JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

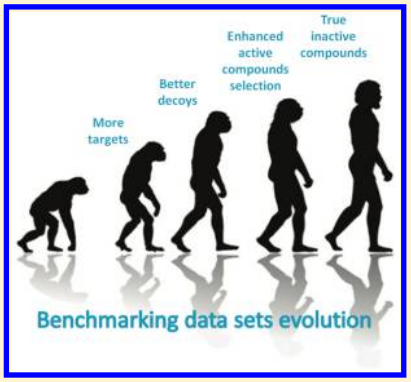
pubs.acs.org/jcim

Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives

Nathalie Lagarde, Jean-François Zagury, and Matthieu Montes*

Laboratoire Génomique, Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France

ABSTRACT: Virtual screening methods are commonly used nowadays in drug discovery processes. However, to ensure their reliability, they have to be carefully evaluated. The evaluation of these methods is often realized in a retrospective way, notably by studying the enrichment of benchmarking data sets. To this purpose, numerous benchmarking data sets were developed over the years, and the resulting improvements led to the availability of high quality benchmarking data sets. However, some points still have to be considered in the selection of the active compounds, decoys, and protein structures to obtain optimal benchmarking data sets.



INTRODUCTION

Drug discovery is a highly time-consuming and expensive process, estimated to last between 12 and 14 years for an approximate total cost of 1 billion dollars that encompasses all steps from the definition of a therapeutic target of interest to the marketing of a new drug.[1,2] The methods used in this process have to be optimized to reduce its cost and duration and to obtain the most promising drug candidate. The "hit" identification step, corresponding to the search of compounds able to interact with the target and to modulate its action, is often achieved by screening large databases. This screening can be either performed *in vitro*, by high throughput screening (HTS) facilities, or *in silico* using virtual ligand screening (VLS) methods. Current drug discovery strategies combine both protocols.[3]

In silico methods are easy to set up, require little financial investment, and are relatively fast, depending on the computational capacity of the research group.[4] These methods are often used as preliminary filters of large compound collections, and only the predicted most promising molecules are tested experimentally.[5]

The predictive power of these methods has to be assessed to ensure their reliability and to guide the choice of the most performing tools and protocols. Ideally, such evaluations should be performed prospectively by confirming experimentally predicted binding affinities values.[6] However, such studies would be very expensive to perform and could be approached with available HTS data that pharmaceutical companies tend to protect for intellectual property concerns.[7]

The evaluation of virtual screening methods is then achieved retrospectively by assessing two main criteria: the enrichment of benchmarking data sets in active compounds and the accuracy of the prediction of their binding mode.[8]

Benchmarking databases gathers, for one or more targets, two types of compounds: active compounds, i.e. compounds with a known and documented activity on the target of interest and inactive compounds. Ideally, inactive compounds should be selected just as active compounds, on the basis of experimental data proving their lack of activity on the target studied. However, compounds that are found to be inactive for a target are seldom described in scientific literature, and benchmarking data sets use instead assumed nonbinding compounds named "decoys".[8]

This review is organized in two parts. The first provides an overview of the main benchmarking data sets developed over the years, focusing on their specificities, the major improvements they brought, and their weakness. The second focuses on the issues that should be considered to improve the quality of the benchmarking data sets and thus enable better evaluations of virtual screening methods.

BENCHMARKING DATA SETS: AN HISTORIC PERSPECTIVE

Numerous benchmarking data sets were developed among the years (Figure 1 and Table 1), but the first was proposed by Bissantz et al. in 2000.[9] This database was constructed for the purpose of evaluating 3 docking softwares (DOCK,[10] FlexX,[11] and GOLD[12]) and 7 scoring functions (Dock, FlexX,[11] GOLD,[12] PMF,[13] Chemscore,[14] Fresno,[15] and Score[16]). Only 2 targets, estrogen alpha receptor (ER_alpha) and Thymidine Kinase (TK), were included in the benchmarking database. For each target, the database gathered 1 PDB structure, 10 active compounds, and 990 decoys randomly selected in the

Received: February 17, 2015
Published: June 3, 2015

ACS Publications © 2015 American Chemical Society

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

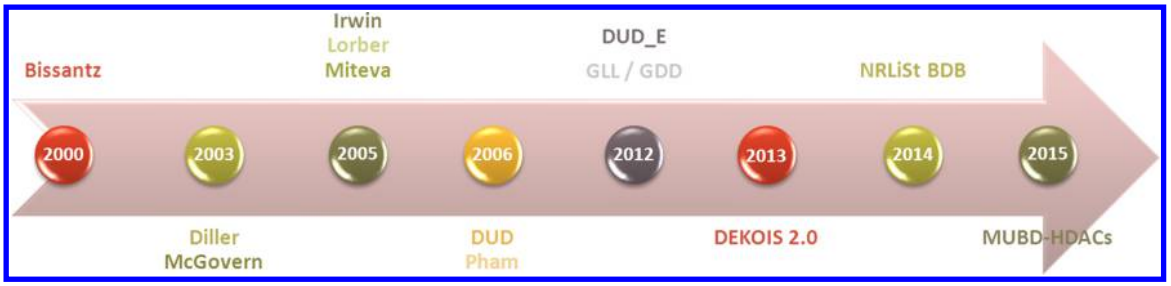


Figure 1. Chronological representation of the main benchmarking data sets developed over the years for the evaluation of virtual screening methods.

Advanced Chemical Directory (ACD),[17] previously filtered to eliminate undesired compounds (chemical reagents, inorganic compounds, and molecules with a molecular weight lower than 250 or higher than 500 Da).

Three years later, McGovern et al.[18] used the MDL Drug Data Report (MDDR)[19] to develop a benchmarking database for 10 enzymatic targets (DHFR, PNP-PO4, PNP+PO4, PARP, Thrombin, GART, SAHH, AR, TS) in the aim to evaluate the impact of the query conformation of the binding site on docking performances. The MDDR provided both the 2200 active compounds (from 35 for the two PNP data sets to 908 for the AR data set) and the 95579 decoys (corresponding to the entire MDDR database filtered to eliminate undesirable compounds). For each target, the data set contained also 3 conformations of the protein: a holo X-ray structure, an apo X-ray structure, and a modeled structure.

The same year, a benchmarking database dedicated to Tyrosine and Serine/Threonine kinase families[20] was constructed to study the impact of using homology models for high throughput docking. Diller et al. chose to focus on the kinase family for three main reasons: 1. their high potential as therapeutic targets, 2. the large number of existing human protein kinases, and 3. the availability of 3D structures resolved by X-ray crystallography. Data sets were formed, for 6 kinases (EGFr, FGFr1, VEGFr1, PDGFrB, P38, and SRC), using 958 kinase inhibitors assembled from the scientific literature (from 46 inhibitors for VEGFr1 to 387 for EGFr), 32000 decoys displaying similar properties, in terms of polarity (number of rotable bonds, hydrogen bond donors and acceptors) and molecular weight, to kinase inhibitors to avoid selection bias on the size of compounds randomly selected from an internal collection and homology models. This represents the first attempt to obtain better decoys than using randomly selected compounds. However, the use of this benchmarking data set was limited since the MDDR was a nonpublic access database.[21]

Starting from the MDDR,[19] Lorber et al.[22] selected active compounds for 7 unrelated targets (DHFR, Neutral Endopeptidase, Thrombin, Thymidylate Synthase, Phospholipase C, Adenosine Kinase, Acetylcholinesterase) to investigate a hierarchical preorganization of multiple conformations of small molecules. A total of 2201 active ligands (from 25 Phospholipase C ligands to 788 Thrombin ligands) were gathered with about 98500 decoys resulting from the filtering of the MDDR database to remove reactive functionalities, aliphatic chains longer than hexane, and molecules containing silicon. For each target, both apo and holo conformations of the protein resolved by X-ray crystallography were provided.

Using the same protocol as McGovern et al.,[18] Irwin and coworkers[23] constructed a benchmarking database to investigate the ability of molecular docking methods to identify potential ligands of metalloenzymes using a "standard" scoring function. Five metalloenzymes (carbonic anhydrase, matrix metalloproteinase, neutral endopeptidase, peptide deformylase and xanthine oxidase) with a sufficient number of ligands and an available X-ray structure of the protein were included in the database. The data sets contained from 26 (peptide deformylase) to 337 active compounds resulting in a total of 862 active molecules combined to the 95579 decoys previously generated.[18]

To evaluate the performances obtained by combining three VLS methods freely available to academic users (FRED,[24] Surflex-dock,[25] and DOCK[26]), 4 targets were selected according to the availability and diversity of active ligands and protein structures (Estrogen Receptor, Thymidine Kinase, Coagulation Factor VIIa, and Neuraminidase).[27] 49 known inhibitors were chosen in the literature (10 active compounds per target for all but coagulation factor VIIa) to constitute the active data sets. The decoys were obtained by filtering the ACD database[17] according to several criteria: molecular weight, number of carbons, rings and rotable bonds, the atomic composition, number of hydrogen bond donor and acceptor, sum of formal charges, XlogP and presence of toxic functional group, resulting in 65611 compounds. For each target, a cocrystallized structure of the protein was extracted from the PDB to complete each data set.

In 2006, Pham and Jain[28] integrated the Bissantz database[9] with 27 new data sets into a new benchmarking database. Active compounds were retrieved from previous work for 2 targets (Poly ADP-Ribose Polymerase PARP[29] and Protein Tyrosine Phosphatase 1b (PTP1b)[30]) or generated using the PDBbind database[31] for the 25 others. Bissantz's decoys[9] were preserved but filtered for presenting less than 15 rotable bonds, and for the 27 new targets, 1000 decoys were randomly selected in the "drug like" portion of the ZINC database.[32] A total of 226 active ligands and 1861 decoys were included in the data sets, and for each target, one PDB structure was selected.

Until this point, most databases had a lot in common. They classically used a database of annotated ligands as a source of both active and decoys compounds, and decoys were selected by filtering the database to remove undesirable compounds or present "drug like" properties. However, Diller et al.[20] proposed a more accurate selection of the decoys by choosing compounds with similar polarity and molecular weight to active compounds to overcome selection bias on the size of the compounds.

To enhance the selection of the decoys, the DUD (Directory of Useful Decoys)[21] was created and quickly became the widely used standard for the evaluation of VLS methods. The major improvements provided by the DUD are a rational decoys selection and a large number (40) of data sets to ensure a complete and robust evaluation of VLS methods. To reduce

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

Table 1. Overview of the Main Benchmarking Data Sets Developed over the Years

benchmarking database	year	no. of data sets	no. of active ligands	no. of decoys	decoys selection
Bissantz et al.[9]	2000	2	20	1980	random selection in the ACD previously filtered to eliminate undesired compounds
McGovern et al.[18]	2003	10	2200	95579	entire MDDR database filtered to eliminate undesirable compounds
Diller et al.[20]	2003	6	958	32000	random selection from an internal collection of compounds previously filtered to keep only molecules presenting similar polarity and molecular weight to actives
Lorber et al.[22]	2005	7	2201	98500	entire MDDR database filtered to eliminate undesirable compounds
Irwin et al.[23]	2005	5	862	95579	entire MDDR database filtered to eliminate undesirable compounds
Miteva et al.[27]	2005	4	49	65611	entire ACD database filtered according to drug-like criteria
Pham and Jain[28]	2006	29	226	1861	861 decoys randomly selected in the ACD previously filtered to eliminate undesired compounds and with less than 15 rotable bonds and 1000 decoys randomly selected in the "drug like" portion of the ZINC database
DUD[21]	2006	40	2950	95316	extracted from the Lipinski-compliant fraction of the ZINC database compounds to be structurally dissimilar to the actives but to present similar physicochemical properties
DUD-E[54]	2012	102	66695	1420433	extracted from the Lipinski-compliant fraction of the ZINC database compounds to be structurally dissimilar to the actives but to present similar physicochemical properties
GLL − GDD[60]	2012	147	25145	980655	extracted from the ZINC database according to their similarity with active compounds in terms of physicochemical properties (molecular weight of ±50, identical number of rotable bond and hydrogen bond donors and acceptors, logP of ±1 and identical formal charge) and to their structural dissimilarity (Tc < 0.75)
DEKOIS 2.0[95]	2013	81	3240	97200	extracted from the ZINC database according to their high physicochemical similarity with active compounds (molecular weight, logP, hydrogen bond donors and acceptors, number of rotatable bonds, positive and negative charge, number of aromatic rings) and the avoidance of latent actives in the decoy set (LADS)
NRLiSt BDB[62]	2014	54	9905	458981	DUD-E decoy generation tool
MUBD-HDACs[63]	2015	14	631	24609	in two steps: a "Preliminary Filtering" extract compounds of the ZINC database according to their similarity with active compounds in terms of physicochemical properties and to their dissimilarity with the structure of the active compounds followed by a "Precise Filtering" in which 39 decoys are selected on the basis of their high similarity in terms of physicochemical properties and their random spatial distribution to each active compound

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

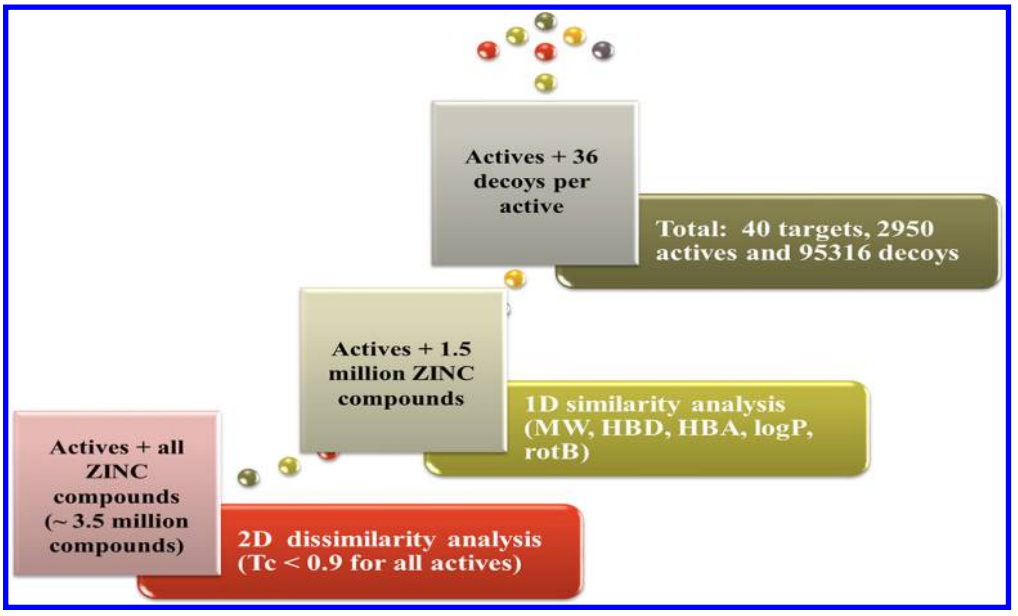


Figure 2. Protocol used for the generation of the decoys of the DUD benchmarking data sets (with MW: Molecular Weight, HBD: Hydrogen Bond Donor, HBA: Hydrogen Bond Acceptor, logP: octanol−water partition coefficient, rotB: number of rotable bonds).

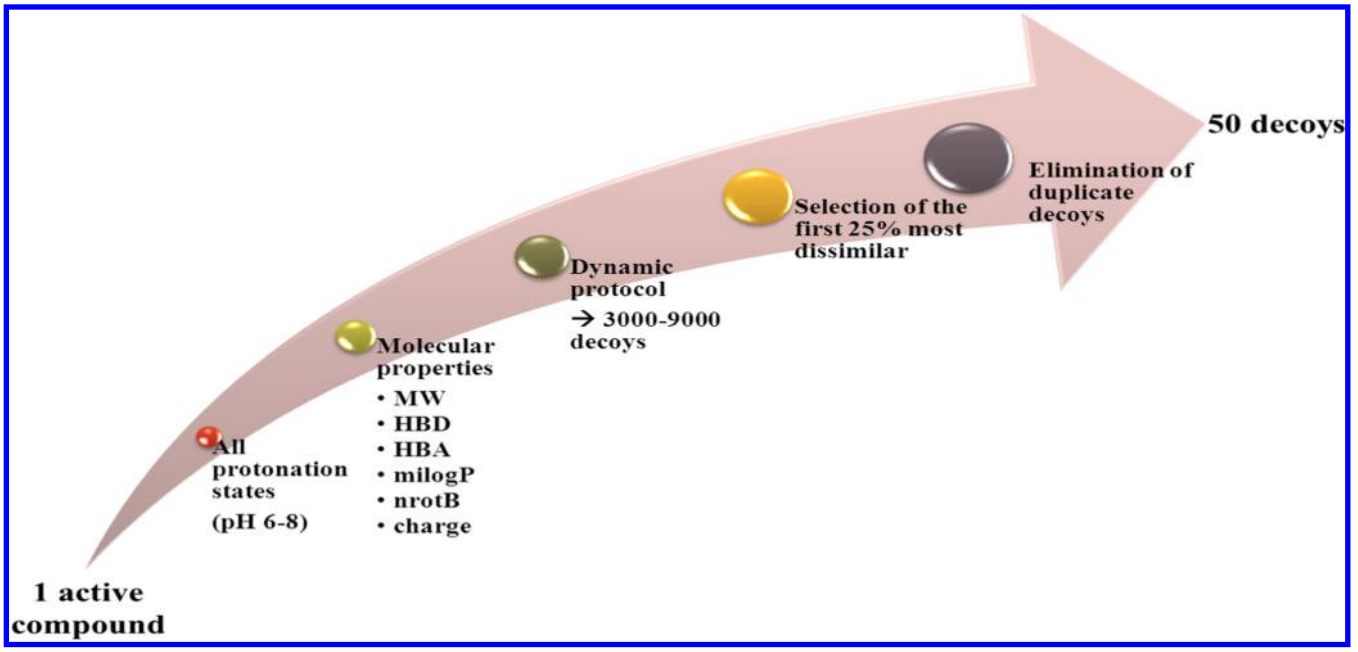


Figure 3. Protocol used for the generation of the decoys of the DUD-E benchmarking data sets (with MW: Molecular Weight, HBD: Hydrogen Bond Donor, HBA: Hydrogen Bond Acceptor, logP: octanol−water partition coefficient, rotB: number of rotable bonds).

selection bias, decoys were chosen to present similar physicochemical properties to active molecules but also to be structurally different in an attempt to select compounds unable to bind to the studied target (Figure 2). Active ligands were retrieved from several bioactivity databases (KiBank,[33] PDBbind database,[31] PubChem[34]) or previous studies.[9,20,35−53] Decoys were extracted from the Lipinski-compliant fraction of the ZINC database compounds[32] in two steps. First, using CACTVS type 2 fingerprints, only 1.5 million ZINC compounds, topologically dissimilar to the active compounds, i.e. with a Tanimoto coefficient inferior to 0.9 with any active compounds, were preserved. Then, among these 1.5 million compounds, 36 decoys were selected for each active ligand, on the basis of the similarity of their physical properties to active compounds, in particular the molecular weight, the number of hydrogen bond acceptors and donors, the number of rotable bonds, the log P, and to a less extent the number of important functional groups (amine, amide, amidine, and carboxylic acid). Finally, the DUD is constituted of 40 data sets gathering 2950 active ligands (from 12 COMT ligands to 416 EGFr ligands), 95316 decoys, and a 3D structure extracted from the PDB for each target.

In 2012, an enhanced version of the DUD, the DUD-E database, was proposed.[54] The DUD-E presented a large number of data sets (102 targets divided in 8 protein categories) and was constructed to address some weaknesses of the ligands and decoys of the original DUD.[21] In particular, several studies pointed out that the charge was not considered

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

during the property-matching selection of the decoys[6,55,56] and that some decoys were able to bind to their corresponding target.[57] A total of 66695 active compounds were selected using the ChEMBL database[58] and included in the DUD-E database (from 50 for GART to 3090 for FXa). The DUD-E decoys generation protocol was slightly different than the one used for the DUD (Figure 3). For each protonation state of each active molecule in a pH range of 6 to 8, 3000 to 9000 DUD-E decoys were extracted from the ZINC database using a dynamic protocol that narrowed or widened the local chemical space around 6 properties of the given active ligand: molecular weight, estimated water-octanol partition coefficient, rotable bonds, hydrogen bond acceptors and donors, and net charge. Using ECFP4 fingerprints[59] and Tanimoto coefficient, a more stringent filtering on topological dissimilarity was applied, and only the first 25% most dissimilar ZINC compounds were retained. After elimination of the duplicate molecules, 50 decoys were randomly selected from the resulting compounds for each active compound, leading to a total of 1420433 decoys.

The G protein-coupled receptors (GPCR) Ligand Library (GLL) and GPCR Decoy Database (GDD)[60] were created following the DUD model[21] to optimize the drug discovery protocols toward this important class of therapeutic targets. The receptors and active compounds were selected by selecting in the GLIDA database the compounds with a biological activity. The particularity of this database lies on the separation of the active compounds in agonist and antagonist ligands data sets. The resulting database contained 25145 agonist and antagonist ligands, extracted from the PubChem database and correctly prepared, for a total of 147 targets. For each active compound in the ZINC database, 39 decoys were chosen according to their similarity with the given active compound in terms of physicochemical properties (molecular weight of ±50, identical number of rotable bond and hydrogen bond donors and acceptors, logP of ±1, and identical formal charge) and to their structural dissimilarity (Tc < 0.75) with the given active compound.

The DEKOIS 2.0[95] proposed 81 benchmarking sets for 80 protein targets, selected according to their interest as therapeutic targets. This database can be used as an extension and a complement to the current available benchmarking data sets since an effort was made to include nonclassical binding sites such as protein−protein interaction targets or allosteric binding sites for example. DEKOIS 2.0 was constructed using an original protocol for the selection of both active compounds and decoys. The active ligands were retrieved from the BindingDB[96] using several filters: exclusion of compounds identified as weak binders, substructure filtering of potential false positive PAINS (pan assay interference compounds), and elimination of ligands with nonspecified stereocenters or reactive groups. Finally, to ensure a good structural diversity of active compounds in the DEKOIS 2.0, for each target, active ligands were divided in 40 clusters using FCFP_6 descriptors, and only the most potent compound of each cluster was included in the data set. The decoy sets were constructed following a similar but improved 4 steps protocol to the one previously presented and known as DEKOIS.[57] In the first step, 15 million ZINC compounds are divided into bins according to 8 physicochemical properties (molecular weight, logP, hydrogen bond acceptors and donors, number of rotatable bonds, positive and negative charge, and aromatic rings). These bins are then associated to create cells (eight-dimensional bins) accounting for each combination of the eight physicochemical properties. In the second step, actives compounds are assigned to its corresponding cells according to its physicochemical properties. In the third step, 1500 decoys are extracted for each active from the corresponding cells. Finally, 30 decoys among the 1500 initial potential decoys are selected according to high physicochemical similarity between actives and decoys evaluating by a physicochemical similarity score (PSS) and the elimination of latent actives in the decoy set (LADS) using a LADS Score based on FCFP_6 fingerprints. For each target, the PDB structure of the protein presenting the best resolution was also provided (holo structures were preferred to apo structures whenever possible).

The DUD, and by extension the DUD-E, are considered as the current widely used standard to evaluate VLS methods. However, in a study of 2013, Ben Nasr et al. pointed out that the quality of these databases could still be enhanced, especially by paying more attention to the selection of the active compounds and by constructing separate data sets according to the pharmacological profile of the ligands.[61] Thus, taking into account these points, in 2014, the Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB),[62] a benchmarking database dedicated to the ligands and structures of nuclear receptors (NRs), was constructed. The NRLiSt BDB gathered one agonist data set and one antagonist data set for 27 human NRs (out of the 48 human NRs known) according to the availability of both ligands and target experimental structures. All 339 human experimental holo structures of these 27 targets were extracted from the PDB and classified according to the pharmacological profile of the cocrystallized ligand (agonist-bound structures, antagonist-bound structures, and other-bound structures). Similarly, 9905 active ligands (7853 agonists and 2052 antagonists) were manually collected during an extensive review of the scientific literature, resulting in 54 data sets of different sizes according to data availability (from 2 CAR antagonists to 1820 PPAR gamma agonists). In this case, bioactivity databases were not used as a reference for the compounds to include in the database but as a source of bibliographic references in which each data was manually checked. 458981 decoys were obtained using the DUD-E decoy generation tool.[54]

In a very recent publication, Xia et. al[63] presented a new benchmarking database suitable for the evaluation of both ligand and structure-based VLS methods. The Maximal Unbiased Benchmarking Data sets for Histone Deacetylases Inhibitors (MUBD-HDACs) targets and ligands were selected according to ChEMBL.[58] A total of 14 data sets were constructed, among which 5 (HDAC5, HDAC6, HDAC9, HDAC10, and HDAC11) were entirely dedicated to the evaluation of ligand-based virtual screening methods since no experimental structure of the protein was available for these targets. Starting from 3036 active ligands, only 631 active compounds were included in the MUBD-HDACs after the analogue excluding process to reduce analogue bias (from 16 molecules for the SIRT3 data set to 180 molecules for the HDAC1 data set). Xia et al. adapted and optimized the DUD decoys selection protocol and proposed a new decoy selection algorithm[64,65] in two steps. In the first step ("Preliminary Filtering"), a classical approach was used to extract compounds of the ZINC database according to their similarity with active compounds in terms of physicochemical properties and according to their dissimilarity with the structure of the active compounds. The enhancement in the decoys selection occurred in the second step called "Precise Filtering" in which 39 decoys

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

were selected on the basis of their high similarity in terms of physicochemical properties and their random spatial distribution to each active compound.

CAVEATS AND IMPROVEMENTS

The evaluation of virtual screening methods is a crucial point to ensure their reliability and to validate the results obtained in a prospective virtual screening protocol. The quality of the evaluation is directly correlated to the availability and to the quality of benchmarking data sets. The scientific literature is full of examples of a successful use of benchmarking data sets to optimize a virtual screening protocol that led to the discovery of new potent compounds for a given target. For structure-based virtual screening methods, benchmarking data sets are used to choose the protocol, conformational search algorithm, and scoring function that should be used to ensure best performances for a given target. For example, Kobayashi et al.[66] used the DHFR (dihydrofolate reductase) DUD data set[21] to choose the best protocol to identify novel potential antibiotics against *Staphylococcus epidermis*. They compared the performances of DOCK 6.4[67] and GOLD 5.0.1[12] using a single conformer compound collection and the performance of GOLD 5.0.1 using multiple conformer compound collection. According to the results of this evaluation, Kobayashi et al. defined a hierarchical virtual screening protocol, starting from the less accurate method (DOCK 6.4) as a primary filter, using GOLD 5.0.1 as a refinement with a single conformation of each compound and finally using GOLD 5.0.1 in multiple conformers mode. Using this hierarchical protocol, a compound, KB1, displaying a strong inhibitory effect on the growth of *Staphylococcus epidermis*, was identified. Similarly, in order to choose the virtual screening protocol to identify new AR (Androgen Receptor) binders, Bobach et al.[68] realized an evaluation of several structure-based virtual screening methods (GOLD 3.1 using ChemScore[14] and GoldScore[12] scoring functions; MOE using different scoring functions;[69] PLANTS using Chemplp[70,71] scoring function). The best performing protocol that used PLANTS and Chemplp scoring function led to the successful identification of 11 new AR ligands. Concerning ligand-based virtual screening methods, and particularly pharmacophore screening methods, the availability of high quality benchmarking data sets is also critical to select the best pharmacophore hypotheses, i.e. those able to distinguish active compounds from inactive ones.[72] As an example, Balaji et al.[73] used the COX-1 (Cyclooxygenase-1) DUD data set[21] to validate a 3D-QSAR pharmacophore model for COX-1 inhibition. Using this validated pharmacophore model in a prospective study combining pharmacophore-based filtering and docking studies, 5 potent hit compounds were identified. All those examples illustrate the importance and the large utilization of benchmarking data sets in drug discovery.

Since the year 2000 and the first data sets proposed by Bissantz et al.,[9] numerous other data sets were constructed and the successive improvements that were performed, especially in terms of 1. the diversity of targets, 2. the diversity of ligands, and 3. the selection of appropriate decoys, led to high quality and reliable benchmarking data sets. In the present work, we proposed some guidelines that could be helpful to identify the optimal benchmarking data set(s) according to the aim(s) of the evaluation. Additionally, since the currently available benchmarking data sets are still not perfect, we highlighted the errors to avoid and the potential improvements that can be performed to constitute better benchmarking data sets according to their key elements: the active ligands, the decoys, and the structure(s) of the target.

Selection of the Optimal Benchmarking Data Set According to the Purpose of the Evaluation. The first part of this review highlighted that a large number of benchmarking data sets are currently available for the evaluation of virtual screening methods, raising the following question: "which benchmarking data set(s) should I use to validate my virtual screening protocol?". This choice is critical since the performances of a virtual screening method can vary considerably with the benchmarking data set used for the study.[74,75] The first element that can guide the selection of a benchmarking data set is the nature of the virtual screening method that will be evaluated. Numerous data sets were developed for the evaluation of structure-based virtual screening methods, in particular the DUD[21] and the DUD_E.[54] Even if these benchmarking data sets were largely used for the evaluation of ligand-based virtual screening methods, some publications pointed out that artificial enrichments in active compounds could be obtained[76−78] and that the construction of ligand-based and structure-based benchmarking data sets should be completely different.[6] These artificial good performances resulted from two different biases: "complexity bias" and "analogue bias".[76] The "complexity bias" lies in the differences in structural complexity between active compounds, that are often highly optimized compounds extracted from the scientific and patent literature, and inactive compounds, that are often less structurally complex.[79] The evaluation of the performance of ligand-based virtual screening methods can then be biased since they can easily discriminate compounds according to their differences in structural complexity. The "analogue bias" results from the presence of a large number of active compounds issued from the same chemical series used as a reference. Some benchmarking data sets were therefore constructed specifically for the evaluation of ligand-based virtual screening methods, containing active compounds that displayed significantly reduced structural complexity and less structural analogues.[76,80,81] Additionally, benchmarking data sets adapted for the evaluation of ligand-based virtual screening methods can be derived from benchmarking data sets designed for the evaluation of structure-based virtual screening methods like the DUD LIB VS 1.0[82] derived from the DUD. In our opinion, even it is not incorrect to use benchmarking data sets constructed for the evaluation of structure-based virtual screening methods to assess the performance of ligand-based virtual screening methods, a more accurate evaluation will be performed using finely tuned data sets. For example, benchmarking data sets for the evaluation of 2D methods should include decoys highly structurally similar to the active compounds, i.e., the exact opposite of the DUD- and DUD_E-like decoys. The ideal solution would be to select data set dedicated to the target on which the prospective study will be conducted, which would allow a fine-tuning of the resulting prospective virtual screening protocol.

Evaluation of the Scaffold Hopping Potential. Scaffold hopping, i.e. the ability to identify active compounds that structurally differ from reference molecules, is one of the major goals of ligand-based virtual screening methods.[79] However, the evaluation of the scaffold hopping potential of virtual screening approaches remains an uneasy task[83] partly because the commonly used benchmarking data sets are not suitable for this purpose. Some efforts were made in the last years to overcome this issue. As an example, Rohrer and Baumann[75]

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

proposed a methodology to characterize the topology of a data set that they applied to extract subsets displaying well-defined distinct topologies from previously available benchmarking data sets. Their work was based on the use of a single scalar describing the properties of the data set such as the number of clusters, their respective densities, and their relative distances. In a similar way, Vogt et al.[84] developed a benchmarking system specifically designed for the evaluation of the scaffold hopping potential with 1675 compounds representing 334 unique scaffolds for 17 targets. Li et al.[83] also developed a useful tool for the evaluation of the scaffold hopping potential, which enabled to quantify the structural distance between different scaffolds, regardless of their chemical composition, topology, or size. They used their tool to identify, within the ChEMBL database,[58] data sets that were suitable to benchmark ligand-based virtual screening methods.[85] All those findings should be taken into account for the construction of benchmarking data sets but not just for the selection of the active compounds as it is currently done but also for the selection of their decoys. Indeed, we assume that the perfect benchmarking data sets for the evaluation of the scaffold hopping potential should display an equilibrated compound repartition between the different scaffolds.

Selection of Active Compounds. The selection of the active compounds is a relatively easy task thanks to the data available in the literature in particular with bioactivity databases. Several studies notified that the bioactivity databases should be used with caution since errors in the data they provided could arise[62,86] (Table 2). A manual curation of all data is then required before inclusion to limit the integration of errors. This represents an extremely time-consuming but necessary process to provide a reliable database. This careful selection of active molecules must include the assessment that they all target the same binding site in the protein. A separation of the active ligands according to their mechanism of action or pharmacological profile is very important to guarantee a high quality evaluation of the methods.[61,62,87] As examples, a separation into distinct data sets of agonist ligands and antagonist ligands for nuclear receptors or allosteric site inhibitors and catalytic site inhibitors for enzymatic targets should be performed.

The diversity of the compounds within each data set is also a possible source of bias since overestimated performances can be obtained with data sets containing numerous highly similar ligands for a given target.[6] However, this "analogue bias" can be either detected by clustering ligands using a Tanimoto-based structure similarity search[6,78] or corrected with modifications of the ROC curves by weighting the detection of the active ligands on the size of their corresponding clusters.[88]

Decoys Selection. The first benchmarking data sets were constructed by randomly selecting decoys in large compound collections such as the ACD[17] or the MDDR.[19] However, it rapidly appeared that artificial good enrichments could be obtained with data sets gathering ligands and decoys significantly different in terms of size[89] and that this observation could be extrapolated to other physicochemical properties such as the number of hydrogen bond acceptors and/or donors, the number of rotable bonds, log P, and net charge.[6,21,54,60] To overcome this problem, Huang et al.[21] proposed to select as decoys compounds that are structurally dissimilar to the active compounds, i.e. compounds that are more likely to be nonbinders. Despite these precautions, some DUD decoys were identified as actual binders.[57] As a remark, such decoys are completely inappropriate for the evaluation of 2D ligand-based methods.[6] The best solution to overcome these problems would be to use true inactive compounds, i.e. compounds that have been experimentally tested and that displayed no ability to bind to the considered target as in the DUD-E database in which 9219 experimental decoys were provided.[54] Similarly, in the Maximum Unbiased Validation (MUV) data sets,[80] experimental decoys were extracted from the results of bioassays available in PubChem for which compounds found to be active and compounds found to be inactive were listed. Another possibility is to use real ligands as decoys for a nonwanted activity as in a recent study[87] where for each target with sufficient data, the agonist ligands of the NRLiSt BDB for a given target were used as decoys for the evaluation of the antagonistic activity and reciprocally the antagonist ligands of the NRLiSt BDB for a given target were used as decoys for the evaluation of the agonistic activity. This could be extended to

Table 2. Classification of the Number of Publications and Compounds Found to Be Wrongly Associated with a Nuclear Receptor during NRLiSt BDB Construction According to the Cause of Error

			cause of error		
target	no. of publications	no. of compds	different isoform	different receptor	other
CAR	2	5		X	
ERα	1	4	X		
	61	410		X	
	14	61			X
ERβ	30	191		X	
	8	34			X
ERRα	1	4	X		
	5	66		X	
FXRα	10	40		X	
	2	2			X
GR	41	218		X	
	23	30			X
LXRα	10	21		X	
	1	4			X
LXRβ	1	9	X		
	7	14		X	
MR	20	96		X	
	1	3			X
PPARα	10	198	X		
	6	12		X	
	6	39			X
PPARβ	32	328	X		
	12	34		X	
	9	16			X
PPARγ	13	108	X		
	16	93		X	
	10	52			X
PR	19	56		X	
	4	4			X
PXR	3	5		X	
	28	65			X
RORα	2	2		X	
RORγ	1	2		X	
RXRα	10	32		X	
	2	37			X
RXRβ	3	13		X	
RXRγ	3	13		X	
total	427	2321			

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

other receptors, for which agonist ligands and competitive antagonist ligands for the same binding site exist. Counter screens studies can also be used as a reliable source of experimental true inactives. As an example, a recent study used 6 nuclear receptors as a counter screen for unwanted activity and evaluated the agonistic and antagonistic activities of 615 known drugs.[90] The results showed that only 4.7% of these 615 drugs presented an agonistic potential and 12.4% an antagonistic potential. The compounds with no agonistic or antagonistic activities could thus be used as experimental decoys for the corresponding targets in the NRLiSt BDB. More recently, Lindh et al.[81] identified among the PubChem BioAssay database, 7 data sets suitable for the validation of both structure and ligand-based virtual screening methods using several quality filters (experimental screening against an identified protein target, number of compounds screened, availability of a crystal structure cocrystallized with a drug-like ligand, ...). A great interest grows for data sets that include experimental data; however, retrieving such information in the scientific literature is still difficult, notably about inactive compounds since negative data are often not published. Another idea was to use virtual decoys that ignore synthetic feasibility to obtain compounds displaying physicochemical properties that were more similar to the properties of the active compounds.[91] Using virtual decoys could solve the problem of active ligands for which too few decoys with similar properties are available. These decoys are generated using a library of chemical building blocks to construct chemically correct molecules that are then filtered to keep only compounds that match the active compound physical properties. Finally, starting from the observation that the majority of the available benchmarking data sets are suitable for the evaluation of structure-based virtual ligand screening methods, Xia et al.[64,65] developed a method to select decoys for the evaluation of ligand-based virtual screening methods. Indeed, DUD decoys are selected to be chemically similar to the active compounds while being as topologically dissimilar as possible to the active compounds. The selection of DUD decoys thus excluded the compounds that would constitute the most challenging decoys for the evaluation of 2D ligand-based methods.[6] The decoys proposed by Xia et al. presented similar physicochemical properties to the active compounds and a MACS fingerprints Tc cutoff of 0.75 (just as DUD- and DUD_E-like decoys), but the particularity of their decoys selection protocol is that the topological dissimilarity of a given decoy to its reference active compound is balanced by a "simsdiff" value that takes into account both the topological similarity between the query active compound and all the other active compounds and the topological similarity of the decoy with all the other active ligands.

Target and Protein Structures Selection. The targets included in a benchmarking database should be chosen to be as diverse as possible in order to cover a wide "target space".[6] Even if the first benchmarking data sets presented less than 10 targets, they grew wider among the years. For example, the DUD-E is composed of a benchmarking data set for 102 targets. Since benchmarking data sets are used to evaluate virtual ligand screening methods, the targets are classically members of important therapeutic families of proteins. However, the main inclusion criterion remains the availability of known active compounds and of experimentally resolved protein structures. Most of the time, benchmarking data sets included only experimentally resolved protein structures, but in some cases, homology models were also included.[20,21] Since homology models are only predictions of the possible 3D structure of a given protein, only experimental structures should be included in benchmarking databases. Accordingly, the PDGFR_beta data set of the DUD data set was not included in the DUD_E database since the only protein structure available for PDGFR_beta was a homology model. Classically only one structure is provided for a given target, often arbitrarily chosen[28] or on the basis of previous docking studies, structure resolution or its absence of major errors.[21] It is important to keep in mind that the starting conformation can impact the docking performances[18,55,61,92,93] and that crystallographic structures are not free of errors.[55] Thus, it is mandatory to carefully inspect the structures to include in the benchmarking data sets the optimal structures to ensure a better evaluation of the methods. Indeed, Warren et al.[94] showed that not all structures are suitable to be used in docking protocol and proposed criteria to judge the quality of protein−ligand structure that should be used for benchmarking data sets construction. Those criteria can be divided into two categories: (i) clerical filters such as the access to electron density and the verification of ligand and atom identity, functional groups, bond order and ionization state, stereochemistry and tautomer states to identify errors that can be corrected and (ii) experimental data quality filters to dismiss structures associated with poor experimental data quality that cannot be corrected (*R*-factor <0.40, density for ligands, noncovalent ligands, no alternate conformations for side chains, no alternate conformations for ligands, active-site crystal contacts >6 Å, complete density in active site, occupancy in active site = 1.0). In our opinion, if only one structure is provided for each data set, this structure should be selected according to stringent criteria to limit artificial poor enrichments issued from a nonoptimal structure. To select the optimal structures among all available, some points have already been highlighted. Holo structures, when available, should be preferred to apo structures since holo structures are associated with better virtual screening performances.[18,94] Lagarde et al.[87] demonstrated that the pharmacological profile of the cocrystallized ligand should be taken into account: agonist-bound structures should be used for virtual screening protocol dedicated to the search of new agonist ligands and conversely when the evaluation is part of a project to identify new antagonist compounds, antagonist-bound structures should be preferred. Ben Nasr et al.[61] proposed simple and useful structure-based guidelines for the selection of the optimal conformation that should be taken into account to decide which protein structures should be included in the benchmarking data sets. To avoid selection bias, we suggest to provide all structures available for each target, similarly to the NRLiSt BDB[62] protein selection protocol in which all holo human structures available were included.

■ CONCLUSION

Benchmarking data sets play a central role in the evaluation of virtual ligand screening methods. This evaluation is important not only to assess the reliability of the results obtained during prospective campaigns but also to guide the selection of the right virtual screening tool(s). Numerous benchmarking data sets have been developed over the years, resulting in an enhancement of the global quality of the evaluations reported. However, some points are critical to ensure the quality of benchmarking data sets and some issues affecting the selection of (1) the active compounds, such as the importance of the

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

manual curation of the data and of the pharmacological profile of the active compounds; (2) the selection of the decoys and particularly the importance of switching from putative to true experimentally tested inactive compounds; and (3) the structure(s) of the target should be corrected or improved to guarantee a high quality evaluation of the methods.

AUTHOR INFORMATION

Corresponding Author

*E-mail: matthieu.montes@cnam.fr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

N.L. is recipient of a CNAM/MNESR ATER fellowship.

ABBREVIATIONS

HTS, high throughput screening; VLS, virtual ligand screening; ER_alpha, estrogen alpha receptor; TK, thymidine kinase; ACD, advanced chemical directory; MDDR, MDL drug data report; PARP, poly ADP-ribose polymerase; PTP1b, protein tyrosine phosphatase 1b; DUD, directory of useful decoys; GPCR, G protein-coupled receptors; GLL, GPCR Ligand Library; GDD, GPCR Decoy Database; NRs, nuclear receptors; NRLiSt BDB, nuclear receptors ligands and structures benchmarking database; MUBD-HDACs, maximal unbiased benchmarking data sets for histone deacetylases inhibitors; MUV, maximum unbiased validation; DHFR, dihydrofolate reductase; AR, androgen receptor; COX-1, cyclooxygenase-1

REFERENCES

(1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9* (3), 203−214.

(2) Morgan, S.; Grootendorst, P.; Lexchin, J.; Cunningham, C.; Greyson, D. The cost of drug development: a systematic review. *Health Policy* **2011**, *100* (1), 4−17.

(3) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1* (11), 882−894.

(4) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153* (Suppl 1), S7−26.

(5) Tanrikulu, Y.; Kruger, B.; Proschak, E. The holistic integration of virtual screening in drug discovery. *Drug Discovery Today* **2013**, *18* (7−8), 358−364.

(6) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 193−199.

(7) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49* (20), 5851−5855.

(8) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection−what can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 213−228.

(9) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43* (25), 4759−4767.

(10) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161* (2), 269−288.

(11) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470−489.

(12) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267* (3), 727−748.

(13) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42* (5), 791−804.

(14) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425−445.

(15) Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.* **1999**, *42* (22), 4650−4658.

(16) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. *J. Mol. Model.* **1998**, *4*, 379.

(17) *Advanced Chemical Directory (ACD) v.2000-1*; Molecular Design Limited: San Leandro.

(18) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46* (14), 2895−2907.

(19) *MDL Drug Data Report (MDDR) v.2000.2*; MDL Inc.: San Leandro.

(20) Diller, D. J.; Li, R. Kinases, homology models, and high throughput docking. *J. Med. Chem.* **2003**, *46* (22), 4638−4647.

(21) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789−6801.

(22) Lorber, D. M.; Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Curr. Top. Med. Chem.* **2005**, *5* (8), 739−749.

(23) Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry* **2005**, *44* (37), 12316−12328.

(24) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68* (1), 76−90.

(25) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46* (4), 499−511.

(26) Makino, S.; Kuntz, I. D. Automated flexible ligand docking method and its application for database search. *J. Comput. Chem.* **1997**, *18*, 1812−1825.

(27) Miteva, M. A.; Lee, W. H.; Montes, M. O.; Villoutreix, B. O. Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J. Med. Chem.* **2005**, *48* (19), 6012−6022.

(28) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49* (20), 5856−5868.

(29) Perkins, E.; Sun, D.; Nguyen, A.; Tulac, S.; Francesco, M.; Tavana, H.; Nguyen, H.; Tugendreich, S.; Barthmaier, P.; Couto, J.; Yeh, E.; Thode, S.; Jarnagin, K.; Jain, A.; Morgans, D.; Melese, T. Novel inhibitors of poly(ADP-ribose) polymerase/PARP1 and PARP2 identified using a cell-based screen in yeast. *Cancer Res.* **2001**, *61* (10), 4175−4183.

(30) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45* (11), 2213−2221.

(31) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47* (12), 2977−2980.

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

(32) Irwin, J. J.; Shoichet, B. K. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177−182.
(33) Zhang, J.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem.* **2004**, *28* (5−6), 401−407.
(34) Bolton, E. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry* **2008**, *4*, 217−241.
(35) Fang, H.; Tong, W.; Shi, L. M.; Blair, R.; Perkins, R.; Branham, W.; Hass, B. S.; Xie, Q.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.* **2001**, *14* (3), 280−294.
(36) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44* (7), 1035−1042.
(37) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45* (3), 549−561.
(38) Wright, L.; Barril, X.; Dymock, B.; Sheridan, L.; Surgenor, A.; Beswick, M.; Drysdale, M.; Collier, A.; Massey, A.; Davies, N.; Fink, A.; Fromont, C.; Aherne, W.; Boxall, K.; Sharp, S.; Workman, P.; Hubbard, R. E. Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem. Biol.* **2004**, *11* (6), 775−785.
(39) Dymock, B. W.; Barril, X.; Brough, P. A.; Cansfield, J. E.; Massey, A.; McDonald, E.; Hubbard, R. E.; Surgenor, A.; Roughley, S. D.; Webb, P.; Workman, P.; Wright, L.; Drysdale, M. J. Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design. *J. Med. Chem.* **2005**, *48* (13), 4212−4215.
(40) Hennequin, L. F.; Thomas, A. P.; Johnstone, C.; Stokes, E. S.; Ple, P. A.; Lohmann, J. J.; Ogilvie, D. J.; Dukes, M.; Wedge, S. R.; Curwen, J. O.; Kendrew, J.; Lambert-van der Brempt, C. Design and structure-activity relationship of a new class of potent VEGF receptor tyrosine kinase inhibitors. *J. Med. Chem.* **1999**, *42* (26), 5369−5389.
(41) Hennequin, L. F.; Stokes, E. S.; Thomas, A. P.; Johnstone, C.; Ple, P. A.; Ogilvie, D. J.; Dukes, M.; Wedge, S. R.; Kendrew, J.; Curwen, J. O. Novel 4-anilinoquinazolines with C-7 basic side chains: design and structure activity relationship of a series of potent, orally active, VEGF receptor tyrosine kinase inhibitors. *J. Med. Chem.* **2002**, *45* (6), 1300−1312.
(42) Sun, L.; Tran, N.; Liang, C.; Tang, F.; Rice, A.; Schreck, R.; Waltz, K.; Shawver, L. K.; McMahon, G.; Tang, C. Design, synthesis, and evaluations of substituted 3-[(3- or 4-carboxyethylpyrrol-2-yl)methylidenyl]indolin-2-ones as inhibitors of VEGF, FGF, and PDGF receptor tyrosine kinases. *J. Med. Chem.* **1999**, *42* (25), 5120−5130.
(43) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46* (26), 5781−5789.
(44) Bohm, M.; Stürzebecher, J.; Klebe, G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42* (3), 458−477.
(45) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.* **2004**, *47* (22), 5541−5554.
(46) Varney, M. D.; Palmer, C. L.; Romines, W. H., III; Boritzki, T.; Margosiak, S. A.; Almassy, R.; Janson, C. A.; Bartlett, C.; Howland, E. J.; Ferre, R. Protein structure-based design, synthesis, and biological evaluation of 5-thia-2,6-diamino-4(3H)-oxopyrimidines: potent inhibitors of glycinamide ribonucleotide transformylase with potent cell growth inhibition. *J. Med. Chem.* **1997**, *40* (16), 2502−2524.
(47) Van Zandt, M. C.; Jones, M. L.; Gunn, D. E.; Geraci, L. S.; Jones, J. H.; Sawicki, D. R.; Sredy, J.; Jacot, J. L.; Dicioccio, A. T.; Petrova, T.; Mitschler, A.; Podjarny, A. D. Discovery of 3-[(4,5,7-trifluorobenzothiazol-2-yl)methyl]indole-N-acetic acid (lidorestat) and congeners as highly potent and selective inhibitors of aldose reductase for treatment of chronic diabetic complications. *J. Med. Chem.* **2005**, *48* (9), 3141−3152.
(48) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure* **2002**, *10* (7), 1013−1023.
(49) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005**, *48* (11), 3714−3728.
(50) Tondi, D.; Morandi, F.; Bonnet, R.; Costi, M. P.; Shoichet, B. K. Structure-based optimization of a non-beta-lactam lead results in inhibitors that do not up-regulate beta-lactamase expression in cell culture. *J. Am. Chem. Soc.* **2005**, *127* (13), 4632−4639.
(51) Wang, J.; Kang, X.; Kuntz, I. D.; Kollman, P. A. Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. *J. Med. Chem.* **2005**, *48* (7), 2432−2444.
(52) Tikhe, J. G.; Webber, S. E.; Hostomsky, Z.; Maegley, K. A.; Ekkers, A.; Li, J.; Yu, X. H.; Almassy, R. J.; Kumpf, R. A.; Boritzki, T. J.; Zhang, C.; Calabrese, C. R.; Curtin, N. J.; Kyle, S.; Thomas, H. D.; Wang, L. Z.; Calvert, A. H.; Golding, B. T.; Griffin, R. J.; Newell, D. R. Design, synthesis, and evaluation of 3,4-dihydro-2H-[1,4]diazepino-[6,7,1-hi]indol-1-ones as inhibitors of poly(ADP-ribose) polymerase. *J. Med. Chem.* **2004**, *47* (22), 5467−5481.
(53) Ealick, S. E.; Babu, Y. S.; Bugg, C. E.; Erion, M. D.; Guida, W. C.; Montgomery, J. A.; Secrist, J. A., III Application of crystallographic and modeling methods in the design of purine nucleoside phosphorylase inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88* (24), 11540−11544.
(54) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582−6594.
(55) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 179−190.
(56) Mysinger, M. M.; Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50* (9), 1561−1573.
(57) Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: demanding evaluation kits for objective in silico screening–a versatile tool for benchmarking docking programs and scoring functions. *J. Chem. Inf. Model.* **2011**, *51* (10), 2650−2665.
(58) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100−1107.
(59) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(60) Gatica, E. A.; Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **2012**, *52* (1), 1−6.
(61) Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J. F.; Montes, M. Multiple structures for virtual ligand screening: defining binding site properties-based criteria to optimize the selection of the query. *J. Chem. Inf. Model.* **2013**, *53* (2), 293−311.
(62) Lagarde, N.; Ben Nasr, N.; Jeremie, A.; Guillemain, H.; Laville, V.; Labib, T.; Zagury, J. F.; Montes, M. NRLiSt BDB, the manually curated nuclear receptors ligands and structures benchmarking database. *J. Med. Chem.* **2014**, *57* (7), 3117−3125.
(63) Xia, J.; Tilahun, E. L.; Kebede, E. H.; Reid, T. E.; Zhang, L.; Wang, X. S. Comparative Modeling and Benchmarking Data Sets for Human Histone Deacetylases and Sirtuin Families. *J. Chem. Inf. Model.* **2015**, *55* (2), 374−388.
(64) Xia, J.; Jin, H.; Liu, Z.; Zhang, L.; Wang, X. S. An unbiased method to build benchmarking sets for ligand-based virtual screening and its application to GPCRs. *J. Chem. Inf. Model.* **2014**, *54* (5), 1433−1450.

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307

(65) Xia, J.; Tilahun, E. L.; Reid, T. E.; Zhang, L.; Wang, X. S. Benchmarking methods and data sets for ligand enrichment assessment in virtual screening. *Methods* **2015**, *71*, 146−157.
(66) Kobayashi, M.; Kinjo, T.; Koseki, Y.; Bourne, C. R.; Barrow, W. W.; Aoki, S. Identification of novel potential antibiotics against Staphylococcus using structure-based drug screening targeting dihydrofolate reductase. *J. Chem. Inf. Model.* **2014**, *54* (4), 1242−1253.
(67) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* **2009**, *15* (6), 1219−1230.
(68) Bobach, C.; Tennstedt, S.; Palberg, K.; Denkert, A.; Brandt, W.; de Meijere, A.; Seliger, B.; Wessjohann, L. A. Screening of synthetic and natural product databases: Identification of novel androgens and antiandrogens. *Eur. J. Med. Chem.* **2015**, *90*, 267−279.
(69) *Molecular Operating Environment (MOE), 2007.1*; Chemical Computing Group Inc.: Montreal, 2007.
(70) Korb, O.; Stützle, T.; Exner, T. E. PLANTS: Application of Ant Colony Optimization to Stucture-Based Drug Design. In *Ant Colony Optimization and Swarm Intelligence*; Dorigo, M., Gambardella, L. M., Birattari, M., Martinoli, A., Poli, R., Stützle, T., Eds.; Berlin/Heidelberg, **2006**; pp 247−258.
(71) Korb, O.; Stützle, T.; Exner, T. E. An ant colony optimization approach to flexible protein-ligand docking. *Swarm Intelligence* **2007**, *1*, 115−134.
(72) Zhang, S.; Tan, J.; Lai, Z.; Li, Y.; Pang, J.; Xiao, J.; Huang, Z.; Zhang, Y.; Ji, H.; Lai, Y. Effective virtual screening strategy toward covalent ligands: identification of novel NEDD8-activating enzyme inhibitors. *J. Chem. Inf. Model.* **2014**, *54* (6), 1785−1797.
(73) Balaji, B.; Hariharan, S.; Shah, D. B.; Ramanathan, M. Discovery of potential and selective COX-1 inhibitory leads using pharmacophore modelling, in silico screening and in vitro evaluation. *Eur. J. Med. Chem.* **2014**, *86*, 469−480.
(74) Muegge, I.; Enyedy, I. J. Virtual screening for kinase targets. *Curr. Med. Chem.* **2004**, *11* (6), 693−707.
(75) Rohrer, S. G.; Baumann, K. Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model.* **2008**, *48* (4), 704−718.
(76) Ripphausen, P.; Wassermann, A. M.; Bajorath, J. REPROVIS-DB: a benchmark system for ligand-based virtual screening derived from reproducible prospective applications. *J. Chem. Inf. Model.* **2011**, *51* (10), 2467−2473.
(77) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: a virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18* (7−9), 529−536.
(78) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 169−178.
(79) Stumpfe, D.; Bajorath, J. Applied Virtual Screening: Strategies, Recommandations, and Caveats. In *Methods and Principles in Medicinal Chemistry. Virtual Screening. Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Weinhem, **2011**; pp 73−103.
(80) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49* (2), 169−184.
(81) Lindh, M.; Svensson, F.; Schaal, W.; Zhang, J.; Skold, C.; Brandt, P.; Karlen, A. Toward a Benchmarking Data Set Able to Evaluate Ligand- and Structure-based Virtual Screening Using Public HTS Data. *J. Chem. Inf. Model.* **2015**, *55* (2), 343−353.
(82) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminf.* **2009**, *1*, 14.
(83) Li, R.; Stumpfe, D.; Vogt, M.; Geppert, H.; Bajorath, J. Development of a method to consistently quantify the structural distance between scaffolds and to assess scaffold hopping potential. *J. Chem. Inf. Model.* **2011**, *51* (10), 2507−2514.
(84) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *J. Med. Chem.* **2010**, *53* (15), 5707−5715.
(85) Li, R.; Bajorath, J. Systematic assessment of scaffold distances in ChEMBL: prioritization of compound data sets for scaffold hopping analysis in virtual screening. *J. Comput.-Aided Mol. Des.* **2012**, *26* (10), 1101−1109.
(86) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating error rates in bioactivity databases. *J. Chem. Inf. Model.* **2013**, *53* (10), 2499−2505.
(87) Lagarde, N.; Zagury, J. F.; Montes, M. Importance of the pharmacological profile of the bound ligand in enrichment on nuclear receptors: toward the use of experimentally validated decoy ligands. *J. Chem. Inf. Model.* **2014**, *54* (10), 2915−2944.
(88) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 141−146.
(89) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793−806.
(90) Fan, F.; Hu, R.; Munzli, A.; Chen, Y.; Dunn, R. T., II; Weikl, K.; Strauch, S.; Schwandner, R.; Afshari, C. A.; Hamadeh, H.; Nioi, P. Utilization of Human Nuclear Receptors as an Early Counter Screen for Off-Target Activity: A Case Study with a Compendium of 615 Known Drugs. *Toxicol. Sci.* **2015**, *45* (2), 283−295.
(91) Wallach, I.; Lilien, R. Virtual decoy sets for molecular docking benchmarks. *J. Chem. Inf. Model.* **2011**, *51* (2), 196−202.
(92) Thomas, M. P.; McInnes, C.; Fischer, P. M. Protein structures in virtual screening: a case study with CDK2. *J. Med. Chem.* **2006**, *49* (1), 92−104.
(93) Giganti, D.; Guillemain, H.; Spadoni, J. L.; Nilges, M.; Zagury, J. F.; Montes, M. Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J. Chem. Inf. Model.* **2010**, *50* (6), 992−1004.
(94) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today* **2012**, *17* (23−24), 1270−1281.
(95) Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0–a public library of challenging docking benchmark sets. *J. Chem. Inf. Model.* **2013**, *53*, 1447−1462.
(96) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−201.

DOI: 10.1021/acs.jcim.5b00090
J. Chem. Inf. Model. 2015, 55, 1297−1307