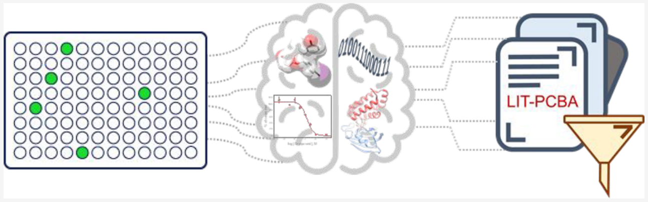
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening

Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 4263−4273 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Comparative evaluation of virtual screening methods requires a rigorous benchmarking procedure on diverse, realistic, and unbiased data sets. Recent investigations from numerous research groups unambiguously demonstrate that artificially constructed ligand sets classically used by the community (e.g., DUD, DUD-E, MUV) are unfortunately biased by both obvious and hidden chemical biases, therefore overestimating the true accuracy of virtual screening methods. We herewith present a novel data set (LIT-PCBA) specifically designed for virtual screening and machine learning. LIT-PCBA relies on 149 dose−response PubChem bioassays that were additionally processed to remove false positives and assay artifacts and keep active and inactive compounds within similar molecular property ranges. To ascertain that the data set is suited to both ligand-based and structure-based virtual screening, target sets were restricted to single protein targets for which at least one X-ray structure is available in complex with ligands of the same phenotype (e.g., inhibitor, inverse agonist) as that of the PubChem active compounds. Preliminary virtual screening on the 21 remaining target sets with state-of-the-art orthogonal methods (2D fingerprint similarity, 3D shape similarity, molecular docking) enabled us to select 15 target sets for which at least one of the three screening methods is able to enrich the top 1%-ranked compounds in true actives by at least a factor of 2. The corresponding ligand sets (training, validation) were finally unbiased by the recently described asymmetric validation embedding (AVE) procedure to afford the LIT-PCBA data set, consisting of 15 targets and 7844 confirmed active and 407,381 confirmed inactive compounds. The data set mimics experimental screening decks in terms of hit rate (ratio of active to inactive compounds) and potency distribution. It is available online at http://drugdesign.unistra.fr/LIT-PCBA for download and for benchmarking novel virtual screening methods, notably those relying on machine learning.

INTRODUCTION

Virtual screening (VS) of compound libraries has established itself, notably in academic settings, as a fast and cost-efficient alternative to high-throughput screening (HTS) for identifying preliminary hits of pharmaceutically interesting targets.[1−3] Because of the availability of hundreds of virtual screening tools,[4] choosing the right method for a specific project often relies on benchmarking studies designed to delineate the context-specific advantages and drawbacks of each method. Many target-specific ligand sets[5−10] and statistical evaluation protocols[11−13] have been reported in the past decade to pinpoint the ability of a VS method to prioritize, for purchase and validation, the shortest possible hit list with an optimal enrichment in true actives. In the early 2000s, such data sets were limited in size due to the paucity of available experimental data. Inactive compounds were notably randomly chosen among drug-like compound databases.[5,14,15] Very soon, it appeared that random selection of presumably inactive compounds (decoys) led to artificially high enrichment values because of a bias in molecular property ranges (e.g., molecular weight) that often differed between active and inactive sets.[16] A first attempt to design a docking-dedicated benchmark set led to the DUD data set,[6] which gathers 2950 ligands of 40 different targets from the literature, seeded among property-matched decoys (36 decoys for each active) from the ZINC archive of commercially available ligands.[17] In DUD, decoys were specifically designed to share physicochemical properties with actives but with a different chemical topology. Despite the caution given to the selection of decoys, independent groups rapidly noticed three major biases for both DUD active and decoy sets: (i) Actives tend to spread over a few dominant scaffolds (so-called "analog bias").[18] (ii) Decoys exhibited molecular net charges different from those of actives.[19] (iii) Decoys were too similar to true actives and likely false negatives.[8] The DUD set was upgraded to a revised version (DUD-E)[10] describing an enhanced and more diverse target space (102 targets), containing 22,886 clustered true actives with known experimental data from the ChEMBL database,[20]

Special Issue: New Trends in Virtual Screening

Received: February 12, 2020
Published: April 13, 2020

ACS Publications

https://dx.doi.org/10.1021/acs.jcim.0c00155

removing all above-cited biases and enhancing the proportion of decoys (50 decoys for each active). The debate on the best protocol to select decoys has led to many contributions[8,9,21] to design alternative decoy sets to that proposed by DUD-E. As an alternative to DUD-E, other sources of active compounds (e.g., PubChem BioAssay[22]) have also been utilized. Noteworthy is the MUV[7] database that provides many advantages: (i) The data set (target, ligand, assay conditions) is publicly available. (ii) Compound collections are drug-like. (iii) Many experimental data were utilized to remove false positives and assay artifacts. (iv) Ligands are selected by a nearest neighbor analysis to permit a spatially unbiased distribution of actives and decoys. Consequently, the MUV data set is considered more challenging than DUD-E.[23]

For many years, the DUD-E has been considered as a gold standard for benchmarking VS and machine learning methods, until recent reports[24−27] warned the community on both obvious and hidden biases in its design. First, Chaput et al.[24] noticed that differences in key molecular properties (polar surface area, hydrogen bond donor count, embranchment count) remain between DUD-E actives and decoys. Moreover, a chemical bias is still present in actives that tend to resemble target-bound PDB ligands, thereby overestimating the real discriminatory accuracy of standard docking methods.[24] Wallach and Heifets described the asymmetric validation embedding (AVE) method[25] to quantify the bias in ligand sets and optimally design training and validation sets. When applied to ligand-based VS methods, all standard benchmark sets (e.g., DUD, DUD-E, MUV) were shown to be massively biased, rewarding memorization rather than learning.[25] The latter danger is even higher for currently popular artificial intelligence methods (e.g., machine learning, deep neural networks)[28] that are hardly interpretable and tightly dependent on the quality of the input data and the way they are split to train and test a model. Two different groups[26,27] just reported hidden biases in the DUD-E data set when applying deep neural networks (DNNs) to either predict binding affinities or classify complexes as active/inactive from X-ray structures or docking poses. Intriguingly, DNNs trained with rigorous cross-validation procedures on simple ligand descriptors were almost as accurate as those trained on protein−ligand attributes, evidencing that deep learning did not learn anything about the physics of protein−ligand interactions. Strikingly, the literature is full of overoptimistic reports describing machine learning models[29−31] with near perfect performances on the above-described data sets, although true VS practitioners have known for a long time that such an accuracy level does not mirror the proportion of experimentally confirmed hits in real prospective VS experiments.

There is more than ever an urgent need to design an unbiased and realistic data set specifically dedicated to virtual screening and machine learning.[27] We herewith present our contribution based on the following seven principles:

(i) The data set should mimic “real-life” screening decks and guide VS methods to discriminate moderately potent actives (primary hits) from inactive compounds.

(ii) The potency of all compounds (actives, inactives) for a particular target should have been determined experimentally in homogeneous conditions.

(iii) The ratio of actives to inactives should reflect hit rates typically observed in HTS campaigns against targets of pharmaceutical interest.[32]

(iv) Actives should be filtered to remove false positives, frequent hitters, assay artifacts, and truly undruggable compounds. In addition, dose−response curves should be available for all actives.

(v) Active and inactive compounds should span common molecular property ranges.

(vi) Potency distribution of confirmed actives should not be biased toward too high affinities and should ideally mimic that observed in HTS decks.

(vii) The data set should be applicable to both ligand-based and structure-based virtual screening.

(viii) Unbiased training and validation sets should be available for machine learning.

We therefore decided to choose the PubChem BioAssay database (PCBA)[22] as the source of experimental bioactivity data. PCBA is an open-access archive hosted by the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), and National Institute of Health (NIH). At the time of writing this manuscript, the database stores over 1 million assay records, out of which 134,000 are annotated by activity type (IC_{50}, EC_{50}, K_d, K_i). It covers about 7200 HTS projects from 80 sources (academic, governmental, pharmaceutical companies) on a chemical repertoire of 2.2 million compounds. The database can be easily queried according to numerous filters and is a premier source of bioactivity data for computer-aided drug discovery.[33]

We hereby describe a workflow for retrieving assays of interest and filtering compounds and targets for bioactivity data acquisition. The retrieved target sets were then utilized for state-of-the-art virtual screening experiments in order to ascertain their suitability. The final data set (LIT-PCBA) contains 15 targets and 7844 actives and 407,381 inactive compounds, with ready-to-use input files (ligands, targets) that have been unbiased for machine learning applications. It is available for download at http://drugdesign.unistra.fr/LIT-PCBA.

■ COMPUTATIONAL METHODS

Data Set. Bioactivity data were retrieved from the PubChem BioAssay database,[22] where all information on true active and true inactive substances for a protein target is provided based on experimental results from confirmatory dose−response bioactivity assays, whose related details including assay principles, general protocols, and other remarks are also given. All data were updated as of December 31, 2018. The “limits” search engine (https://www.ncbi.nlm.nih.gov/pcassay/limits) was used to filter the PubChem BioAssays resource by various options, with “Activity Outcome” set as “Active”, “Substance Type” set as “Chemical”, and “Screening Stage” defined as “Confirmatory, Dose-Response”. Here, 149 assays targeting a single protein target, operated on at least 10,000 substances, and leading to at least 50 confirmed actives were first retained. The experimental screening data were kept if the target was characterized by at least one Protein Data Bank (PDB)[34] entry, in complex with a ligand of the same phenotype (i.e., inhibitor, agonist, or antagonist) as that of the tested active substances of the corresponding bioactivity assay. Altogether, 21 raw HTS data tables were directly retrieved as csv files from the PubChem BioAssay website as well as actives and inactives in separate sd files. The PDB resource was then browsed by Uniprot identifier (Uniprot ID)[35] to retrieve the corresponding PDB entries in the suitable ligand-bound form.

Template Structure Preparations for Each Target Set. Protein−ligand complexes (in pdb file format) corresponding to the chosen target sets were processed as follows. For each PDB entry, explicit hydrogen atoms were added with Protoss[36] to any molecule (protein, cofactor, prosthetic group, ion, ligand, water). The output pdb file was then visualized in Sybyl-X 2.1.1.[37] A water molecule was kept under two conditions: (i) It was found at the binding site of the ligand; i.e., the distance between the oxygen atom of the water molecule and at least one heavy atom of the cocrystallized ligand was not greater than 5 Å. (ii) It engaged in no fewer than three hydrogen bonds with the protein and/or the ligand, at least two of which were with the protein. Hydrogen bonds must satisfy the following criteria: The donor−acceptor distance must not exceed 3.5 Å, and the angle formed by the donor, hydrogen atom, and acceptor (the vertex of the angle was positioned at the hydrogen atom) must be higher than 120°. The protonated ligand and protein (including all remaining bound water molecules, cofactors, prosthetic groups, and ions) were saved separately in a mol2 file format with Sybyl-X 2.1.1.[37]

In case more than 20 ligand-bound protein entries were available for each target, all protein−ligand structures were clustered according to the diversity of observed protein−ligand interaction patterns. Protein−ligand interaction patterns were computed as graphs with IChem[38] as previously described,[39] and target-specific interaction pattern similarity matrices were computed using the GRIMscore metric.[39] Each matrix was then used as input for an agglomerative nesting clustering using the agnes function in R v.3.5.2, the Ward clustering method, a Euclidean distance matrix, and a total number of clusters fixed to 15. For each cluster, the PDB entry with the highest resolution was chosen as the protein−ligand PDB template for the corresponding target set.

Determination of Filtering Rules for True Active and True Inactive Substances of Each Target Set. Metadata on each substance (true active and true inactive) included in each selected target set were collected directly from the website of the PubChem BioAssay database including the substance identifier (SID), the activity label (active or inactive), the phenotype (inhibitor, agonist, or antagonist), the potency (EC_{50} or IC_{50}, in μM), and the Hill slope for the dose−response curve of each true active. The frequency of hits (FoH) for each true active was computed as the ratio of the number of PubChem bioassays in which a substance was confirmed as a true active to the number of assays in which it was tested. Additional molecular properties (molecular weight, AlogP, total formal charge, number of rotatable bonds, number of hydrogen bond donors and acceptors) were computed in Pipeline Pilot v.19.1.0.1964.[40]

For each target set, all true actives and true inactives were then filtered according to four steps:

Step 1: Organic compound filter. Molecules bearing at least one atom other than H, C, N, O, P, S, F, Cl, Br, and I were removed.

Step 2: False positives filter (this particular step was applied only to true active substances).

- Step 2a: 0.5 < Hill slope h < 2.0
- Step 2b: FoH < 0.26[7]
- Step 2c: Aggregator/autofluorescence, luciferase filter. All true actives reported as actives in PubChem aggregation (actives in AID 585 or AID 485341 but not in AID 584 and AID 485294), luciferase inhibition (AID 411), or autofluorescence (AID 587, AID 588, AID 590, AID 591, AID 592, AID 593, AID 594) bioassays were eliminated.

Step 3: Molecular property range filter. Remaining actives and inactives were kept if

- 150 < Molecular weight < 800 Da
- −3.0 < AlogP < 5.0
- Number of rotatable bonds < 15
- H-bond acceptor count < 10; H-bond donor count < 10;
- −2.0 < total formal charge < + 2.0

Step 4: 3D conversion and normalization filter. The two-dimensional (2D) sd files of the remaining compounds (actives, inactives) were converted into a 3D sd file format using default settings of Corina v. 3.4.[41] Last, compounds were standardized and ionized at physiological pH with Filter 2.5.1.4.[42]

2D Similarity Search. Extended-connectivity circular-ECFP4 fingerprints[43] were computed for PubChem compounds and PDB ligands in PipelinePilot. Pairwise similarity of PubChem compounds to PDB ligands was estimated by the Tanimoto coefficient (Tc), thereby leading to a PDB ligand-specific hit list sorted by decreasing Tc value. The areas under the ROC (receiver operating characteristic)[11] and BEDROC (Boltzmann-enhanced discrimination of ROC)[12] curves (α = 20) along with the enrichment in true actives at a constant 1% false positive rate over random picking (EF1%) were calculated for each separate hit list. The same procedure was applied by fusing all lists and keeping the maximal Tc value for each compound.

3D Similarity Search. For each target set, a maximal number of 200 conformers were generated for every PubChem compound with standard settings of Omega2 v.2.5.1.4.[44] All conformers were then compared to the query (PDB ligand) with ROCS v.3.2.0.4.[45] The best matching conformer was selected for every ligand according to the TanimotoCombo similarity score,[13] and all molecules of each target set were sorted based on this same value in descending order. ROC AUC, BEDROC AUC, and EF1% values were calculated as described above.

Molecular Docking. Starting from the mol2 structure of a fully processed template protein (including remaining bound water molecules after preparation) and that of its cocrystallized ligand, a protomol representing the ligand-binding site was generated from protein-bound ligand atomic coordinates using the default settings of Surflex-Dock v.3066.[46] All molecules in the relevant target set were docked into the protomol with the "*−pgeom*" option of the docking engine. The best-ranked pose according to docking scores (pK_d values) was retained for each molecule, and all ligands of the set were then sorted based on this value in descending order. ROC AUC, BEDROC AUC, and EF1% values were calculated as described above.

Target Set Unbiasing. For each target set, unbiasing of the training and validation sets was done using the previously described asymmetric validation embedding (AVE) method,[25] which systematically measures the pairwise distance in chemical space between molecules belonging to four sets of compounds (training actives, training inactives, validation actives, validation inactives). Using circular ECFP4 fingerprints[43] as chemical descriptors and a training to validation ratio of 3, a maximal number of 300 iteration steps of the AVE genetic algorithm

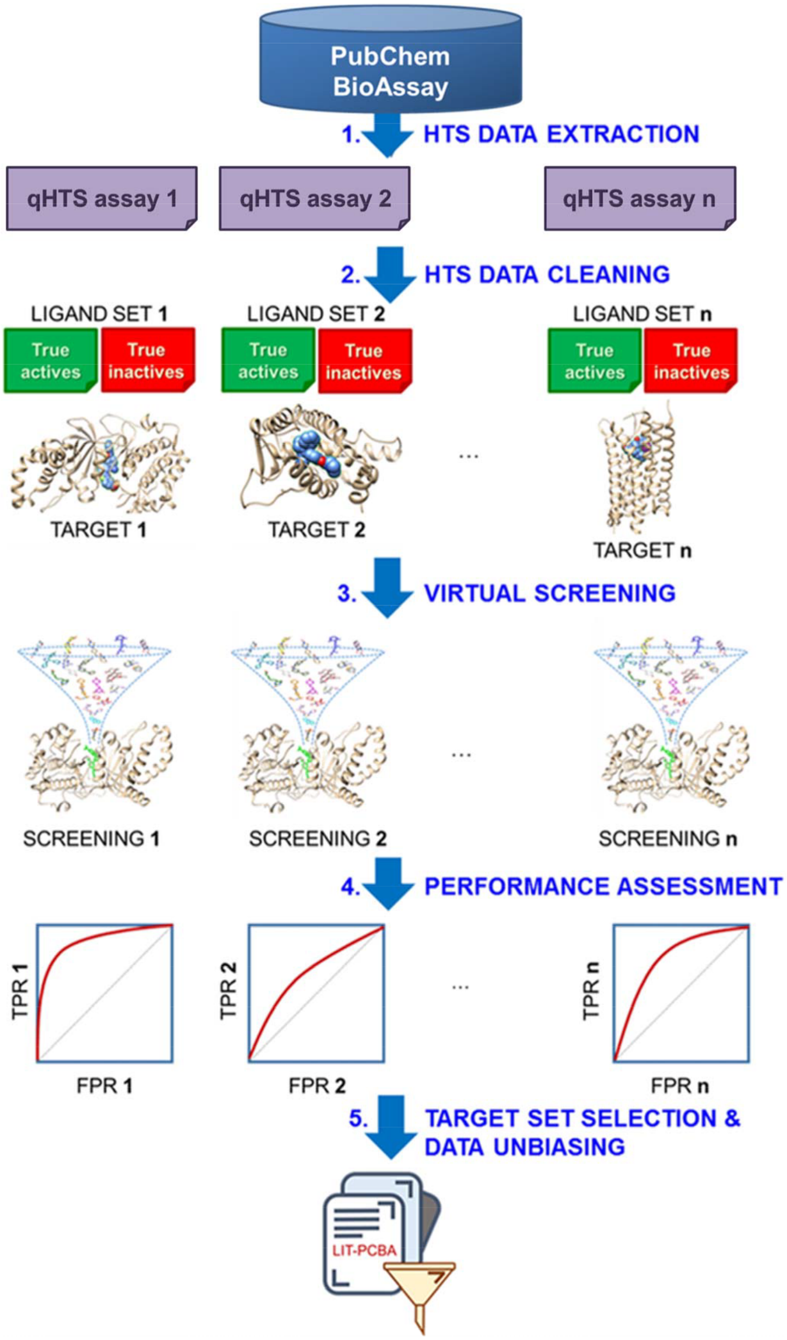


Figure 1. Workflow for setting up the LIT-PCBA data set. (1) Data retrieval from the PubChem BioAssay database according to user-defined filters (Activity outcome: active, count of tested substances ≥10,000, count of active substances ≥50. Substance type: chemical. Screening stage: confirmatory, dose−response. Target: single. Target type: protein target). (2) Data cleaning: removal of inorganic compounds, false positives, frequent hitters, assay artifacts, and compounds with extreme molecular properties. Selection of target sets for which the target has a representative structure in the Protein Data Bank, cocrystallized with a ligand of the same phenotype (e.g., inhibitor, agonist, antagonist) as that of active compounds in the corresponding bioassay. (3) Virtual screening (VS) of cleaned HTS target sets with three methods (2D similarity, 3D shape similarity, docking). (4) Performance assessment of the three VS methods on all cleaned target sets (ROC, BEDROC, EF1%). (5) Selection of target sets for which at least one VS method achieved an enrichment in true positives higher than 2.0. AVE[25] unbiasing of the corresponding ligand sets and definition of training and validation sets for machine learning.

were run to select training and validation molecules while minimizing the overall bias B (B ∈ [0, 1]) of the target set. Convergence was reached when the bias value B was lower than 0.01. To enable the script processing large sets of compounds, the bias removing script (remove_AVE_bias.py) originally proposed by Wallach and Heifets[25] was modified to enable a faster calculation on multiple cores.

■ RESULTS AND DISCUSSION

The aim of the present study is to design an unbiased data set dedicated to virtual screening as well as machine learning, along four main ideas:

(1) Experimental binding data should be available for all compounds, including inactives. Each true active should have been confirmed by a full dose−response curve.

Table 1. List of 21 Selected PubChem Bioassays

Target ID	Name	Assay AID[a]	Substances[b] Tested	Actives	Phenotype	PDB entries
ADRB2	Beta2 adrenergic receptor	492947	331,108	80	Agonist	8
ALDH1	Aldehyde dehydrogenase 1	1030	220,402	16117	Inhibitor	8
ARO1	Aromatase	743083	10,486	905	Inhibitor	3
ESR1_ago	Estrogen receptor alpha	743075	10,486	589	Agonist	15
ESR1_ant	Estrogen receptor alpha	743080	10,486	477	Antagonist	15
FEN1	Flap endonuclease 1	588795	391,275	1368	Inhibitor	1
GBA	Glucocerebrosidase	2101	326,770	299	Inhibitor	6
GLP1R	Glucagon-like peptide-1 receptor	624417	408,352	6432	Inverse agonist	2
GLS	Glutaminase	624170	409,400	846	Inhibitor	11
IDH1	Isocitrate dehydrogenase	602179	390,606	365	Inhibitor	14
KAT2A	Histone acetyltransferase KAT2A	504327	387,485	817	Inhibitor	3
L3MBTL1	Lethal(3)malignant brain tumor-like protein isoform I	485360	225,505	1495	Inhibitor	1
MAPK1	Mitogen-activated protein kinase 1	995	72,004	711	Inhibitor	15
MTORC1	Mechanistic target of rapamycin	493208	43,989	342	Inhibitor	11
OPRK1	Kappa opioid receptor	1777	284,220	51	Agonist	1
PKM2	Pyruvate kinase muscle isoform 2	1631	264,516	892	Agonist	9
PPARG	Peroxisome proliferator-activated receptor gamma	743094	10,486	78	Agonist	15
RORC	Retinoic acid-related orphan receptor gamma	2551	309,031	16824	Inhibitor	15
THRB	Thyroid hormone receptor	1469	282,587	183	Inhibitor	1
TP53	Cellular tumor antigen p53	651631	10,488	602	Agonist	6
VDR	Vitamin D receptor	504847	401,452	3735	Antagonist	2

[a]Full details for each assay are available at https://pubchem.ncbi.nlm.nih.gov/bioassay/AID. [b]Structures deposited by individual data contributors. Unique chemical structures are called compounds.

(2) The target should be a single protein, for which a high-resolution X-ray structure is available in the PDB. Moreover, the target should have been crystallized at least once, with a ligand exhibiting a phenotype (e.g., inhibitor, full agonist, neutral antagonist) identical to that of active compounds in the corresponding bioassay.

(3) PubChem target sets should be suitable for virtual screening. Performance of three nonorthogonal VS methods (2D fingerprint similarity, 3D shape similarity, molecular docking) was assessed to select target sets for which at least one of the three VS methods achieves an enrichment in true positives higher than 2, in other words, twice better than random picking.

(4) The finally selected target sets should be as unbiased as possible, when comparing true actives to true inactives in chemical space, as well as by splitting the data in training and validation sets.

To this end, we designed a computational workflow (Figure 1) that will be presented and discussed, step-by-step in the following sections.

HTS Data Extraction. PubChem (https://pubchem.ncbi.nlm.nih.gov) is a public repository for information on 91 million chemical substances and 268 million biological activities, launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the U.S. National Institutes of Health (NIH). The PubChem BioAssay resource[22] was queried to retrieve 149 assays according to multiple queries (see Data Set section of Computational Methods). To ascertain that the data set will be further suitable for either ligand-based or structure-based VS, we checked that each single protein target not only had a representative structure in the PDB but was also cocrystallized with a ligand sharing the same phenotype or function with the true actives. This sanity check enables the selection of the right activation state (e.g., for G-protein coupled receptors) and the right binding site for docking. Of course, we cannot ensure at this step that all true actives share the same binding site with all PDB ligand templates. However, it enables a first filter to avoid comparing ligands with known opposite or different functions. To control the bioactivity of each compound, only confirmatory dose−reponse screening assays were kept. A total of 21 assays (Table 1) performed on isolated enzymes ($n = 6$), soluble protein−protein interactions ($n = 4$), and target-expressing cells ($n = 11$), using four different readouts (fluorescence intensity, fluorescence polarization, luminescence, alpha screen), were finally saved. Except for five screens for which only 10,000 compounds have been tested, most assays have been run on a large number of compounds (from 200,000 to 400,000). Importantly, each assay has already been analyzed in detail, notably regarding the activity threshold qualifying a compound as active, that we did not modify and that is target-dependent. Moreover, compounds whose activity outcome was qualified as inconclusive were removed from the final hit list.

Corresponding targets are single proteins representing 11 families of pharmaceutical interest, including nuclear hormone receptors ($n = 5$), protein kinases ($n = 3$), and G protein-coupled receptors ($n = 3$). Most target sets describe compounds tested for an inhibitory activity against a protein target (13 target sets). Overall, 162 structures of protein−ligand complexes in the pdb format were chosen as templates for the 21 target sets (Table 1). More information on each selected PubChem BioAssay (brief assay description, readout, format, PDB templates) can be found in Table S1.

HTS Data Cleaning. All active and inactive compounds were next submitted to a series of filters (see Computational Methods) aimed at removing inorganic compounds (step 1), frequent hitters and assay artifacts (step 2),[7] compounds exhibiting molecular properties outside predefined ranges (step 3), and last, molecules for which either 2D to 3D conversion and ionization at pH 7.4 failed (step 4). It can be observed that nearly 60% of true active substances were removed during the filtering steps (see Tables S2 and S3, for exhaustive statistics), with step 2a eliminating the most true actives (Figure 2). This step is aimed at ruling out actives exhibiting very strong binding cooperativity and multiple binding sites.[47] True inactive substances, on the other hand, were not subjected to the three filtering substeps 2a, 2b, and 2c, thus lost much fewer members than the true actives, with over 90% of substances still remaining in the end.

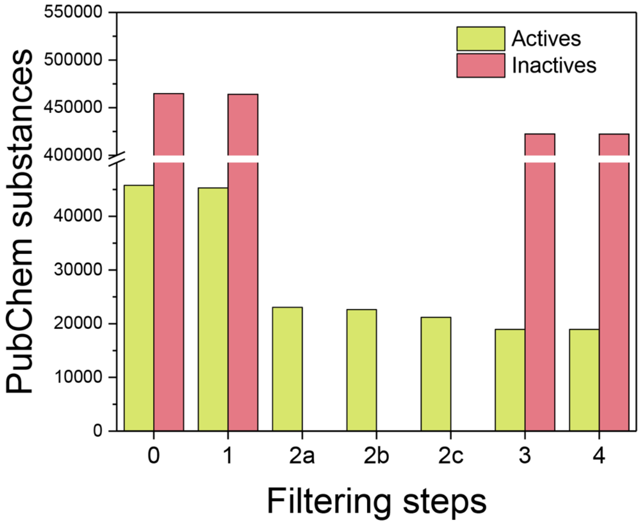


Figure 2. Total number of actives and inactives remaining after each filtering step was applied to the 21 selected target sets from PubChem bioassays: step 1, inorganic molecules; step 2a, compounds with Hill number $n_H < 0.5$ or $n_H > 2$; step 2b, frequency of hits FoH ≥ 0.26; step 2c, assay artifacts interfering with the readout (10,892 substances classified as aggregators or autofluorescent molecules or luciferase inhibitors); step 3, compounds with extreme molecular properties; step 4, 3D conversion and ionization failures. Steps 2a, 2b, and 2c were not applied to true inactives.

The filtering steps highlight the importance of removing assay artifacts in the composition of active substances. These steps not only prevented false positives that could affect subsequent screening performances but also significantly reduced the number of true actives in comparison with true inactives, thus bringing hit rates closer to that typically observed in experimental screening decks[32] but lower (in 15/21 cases) than that of artificially constructed data sets routinely used in cheminformatics (Figure 3A).

We next looked at the potency distribution of true actives (Figure 3B) in our data set with respect to that of the DUD-E and ChEMBL.[20] We can observe different potency distributions for DUD-E actives (n = 67,659; median potency = 7.46 ± 0.96) and for LIT-PCBA actives (n = 19,985; median potency = 5.22 ± 0.54). The micromolar potencies observed for most LIT-PCBA actives reflect affinities typically observed in HTS campaigns. Conversely, DUD-E actives tend to be much more potent (submicromolar in most cases) and consequently easier to be picked, thereby overestimating the real benefit of VS methods. At the individual target set level, the same trend applies when comparing the potencies of LIT-PCBA and ChEMBL ligands for 19 common target sets (Figure S1). Importantly, we believe that the enhanced difficulty proposed by our data set may enable a better discrimination of VS methodologies.

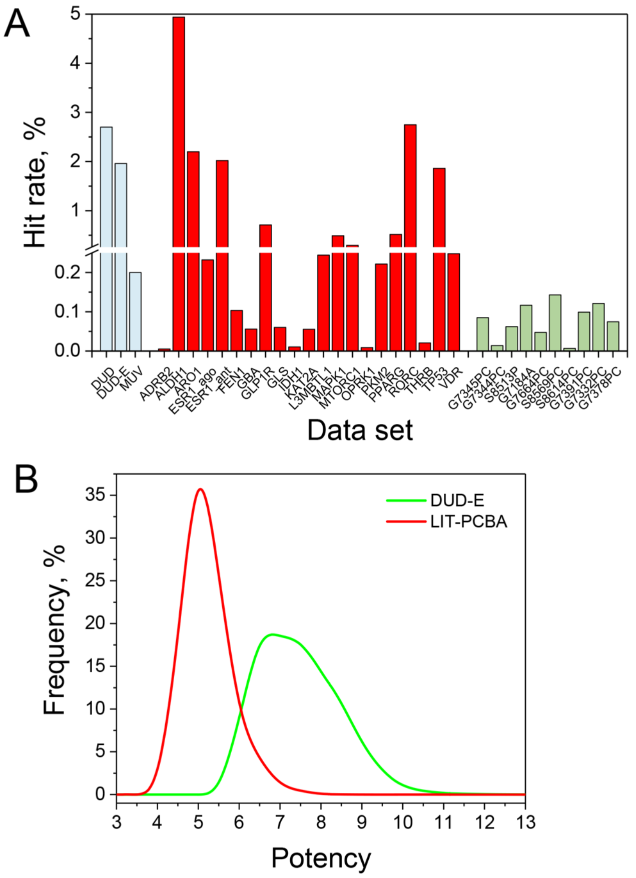


Figure 3. Properties of LIT-PCBA and standard data sets. (A) Confirmed hit rates for the LIT-PCBA data set (red bars), standard cheminformatics data sets (DUD,[6] DUD-E,[10] MUV;[7] blue bars), and a representative sample of 10 high-throughput screens from a major pharmaceutical company (green).[32] (B) Potency distribution of actives in the LIT-PCBA (red) and DUD-E (green) data sets. Potency is expressed as pIC_{50}, pEC_{50}, pK_i, or pK_d.

Virtual Screening and Performance Assessment. The suitability of the 21 fully processed target sets for virtual screening was next assessed by three standard methodologies: 2D fingerprint similarity, 3D shape similarity, and molecular docking. The aim of the computational experiment was not to compare the virtual screening accuracies of all methods but to check which of the 21 target sets may be unsuitable for virtual screening purposes. Hence, there is no guarantee that PubChem and PDB template ligands are strictly comparable (e.g., share the same binding site and molecular mechanism of action). Ligand-based VS will rapidly assess whether obvious biases are present in the ligand sets in terms of either 2D or 3D topologies. In addition, docking will ascertain if PubChem ligands share binding sites and interaction patterns with PDB templates. In each VS, all available PDB ligand or PDB target templates were iteratively used as reference, thereby generating as many hit lists as the available 162 templates. This exhaustive approach, albeit cumbersome, enables the selection of all references and takes into account the known chemical diversity of target-bound ligands (ligand-based VS) or the known conformational space accessible to the target of interest (docking). In addition, a target-based "max-pooling" approach was followed by merging all VS data related to any LIT-PCBA ligand, whatever the corresponding template, and retaining the highest value (2D similarity, 3D similarity, docking score) per ligand.

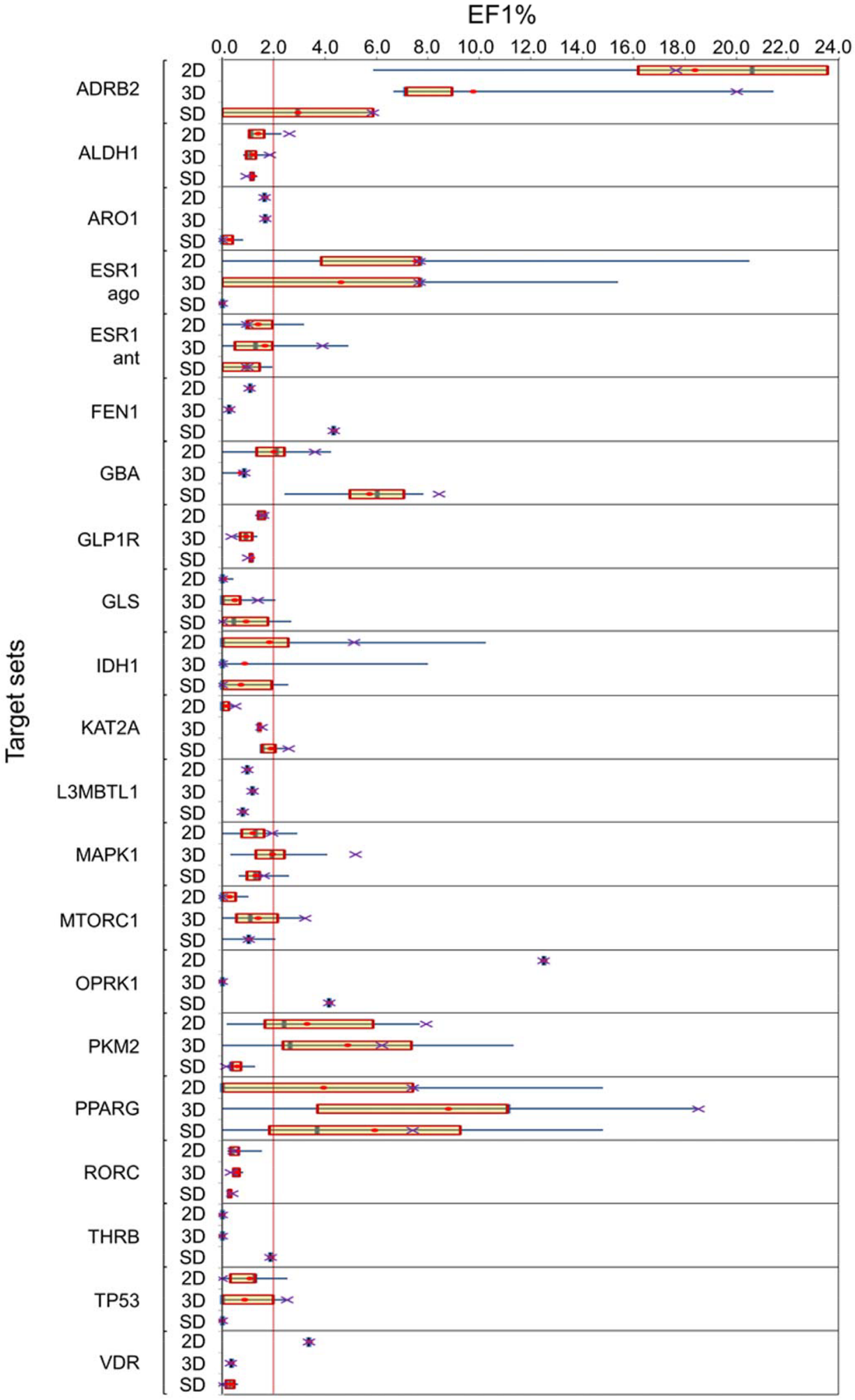


Figure 4. Performance of three different virtual screening methods (2D, ECFP4 fingerprint similarity; 3D, 3D shape similarity; SD, molecular docking with Surflex-Dock) on 21 fully processed target sets. The graphs represent the distribution of EF1% values (enrichment in true actives at a constant 1% false positive rate over random picking) obtained after screening. The boxes delimit the first and third quartiles, and the whiskers delimit the minimum and the maximum values. The median and the mean values are indicated by a green vertical line and a red dot located in each box, respectively. In cases where there is only one PDB template for a target set, or all templates gave the same EF1% value, the boxes are shrunk down into a single line. The purple crosses represent the EF1% values obtained by the max-pooling approach.

Statistical analyses of the data were primarily focused on enrichment factors in true actives at a constant 1% false positive rate (EF1%) as it mirrors the expectation of prospective VS practices. In addition, areas under the ROC curves have also been calculated and are given in Tables S4−S6.

As to be expected, inspection of observed enrichment in true actives for all 21 target sets clearly shows that the EF1% values may vary quite significantly according to the chosen template. In many instances, enrichment close to or even poorer than that obtained by random picking (EF1% = 1.0) is observed (Figure

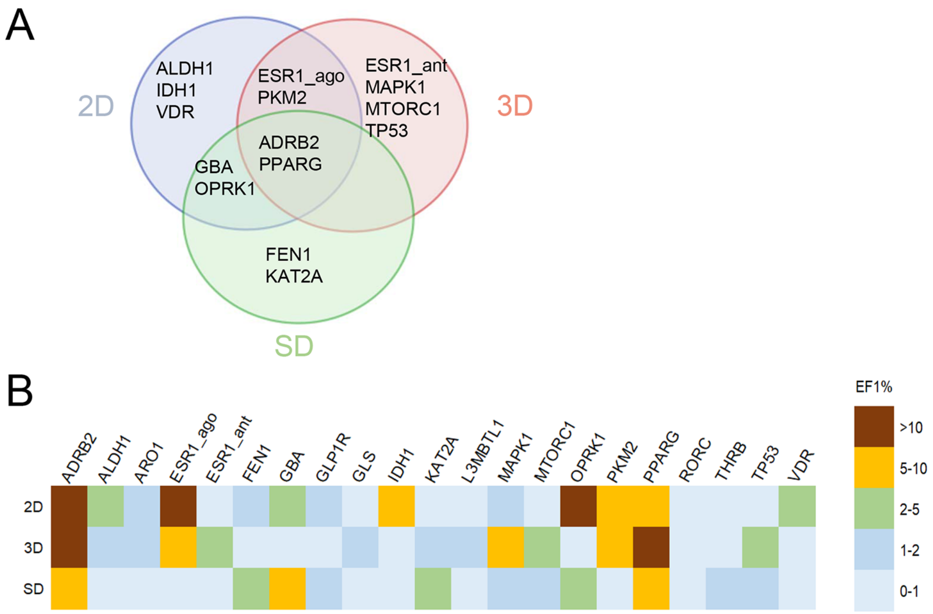


Figure 5. Comparative performance of three VS protocols (2D, ECFP4 fingerprint similarity search; 3D, shape similarity search; SD, molecular docking) for the 21 target sets, processed by a max-pooling approach. (A) Venn diagram of target sets for which an EF1% higher than 2.0 is observed. (B) Heatmap representing fused values of EF1% obtained for each of the 21 fully processed target sets by the three virtual screening methods. Abbreviations of target sets are indicated above the heat map.

4). We considered as acceptable any VS protocol yielding an EF1% value higher than 2, in other words, at least twice better than random picking. At this threshold, ligand-based methods clearly outperform docking (Figure 4). Interestingly, only five out the 63 VS assays, all concerning ligand-based approaches, led to enrichment higher than 10. This result highlights the particular challenge of screening the current data set that we attribute to two main reasons: (i) the apparent absence of obvious biases in the distribution of PubChem actives with respect to PDB templates in ligand space and (ii) the potency distribution of PubChem actives centered on micromolar hits.

Final Target Set Selection and Unbiasing. In order to facilitate the analysis, we from hereon discuss the results obtained by fusing, for each VS method, all data across all available target-specific templates ("max-pooling" approach). This strategy was supported by two main reasons: (i) The fused approach provides enrichments usually close to that obtained with the best possible template (Tables S4−S6). (ii) It enables the definition of a single hit list for each VS run while considering all templates. Fifteen out of the initial 21 target sets can be considered to be suited (EF1% > 2) for at least one of the three VS methods (Figure 5).

The current VS exercise suggests that six target sets (GLS, GLP1R, ARO1, THRB, RORC, L3MBTL1) are not adequate for VS purposes since none of the three VS methods is able to clearly distinguish confirmed actives from inactive compounds when the max-pooling approach was applied (EF1% < 2.0) (Figures 4 and 5). Moreover, for the five target sets among them (GLS as the only exception), the template-based scoring approach did not give any EF1% value above 2.0 either. Reasons for failures in screening these targets were (i) the promiscuity of the binding site toward many low-affinity chemotypes (e.g., ARO1), (ii) the presence of nonoverlapping binding sites (orthosteric vs allosteric) for PDB templates and PubChem actives (e.g., GLP1R, GLS, RORC), and (iii) the availabililty of a single PDB template (e.g., L3MBTL1, THRB).

Two target sets (ADRB2, PPARG) seem easier to handle since any of the three VS methods could successfully retrieve true actives with enrichments higher than 5. In four cases (GBA, OPRK1, PKM2, ESR1_ago), two VS methods succeeded. Last, only one VS method was able to perform correctly for nine sets (ALDH, IDH1, VDR, MTORC1, MAPK1, ESR1_ant, TP53, FEN1, KAT2A; Figure 5). This result is in agreement with many previous studies[48−50] suggesting that VS methodologies are orthogonal and is reassuring as it highlights the absence of obvious biases in either 2D molecular graph or 3D shape of LIT-PCBA compounds. It can therefore be implied that the remaining true actives (besides the ADRB2 and PPARG sets) do not resemble their corresponding PDB template ligands in both 2D and 3D shapes; meaning similarities between them, if there were any, did not significantly contribute to improving virtual screening performances, notably in early enrichment of true actives.

For each of the remaining 15 target sets, we ensured that the chemical diversity of PDB template ligands was not biasing our analysis. A first comparison of the number of Bemis−Murcko frameworks[51] to the total number of templates indicates that a wide variety of chemotypes are indeed available among the chosen PDB template ligands (Table S7). A self-similarity plot of templates (Tanimoto coefficient on MDL public keys) confirms this observation and shows, for most of the target sets (MTORC1 being an exception), a large chemical diversity (Figure S2).

The 15 target sets were last unbiased by the AVE method[25] to propose optimal training and validation sets for machine learning applications. In brief, a genetic algorithm (GA) is used to select four subsets of active and inactive compounds for training and validation sets, based on pairwise distance in chemical space (ECFP4 circular fingerprints) between the above-described four ligand subsets. The objective function of the GA (bias value) gears the splitting procedure to select

Table 2. Final List of 15 Targets Sets of LIT-PCBA Database

		AVE		Actives		Inactives		knn1[a]
Target	Target name	Bias	Iterations	Validation	Training	Validation	Training	ROC AUC
ADRB2	Beta2 adrenergic receptor	0.003	2	4	13	78,120	234,363	0.500
ALDH1[b]	Aldehyde dehydrogenase 1	0.092	195	1344	4032	25,868	77,606	0.556
ESR1_ago	Estrogen receptor alpha	0.001	1	3	10	1395	4188	0.499
ESR1_ant	Estrogen receptor alpha	0.006	9	25	77	1237	3711	0.517
FEN1	Flap endonuclease 1	0.076	39	92	277	88,850	266,552	0.499
GBA	Glucocerebrosidase	0.005	9	41	125	74,013	222,039	0.524
IDH1	Isocitrate dehydrogenase	0.001	4	9	30	90,512	271,537	0.500
KAT2A	Histone acetyltransferase KAT2A	0.001	5	48	146	87,137	261,411	0.500
MAPK1	Mitogen-activated protein kinase 1	0.000	8	77	231	15,657	46,972	0.505
MTORC1	Mechanistic target of rapamycin	0.001	7	24	73	8243	24,729	0.499
OPRK1	Kappa opioid receptor	0.000	3	6	18	67,454	202,362	0.500
PKM2	Pyruvate kinase muscle isoform 2	0.009	28	136	410	61,380	184,143	0.507
PPARG	Peroxisome proliferator-activated receptor γ	0.000	4	6	21	1302	3909	0.500
TP53	Cellular tumor antigen p53	0.008	29	19	60	1042	3126	0.491
VDR[b]	Vitamin D receptor	0.044	62	165	498	66,635	199,906	0.499

[a] Area under the ROC curve for a binary classification of validation compounds (active, inactive) based on a one-nearest-neighbor similarity search (ECFP4 fingeprints) model trained on target-specific training sets. [b] The size of the target set was reduced by 25% at the unbiasing stage due to the large number of remaining true actives.

training and validation sets for which distances in chemical space are homogeneously distributed when comparing training actives, validation actives, training inactives, and validation inactives. For 14 out of 15 target sets, just a few iterations (<100) of the GA were necessary to unbias the corresponding target sets with low bias values (Table 2). Interestingly, an optimal splitting was achieved without removing a single compound from 13 out of the 15 initial PubChem compound collections, thereby suggesting that the latter input did not exhibit major biases. The final AVE-unbiased LIT-PCBA data set covers 15 target sets, 7844 unique actives, and 407381 unique inactives (Table 2).

For two target sets (ALDH1, VDR), the high number of true actives forced us to reduce by 25% the size of the data set in order to reach completion of the GA search. In both cases, care was taken to keep the hit rate unchanged after data reduction. A one-nearest neighbor (knn1) binary classification of the 15 validation sets, still using ECFP4 fingerprints as descriptors, led to ROC area under the curve values close to random (0.500) and thereby supports the bona fide debiasing of all corresponding target sets. Analyses of the three baseline VS experiments for the AVE validation sets only (Table S8) confirm the very challenging nature of the data set as the performance drops for many target sets, notably those with a low number of actives (e.g., ADRB2, IDH1) or few PDB template ligands (e.g., OPRK1). As previously indicated, the baseline VS protocol was just intended to remove PubChem HTS data unsuitable for virtual screening applications and is not indicative of the performance of modern machine learning approaches. We however recommend the application of such methods to target sets exhibiting enough true actives to train on (ALDH1, FEN1, GBA, KAT2A, MAPK1, PKM2, VDR; Table 2).

CONCLUSION

A rigorous ligand data set preparation is necessary to benchmark virtual screening and/or machine learning methods. Since the body of known experimental data is continuously increasing, such benchmark data sets need periodical revisions to remove both obvious and hidden biases that are inherent to human decision making. Otherwise, errors are propagated across the literature and prevent a true comparison of novel methodological developments. Several recent reports[24−27] unambiguously demonstrated that the cheminformatics community is currently facing this situation, leading notably to overoptimistic reports on the real benefit of artificial intelligence methods (e.g., deep neural networks) when applied to structure-based ligand design. We herewith present LIT-PCBA as a novel generation of virtual screening benchmarking data sets, specifically designed to reveal the true potential of computational methods in virtual screening exercises. The data set has been designed from dose−response PubChem bioassays for which active and inactive compounds are unambiguously defined. Importantly, a careful examination of metadata enabled the removal of assay artifacts, frequent hitters, and false positives. LIT-PCBA consists of 15 target sets covering a wide diversity of ligands and target proteins. Preliminary virtual screening attempts with state-of-the-art methods (2D similarity, 3D shape matching, docking) suggest that the data set is very challenging, notably because potency distribution biases among labeled active compounds are absent. Last, a recently described unbiasing procedure[25] was applied to LIT-PCBA to enable an optimal distribution of training and validation compounds for machine learning. We do believe that the particular challenge brought by this data set will enable a clearer appreciation of modern artificial intelligence methods in structure-based virtual screening scenarios. The LIT-PCBA data set is freely accessible at http://drugdesign.unistra.fr/LIT-PCBA.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00155.

Distribution of potencies (in pIC_{50}, pEC_{50}, pK_i, pK_d) for confirmed actives of the LIT-PCBA, DUD-E and ChEMBL ligands; self-similarity matrix of PDB template ligands; description of 21 selected PubChem BioAssays; number of confirmed active compounds remaining after each filtering step; number of inactive compounds remaining after each filtering step; virtual screening results obtained from 2D ECFP4 similarity search on 21 fully processed selected target sets; virtual screening results obtained from 3D shape similarity search on 21 fully processed selected target sets; virtual screening results obtained by molecular docking on 21 fully processed selected target sets; chemical diversity of PDB template ligands, assessed by the number of unique Bemis−Murcko frameworks; virtual screening results (EF1%) obtained by 2D ECFP4 similarity search, 3D shape similarity search, and molecular docking on 15 validation sets after debiasing with AVE (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Didier Rognan − *Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France;* orcid.org/0000-0002-0577-641X; Phone: +33 3 68 85 42 35; Email: rognan@unistra.fr; Fax: +33 3 68 85 43 10

Authors

Viet-Khoa Tran-Nguyen − *Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France;* orcid.org/0000-0001-7497-333X

Célien Jacquemard − *Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00155

Funding

The authors are thankful for the financial support from the doctoral school of chemical science (EDSC, University of Strasbourg) for a Ph.D. grant to V-K.T-N.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) is acknowledged for allocation of computing time and excellent support. We sincerely thank Prof. M. Rarey (University of Hamburg, Germany) and Dr. I. Wallach (Atomwise, U.S.A) for providing an executable version of Protoss and of the original AVE unbiasing Python script, respectively.

■ REFERENCES

(1) Rognan, D. The Impact of in Silico Screening in the Discovery of Novel and Safer Drug Candidates. *Pharmacol. Ther.* **2017**, *175*, 47−66.

(2) Wingert, B. M.; Camacho, C. J. Improving Small Molecule Virtual Screening Strategies for the Next Generation of Therapeutics. *Curr. Opin. Chem. Biol.* **2018**, *44*, 87−92.

(3) Perez-Sianes, J.; Perez-Sanchez, H.; Diaz, F. Virtual Screening Meets Deep Learning. *Curr. Comput.-Aided Drug Des.* **2018**, *15*, 6−28.

(4) Gimeno, A.; Ojeda-Montes, M. J.; Tomas-Hernandez, S.; Cereto-Massague, A.; Beltran-Debon, R.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int. J. Mol. Sci.* **2019**, *20*, 1375.

(5) 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*, 4759−4767.

(6) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(7) 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*, 169−184.

(8) 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*, 2650−2665.

(9) Wallach, I.; Lilien, R. Virtual Decoy Sets for Molecular Docking Benchmarks. *J. Chem. Inf. Model.* **2011**, *51*, 196−202.

(10) 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*, 6582−6594.

(11) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.

(12) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(13) Empereur-Mot, C.; Guillemain, H.; Latouche, A.; Zagury, J. F.; Viallon, V.; Montes, M. Predictiveness Curves in Virtual Screening. *J. Cheminf.* **2015**, *7*, 52.

(14) 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*, 2895−2907.

(15) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856−5868.

(16) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 793−806.

(17) Irwin, J. J.; Shoichet, B. K. Zinc - a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(18) 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*, 169−178.

(19) 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*, 179−190.

(20) 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*, D1100−1107.

(21) Réau, M.; Langenfeld, L.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*, 11.

(22) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. Pubchem Bioassay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955−D963.

(23) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods against the Muv Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 2168−2178.

(24) Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminf.* **2016**, *8*, 56.

(25) Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather Than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916−932.

(26) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the Dud-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* **2019**, *14*, No. e0220113.

(27) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947−961.

(28) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520−10594.

(29) Wallach, I.; Dzamba, M.; Heifets, A. Atomnet: A Deep, Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. *arXiv*, arXiv:1510.02855, October 10, **2015**, ver. 1.

(30) Pereira, J. C.; Caffarena, E. R.; Dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* **2016**, *56*, 2495−2506.

(31) Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. *J. Chem. Inf. Model.* **2018**, *58*, 2319−2330.

(32) Posner, B. A.; Xi, H.; Mills, J. E. Enhanced Hts Hit Selection Via a Local Hit Rate Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2202−2210.

(33) Kim, S. Getting the Most out of Pubchem for Virtual Screening. *Expert Opin. Drug Discovery* **2016**, *11*, 843−855.

(34) 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.

(35) The Uniprot Consortium. Uniprot: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699.

(36) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminf.* **2014**, *6*, 12.

(37) *Certara USA, Inc.* https://www.certara.com/ (accessed April 2020).

(38) Da Silva, F.; Desaphy, J.; Rognan, D. Ichem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507−510.

(39) 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.

(40) Dassault Systèmes, Biovia Corp. https://www.3dsbiovia.com/ (accessed April 2020).

(41) Molecular Networks Gmbh. https://www.mn-am.com/ (accessed April 2020).

(42) Openeye Scientific Software. https://www.eyesopen.com/ (accessed April 2020).

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

(44) Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with Omega: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(45) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(46) Jain, A. N. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281−306.

(47) Prinz, H. Hill Coefficients, Dose-Response Curves and Allosteric Mechanisms. *J. Chem. Biol.* **2010**, *3*, 37−44.

(48) Kruger, D. M.; Evers, A. Comparison of Structure- and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2010**, *5*, 148−158.

(49) Tian, S.; Sun, H. Y.; Li, Y. Y.; Pan, P. C.; Li, D.; Hou, T. J. Development and Evaluation of an Integrated Virtual Screening Strategy by Combining Molecular Docking and Pharmacophore Searching Based on Multiple Protein Structures. *J. Chem. Inf. Model.* **2013**, *53*, 2743−2756.

(50) Tran-Nguyen, V. K.; Da Silva, F.; Bret, G.; Rognan, D. All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception, and Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 573−585.

(51) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.