True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better

Viet-Khoa Tran-Nguyen, Guillaume Bret, Didier Rognan

True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better.

Viet-Khoa Tran-Nguyen[†], Guillaume Bret[†] and Didier Rognan[†]*

[†] Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France.

* To whom correspondence should be addressed (phone: +33 3 68 85 42 35, fax: +33 3 68 85 43 10, email: rognan@unistra.fr)

ABSTRACT

Hundreds of fast scoring functions have been developed over the last 20 years to predict binding free energies from the three-dimensional structures of protein-ligand complexes. Despite numerous statistical promises, we believe that none of them has been properly validated for daily prospective high-throughput virtual screening studies, mostly because *in silico* screening challenges usually employ artificially-built and biased datasets. We herewith carry out a fully unbiased evaluation of four scoring functions (Pafnucy, $\Delta_{vina}RF_{20}$, IFP, GRIM) on an in-house developed data collection of experimental high-confidence screening data (LIT-PCBA) covering about 3 million data points on 15 diverse pharmaceutical targets. All four scoring functions were applied to rescore the docking poses of LIT-PCBA compounds in conditions mimicking exactly standard drug discovery scenarios, and were compared in terms of propensity to enrich true binders in the top 1%-ranked hit lists. Interestingly, rescoring based on simple interaction fingerprints or interaction graphs outperforms state-of-the-art machine learning and deep learning scoring functions in most cases. The current study notably highlights the strong tendency of deep learning methods to predict affinity values within a very narrow range centered on the mean value of samples used for training. Moreover, it suggests that the knowledge of preexisting binding modes is the key to detecting the most potent binders.

INTRODUCTION

Predicting binding free energies from the three-dimensional (3D) structure of a bimolecular complex remains a grand challenge of computational chemistry.[1] In the field of drug discovery, this problem is of paramount importance since billions of "make-on-demand" compounds can nowadays be screened either experimentally or virtually at unprecedented speeds.[2-3] To limit efforts in human and financial resources, virtual screening[4] has gained considerable importance in the last decade with three major objectives: (i) prioritize binding modes to a target of interest, (ii) rank compounds by decreasing binding free energy (affinity) values, and (iii) guide hit-to-lead optimization. Molecular docking[5] remains the computational method of choice to answer these three questions, notably when it comes to addressing a large chemical space. It consists in simultaneously solving two issues: finding the best orientation of a ligand inside a protein cavity, and predicting its binding free energy or free energy difference in comparison to a reference ligand. Hundreds of potential algorithmic solutions have been proposed and surveyed in numerous reviews.[6-9] The drug discovery community now agrees on the statement that most docking algorithms are generally able, if a biologically relevant target's 3D structure is available and if the docked ligand obeys drug-likeness rules, to propose a docking pose close to the experimental binding mode.[10] The major issue is still to prioritize near-native binding modes from irrelevant poses, and consequently to rank potential ligands from a large chemical space. On the one hand, free energy perturbation methods have reached the speed and an accuracy level necessary to properly rank a limited set of congeneric ligands for a wide array of targets.[11] On the other hand, fast empirical scoring functions (SFs) can score billions of docking poses but with very limited accuracy.[12] Post-processing docking poses by alternative scoring schemes has therefore been the subject of intense research during the last decade.[13-14] Notably, machine learning (ML)[15] has gained considerable popularity, since the corresponding algorithms (e.g. support vector machines, random forests, decision trees, deep neural networks) can theoretically delineate subtle non-linear relationships in a vast hyperparameter space describing experimentally solved protein-ligand complexes. Unfortunately, the true benefit of ML-based SFs remains a matter of debate,[16-18] notably

because of a lack of unbiased and shared data/protocols to rigorously compare such methods. Noticeable attempts to organize international docking/scoring contests (CSAR, CASF, D3R, CELPP) [19-23, 10, 24] have flourished and helped the community to better define the remaining challenges involving:[25] (i) the ability to discriminate near-native docking poses from irrelevant ones (docking power), (ii) the propensity to predict experimental affinities (scoring power), (iii) the capability of ranking known ligands by decreasing binding free energy values (ranking power); and (iv) the ability to detect true binders among a large set of compounds (screening power). The impact of these challenges on the development of better predictive SFs remains, however, disappointingly limited for multiple reasons. First, many algorithm developers are not aware of these initiatives or do not participate in these contests, leaving their methods insufficiently validated. Consequently, virtual screening (VS) practitioners observe a huge gap between the claimed accuracy level in retrospective studies and real-life performances in prospective VS campaigns.[26] Last, recent studies have brought to light unintentional biases in datasets (e.g. DUD, DUD-E) classically used in VS challenges,[27-28] leading to an overestimation of virtual screening accuracy in all existing studies.

To foster unbiased VS comparisons, we recently developed a novel dataset (LIT-PCBA),[26] specifically designed to evaluate the accuracy of VS methods, consisting of 15 target sets, 8020 true actives and 2,675,399 true inactive compounds. LIT-PCBA differs from all existing benchmark sets in the following key properties: (i) the dataset mimics real-life screening decks as it has been designed to discriminate moderately potent actives (primary hits) from inactive compounds; (ii) the potency of all compounds (actives, inactives) for a particular target has been determined experimentally under homogeneous conditions; (iii) the active-to-inactive ratio (0.1-0.5%) reflects hit rates typically observed in high-throughput screening (HTS) campaigns against targets of pharmaceutical interest; (iv) the actives have been filtered to remove false positives, frequent hitters, assay artifacts and truly undruggable molecules; (v) dose-response curves are available for all actives.

Preliminary VS experiments logically provided evidence that LIT-PCBA is very challenging whatever the computational screening method,[26] thereby offering an opportunity to estimate the true virtual screening accuracy of state-of-the-art SFs. In the current study, we specifically evaluated four SFs in terms of VS accuracy on the same set of LIT-PCBA docking poses **(Figure 1**).

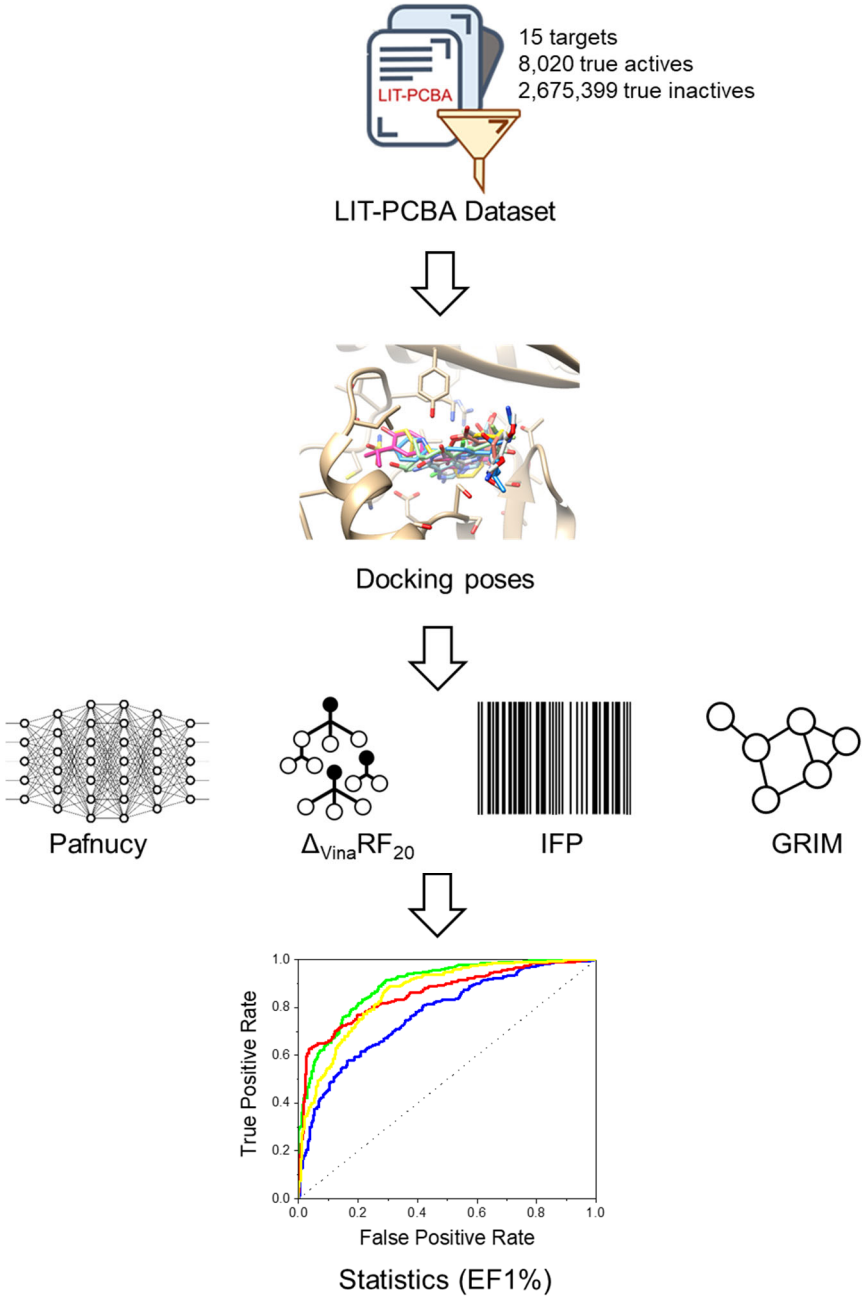

Figure 1. Overall flowchart of the current study. LIT-PCBA compounds were docked into their respective 15 targets and rescored according to four scoring fucntions: two machine learning ($\Delta_{vina}RF_{20}$, Pafnucy) and two knowledge-based topological functions (IFP, GRIM). For each target, compounds were ranked by decreasing scores and the virtual screening accuracy of each rescoring method was analysed statistically (enrichment in true actives at 1% false positive rate, EF1%)

The first two SFs (Pafnucy, $\Delta_{vina}RF_{20}$) are representative of recently-developed ML-based methods while the two others (IFP, GRIM) illustrate simpler knowledge-based topological SFs. In other words, the first two methods predict absolute binding free energies whereas the last two approaches rank compounds irrespective of energetic criteria. Each rescoring method is briefly described below.

Pafnucy[29] was selected as a representative of recent deep neural network (DNN)-based SFs. By mimicry to image recognition, the protein-ligand complex is voxelized in a 1 Å-resolution 20 Å-wide 3D cubic grid, with each voxel storing 19 features of the complex (e.g. physicochemical and pharmacophoric properties, interaction energies). In Pafnucy's architecture, the protein-ligand complex is described by a 4D tensor processed by three convolutional layers and three dense layers to predict the binding affinity. The DNN achieved a scoring performance on par with the best ML or DNN-based SFs on the standard PDBbind 2016 benchmark dataset.[30]

The $\Delta_{vina}RF_{20}$ scoring function[31] was chosen as a prototypical ML-based SF achieving top performances, among a wide list of competing functions, in several challenges (scoring, docking, ranking, virtual screening) using three benchmark sets (CASF-2007, CASF-2013, CASF-2016).[31, 25] In brief, a random forest (RF) algorithm is applied to predict the differences between experimental affinities and those calculated by Autodock Vina's scoring function[32] from a set of 20 molecular features. The RF correction term is then added to the native AutoDock Vina score to predict the absolute binding free energy of a ligand for a given protein.

The IFP[33] method relies on the similarity of protein-ligand interactions between a docking pose and any given template (e.g., the X-ray structure of the cognate protein with a known active ligand). In the first step, an interaction fingerprint for each docking pose is generated as a fixed-length bit string registering the presence or the absence of seven non-covalent interactions with user-defined cavity-lining amino acids (including co-factors, ions and water molecules if desired). Docking poses and consequently ligands are then sorted by decreasing interaction fingerprint similarity levels to that given by the template, expressed by a simple Tanimoto coefficient.[33] In the hands of independent

research groups, IFP rescoring has demonstrated its superiority to conventional energy-based SFs in several retrospective virtual screening studies[33-36] and has repeatedly enabled the discovery of experimentally verified hits for a wide array of protein targets.[37-40]

Last, GRIM[41] focuses on coordinate-frame invariant protein-ligand interaction patterns. The interaction pattern is formalized as a graph whose nodes are placed on so-called interaction pseudoatoms (IPAs) featuring for each non-covalent interaction a ligand-interacting atom, a protein-interacting atom, and the barycenter of these two atoms. A graph is first computed for a reference template, which is usually an X-ray protein-ligand structure, and another graph is then created for a given docking pose. A clique detection algorithm is used to find the maximal common subgraph between the above two graphs. The similarity of the two interaction patterns is expressed by a composite score (GRIMscore) featuring the number of matched IPAs, the quality and the root-mean-square deviation of the matched clique.[41] In comparison to interaction fingerprints, interaction pattern graphs are not restricted to a fixed list of binding site atoms such that pairwise comparisons are also possible for binding cavities of different sizes. GRIM rescoring has been proven effective in predicting, with a high accuracy level, the binding modes of various ligands before the release of experimental crystallographic structures in international docking/scoring contests.[42-43]

COMPUTATIONAL METHODS

LIT-PCBA dataset. The full dataset[26] was downloaded from http://drugdesign.unistra.fr/LIT-PCBA and used with no modifications. For each of the 15 targets, at least four files are provided: the protein PDB X-ray structure (MOL2 file format), the bound PDB ligand (MOL2 file format), a list of true actives (SMILES string), and a list of true inactives (SMILES string). In case several PDB templates are available, several protein and ligand input files are given. Ligand and protein input coordinates (protonation, tautomerism, generation of ligand 3D structures; inclusion of water molecules, co-factors and ions in protein structures) were prepared as previously described.[26]

Docking poses. Starting from the mol2 structure of the template protein and that of its co-crystallized ligand, a protomol representing the ligand-binding site was generated from protein-bound ligand atomic coordinates using default settings of Surflex-Dock v.3066.[44] All molecules in the relevant target set were docked into the protomol using the "*–pgeom*" parameter (geometric docking search) of the docking engine. The best 20 poses ranked by decreasing pK_d values were retained for further rescoring. Surflex-Dock[44] uses an empirically derived scoring function based on the binding affinities of protein-ligand complexes coupled with their crystallographically determined structures. The function's primary terms involve hydrophobic and polar complementarity, with additional terms for entropy and solvation effects.

Deep learning rescoring (Pafnucy). The package was downloaded from https://gitlab.com/cheminfIBB/pafnucy. In the first step, 3D grids were prepared for each protein-ligand complex in MOL2 file format, to create an HDF file with atoms' coordinates and features. In the second step, the recommended model (batch5-2017-06-05T07:58:47-best) was used to rescore each protein-ligand complex, expressing results in pK_d unit.

Machine learning rescoring ($\Delta_{vina}RF_{20}$). DeltaVina[31] was downloaded from https://github.com/chengwang88/deltavina and directly used to rescore the interactions between the above-described protein coordinates and LIT-PCBA docked poses in MOL2 file format, outputting results in pK_d unit.

Rescoring by Protein-Ligand Interaction Fingerprint (IFP) Similarity. The IFP module[33] of the IChem v5.2.9 package[45] was employed to compute the similarity between the IFP recorded for each docked LIT-PCBA ligand and that of a protein-ligand PDB template. The mol2 structure of the template binding site and the multi-mol2 files containing the merged docking poses issued by Surflex-Dock were used as input for IChem processing. The binding site refers here to amino acid residues (plus water molecules, ions and co-factors) of the protein having at least one heavy atom within 5.0 Å from any heavy atom of the co-crystallized template ligand. The IFP similarity between the template and each docking pose was expressed by a Tanimoto coefficient.

Rescoring by Interaction Graph-Matching (GRIM). The GRIM module[41] of the IChem v.5.2.9 package[45] was employed to post-process Surflex-Dock docking poses, using the same input files as those described for IFP rescoring. Output was expressed by decreasing GRIMscores.[41]

Statistics. The enrichment in true active molecules at a constant 1% false positive rate over random picking (EF1%) was calculated for each separate hit list. The same procedure was carried out by fusing all lists for a given target (in case of multiple PDB templates) and keeping the highest score value for each compound ("max-pooling" approach) and each specific scoring function. In other words, all ligands from a target set (e.g. ADRB2) are first ranked by decreasing scores, thereby generating 4 hit

lists (Pafnucy, $\Delta_{vina}RF_{20}$, IFP, GRIM) for each PDB template. For each scoring function (e.g. Pafnucy), the lists originating from all templates (8 in the ADRB2 case) are then merged, retaining the highest score for each ligand.

RESULTS AND DISCUSSION

We recently proposed a novel dataset (LIT-PCBA)[26] for benchmarking virtual screening methods in a truly unbiased manner. Since the dataset is applicable to both ligand-based and structure-based methods, it can be used to compare state-of-the-art SFs of diverse physicochemical backgrounds. In the current study, all LIT-PCBA ligands (true actives as well as true inactives) were docked to their cognate targets with a single tool, thereby offering the possibility to compare, head to head, four representative scoring schemes in a virtual screening exercise very close to daily encountered drug discovery scenarios.

Simple rescoring methods constantly outperform machine leaning and deep learning methods. EF1% values were first computed from pooled lists in which the highest score was assigned to a given ligand bound to its target, whatever the target PDB structure (some targets have up to 15 input X-ray structures) and the number of docking poses (**Table 1**, **Figure 2**). The obtained results confirm that the dataset is really challenging since EF1% enrichment factors (ranging on average from 2 to 7, **Table 1**; see all values in **Tables S1-S5**) are much lower than those obtained in previous benchmarking studies on easier but unfortunately biased datasets.[46-47] All four investigated SFs clearly outperformed the native Surflex-Dock function in enriching true binders among the top scorers.

To verify whether the Surflex-Dock docking engine was responsible for hidden biases or not (e.g. if better performances of Surflex-Dock in reproducing crystallographic structures translate into higher EF1% values), we redocked all 129 LIT-PCBA template ligands to their cognate targets (**Table S6**). For 108 out of the 129 ligands, Surflex-Dock is able to generate at least a docking pose that does not

deviate more than 2.0 Å from the corresponding X-ray pose (rmsd ≤ 2.0 Å). Considering the top-ranked pose only, this ratio is, as expected, lower (78/129). The top-ranked pose is rarely the one with the lowest rmsd value to the X-ray pose. The average rmsd of the top poses (2.87 Å) is indeed much higher than that of the absolute lowest rmsd poses (1.41 Å). The benefit of rescoring all docking poses over just picking the top-ranked solution is independent of the target set, and observed for all targets. To get the maximal benefit of rescoring, all poses should therefore be considered. Rescoring the sole top-ranked docking pose by alternative methods, as proposed by many recent deep learning models,[48-50] should be avoided in the light of this study. Altogether, we could not see any relationship between the self-docking performance of our docking engine and the observed EF1% values for the 15 LIT-PCBA targets. We therefore conclude that the docking performance of Surflex-Dock has not brought obvious biases in our data collection.

Four target sets (IDH1, MTORC1, TP53, VDR) are really challenging since none of the rescoring methods was able to yield EF1% values above 5 (**Table 1**; **Figure S1**). The main reason for this failure is the weak potency of some actives (notably for IDH1 and VDR ligands). In addition, we cannot rule out the possibility that the Surflex-Dock pose sampler has been unable to propose at least one native-like pose for these ligands. The inability to recover true LIT-PCBA actives might also come from possible binding of these molecules to different pockets other than those investigated in the present study (e.g. allosteric ones). In most cases, such additional pockets could be detected at the surface of the corresponding target structures. Last, the bad performance observed for the TP53 set may also be explained by the lack of a full-length PDB structure. For the remaining 11 targets, a clear trend is observed, with simple interaction-based SFs (IFP, GRIM) outperforming machine learning/deep learning functions (Pafnucy, $\Delta_{vina}RF_{20}$). The particular advantage of an SF can be target-dependent but target-averaged EF1% values unambiguously demonstrate equal performances of IFP and GRIM, superior to that of Pafnucy and $\Delta_{vina}RF_{20}$ (**Table 1**). Paired samples t-tests at a 0.05 significance level show that Pafnucy and $\Delta_{vina}RF_{20}$ score distributions are statistically not different (p-value = 0.82), as well as IFP and GRIM score distributions (p-value = 0.92). However, the GRIM score distribution was

found significantly different from those of both ML SFs (p-values of 0.012 and 0.0085 against Pafnucy and $\Delta_{vina}RF_{20}$, respectively). Considering EF1% values of increasing thresholds (2, 5, 10) corresponding to virtual screening results of increasing accuracy levels (EF1%>2: moderate, EF1%>5: good, EF1%>10: excellent), GRIM came up as the most successful method when applied to the LIT-PCBA dataset, closely followed by IFP and then by Pafnucy and $\Delta_{vina}RF_{20}$, which, although different in their conception and physical background, exhibit almost identical behaviors (**Figure 2**, **Table 1**). The apparent superiority of interaction-based SFs is not explained by the number of existing protein-ligand PDB templates from which similarity values are inferred, since target sets with a single PDB template used for docking (e.g. FEN1, OPRK1) are also better handled. Likewise, we could not depict any relationship between VS performances and the number of true actives for all investigated targets. Last, previous comparisons of true actives to PDB ligands used as templates did not detect high 2D fingerprint or 3D shape-based pharmacophore similarities,[26] therefore excluding the possibility that rescoring based on interaction fingerprints or interaction graphs could be dominated by simple ligand neighborhoods. Although simplistic in their formulation, selecting hits based on protein-ligand interaction similarity to existing templates can therefore be considered a very robust approach, if at least one holo-protein structure and a binding pocket is known. Two state-of-the-art ML SFs still provide a significant enrichment in true actives in comparison to both random picking and docking scores, but disappointingly exhibit a lower performance than IFP/GRIM, although being trained on a much larger body of experimental protein-ligand complexes.[31, 29] IFP/GRIM can be considered as target-focused SFs where preliminary data on co-crystallized ligands with the target of interest are used as a guide to rank poses for new ligands. This is the main difference to general-purpose ML SFs that have been parameterized from a limited target space (usually the PDBbind dataset)[51] and therefore lack the close (ligand) neighborhood relationships that we believe to be important to discriminate realistic poses from irrelevant ones. It is, however, very disappointing to see that the enhanced ranking accuracy (ability to predict experimental affinities) reported for these new scoring functions[31, 25] does not translate into an enhanced virtual screening accuracy level, suggesting either some overtraining or more likely a narrow applicability

domain that is limited to the PDBbind target space. In the face of a very challenging task, a much broader and more diverse target space is required to properly train ML SFs in order to reach the simplicity and interpretability of interaction-based SFs.

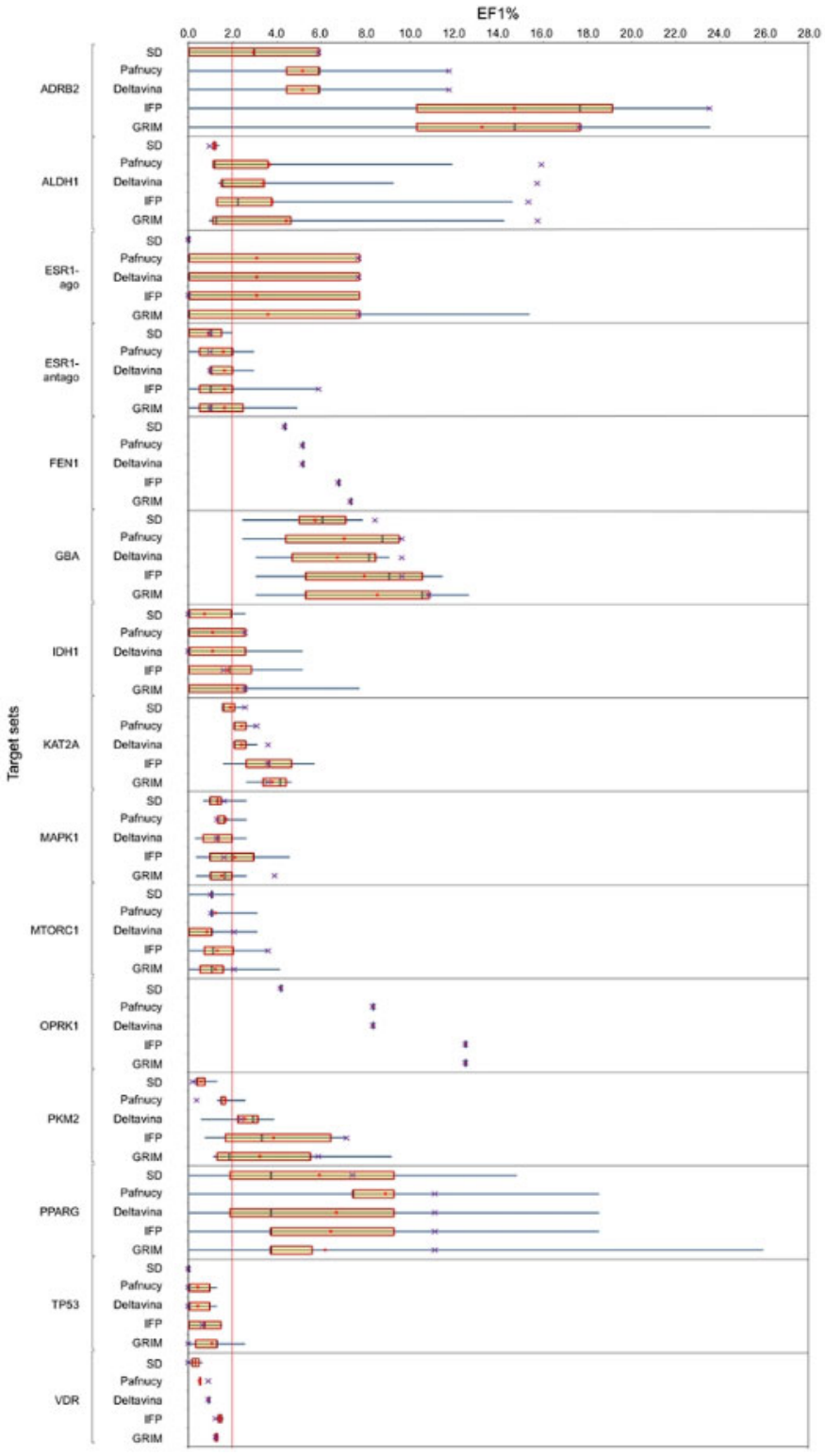

Figure 2. Performance of four rescoring methods (Pafnucy, $\Delta_{vina}RF_{20}$, IFP, GRIM) on 15 LIT-PCBA target sets, in comparison to those of Surflex-Dock's native scoring function (SD). The graph represents the distribution of EF1% values (enrichment in true actives at a constant 1% false positive rate over random picking) obtained after virtual screening. The boxes delimit the 1st and the 3rd quartiles, 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.

Table 1. Virtual screening accuracy of five scoring functions in scoring LIT-PCBA docking poses. Accuracy is expressed as the enrichment factor in true positives at a 1% false positive rate (EF1%). Enrichment factors are highlighed in yellow and green in cases of acceptable (EF1%>2) or high (EF1%>10) performances, respectively. Target-specific EF1% values corresponding to the most accurate function are underlined. All enrichment factors were calculated from pooled hit lists ("max-pooling" approach keeping the highest score for a given protein-ligand complex).

Target set	Scoring function					PDB	Number	EF1%
	Surflex	Pafnucy	$\Delta_{vina}RF_{20}$	IFP	GRIM	Templates	of Actives	max
ADRB2	5.88	11.76	11.76	23.53	17.65	8	17	100.00
ALDH1	0.96	15.93	15.74	15.35	15.76	8	7168	20.24
ESR1-ago	0	7.69	7.69	7.69	7.69	15	13	100.00
ESR1-ant	0.98	0.98	0.98	5.88	0.98	15	102	49.50
FEN1	4.34	5.15	5.15	6.78	7.32	1	369	100.00
GBA	8.43	9.64	9.64	9.64	10.84	6	166	100.00
IDH1	0	2.56	0	1.61	2.56	14	39	100.00
KAT2A	2.58	3.09	3.61	3.61	3.61	3	194	100.00
MAPK1	1.62	1.3	1.3	1.62	3.9	15	308	100.00
MTORC1	1.03	1.03	2.06	3.61	2.06	11	97	100.00
OPRK1	4.17	8.33	8.33	12.5	12.5	1	24	100.00
PKM2	0.18	0.37	2.38	7.14	5.86	9	546	100.00
PPARG	7.41	11.11	11.11	11.11	11.11	15	27	100.00
TP53	0	0	0	0.66	0	6	79	53.75
VDR	0	0.91	0.91	1.24	1.24	2	884	100.00
Average	2.51	5.32	5.38	7.46	6.87			
EF1% > 2	6	9	10	11	12			
EF1% > 5	3	7	7	9	8			
EF1% > 10	0	3	3	4	5			

Do different scoring functions retrieve the same hits? The previous analysis suggests that Pafnucy and $\Delta_{vina}RF_{20}$, although different in their conception and physical background, often provide almost identical enrichment factors (**Table 1**). We next examined whether the same set of actives tend to be retrieved by ML SFs on the one hand, and by interaction-based functions on the other hand. To this end, we counted the numbers of true actives uniquely found among the top 1% scored compounds by

each scoring function and each possible combination of two up to four rescoring schemes (**Figure 3**). To avoid biasing the statistics from ALDH1 inhibitors, which provide most of the recovered true actives (2671 out of 2949 ligands), these ligands were discarded from the current analysis. Interestingly, more than 50% of the hits were uniquely found by a single function, evidencing a rather orthogonal selection of compounds by the herein investigated SFs. As to be expected, IFP and GRIM interaction-based SFs, sharing the same roots, select a large number of identical ligands. However, each of these interaction-based SFs also retrieves some unique hits of its own, most likely by different treatment of apolar interactions, with the latter being less weighted in GRIM.[41] Interestingly, very few compounds were selected by any other possible combination of two scoring functions. Clearly, ML scoring functions (Pafnucy and $\Delta_{vina}RF_{20}$) do not select the same sets of hits as those issued by interaction-based functions, and they also pick different sets of ligands (**Figure 3**). The observed trends were not significantly altered by looking at all target sets including ALDH1 inhibitors (**Figure S2**).

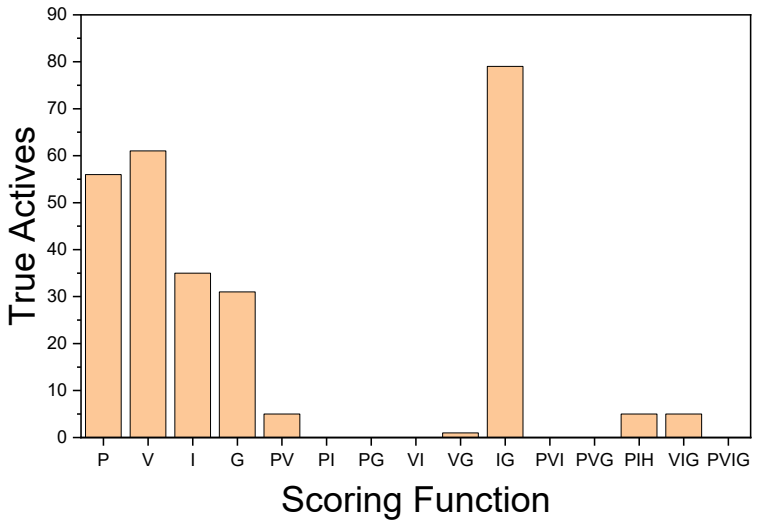

Figure 3. Number of unique true actives retrieved among the top 1% scored compounds by individual scoring functions or all possible scoring combinations (P: Pafnucy, V: $\Delta_{vina}RF_{20}$, I: IFP, G: GRIM). ALDH1 inhibitors have been discarded from the current analysis.

ML-based SFs logically select compounds based on their predicted biding affinity, irrespectively of known interaction patterns. Conversely, interaction-based SFs retrieve true actives which are not

necessarily chemically similar to known PDB ligands, but present key non-covalent interactions with key active site residues already observed in co-crystallized ligands.

Beware of score distributions. We previously described the tendency of ML SFs to predict binding affinities within a very narrow range centered on the mean values of samples on which they have been trained.[16] We therefore plotted the predicted values (pK_d for Pafnucy and $\Delta_{vina}RF_{20}$, similarity scores for IFP and GRIM) for the entire LIT-PCBA set (**Figure 4**).

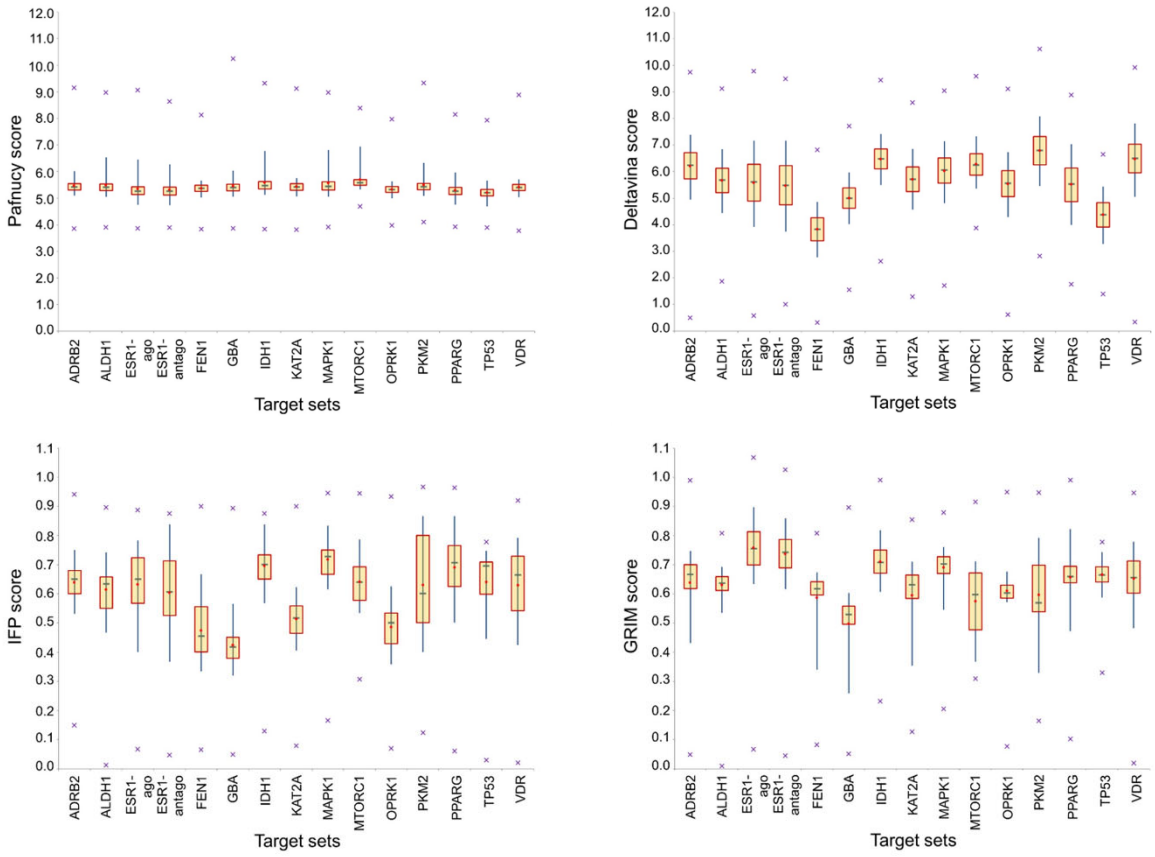

Figure 4. Distribution of Pafnucy, $\Delta_{vina}RF_{20}$, IFP, and GRIM scores on 15 target sets of the LIT-PCBA data collection using the "max-pooling" approach. The boxes delimit the 1st and the 3rd quartiles. The whiskers delimit the 5th and the 95th percentiles. The crosses indicate the minimal and the maximal scores. The median and the mean scores are indicated by a green vertical line and a red dot located in each box, respectively.

Analogously to our previous observation on ML SFs,[16] Pafnucy-predicted affinities are almost target-independent, clearly distributed within a very tiny range (1.5 pK_d unit), with almost 90% of predictions lying between pK_d values of 5 and 6.5, and no predicted value below 3.7 (**Figure 4**). Conversely, affinity values predicted by the other three SFs are much more widely spread, with a mean value that is clearly

target-dependent, and both low minimal and high maximal values (**Figure 4**). The above-described Pafnucy behavior has also been observed for in-house developed DNNs on the PDBbind dataset (data not shown), remains highly suspicious and likely to indicate overtraining and a very tiny applicability domain outside the training space. We therefore strongly advise considering the width of predicted affinity distributions when it comes to benchmarking novel SFs.

Retrieving highly potent compounds. We next checked the ability of the SFs studied herein to predict highly potent binders by plotting the predicted scores versus the experimental affinities of 93 highly potent LIT-PCBA ligands (pK_d > 7.0) spread over all targets. Hence, the above-reported tendency of Pafnucy to predict affinities within a very narrow range can still be acceptable on the condition that high scores are assigned to highly potent ligands that any VS method must prioritize for experimental validation. Both ML SFs are disappointingly unable to prioritize these highly potent ligands (**Figure 5**). Considering a predicted affinity threshold of 7.0, only eight and 13 out of 93 strong binders would have been retrieved by Pafnucy and $\Delta_{vina}RF_{20,}$ respectively. No correlation could be observed between the predicted and experimental scores for these two SFs (**Figure 5**). Conversely, a still imperfect but clear trend is observed for the two interaction-based SFs to correlate predictions with experiments (Pearson's r of 0.39 and 0.43 for IFP and GRIM rescoring, respectively; **Figure 5**). Applying cut-off values (0.70 for IFP and GRIM similarity scores) higher than those recommended (0.60 for IFP,[33, 37-40] 0.65 for GRIM[41-43]), a large majority of strong binders (60 and 61 after IFP and GRIM rescoring, respectively) would have been retrieved for experimental conformation. We acknowledge that the numbers of false positives at these cut-off values remain very high (97,272 and 44,858 for IFP and GRIM rescoring, respectively) out of a total of 2.8 million inactive compounds. The above reported trend is independent of the affinity threshold used to select potent ligands. Hence, retrieving all submicromolar ligands (pK_d > 6.0) leads to the same conclusion. The observed correlation, although less clear, is the same: IFP and

GRIM similarity scores still correlate better with experimental affinities than those predicted by Pafnucy and DeltaVina (**Figure S3).**

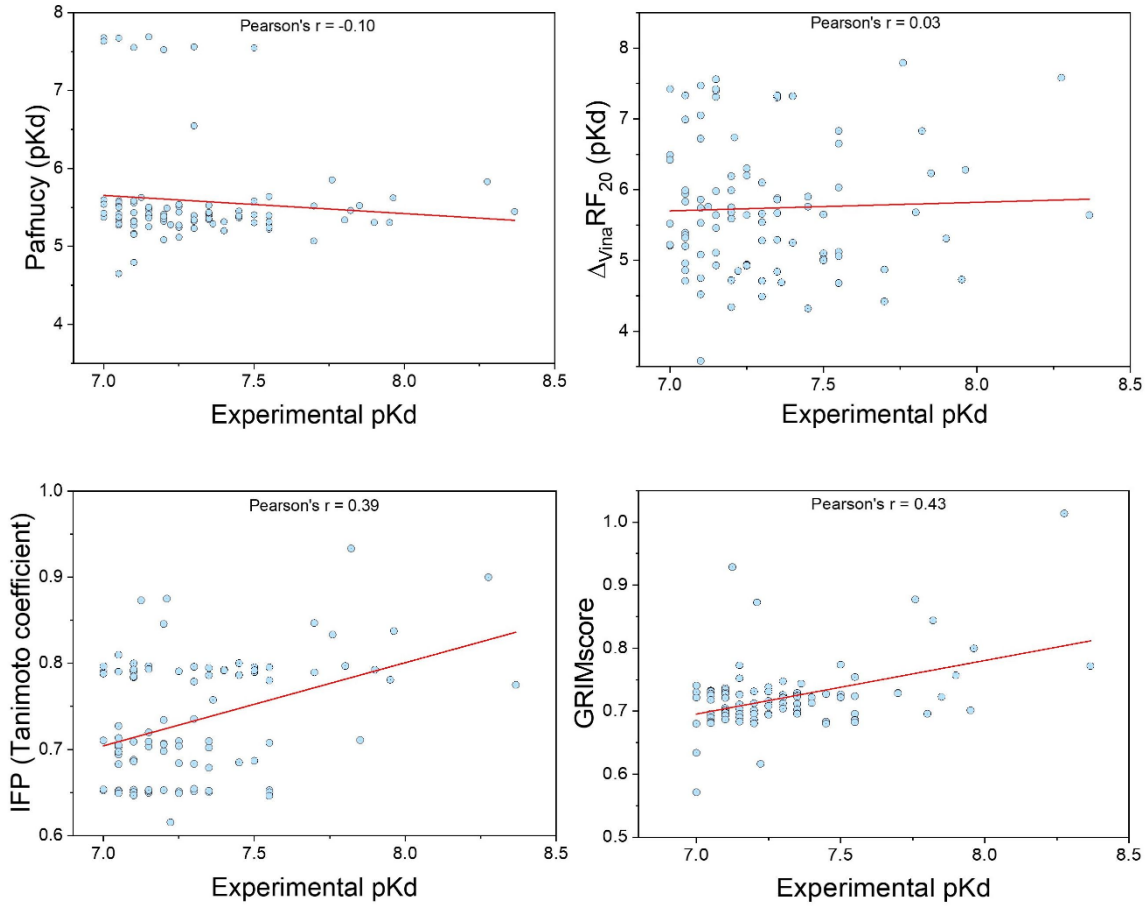

Figure 5. Predicted scores vs. experimental affinities for a subset of 93 highly potent LIT-PCBA ligands ($pK_d > 7.0$)

CONCLUSIONS

The previous development of a dataset gathering 3 million high-confidence affinity data points for 15 targets offered an opportunity to compare state-of-the-art scoring functions with diverse physicochemical backgrounds in a realistic structure-based virtual screening exercise. Although all rescoring schemes proved to rescue true binders omitted by top-ranked docking poses, simplistic knowledge-based scoring functions measuring the similarity of protein-ligand interactions to those of PDB templates appear to systematically outperform two modern machine learning methods in enriching a small-sized hit list in true binders, as well as prioritizing the most potent actives.

Of course, this conclusion only applies to the herein investigated rescoring methods and experimental dataset. Given that both machine learning methods are true representatives of the current state-of-the-art techniques, and are considered among the best existing algorithmic solutions to predict binding free energies,[31, 29, 25] we believe that the above conclusions are likely to apply to all scoring functions with similar physicochemical principles. Importantly, the current study highlights three basic but important rules often neglected before structure-based virtual screening methods are developed or applied. First, the accuracy of any scoring function in predicting known experimental affinities is not indicative of its virtual screening power. Second, assessing the real virtual screening accuracy of a scoring function requires a careful and unbiased examination on a test set mimicking true compound screening collections and not artificially assembled ligands/decoy sets. Last, rescoring virtual hits by mimicry to existing binding modes (already depicted in experimentally-determined protein-ligand structures) is a simple, robust and efficient approach to secure the chance of identifying potent binders.

ACKNOWLEDGMENTS

The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) is acknowledged for the allocation of computing time and excellent support. We sincerely thank Prof. M. Rarey (University of Hamburg, Germany) for providing an executable version of Protoss.

FUNDING

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

Supporting Information.

Enrichment factor in true positives at a 1% false positive rate (EF1%) obtained by the native Surflex-dock scoring on the docking poses issued by Surflex-Dock across all 15 target sets of the LIT-PCBA data collection; Enrichment factor in true positives at a 1% false positive rate (EF1%) obtained by Pafnucy rescoring on the docking poses issued by Surflex-Dock across all 15 target sets of the LIT-PCBA data collection; Enrichment factor in true positives at a 1% false positive rate (EF1%) obtained by $\Delta_{Vina}RF_{20}$ rescoring on the docking poses issued by Surflex-Dock across all 15 target sets of the LIT-PCBA data collection; Enrichment factor in true positives at a 1% false positive rate (EF1%) obtained by IFP rescoring on the docking poses issued by Surflex-Dock across all 15 target sets of the LIT-PCBA data collection; Enrichment factor in true positives at a 1% false positive rate (EF1%) obtained by GRIM rescoring on the docking poses issued by Surflex-Dock across all 15 target sets of the LIT-PCBA data collection. Root-mean-square deviations (in Å, heavy atoms only) of Surflex-Dock docking poses to X-ray structures of the 129 LIT-PCBA templates. Enrichment curves in true positives by docking-based virtual screening of LIT-PCBA ligands to their 15 targets. Number of unique true actives (including

ALDH1 inhibitors) retrieved among the top 1% scored compounds by individual scoring functions or all possible scoring combinations. Predicted scores vs. experimental affinities for a subset of 288 LIT-PCBA ligands ($pK_d > 6.0$).

This material is available free of charge via the Internet at http://pubs.acs.org.

Data and Software Availability

Data.

The LIT-PCBA dataset is available at http://drugdesign.unistra.fr/LIT-PCBA. For each of the 15 targets, at least four input files are provided: the protein PDB X-ray structure (MOL2 file format), the bound PDB ligand (MOL2 file format), a list of true actives (SMILES string), and a list of true inactive compounds (SMILES string). All PubChem BioAssay activity data (substance identifier SID, compound identifier CID, EC50/IC50, pEC50/pIC50) are available as an Excel spreadsheet at http://drugdesign.unistra.fr/LIT-PCBA/Files/LIT-PCBA_bioactivities.xlsx. In case several PDB templates are available, several protein and ligand input files are given.

Software.

An academic license for Surflex-Dock version 3.3066 was obtained from Biopharmics LLC (https://www.biopharmics.com/downloads/). Default settings were used to generate protomol files from protein-bound ligand coordinates, and to dock LIT-PCBA compounds with the exception of the "–pgeom" parameter used to refine docking poses.

Pafnucy version 1.0 was downloaded from https://gitlab.com/cheminfIBB/pafnucy, and used with default settings. Rescoring was performed using the recommended model batch5-2017-06-05T07:58:47-best.

DeltaVinaRF20 (no version specified) was downloaded from https://github.com/chengwang88/deltavina and used with default parameters.

IChem (version 5.2.9) was downloaded from http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html, and used with default settings of IFP and GRIM rescoring.

REFERENCES

1. Mobley, D. L.; Gilson, M. K., Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.,* **2017,** *46*, 531-558.
2. Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J., Ultra-Large Library Docking for Discovering New Chemotypes. *Nature,* **2019,** *566*, 224-229.
3. Gorgulla, C.; Boeszoermenyi, A.; Wang, Z. F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H., An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature,* **2020,** *580*, 663-668.
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*.
5. 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*, 269-288.
6. Pagadala, N. S.; Syed, K.; Tuszynski, J., Software for Molecular Docking: A Review. *Biophys. Rev.,* **2017,** *9*, 91-102.
7. Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, O.; McCammon, J. A.; Miao, Y.; Smith, J. C., Ensemble Docking in Drug Discovery. *Biophys. J.,* **2018,** *114*, 2271-2278.
8. Pinzi, L.; Rastelli, G., Molecular Docking: Shifting Paradigms in Drug Discovery. *Int. J. Mol. Sci.,* **2019,** *20*.
9. Torres, P. H. M.; Sodero, A. C. R.; Jofily, P.; Silva-Jr, F. P., Key Topics in Molecular Docking for Drug Design. *Int. J. Mol. Sci.,* **2019,** *20*.
10. Wagner, J. R.; Churas, C. P.; Liu, S.; Swift, R. V.; Chiu, M.; Shao, C.; Feher, V. A.; Burley, S. K.; Gilson, M. K.; Amaro, R. E., Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *Structure,* **2019,** *27*, 1326-1335 e1324.

11. Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R., Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.,* **2015,** *137*, 2695-2703.

12. Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E., Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.,* **2018,** *9*, 1089.

13. Rognan, D.; Desaphy, J., Molecular Interaction Fingerprints. In *Scaffold Hopping in Medicinal Chemistry*, Brown, N., Ed. Wiley-VCH Verlag GmbH & Co KGaA: 2013; pp 215-230.

14. Li, H. J.; Sze, K. H.; Lu, G.; Ballester, P. J., Machine-Learning Scoring Functions for Structure-Based Drug Lead Optimization. *Wires Comput. Mol. Sci.,* **2020,** *10*.

15. 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.

16. Gabel, J.; Desaphy, J.; Rognan, D., Beware of Machine Learning-Based Scoring Functions-on the Danger of Developing Black Boxes. *J. Chem. Inf. Model.,* **2014,** *54*, 2807-2815.

17. Khamis, M. A.; Gomaa, W., Comparative Assessment of Machine-Learning Scoring Functions on PDBbind 2013. *Eng. Appl. Artif. Intel.,* **2015,** *45*, 136-151.

18. Shen, C.; Hu, Y.; Wang, Z.; Zhang, X.; Pang, J.; Wang, G.; Zhong, H.; Xu, L.; Cao, D.; Hou, T., Beware of the Generic Machine Learning-Based Scoring Functions in Structure-Based Virtual Screening. *Brief. Bioinform.,* **2020**.

19. Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R., Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.,* **2009,** *49*, 1079-1093.

20. Dunbar, J. B., Jr.; Smith, R. D.; Yang, C. Y.; Ung, P. M.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A., CSAR Benchmark Exercise of 2010: Selection of the Protein-Ligand Complexes. *J. Chem. Inf. Model.,* **2011,** *51*, 2036-2046.

21. Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B., Jr.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K., D3R Grand Challenge 2015: Evaluation of Protein-Ligand Pose and Affinity Predictions. *J. Comput Aided Mol. Des.,* **2016,** *30*, 651-668.

22. Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E., D3R Grand Challenge 2: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided Mol. Des.,* **2018,** *32*, 1-20.

23. Gaieb, Z.; Parks, C. D.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Lambert, M. H.; Nevins, N.; Bembenek, S. D.; Ameriks, M. K.; Mirzadegan, T.; Burley, S. K.; Amaro, R. E.; Gilson, M. K., D3R Grand Challenge 3: Blind Prediction of Protein-Ligand Poses and Affinity Rankings. *J. Comput. Aided Mol. Des.,* **2019,** *33*, 1-18.

24. Parks, C. D.; Gaieb, Z.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Jansen, J. M.; McGaughey, G.; Lewis, R. A.; Bembenek, S. D.; Ameriks, M. K.; Mirzadegan, T.; Burley, S. K.; Amaro, R. E.; Gilson, M. K., D3R Grand Challenge 4: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided Mol. Des.,* **2020,** *34*, 99-119.

25. Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R., Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.,* **2019,** *59*, 895-913.

26. Tran-Nguyen, V. K.; Jacquemard, C.; Rognan, D., LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.,* **2020,** *60*, 4263-4273.

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

28. 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.

29. Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P., Development and Evaluation of a Deep Learning Model for Protein-Ligand Binding Affinity Prediction. *Bioinformatics,* **2018,** *34*, 3666-3674.

30. Rognan, D., Modeling Protein-Ligand Interactions: Are We Ready for Deep Learning? In *Systems Medicine: Integrative, Qualitative and Computational Approaches.*, Wolkenhauer, O., Ed. Elsevier: 2019; Vol. 2, pp 163-173.

31. Wang, C.; Zhang, Y. K., Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.,* **2017,** *38*, 169-177.

32. Trott, O.; Olson, A. J., Software News and Update Autodock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.,* **2010,** *31*, 455-461.

33. Marcou, G.; Rognan, D., Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.,* **2007,** *47*, 195-207.

34. de Graaf, C.; Rognan, D., Selective Structure-Based Virtual Screening for Full and Partial Agonists of the Beta2 Adrenergic Receptor. *J. Med. Chem.,* **2008,** *51*, 4978-4985.

35. Venhorst, J.; Núñez, S.; Terpstra, J. W.; Kruse, C. G., Assessment of Scaffold Hopping Efficiency by Use of Molecular Interaction Fingerprints. *J. Med. Chem.,* **2008,** *51*, 3222-3229.

36. Chalopin, M.; Tesse, A.; Martínez, M. C.; Rognan, D.; Arnal, J. F.; Andriantsitohaina, R., Estrogen Receptor Alpha as a Key Target of Red Wine Polyphenols Action on the Endothelium. *PloS one,* **2010,** *5*, e8554.

37. de Graaf, C.; Kooistra, A. J.; Vischer, H. F.; Katritch, V.; Kuijer, M.; Shiroishi, M.; Iwata, S.; Shimamura, T.; Stevens, R. C.; de Esch, I. J.; Leurs, R., Crystal Structure-Based Virtual Screening for Fragment-Like Ligands of the Human Histamine H(1) Receptor. *J. Med. Chem.,* **2011,** *54*, 8195-8206.

38. de Graaf, C.; Rein, C.; Piwnica, D.; Giordanetto, F.; Rognan, D., Structure-Based Discovery of Allosteric Modulators of Two Related Class B G-Protein-Coupled Receptors. *ChemMedChem,* **2011,** *6*, 2159-2169.

39. Pallandre, J. R.; Borg, C.; Rognan, D.; Boibessot, T.; Luzet, V.; Yesylevskyy, S.; Ramseyer, C.; Pudlo, M., Novel Aminotetrazole Derivatives as Selective Stat3 Non-Peptide Inhibitors. *Eur. J. Med. Chem.,* **2015,** *103*, 163-174.

40. Rivat, C.; Sar, C.; Mechaly, I.; Leyris, J. P.; Diouloufet, L.; Sonrier, C.; Philipson, Y.; Lucas, O.; Mallie, S.; Jouvenel, A.; Tassou, A.; Haton, H.; Venteo, S.; Pin, J. P.; Trinquet, E.; Charrier-Savournin, F.; Mezghrani, A.; Joly, W.; Mion, J.; Schmitt, M.; Pattyn, A.; Marmigere, F.; Sokoloff, P.; Carroll, P.; Rognan, D.; Valmier, J., Inhibition of Neuronal Flt3 Receptor Tyrosine Kinase Alleviates Peripheral Neuropathic Pain in Mice. *Nat. Commun.,* **2018,** *9*, 1042.

41. 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.

42. Slynko, I.; Da Silva, F.; Bret, G.; Rognan, D., Docking Pose Selection by Interaction Pattern Graph Similarity: Application to the D3R Grand Challenge 2015. *J. Comput. Aided Mol. Des.,* **2016,** *30*, 669-683.

43. da Silva Figueiredo Celestino Gomes, P.; Da Silva, F.; Bret, G.; Rognan, D., Ranking Docking Poses by Graph Matching of Protein-Ligand Interactions: Lessons Learned from the D3R Grand Challenge 2. *J Comput Aided Mol. Des.,* **2018,** *32*, 75-87.

44. 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.

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

46. 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.

47. Cleves, A. E.; Jain, A. N., Structure- and Ligand-Based Virtual Screening on DUD-E(+): Performance Dependence on Approximations to the Binding Pocket. *J. Chem. Inf. Model,* **2020,** *60*, 4296-4310.

48. 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.

49. Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R., Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.,* **2017,** *57*, 942-957.

50. Brown, B. P.; Mendenhall, J.; Geanes, A. R.; Meiler, J., General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps. *J. Chem. Inf. Model.,* **2021,** *61*, 603-620.

51. Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R., PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics,* **2015,** *31*, 405-412.

For Table of Contents use only

True accuracy of fast scoring functions to predict high-throughput screening data from docking poses:

The simpler the better.

Viet-Khoa Tran-Nguyen, Guillaume Bret and Didier Rognan