforementioned energy-based docking and IFP-based scores.

The presented methods are strongly intertwined with a docking (inverse screening) procedure: Two IFPs can only be compared if they have one constant component (LIFT: same ligand in two different structures, or IFPTarget: same structure with two different ligands) because otherwise the IFP lengths and order differ. Here, the third category of ligand and protein invariant fingerprints, such as TIFP by Desaphy et al.,[111] could find a remedy, but has, to the knowledge of the authors, not yet been used for target prediction.

Consideration of Target Flexibility in Structure-Based Methods. Proteins are flexible, existing in transient conformational states, whereby only a subset may be receptive to ligand binding. Such flexibility is to some extent implicitly considered by the coarse-grained representation of binding sites in the encoding step, such as binned distances (e.g., RAPMAD and FuzCav) and fuzzified graphs (PoLiMorph[114]), as well as by including tolerances during the matching step. Small side-chain flexibility can be explicitly included by, e.g., representing rotatable hydrophilic interactions (TrixP) or "on-the-fly" conformational sampling of side chains (FLAP and BioGPS[118]). Instead of conformational sampling, different parts of the binding site can be investigated separately from each other in order to spot local similarities. Some methods therefore allow for partial shape matching (TrixP) or local examination of binding site segments (ProBis). Inverse screening methods usually treat the target structure as rigid body, while considering ligand flexibility by conformational sampling of the ligand (e.g., iRAISE and INVDOCK).

However, information on protein flexibility can be enriched by including protein ensembles in screening databases, either derived from a set of experimentally determined structures or from molecular dynamics (MD) simulations. The former approach is to some extent integrated whenever methods are built upon a database containing multiple structures per protein (e.g., scPDB-based target databases); however, so far, those structures have not been statistically evaluated as one protein ensemble. Furthermore, such PDB-derived protein ensembles can only cover protein classes with high coverage. Methods describing binding site changes based on MD simulations, as described in TRAPP[119] for transient pockets, are already available but have not been integrated yet into a workflow for target prediction.

DISCUSSION AND FUTURE DIRECTIONS

Since without sufficient data computational target prediction would not be possible at all, we first discuss the beauty and peril of current data sources. We then cover challenges in target ranking and method validation as well as directions on how to overcome them.

Data. Usage of *in silico* techniques for target prediction has been enabled in the first place by the rapidly increasing amount of *available structural, chemical, and biological data*. In this respect, the increasing availability of open access databases for drug discovery should be appreciated, with the PDB,[75] ChEMBL,[18] PubChem,[19] and DrugBank[120] databases being arguably the most well known. While the speed of computation has increased at a phenomenal rate with transistor counts roughly doubling every two years[121] (slowing down in recent years[122]), data availability and quality still form the bottleneck.[20,123] Given more data, more intricate methods can be applied, which should result in higher quality predictions.[21] This does not only concern bioactivity data but also structural information on proteins.[75]

In *ligand-based methods*, the large amount of available bioactivity data is used for model training. Lack of data here typically means that there are not enough experimentally derived activities of compounds for a given target. One way to overcome this is using computational target prediction to fill in the expected bioactivities for proteins that were not experimentally tested.[54,124] However, even if sufficient data is available, this does not directly mean the *data quality* is adequate. It has been shown that the experimental error in bioactivity databases can be substantial.[33,125] In public data, experimental activities are not derived following the same standard operating procedure or are even from the same lab or assay. This leads to a relatively large experimental error in the data (on average 0.47 log units for mixed pK_i data),[33] which is reflected in the prediction accuracy of the models. Data quality and bias each determine the applicability domain of a model and should therefore be addressed early on by comparing the similarity between training and screening compounds. For instance, models trained on smaller or more hydrophobic molecules may not be able to make reliable predictions for larger or more hydrophilic compounds. Furthermore, high chemical similarity within the training set leads to a *bias toward a similar group of compounds*. Therefore, a wide diversity in chemical space is more favorable than a large compound set encompassing a congeneric series of ligands. Models trained only on close analogues cannot predict activities of very dissimilar compounds reliably. In summary, in order to build reliable models, important factors to check are the amount of data and heterogeneity (as discussed here), as well as the bias toward (in)actives (see Pitfalls Defining an "Active" Class section) and toward certain targets (see Target Ranking section).

Structure-based methods build on the structural arrangement of binding site atoms, experimentally derived from currently mostly X-ray crystallography. Such structural arrangements are (i) less reliable with decreasing resolution and (ii) represent only a static (and maybe even artificial) conformational state. The former is usually addressed with resolution thresholds (e.g., <3 Å in case of the scPDB), whereas the latter is sometimes considered with conformational sampling (see Consideration of Target Flexibility in Structure-Based Methods section). Furthermore, using structure-based methods, only targets with available structures can be queried, introducing a *bias toward structurally known targets*. Currently, most methods rely only on the available structures in the PDB. While there are over 140,000 protein structures deposited in the PDB (accessed in November 2018), they only cover at most 30% of the human proteome and 50% of known human drug targets,[126] with protein classes being differently well represented. Homology modeling is a possibility to infer lacking information from determined structures of homologous proteins. Somody et al.[126] have shown that given a sequence identity of ≥30% (as generally accepted lower limit for homology modeling) the structural coverage of the modeled human proteome could approach 70% (that of known human drug targets 95%). While large scale homology models have been used, e.g., for kinome-wide druggability predictions,[127] they have not been widely used yet for target prediction. It should be noted that the higher the sequence identity is, the more reliable the homology models are for structural modeling purposes. Furthermore, target-focused methods such as inverse screening and binding site comparison only require 3D target structures and binding site locations, whereas interaction-focused methods require ligand−target complex information, limiting their applicability. To overcome this, such interactions can be predicted: For instance, interaction fingerprint comparison can be coupled with inverse docking, and reverse pharmacophore screening can be based on target-focused pharmacophore methods such as T^{2}F-Pharm[128] that generate pharmacophores from apo-structures. However, it is important to note that such ligand- as well as structure-based models-based-on-models approaches may introduce noise to the predictions.

Target Ranking. Results from computational target prediction are highly dependent on the scoring function(s) used for target ranking. If two objects of the same type—for example, two small molecules or two protein binding sites—are compared, similarity of the query to the database can directly be inferred from the commonalities or mutual overlap between the objects and ranked accordingly. In contrast, if the objects to be compared are of different types, target ranking becomes more complex. For example, this is the case when the most likely targets are predicted for a small molecule based on individual machine learning models per target (ligand-based methods) or based on inverse screening against a target database (structure-based methods). While it is already challenging to predict the correct activity or binding energy of a ligand against one target, in panel predictions, the ligand is scored individually against multiple targets, requiring inter-target ranking. This is especially ambitious since the predictions are influenced by different forms of bias present in the data. Typically, some protein classes (e.g., kinases or G protein-coupled receptors) have been very well explored, whereas others have been explored less thoroughly (e.g., transporters). This means that more ligands are known for these proteins (ligand-based methods) or more structures have been elucidated (structure-based methods). Thus, the chemical or structural space is better covered, and they might score better compared to less explored chemical or structural spaces. Another form of bias influencing target ranking can be the average molecular weight of ligands for certain protein classes. For example, the molecular weight of class B GPCRs is much higher than that of other proteins such as kinases. The higher molecular weight leads to the presence of more chemical substructures in the fingerprint vector and can increase the amount of predicted targets for these ligands.[58]

In an effort to reduce the effect of these biases on ligand-based prediction probability, raw probabilities can be converted to a z-score.[53] In this method, for all molecules in the training set, a prediction score is obtained for all proteins in the training set. Subsequently, for each protein, a mean probability and standard deviation of this probability can be derived and converted into a z-score. By applying the same z-scoring for novel compounds rather than the raw probability, the predictions are converted to a number of standard deviations over or under the mean for that particular protein. This method has been shown to be more robust than using the raw probability.[58] Similarly, in structure-based inverse screening, the interaction score of the ligand with each target is compared with the interaction score distribution from a set of reference ligands of the respective target complex structures, taken from X-ray structures or determined by docking.[81,83,84]

Validation Strategies. The performance of *ligand-based models* should always be estimated using external test sets to minimize overfitting (besides cross-validation). If test sets are composed randomly, this may lead to overoptimistic performance values as similar ligands may be present in both training and test sets, resulting in "easy" predictions. In order to overcome this effect, cluster splits, where the whole cluster of similar molecules is either contained in the test or training set, or temporal splits, where data from the most recent years is used for testing, can be applied.[129] Predictive performances of ligand-based models can be estimated by metrics such as R^2 and Q^2 as well as error-based metrics such as the root-mean-square error (RMSE) and mean absolute error (MAE). It is debatable what the best metric is to indicate model performance as this is dependent on the data and validation method. Generally, performance can be better estimated when multiple metrics are considered.[130]

Evaluating the performance of *structure-based methods* is based on diverse strategies. Binding site comparison methods, for instance, often screen a query target against a set of true (well-studied protein class with subclass classification) and decoy targets, whereas inverse screening methods often test only one or few query ligands in a set of true (known targets of the ligand) and decoy targets. Evaluation metrics are, for instance, the percentage of true targets in the top x% of the ranked hit list, the so-called enrichment factor (EF), and the area under the curve (AUC). While different sizes and compositions of benchmark data sets and the diverse use of performance metrics hamper a direct comparison between methods, efforts to unify benchmarking have been made. Since binding site comparison is a long-established approach with many published methods, proposed data sets have often been reused. Such an example is the data set compilation by Weill and Rognan,[99] encompassing a set of similar and dissimilar structure pairs as well as sets focused on kinases and serine endopeptidases (all scPDB-based). Also concentrating on similar and dissimilar pairs, Ehrt et al.[131] have recently proposed a collection of new and reused data sets (ProSPECCTs) to test different performance aspects, which the authors applied to multiple binding site methods to establish guidelines for their application scope. For inverse screening methods, Schomburg et al.[83] proposed two data sets together with evaluation strategies: a small data set consisting of three target classes for detailed proof-of-concept and selectivity studies and a large data set with about 8000 protein structures and over 70 drug-like ligands. In addition to the widely used EF and AUC, the authors propose performance metrics capable of measuring the early enrichments, i.e., BEDROC (Boltzman-enhanced discrimination of ROC) and NSLR (normalized sum of logarithmic ranks).

■ CONCLUSION

Drug target identification is one of the most important, but also most complex, aspects of preclinical drug development. In this respect, computational target prediction is a highly valuable tool to identify the most probable targets for a compound under investigation. Such tools can guide wet lab experiments by suggesting potential targets for orphan compounds, supply tool compounds for functional analyses of poorly understood proteins, and thus help to decipher the mode-of-action of a protein under investigation. Furthermore, desired as well as undesired multitarget drug effects can be rationalized by computational (off-)target predictions, and known drugs can potentially be repositioned based on these forecasts.

Computational target prediction methods rely on the general assumption that similar molecules/structures will have similar interactions or interaction patterns. Exceptions are so-called activity cliffs, describing that small changes can cause large differences in activity.[29] Depending on the research question and the data available, ligand- or structure-based target prediction methods can be applied. In ligand-based methods, potential targets can either be inferred from the most similar known ligands or through elaborated machine learning models. The latter require sufficient and well annotated data in order to train proper models. Structure-based approaches compare a query protein based on their binding sites or interaction fingerprints to a panel of protein structures or screen a query compound against these panels using a docking or pharmacophore screening engine. It should be noted that usually ligand-centric methods are faster than structure-centric methods, especially when structural alignment or pose prediction is evoked. The former provides more quantitative information such as predicted bioactivities that can directly be associated with experimental values, whereas the latter can give additional information about the binding pose of ligands to potential targets. It should be noted that most methods do not consider alternate binding pockets on a single protein or the effect of protein complex formation. Although protein function or (de)activation through allosteric modulation can occur, most target prediction methods are based on the assumption that all ligands are orthosteric binders.

In our opinion, future progress needs to promote data coverage from both the ligand and protein point of view, e.g., annotation of non-biased bioactivities (reporting inactives) and deposition of novel structures or the same protein structures, but with different ligands to provide a better view on the dynamics of the ligand binding site (high-throughput crystallization). Furthermore, protein flexibility modeling and inter-target ranking are equally important matters to address. Moreover, new methods should be evaluated on standardized benchmarking data sets and performance metrics, as well as made accessible to the community in order to improve predictability, reliability, and reproducibility. Finally, holistic approaches should and will gain momentum, integrating multiple types of data, e.g., coupling chemical and structural space with information on the proteome level and pathways, linking cellular and molecular scales.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: gerard@lacdr.leidenuniv.nl (G.J.P. van Westen).
*E-mail: andrea.volkamer@charite.de (A. Volkamer).

ORCID

Dominique Sydow: 0000-0003-4205-8705
Lindsey Burggraaff: 0000-0002-2442-0443
Herman W. T. van Vlijmen: 0000-0002-1915-3141
Adriaan P. IJzerman: 0000-0002-1182-2259
Gerard J. P. van Westen: 0000-0003-0717-1817
Andrea Volkamer: 0000-0002-3760-580X

Author Contributions

‖D. Sydow and L. Burggraaff have shared cofirst authorship.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

G.J.P. van Westen thanks the Dutch Scientific Council (NWO) and Stichting Technologie Wetenschappen (STW) for funding (VENI Grant 14410). A. Volkamer and D. Sydow thank the Deutsche Forschungsgemeinschaft (DFG, grant VO 2353/1-1) and the Bundesministerium für Bildung und Forschung (BMBF, grant 031A262C) for funding.

REFERENCES

(1) Jenkins, J. L.; Bender, A.; Davies, J. W. In Silico Target Fishing: Predicting Biological Targets from Chemical Structure. *Drug Discovery Today: Technol.* **2006**, *3*, 413−421.

(2) Hart, C. P. Finding the Target After Screening the Phenotype. *Drug Discovery Today* **2005**, *10*, 513−519.

(3) Lee, J.; Bogyo, M. Target Deconvolution Techniques in Modern Phenotypic Profiling. *Curr. Opin. Chem. Biol.* **2013**, *17*, 118−126.

(4) Niphakis, M. J.; Cravatt, B. F. Enzyme Inhibitor Discovery by Activity-Based Protein Profiling. *Annu. Rev. Biochem.* **2014**, *83*, 341−377.

(5) Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol. Inf.* **2010**, *29*, 176−187.

(6) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(7) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204.

(8) Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.* **2014**, *57*, 7874−7887.

(9) Morphy, R.; Kay, C.; Rankovic, Z. From Magic Bullets to Designed Multiple Ligands. *Drug Discovery Today* **2004**, *9*, 641−651.

(10) AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring Polypharmacology Using a ROCS-Based Target Fishing Approach. *J. Chem. Inf. Model.* **2012**, *52*, 492−505.

(11) Bender, A.; Scheiber, J.; Glick, M.; Davies, J.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* **2007**, *2*, 861−873.

(12) Oprea, T. I.; et al. Drug Repurposing from an Academic Perspective. *Drug Discovery Today: Ther. Strategies* **2011**, *8*, 61−69.

(13) Keiser, M. J.; et al. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175−181.

(14) Ashburn, T. T.; Thor, K. B. Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nat. Rev. Drug Discovery* **2004**, *3*, 673−683.

(15) Berger, A. B.; Vitorino, P. M.; Bogyo, M. Activity-Based Protein Profiling: Applications to Biomarker Discovery, in Vivo Imaging and Drug Discovery. *Am. J. PharmacoGenomics* **2004**, *4*, 371−381.

(16) Schirle, M.; Bantscheff, M.; Kuster, B. Mass Spectrometry-Based Proteomics in Preclinical Drug Discovery. *Chem. Biol.* **2012**, *19*, 72−84.

(17) van Esbroeck, A. C. M.; et al. Activity-Based Protein Profiling Reveals Off-Target Proteins of the FAAH Inhibitor BIa 10−2474. *Science* **2017**, *356*, 1084−1087.

(18) 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*, D1100−D1107.

(19) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(20) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminf.* **2017**, *9*, 17.

(21) Rognan, D. Chemogenomic Approaches to Rational Drug Design. *Br. J. Pharmacol.* **2007**, *152*, 38−52.

(22) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(23) 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.

(24) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(25) Hawkins, P. C. D.; Stahl, G. In *Computational Methods for GPCR Drug Discovery*; Heifetz, A., Ed.; Springer: New York, 2018; pp 365−374.

(26) Shin, W.-H.; Zhu, X.; Bures, G. M.; Kihara, D. Three-Dimensional Compound Comparison Methods and Their Application in Drug Discovery. *Molecules* **2015**, *20*, 12841−62.

(27) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(28) Hu, Y.; Bajorath, J. High-Resolution View of Compound Promiscuity. *F1000Research* **2013**, *2*, 144.

(29) Bajorath, J. Representation and Identification of Activity Cliffs. *Expert Opin. Drug Discovery* **2017**, *12*, 879−883.

(30) Subramanian, A.; et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171*, 1437−1452.

(31) De Wolf, H.; Cougnaud, L.; Van Hoorde, K.; De Bondt, A.; Wegner, J. K.; Ceulemans, H.; Göhlmann, H. High-Throughput Gene Expression Profiles to Define Drug Similarity and Predict Compound Activity. *Assay Drug Dev. Technol.* **2018**, *16*, 162−176.

(32) Simm, J.; et al. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chem. Biol.* **2018**, *25*, 611−618.

(33) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data - a Statistical Analysis. *PLoS One* **2013**, *8*, No. e61007.

(34) Dalke, A. The FPS Fingerprint Format and Chemfp Toolkit. *J. Cheminf.* **2013**, *5*, P36.

(35) Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A Web Server for Target Prediction of Bioactive Small Molecules. *Nucleic Acids Res.* **2014**, *42*, W32−W38.

(36) Nickel, J.; Gohlke, B.-O.; Erehman, J.; Banerjee, P.; Rong, W. W.; Goede, A.; Dunkel, M.; Preissner, R. SuperPred: Update on Drug Classification and Target Prediction. *Nucleic Acids Res.* **2014**, *42*, W26−W31.

(37) Kauvar, L. M. Affinity Fingerprinting. *Nat. Biotechnol.* **1995**, *13*, 965−966.

(38) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Identifying Compound-Target Associations by Combining Bioactivity Profile Similarity Search and Public Databases Mining. *J. Chem. Inf. Model.* **2011**, *51*, 2440−2448.

(39) Lamb, J.; et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **2006**, *313*, 1929−1935.

(40) 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*, 197.

(41) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(42) 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*, 3389−3402.

(43) Pearson, W. R. Empirical Statistical Estimates for Sequence Similarity Searches. *J. Mol. Biol.* **1998**, *276*, 71−84.

(44) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature* **2012**, *486*, 361.

(45) Czodrowski, P.; Bolick, W.-G. OCEAN: Optimized Cross REActivity EstimatioN. *J. Chem. Inf. Model.* **2016**, *56*, 2013−2023.

(46) Cortes Cabrera, A.; Lucena-Agell, D.; Redondo-Horcajo, M.; Barasoain, I.; Díaz, J. F.; Fasching, B.; Petrone, P. M. Aggregated Compound Biological Signatures Facilitate Phenotypic Drug Discovery and Target Elucidation. *ACS Chem. Biol.* **2016**, *11*, 3024−3034.

(47) Grover, A.; Grover, M.; Sharma, K. A Practical Overview of Quantitative Structure-Activity Relationship. *World J. Pharm. Pharm. Sci.* **2016**, *5*, 427−437.

(48) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(49) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(50) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463−4470.

(51) Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513−530.

(52) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chemical Science* **2019**, *10*, 1692.

(53) Mugumbate, G.; Abrahams, K. A.; Cox, J. A. G.; Papadatos, G.; van Westen, G.; Lelièvre, J.; Calus, S. T.; Loman, N. J.; Ballell, L.; Barros, D.; Overington, J. P.; Besra, G. S. Mycobacterial Dihydrofolate Reductase Inhibitors Identified Using Chemogenomic Methods and in Vitro Validation. *PLoS One* **2015**, *10*, No. e0121492.

(54) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of Activity Spectra for Biologically Active Substances. *Bioinformatics* **2000**, *16*, 747−748.

(55) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 885−896.

(56) Afzal, A. M.; Mussa, H. Y.; Turner, R. E.; Bender, A.; Glen, R. C. A Multi-Label Approach to Target Prediction Taking Ligand Promiscuity into Account. *J. Cheminf.* **2015**, *7*, 24.

(57) Anger, L. T.; Wolf, A.; Schleifer, K.-J.; Schrenk, D.; Rohrer, S. G. Generalized Workflow for Generating Highly Predictive in Silico Off-Target Activity Models. *J. Chem. Inf. Model.* **2014**, *54*, 2411−2422.

(58) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminf.* **2017**, *9*, 45.

(59) Zhang, Q.; Hughes-Oliver, J. M.; Ng, R. T. A Model-Based Ensembling Approach for Developing QSARs. *J. Chem. Inf. Model.* **2009**, *49*, 1857−1865.

(60) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. *MedChemComm* **2011**, *2*, 16−30.

(61) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. a Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481−2491.

(62) van Westen, G. J. P.; Swier, R. F.; Cortes-Ciriano, I.; Wegner, J. K.; Overington, J. P.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. *J. Cheminf.* **2013**, *5*, 42.

(63) van Westen, G. J. P.; van den Hoven, O. O.; van der Pijl, R.; Mulder-Krieger, T.; de Vries, H.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. *J. Med. Chem.* **2012**, *55*, 7010−7020.

(64) van Westen, G. J. P.; Wegner, J. K.; Geluykens, P.; Kwanten, L.; Vereycken, I.; Peeters, A.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. *PLoS One* **2011**, *6*, No. e27518.

(65) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9*, 5441−5451.

(66) Oprea, T. I.; Nielsen, S. K.; Ursu, O.; Yang, J. J.; Taboureau, O.; Mathias, S. L.; Kouskoumvekaki, l.; Sklar, L. A.; Bologa, C. G. Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. *Mol. Inf.* **2011**, *30*, 100−111.

(67) Lo, Y.-C.; Senese, S.; Damoiseaux, R.; Torres, J. Z. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. *ACS Chem. Biol.* **2016**, *11*, 2244−2253.

(68) 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*, W39−W45.

(69) Zhao, S.; Li, S. Network-Based Relating Pharmacological and Genomic Spaces for Drug Target Identification. *PLoS One* **2010**, *5*, No. e11764.

(70) Chen, X.; Liu, M.-X.; Yan, G.-Y. Drug-Target Interaction Prediction by Random Walk on the Heterogeneous Network. *Mol. BioSyst.* **2012**, *8*, 1970−1978.

(71) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of DrugâĂŞtarget Interaction Networks from the Integration of Chemical and Genomic Spaces. *Bioinformatics* **2008**, *24*, i232−i240.

(72) OpenEye Scientific Software, ROCS. https://www.eyesopen.com/rocs (accessed on 2018−11−01).

(73) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471−490.

(74) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. *J. Cheminf.* **2015**, *7*, 51.

(75) 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*, 235−242.

(76) Volkamer, A.; von Behren, M. M.; Bietz, S.; Rarey, M. *Applied Chemoinformatics*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2018; pp 283−311.

(77) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: A 3D-Database of Ligandable Binding Sites-10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399−D404.

(78) Xu, X.; Huang, M.; Zou, X. Docking-Based Inverse Virtual Screening: Methods, Applications, and Challenges. *Biophys. Rep.* **2018**, *4*, 1−16.

(79) Huang, H.; Zhang, G.; Zhou, Y.; Lin, C.; Chen, S.; Lin, Y.; Mai, S.; Huang, Z. Reverse Screening Methods to Search for the Protein Targets of Chemopreventive Compounds. *Front. Chem.* **2018**, *6*, 138.

(80) Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: A Web Server for Identifying Drug Targets with Docking Approach. *Nucleic Acids Res.* **2006**, *34*, W219−W224.

(81) Wang, J.-C.; Chu, P.-Y.; Chen, C.-M.; Lin, J.-H. idTarget: A Web Server for Identifying Protein Targets of Small Chemical Molecules with Robust Scoring Functions and a Divide-And-Conquer Docking Approach. *Nucleic Acids Res.* **2012**, *40*, W393−W399.

(82) Chen, Y. Z.; Zhi, D. G. Ligand-Protein Inverse Docking and Its Potential Use in the Computer Search of Protein Targets of a Small Molecule. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 217−226.

(83) Schomburg, K. T.; Bietz, S.; Briem, H.; Henzler, A. M.; Urbaczek, S.; Rarey, M. Facing the Challenges of Structure-Based Target Prediction by Inverse Virtual Screening. *J. Chem. Inf. Model.* **2014**, *54*, 1676−1686.

(84) Santiago, D. N.; Pevzner, Y.; Durand, A. A.; Tran, M.; Scheerer, R. R.; Daniel, K.; Sung, S.-S.; Lee Woodcock, H.; Guida, W. C.; Brooks, W. H. Virtual Target Screening: Validation Using Kinase Inhibitors. *J. Chem. Inf. Model.* **2012**, *52*, 2192−2203.

(85) Ellingson, S. R.; Smith, J. C.; Baudry, J. VinaMPI: Facilitating Multiple Receptor High-Throughput Virtual Docking on High-Performance Computers. *J. Comput. Chem.* **2013**, *34*, 2212−2221.

(86) Li, G.-B.; Yu, Z.-J.; Liu, S.; Huang, L.-Y.; Yang, L.-L.; Lohans, C. T.; Yang, S.-Y. IFPTarget: A Customized Virtual Target Identification Method Based on Protein-Ligand Interaction Fingerprinting Analyses. *J. Chem. Inf. Model.* **2017**, *57*, 1640−1651.

(87) Scafuri, B.; Marabotti, A.; Carbone, V.; Minasi, P.; Dotolo, S.; Facchiano, A. A Theoretical Study on Predicted Protein Targets of Apple Polyphenols and Possible Mechanisms of Chemoprevention in Colorectal Cancer. *Sci. Rep.* **2016**, *6*, 32516.

(88) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998**, *70*, 1129−1143.

(89) Schuster, D. 3D Pharmacophores As Tools for Activity Profiling. *Drug Discovery Today: Technol.* **2010**, *7*, e205−e211.

(90) Wang, X.; Shen, Y.; Wang, S.; Li, S.; Zhang, W.; Liu, X.; Lai, L.; Pei, J.; Li, H. PharmMapper 2017 Update: A Web Server for Potential Drug Target Identification with a Comprehensive Target Pharmacophore Database. *Nucleic Acids Res.* **2017**, *45*, W356−W360.

(91) Liu, B.; Fu, X.-Q.; Li, T.; Su, T.; Guo, H.; Zhu, P.-L.; Tse, A. K.-W.; Liu, S.-M.; Yu, Z.-L. Computational and Experimental Prediction of Molecules Involved in the Anti-Melanoma Action of Berberine. *J. Ethnopharmacol.* **2017**, *208*, 225−235.

(92) Kellenberger, E.; Schalon, C.; Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209−220.

(93) Ehrt, C.; Brinkjost, T.; Koch, O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.* **2016**, *59*, 4121−4151.

(94) Schmitt, S.; Kuhn, D.; Klebe, G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002**, *323*, 387−406.

(95) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 137−145.

(96) Moriaud, F.; Richard, S. B.; Adcock, S. A.; Chanas-Martin, L.; Surgand, J.-S.; Ben Jelloul, M.; Delfaud, F. Identify Drug Repurposing Candidates by Mining the Protein Data Bank. *Briefings Bioinf.* **2011**, *12*, 336−340.

(97) Konc, J.; Janežič, D. ProBiS Algorithm for Detection of Structurally Similar Protein Binding Sites by Local Structural Alignment. *Bioinformatics* **2010**, *26*, 1160−1168.

(98) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287−2299.

(99) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.

(100) Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M.; Baroni, M.; Cruciani, G.; Perruccio, F.; Mason, J. S. High-Throughput Virtual Screening of Proteins Using GRID Molecular Interaction Fields. *J. Chem. Inf. Model.* **2010**, *50*, 155−169.

(101) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339*, 607−633.

(102) von Behren, M. M.; Volkamer, A.; Henzler, A. M.; Schomburg, K. T.; Urbaczek, S.; Rarey, M. Fast Protein Binding Site Comparison Via an Index-Based Screening Technology. *J. Chem. Inf. Model.* **2013**, *53*, 411−422.

(103) Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large-Scale Mining for Similar Protein Binding Pockets: With RAPMAD Retrieval on the Fly Becomes Real. *J. Chem. Inf. Model.* **2015**, *55*, 165−179.

(104) Brakoulias, A.; Jackson, R. M. Towards a Structural Classification of Phosphate Binding Sites in Protein-Nucleotide Complexes: An Automated All-Against-All Structural Comparison Using Geometric Matching. *Proteins: Struct., Funct., Genet.* **2004**, *56*, 250−260.

(105) Aung, Z.; Tong, J. C. BSAlign: A Rapid Graph-Based Algorithm for Detecting Ligand-Binding Sites in Protein Structures. *Genome Inform* **2008**, *21*, 65−76.

(106) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A Simple and Fuzzy Method to Align and Compare Druggable Ligand-Binding Sites. *Proteins: Struct., Funct., Genet.* **2008**, *71*, 1755−1778.

(107) De Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of Protein Kinase Inhibitors to Synapsin I Inferred from Pair-Wise Binding Site Similarity Measurements. *PLoS One* **2010**, *5*, No. e12214.

(108) Salentin, S.; Haupt, V. J.; Daminelli, S.; Schroeder, M. Polypharmacology Rescored: Protein-Ligand Interaction Profiles for Remote Binding Site Similarity Assessment. *Prog. Biophys. Mol. Biol.* **2014**, *116*, 174−186.

(109) Cao, R.; Wang, Y. Predicting Molecular Targets for Small-Molecule Drugs with a Ligand-Based Interaction Fingerprint Approach. *ChemMedChem* **2016**, *11*, 1352−1361.

(110) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(111) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623−637.

(112) Da, C.; Kireev, D. Structural Protein-Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* **2014**, *54*, 2555−2561.

(113) Kinoshita, K.; Nakamura, H. Identification of Protein Biochemical Functions by Similarity Search Using the Molecular Surface Database EF-Site. *Protein Sci.* **2003**, *12*, 1589−1595.

(114) Reisen, F.; Weisel, M.; Kriegl, J. M.; Schneider, G. Self-Organizing Fuzzy Graphs for Structure-Based Comparison of Protein Pockets. *J. Proteome Res.* **2010**, *9*, 6498−6510.

(115) Yeturu, K.; Chandra, N. PocketMatch: A New Algorithm to Compare Binding Sites in Protein Structures. *BMC Bioinf.* **2008**, *9*, 543.

(116) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031−2043.

(117) Liu, T.; Altman, R. B. Using Multiple Microenvironments to Find Similar Ligand-Binding Sites: Application to Kinase Inhibitor Binding. *PLoS Comput. Biol.* **2011**, *7*, No. e1002326.

(118) Siragusa, L.; Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. BioGPS: Navigating Biological Space to Predict Polypharmacology, Off-Targeting, and Selectivity. *Proteins: Struct., Funct., Genet.* **2015**, *83*, 517−532.

(119) Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C. TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins. *J. Chem. Inf. Model.* **2013**, *53*, 1235−1252.

(120) 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*, D668−D672.

(121) Hilbert, M.; López, P. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science* **2011**, *332*, 60−65.

(122) Eeckhout, L. Is Moore's Law Slowing Down? What's Next? *IEEE Micro* **2017**, *37*, 4−5.

(123) Jasial, S.; Hu, Y.; Bajorath, J. Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping. *J. Chem. Inf. Model.* **2016**, *56*, 300−307.

(124) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? *J. Chem. Inf. Model.* **2006**, *46*, 2445−2456.

(125) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. *J. Chem. Inf. Model.* **2013**, *53*, 2499−2505.

(126) Somody, J. C.; MacKinnon, S. S.; Windemuth, A. Structural Coverage of the Proteome for Pharmaceutical Applications. *Drug Discovery Today* **2017**, *22*, 1792−1799.

(127) Volkamer, A.; Eid, S.; Turk, S.; Jaeger, S.; Rippmann, F.; Fulle, S. Pocketome of Human Kinases: Prioritizing the ATP Binding Sites of (Yet) Untapped Protein Kinases for Drug Discovery. *J. Chem. Inf. Model.* **2015**, *55*, 538−549.

(128) Mortier, J.; Dhakal, P.; Volkamer, A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. *Molecules* **2018**, *23*, 1959.

(129) Sheridan, R. P. Time-Split Cross-Validation As a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.

(130) Gramatica, P.; Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. *J. Chem. Inf. Model.* **2016**, *56*, 1127−1131.

(131) Ehrt, C.; Brinkjost, T.; Koch, O. A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (ProSPECCTs). *PLoS Comput. Biol.* **2018**, *14*, No. e1006483.