The impact of compound library size on the performance of scoring functions for structure-based virtual screening

Louison Fresnais[1] and Pedro J. Ballester[1*]

[1] Centre de Recherche en Cancérologie de Marseille (CRCM), Inserm, U1068, Marseille, F-13009, France. CNRS, UMR7258, Marseille, F-13009, France. Institut Paoli-Calmettes, Marseille, F-13009, France. Aix-Marseille University, UM 105, F-13284, Marseille, France.

Abstract

Motivation: Larger training datasets have been shown to improve the accuracy of Machine Learning (ML)-based Scoring functions (SFs) for Structure-Based Virtual Screening (SBVS). In addition, massive test sets for SBVS, known as ultra-large compound libraries, have been demonstrated to enable the fast discovery of selective drug leads with at least nanomolar potency. This proof-of-concept was carried out on two targets using a single docking tool along with its SF. It is thus unclear whether this high level of performance would generalise to other targets, docking tools and SFs.

Results: We found that screening a larger compound library results in more potent actives being identified in all six additional targets using a different docking tool along with its classical SF. Furthermore, we established that a way to improve the potency of the retrieved molecules further is to rank them with more accurate ML-based SFs (we found this true in four of the six targets). A three-fold increase in average hit rate across targets was also achieved by the ML-based SFs. Lastly, we observed that classical and ML-based SFs often find different actives, which supports using both types of SFs on those targets.

Contact: pedro.ballester@inserm.fr

Supplementary information: an online-only supplementary results file is enclosed.

Introduction

The primary goal of Virtual Screening (VS)(Schneider, 2010; Li, K.-S. Leung, *et al.*, 2016; Singh *et al.*, 2020) is to retrieve a small subset of molecules with the highest possible proportion of actives in the screened library. When a 3D structure of the protein target is available and the binding site is known, this problem is more specifically called Structure-Based Virtual Screening (SBVS). An active is here a small organic molecule modulating the

molecular function of a protein via a non-covalent bond between both molecules. Docking is commonly used for SBVS(dos Santos *et al.*, 2018). The advantages of SBVS include the possibility of discovering a high proportion of actives with novel chemical scaffolds in a fast and cost-effective manner(Lagarde *et al.*, 2018; Ballester *et al.*, 2012; Elmessaoudi-Idrissi *et al.*, 2018).

Classical SFs can be divided into three groups: force-field, knowledge-based and empirical. These SFs assume a linear functional form between structure-derived features describing the protein-ligand complex and its binding affinity(Ain *et al.*, 2015). Approaches based on Machine Learning (ML) do not assume any predetermined form, which is learnt instead from training data. ML-based SFs are thus capable of implicitly capturing intermolecular binding interactions that are hard to model explicitly(Ballester and Mitchell, 2010). On popular benchmarks like CASF-2007(Cheng *et al.*, 2009), Random Forest (RF)-based RF-Score v3's correlation with measured affinity was R_p=0.803 five years ago(Li *et al.*, 2015), whereas RF-Score v3 supplemented with ligand features now reaches R_p=0.836(Boyles *et al.*, 2019). The latter is the best accuracy so far on this benchmark relevant for Structure-Based Lead Optimisation (SBLO). By contrast, 16 classical SFs obtained a lower R_p ranging from 0.216 to 0.644 on the same test set 10 years ago. No classical SF has managed to improve these results since then. Introductions to ML-based SFs can be found in these publications (Ain *et al.*, 2015; Yang *et al.*, 2019; Shen *et al.*, 2019; Li *et al.*, 2020; Bitencourt-Ferreira *et al.*, 2019; Wójcikowski *et al.*, 2019).

Applied to SBVS, SFs must be able to predict high affinity for ligands (actives) and low affinity for non-ligands (inactives). SFs for binding affinity prediction are suboptimal for SBVS due to being exclusively trained on crystal structures of complexes, as data instances do not contain much information about docking pose errors or the chemical diversity of inactives. Thus, a Random Forest (RF)-based SF for SBVS, called RF-Score-VS (Wójcikowski *et al.*, 2017), was trained on synthetic training data (i.e. ligands along with many more assumed non-ligands docked to the target). RF-Score-VS obtained a 3.2-times higher average hit rate than the classical SF DOCK3.6(Coleman *et al.*, 2013) at the top 1% of ranked DUD-E molecules(Mysinger *et al.*, 2012). This level of performance has been independently confirmed, and also surpassed by another RF-based SF called SIEVE-Score(Yasuo and Sekijima, 2019). Being a target-specific SF, SIEVE-Score is built per target in order to exploit its specific binding site characteristics. In contrast, RF-Score-VS

was trained on complexes with a range of targets(Wójcikowski *et al.*, 2017). Such generic SF can hence be applied to molecules complexed with any of these targets without retraining.

Unlike classical SFs, ML-based SFs provide more accurate binding affinity prediction owing to larger training sets(Li *et al.*, 2019). The enrichment ability of ML-based SFs for SBVS also improve with more training instances(Li *et al.*, 2011), with the largest training sets to date approaching one million instances(Wójcikowski *et al.*, 2017). Big data approaches are also being used to construct massive test sets (Hoffmann and Gastreich, 2019). A recent study (Lyu *et al.*, 2019) demonstrated the benefits of docking with DOCK3.7 such ultra-large libraries of purchasable molecules on two targets (99 and 138 million against AmpC β-lactamase and D_4 dopamine receptor, respectively). For D_4 dopamine target, these authors showed that novel drug leads with at least low-nanomolar potency and the sought selectivity over off-targets can be discovered in a couple of weeks by simply docking libraries of unprecedented size and structural diversity. By contrast, such classical SFs lead to much weaker compounds of moderate chemical diversity on much smaller libraries, e.g. 1.7 to 100 μM from docking 235,000 molecules(Doman *et al.*, 2002) or 1 to 70 μM from docking 180,000 molecules(Park *et al.*, 2011). On the second target, the employed classical SF (DOCK3.7) performed much worse: only 5 of the 44 SBVS hits that were tested *in vitro* had any AmpC activity (the best being a modest IC_{50} of 10.28 μM). Despite this low hit rate and owing to their novelty in terms of chemical structure, the optimisation of these AmpC actives led to a novel AmpC drug lead with a potent IC_{50} of 270 nM.

In this paper, we will investigate whether larger test sets also result in higher potencies of the retrieved molecules in six additional targets. This will shed light into the generality of the study(Lyu *et al.*, 2019). We will also assess whether these results can also be reached using another freely-available classical SF. Lastly, for the same library size, we will investigate whether the retrieved potencies and hit rates of the classical SF can be improved by employing ML-based SFs instead.

Materials and methods

The study requires the use of two different benchmarks. These previously assembled datasets will be used for testing the SFs across targets on ligand libraries of increasing size.

For ML-based SFs, one benchmark will be used exclusively for training and the other for testing them. The latter requires identifying which targets are common to both benchmarks, removing from the test set of each target those molecules that are in the training set of that target and gathering the potency measurements of active molecules. This section also explains how molecules are docked and re-scored, along with how the resulting predictions are assessed. All experiments were carried out on a workstation with 20 dual-thread CPU cores (Intel® Xeon(R) CPU E5-2660 v3 @ 2.60GHz) and 62.8 GB of RAM memory.

Identifying targets in common between DUD-E and DEKOIS2.0 benchmarks

Evaluating SFs on the DUD-E benchmark alone is known to result in overestimation of their performance due to benchmark biases (Chaput *et al.*, 2016). Here we take several actions to alleviate this problem. First, we reduce the process of tailoring the SF to the benchmark to a minimum (e.g. we do not search for the preprocessing steps that improve performance on the test set such as ligand and target preparation options). Second, we do not tune the hyperparameters of the ML algorithm to training data in the case of ML-based SFs (we used RF with the default values of the scikit-learn implementation except for 500 as the number of trees in the forest and 100 as the number of features to consider when looking for the best split at each node). Third, as in previous studies (Wójcikowski *et al.*, 2017; Yasuo and Sekijima, 2019; Chen *et al.*, 2019), ML-based SFs are trained on data from the DUD-E benchmark and tested on a second benchmark: DEKOIS2.0(Bauer *et al.*, 2013). In this way, the exploitation of DUD-E bias by the ML algorithm is avoided, as DEKOIS2.0 actives and decoys were selected in a different manner than those in DUD-E.

Using DUD-E data as the training set and DEKOIS2.0 as test set per-target requires identifying which targets are in common between both benchmarks. This is not trivial as the same target can have different names depending on the benchmark. Four targets were matched by PDB ID, as they use the same PDB structure on both benchmarks. DUD-E targets (ppara, aa2ar, parp1, and hdac2) use the same PDB IDs (2p54, 3eml, 3l3m, and 3max) as DEKOIS2.0 targets (PPARa, A2A, PARP1, and HDAC2). 27 targets, including the four targets having identical PDB ID, were matched by having 100% sequence identity (these common targets and their respective PDB IDs are shown in Table S4). To do so, we downloaded the list of all PDB IDs in the PDB database clustered with a 100% sequence identity threshold (bc-100.out) from the RCSB PDB FTP server (ftp://resources.rcsb.org/sequence/clusters/). Then we parsed this file to determine which pairs of DUD-E and DEKOIS2.0 PDB ID are in the same cluster. After matching the targets

from the two benchmarks, we employed the PDB structure of each benchmark both to make it harder for docking and to account for the fact that proteins with 100% sequence identity can have different binding pocket shapes due to hosting different ligands.

Removing DEKOIS2.0 molecules that are also in DUD-E

A requirement for the evaluation of any SF is to mimic realistic scenarios as much as practically possible. In particular, a protein-ligand complex in the training set cannot also be in the test set. Otherwise, we would not be estimating the generalisation ability of the SF to other complexes, but how well it memorises training data. Thus, for each of the 27 targets in common, we removed those molecules in the training set (DUD-E) that were also in the test set (DEKOIS2.0). In particular, any test set molecule with a Tanimoto score on Morgan fingerprints greater or equal to 0.99 to any training set molecule was removed, regardless of whether molecules are actives or decoys.

Retrieving measured potencies of DUD-E actives from ChEMBLdataset

To evaluate how well the SFs predict the binding affinities of docked molecules, we need to consider the measured potencies for each active molecule. This is almost never done, as SFs on DUD-E and DEKOIS2.0 are evaluated as binary classifiers only. For each DEKOIS2.0 and DUD-E target, measured potencies are provided for each active (as annotations of the SDF file in DEKOIS2.0 and in the actives_nM_combined.ism file in DUD-E). DUD-E actives were selected from the ChEMBL database (Gaulton *et al.*, 2017). As this database is regularly updated with new bioactivity determinations, we retrieved all the bioactivities of each DUD-E active from ChEMBL to employ more robust estimates than simply the single value provided with DUD-E. As done before (Wójcikowski *et al.*, 2017), we retrieve all IC50s, Kd and Ki measurements for the considered ligand-target association and retained the median as its active concentration. Potency is quantified as the negative-log-transformed active concentration in molar units of the drug when binding to the target. For example, if the active concentration is 1 μM, then the corresponding potency is 6. Likewise, 1 nM active concentration corresponds to a potency of 9.

For some molecules, the difference between several measured potency values is substantial (Kruger and Overington, 2012). To avoid this, we removed those molecules with such unreliable values for each target. More concretely, if the difference between the highest and lowest measured potencies for a molecule-target pair is more than two orders of magnitude, we do not consider that pair (i.e. the molecule is not considered neither as an active nor an

inactive/decoy for that target). We found (Table S5) that shows very few actives had to be removed in the end for this reason.

Retaining targets with sufficiently large test sets

As each DEKOIS2.0 target only has 40 actives, we did not retain a target if more than 10% of its actives were removed by this chemical similarity filter. Out of the 27 targets in common, only 8 were thus retained due to having at least 37 actives: ACES (Acetylcholinesterase), ADRB2 (Beta-2 adrenergic receptor), EGFR (Epidermal growth factor receptor erbB1), HMDH (HMG-CoA reductase), KIF11 (Kinesin-like protein 1), PPARA (Peroxisome proliferator-activated receptor alpha), PPARG (Peroxisome proliferator-activated receptor gamma) and RXRA (Retinoid X receptor alpha). From this list, we further discarded EGFR and KIF11 because of having multiple molecules with activity worse than 30 μM annotated as actives (these molecules would be decoys according to DUD-E and we need to have the same threshold definition). We will hence be analysing here members of the following target classes: enzyme (ACES, HMDH), G protein-coupled receptor (ADRB2) and nuclear receptor (PPARA, PPARG, RXRA). Note that the acronyms of two targets in DUD-E are different from those in DEKOIS2.0: ACHE instead of ACES and HMGR instead of HMDH. Further details of these targets can be found in the benchmark websites (http://dude.docking.org/targets, http://www.dekois.com/).

Docking compound libraries against the six targets using SMINA

Ligand preparation, target preparation and SMINA(Koes *et al.*, 2013) run settings are specified in Table S1 (DUD-E) and Table S2 (DEKOIS2.0). The binding pocket location and extension to dock DUD-E molecules is defined by SMINA with the autobox-ligand parameter using the bound ligand provided with the 3D structure of each target. To dock DEKOIS2.0 molecules, we used the same preparation procedure except that the ligand to specify the binding site of the target is not specified in this benchmark. Thus, we used Fpocket (Le Guilloux *et al.*, 2009) to rank all possible binding pockets found on the structure of the target and select the most likely pocket along with its extracted coordinates required to specify the search space for SMINA when docking DEKOIS2.0 molecules. We visually verified with Chimera (Pettersen *et al.*, 2004) that most likely pocket returned by Fpocket is the same as that used to dock DUD-E molecules in all six targets. While we could have used a superposition with DUD-E ligand-bound PDB structures to manually set the search space of DEKOIS2.0 PDB structures, we wanted to assess how Fpocket could

perform in the absence of this information and all six binding site predictions of this method were found to be correct.

Stratified sampling

To study the impact of compound library size on SBVS, we consider the benchmark with the highest number of actives (DUD-E) and took a stratified sample of size 25% of the target's actives as follows. Actives were sorted by decreasing potency and every fourth active down this ranked list was retained for the sample (for each of the selected actives, its decoys were retained too). There are four ways to implement this sampling protocol: taking the first sample from either the first, second, third or fourth most potent actives. We will consider all four 25% stratified datasets per target. In addition, we took a second sample per target of size 50% of its actives in an analogous manner.

Processed datasets

First, we have three increasingly larger sets of molecules or libraries per target: 25%DUD-E (25% of the DUD-E molecules by stratified sampling), 50%DUD-E (50% of the DUD-E molecules by stratified sampling) and DUD-E (the complete DUD-E dataset). Each of these three libraries has been docked into their corresponding target, thus giving rise to 18 sets of docking poses, each pose with its SMINA score. Second, we have prepared a fourth library per target, DEKOIS2.0, where any DEKOIS2.0 molecule in common with those in DUD-E has been removed, as previously explained.

Rescoring with RF-Score-VS

The second version (v2) of RF-Score-VS is currently one of the best performing SFs for SBVS (Wójcikowski *et al.*, 2017) and is freely available (https://github.com/oddt/rfscorevs_binary). As this ML-based SF was trained on the entire DUD-E dataset (Wójcikowski *et al.*, 2017), it can be applied to re-score molecules docked to any DUD-E target (thus, we will refer to this SF as RF-Score-VS_G, where G stands for generic). As compiled in Table S3, this SF inputs two files: one with the 3D poses of molecules docked to the target and the other with the PDB structure that was employed for docking those molecules. We also generated a target-specific version of RF-Score-VS v2 for each target (RF-Score-VS_TS, where TS stands for Target Specific) as explained in the supplementary information file. With both SFs and for each molecule-target pair, we re-scored the best pose of the molecule (i.e. that with the most negative free energy of binding as predicted by SMINA) as it is customary (Li, K.-S. K.-S. Leung, *et al.*, 2016). RF-Score-

VS v2 must not be confused with RF-Score v2 (Ballester *et al.*, 2014), which was not designed to be applied to SBVS, but to SBLO.

SBVS performance metrics

Enrichment Factor in the top 1% (EF1%) indicates how many times more active compounds were found in the top 1% of the ranked compound library than would be expected by random selection. Hit Rate in the top 1% (HR1%) is the proportion of actives found in the top 1% of the ranked library. By definition, EF1% is HR1% divided by HR100%, the latter corresponding to the proportion of actives in the full library.

The AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristics) curve displays true positive rate against false positive rate across different percentages of the ranked compound library. An AUC-ROC, or simply AUC in this context, of 0.5 corresponds to a random ordering of the library, whereas an AUC of 1 indicates that the library was ranked in an optimal manner (all actives ranked first, followed by the inactives or decoys, i.e. perfect classification). Thus, the higher AUC, the better the SF is at discriminating between actives and inactives. In this article, we used the rocker package (Lätti *et al.*, 2016) to plot ROC curves and calculate their AUC values.

Statistical hypothesis testing

As the distribution of potencies across molecules was often found to be skewed, the normality assumption does not hold. Thus, instead of using parametric tests based on this assumption, we used a suitable non-parametric test: Wilcoxon-Mann-Whitney test. This test also has the advantage of being more robust to assumptions when comparing the potencies of actives and decoys, as it operates with ranks not actual potency values. For example, the default potency assumed for decoys by DUD-E is 4.52 (exactly 30 μM), but decoys will generally have much lower true potencies.

Results and discussion

Does screening larger compound libraries result in discovering more potent actives?

To answer this question, we compare the top potencies of the actives retrieved by SMINA per target and screening library (Figure 1). In each target, the top potencies from the largest library are on average higher than those from the smallest library (the difference is significant in all targets but PPARA). The number of actives employed for the target does

not seem to influence this outcome. Considering larger compound libraries is therefore expected to result in the discovery of more potent actives for these targets.

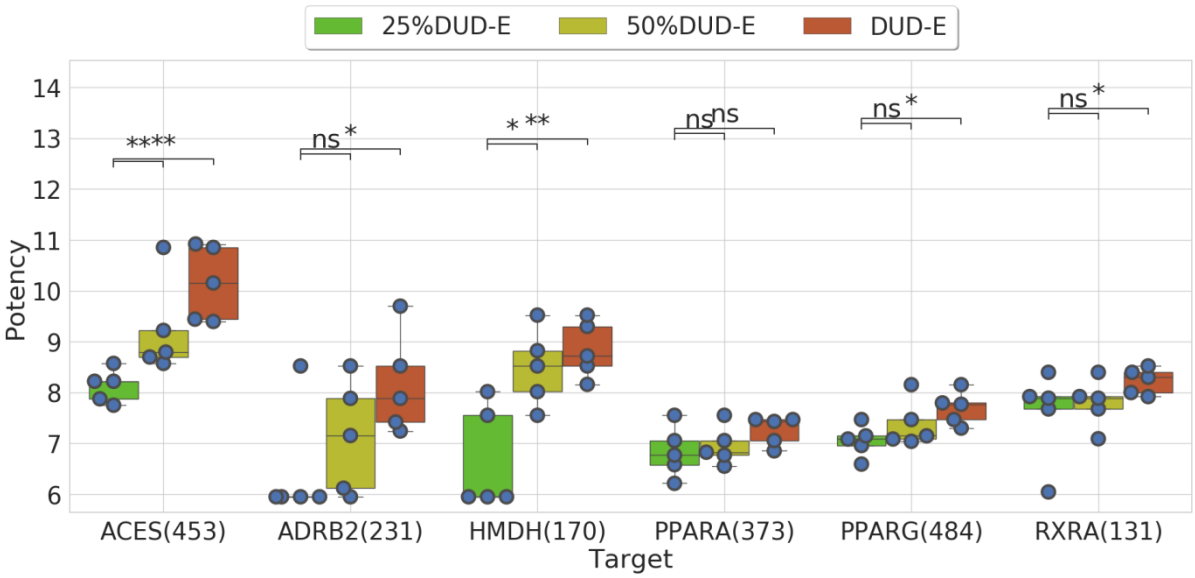

Figure 1. Measured binding affinities of the five most potent actives in the top 1% of SMINA-ranked molecules per screening library and target. The number of actives in DUD-E per target is reported between parentheses. For each target-library dataset, a boxplot with the most potent actives retrieved by SMINA is shown. Per target, we statistically test whether the potencies of the top five actives retrieved from larger libraries are higher than those retrieved from the smallest library, 25%DUD-E (one-sided Wilcoxon-Mann-Whitney test; *: 0.01<pvalue<0.05, **: 0.001<p-value<0.01, ns: p-value>0.05). For quick reference, potencies 6 and 9 correspond to 1 μM and 1 nM IC50/Kd/Ki, respectively.

Note that, in the ADRB2 target, SMINA only retrieves 1 of the 58 actives contained in the smallest library (25%DUD-E). Thus, to permit comparison to other target-library pairs, we completed the boxplot of ADRB2-25%DUD-E with four molecules with the highest activity that a decoy can have, i.e. the activity threshold, which is 30 μM in DUD-E (Mysinger *et al.*, 2012). This procedure had also to be applied to HMDH with 25% of DUD-E, but not on the other 16 target-library pairs as all retrieved at least five actives (each pair corresponds to a boxplot in Figure 1).

In Figure 1, 25%DUD-E was generated by stratified sampling of the entire screening library (DUD-E) starting at the second best-ranked active by SMINA. If we start at the first best-ranked active, the most potent active in 25%DUD-E will be the same as that in DUD-E, which is not realistic (larger libraries should contain more potent actives). In any case, we repeated the analysis with the other three possible 25%-sampled libraries and the same trend

was observed (data not shown). This is understandable since the molecule predicted to have the best potency by SMINA will generally not be the molecule with the best-measured potency due to limitations of predictions.

Does virtual screening performance vary with library size?

For each library, Figure 2 shows that HR1% fluctuates between almost 10% to over 50% across targets. Only relative performance must be considered here and not the absolute values, as it has been argued that DUD-E overestimates the performance of classical SFs in about half of its targets (Chaput *et al.*, 2016). There is small variability in HR1%, and thus in EF1%, within a target compared to some inter-target differences. Furthermore, while the largest library (DUD-E) obtains the worse performance for some targets (ADRB2, PPARA, PPARG, RXRA), it also obtains the best performance in other targets (ACES, HMDH). Therefore, the performance of SMINA does seem to vary with library size in these targets. We think that this is connected to the proportion of actives for a given target and the ability of SMINA to discriminate them remaining the same as the screened library grows.

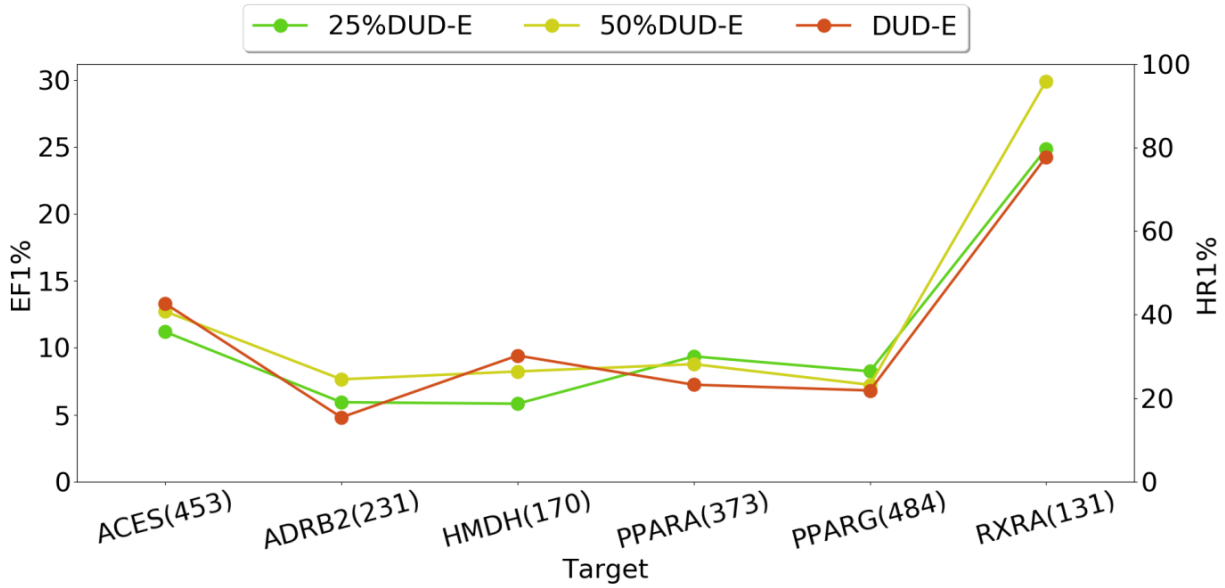

Figure 2. EF1% and HR1% of SMINA on each target-library pair.

Does re-scoring with machine-learning scoring functions improve SBVS performance?

We next consider two ML-based SFs, RF-Score-VS_G and per-target-trained RF-Score-VS_TS, to investigate whether these retrieve more potent actives than SMINA on the same library (Figure 3). Since these ML-based SFs are trained on DUD-E datasets, we can only assess their predictions on DEKOIS2.0 datasets as independent test sets. Unlike DUD-E, the number of actives and decoys is the same for each DEKOIS2.0 target. Small deviations from the provided 40 actives in Figure 3 are due to removing molecules in common with DUD-E.

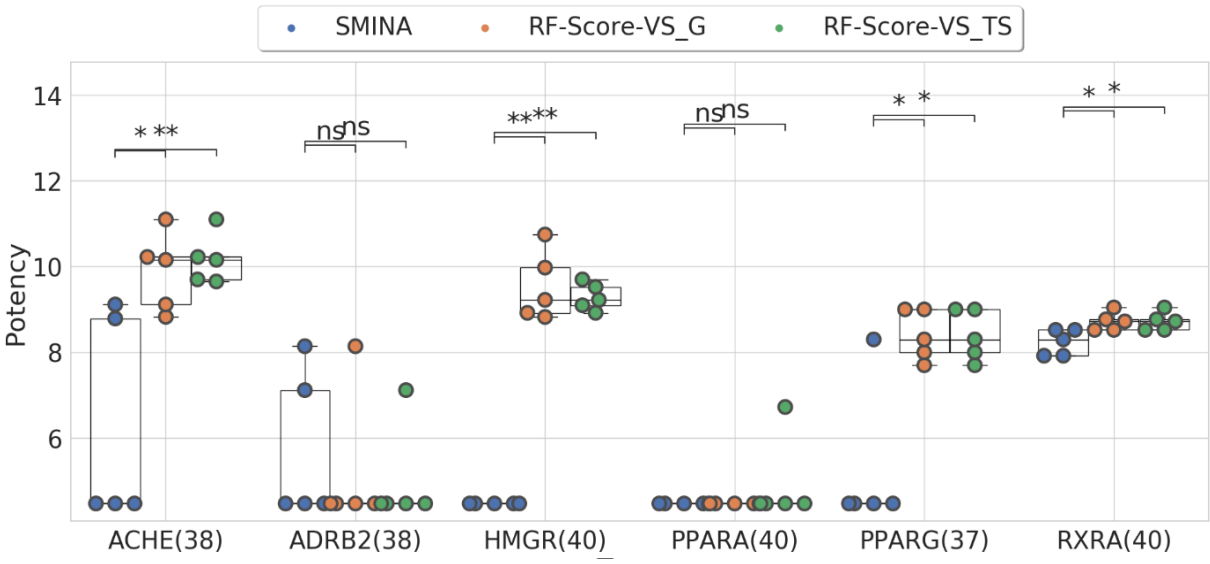

Figure 3. Measured binding affinities of the five most potent actives for each target at the top 1% of the ranked molecules, scored either with SMINA or RF-Score-VS_G or RF-Score-VS_TS. The statistical test used: Wilcoxon-Mann-Whitney, one-sided. We used the one-sided test with the hypothesis that results from the SMINA scoring function are poorer than those obtained with RF-Score-VS. *: 0.01<pvalue<0.05, **: 0.001<p-value<0.01, ns: p-value>0.05.

Figure 3 shows that the top potencies retrieved by the ML-based SFs are significantly higher than those from SMINA in four of the six targets (ACHE, HMGR, PPARG and RXRA). The different can be as high as 6 orders of magnitude (HMGR). Results on the other two targets (ADRB2, PPARA) are inconclusive. Both RF-Score-VS_G and RF-Score-VS_TS obtained similar top potencies.

Figure 4 compares the same SFs in terms of EF1% and HR1%. ML-based SFs strongly improve the performance of SMINA in half of the targets (ACHE, HMGR and PPARG), while obtaining similar performance in the rest of targets. Both RF-Score-VS_G and RF-

Score-VS_TS obtained similar EF1% and HR1% across targets. Comparing the SMINA performance on DEKOIS2.0 (Figure 4, blue) with that of the same SF in DUD-E (Figure 2, red), it is clear that DEKOIS2.0 is a more demanding benchmark than DUD-E. Indeed, whereas the EF1% averaged over the six DUD-E datasets is 10.98, the corresponding EF1% with DEKOIS2.0 is only 6.61.

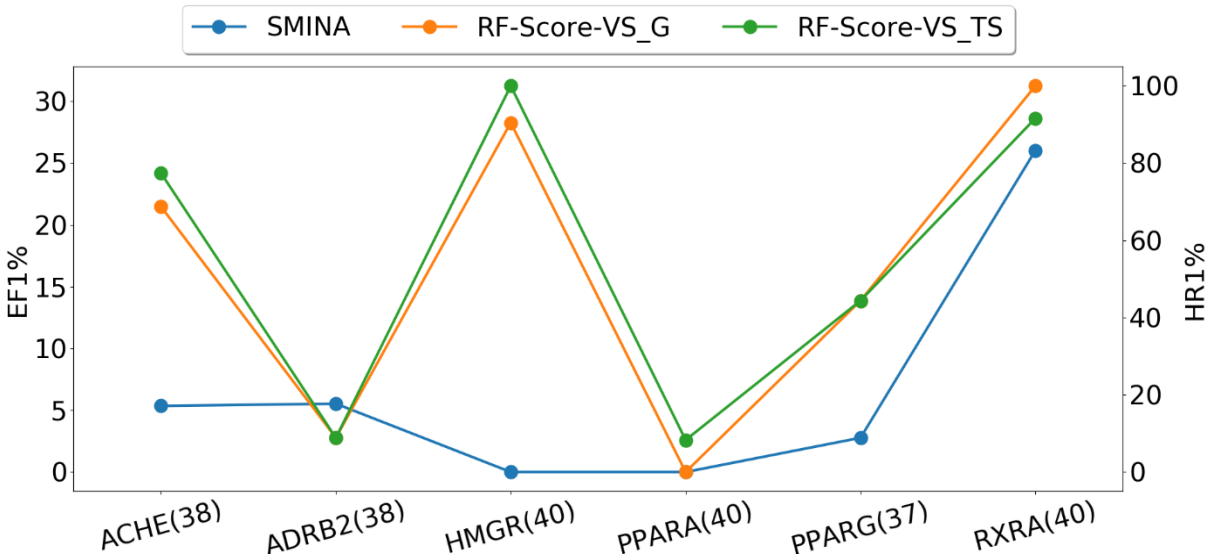

Figure 4. Docking enrichment for the top percent of ranked molecules for each target in the DEKOIS2.0 dataset at the top 1% of the ranked molecules scored either with SMINA or RF-Score-VS_G or RF-Score-VS_TS. For each target, we consider all the DEKOIS2.0 molecules after removing DUD-E-common molecules.

Table S6 compiles all the results on the DEKOIS2.0 test sets. RF-Score-VS_TS achieve the best results overall, closely followed by RF-Score-VS_G, both strongly outperforming SMINA on average. We calculated the AUC values for each case, with Figure S1 showing the corresponding ROC curves of the three SFs in each target. AUC only slightly disagrees with EF1% in one target (RXRA, where RF-Score-VS_G is slightly better than RF-Score-VS_TS in terms of EF1%, the other way around if we consider AUC instead).

Table S6 shows that EF1% averaged over these six DEKOIS2.0 targets are 6.61, 16.27 and 17.21 for SMINA, RF-Score-VS_G and RF-Score-VS_TS, respectively. That is, the EF1% of RF-Score-VS_G and RF-Score-VS_TS are 2.46 and 2.60 times higher than that of SMINA. A previous study compared the EF1% of SMINA versus that of RF-Score-VS_G over 76 DEKOIS2.0 targets obtaining 3.95 and 9.84 (Wójcikowski *et al.*, 2017). On these targets, the EF1% of RF-Score-VS_G is hence 2.49 times higher than that of SMINA. However, an EF analysis with the DEKOIS2.0 benchmark was only done for all targets aggregated, unlike

here where it is also done per target. Furthermore, RF-Score-VS_TS was not evaluated in any way in this benchmark. Moreover, the ability of a SF to retrieve the most potent DEKOIS2.0 compounds was not analysed at all. Independently, other authors (Chen *et al.*, 2019) have evaluated several classical SFs and RF-Score-VS_G over the 55 DEKOIS2.0 targets hardest for RF-Score-VS_G (i.e. those targets that are neither in the training set nor are highly similar to those in the training set). Classical SFs are DLIGAND, DLIGAND2 and Vina (Trott and Olson, 2010), which shares the same functional form with SMINA. The EF1% of RF-Score-VS_G averaged over these 55 targets is 3.15, 1.88 and 1.92 times higher than that of DLIGAND, DLIGAND2 and Vina. Since HR1% is by definition proportional to EF1%, these results also mean that the ML-based SFs are able to discover between 1.88 and 3.15 more actives than the classical SFs. All these analyses characterise the true performance of ML-based SFs, as each SF was trained on a benchmark and tested on a different benchmark.

RF-Score-VS_TS was able to find an active for PPARA, the only target were both SMINA and RF-Score-vs_G failed. This is what is expected from a target-specific SF, to exploit target-specific data so that the embedded feature selection of RF prioritizes features that are important for binding to that particular target. It is likely that constructing target-specific features will further improve performance, as seen elsewhere (Yasuo and Sekijima, 2019). Conversely, RF-Score-vs_G was slightly better than RF-Score-VS_TS on RXRA. This must be due to part of RF-Score-vs_G's multi-target training data being informative for RXRA.

Do different SFs discover different actives even if they obtain a similar hit rate?

To re-score with RF-Score-vs_G or RF-Score-VS_TS, a docked pose of the molecule against the target must be generated (here with SMINA). Therefore, one always counts per target with a ranking of molecules by SMINA in addition to those from re-scoring. It is well-known that different ligand-based chemical similarity search methods tend to identify different actives and thus applying more than one method prospectively is recommended whenever possible (Sheridan and Kearsley, 2002). Here we investigate whether this is worth in the context of these three SFs for structure-based virtual screening (Figure 5).

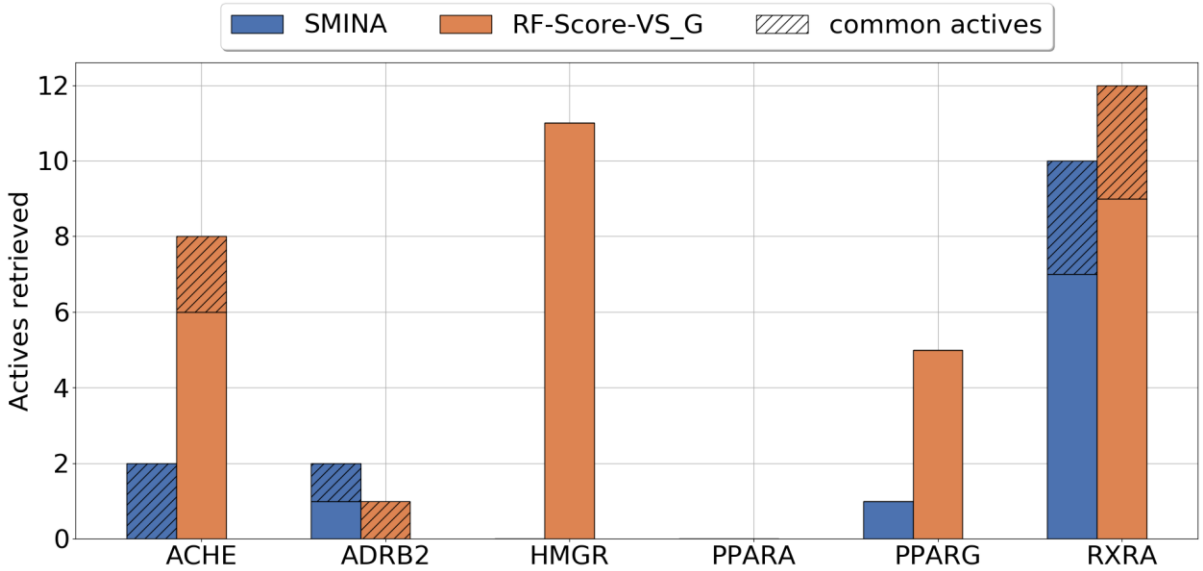

Figure 5. Number of actives found by SMINA and RF-Score-VS_G at the top 1% of the DEKOIS2.0 test sets. Cross-hatching shows the number of actives found by both SFs.

Based on these retrospective results, which SF should be used in prospectively for each target? Despite being outperformed by RF-Score-VS_G on RXRA, SMINA found seven additional actives. Therefore, both SFs should be used prospectively on this target to identify additional actives. However, SMINA did not retrieve any additional active on ACHE or any active at all on HMGR, whereas RF-Score-VS_G obtained high hit rates on these targets (Table S6). Furthermore, SMINA only retrieved one active on PPARG, four less than RF-Score-VS_G. It is thus advisable to only use RF-Score-VS_G prospectively on these three targets. Conversely, SMINA is the only SF that should be applied to ADRB2 in a prospective manner, since no further active was found by RF-Score-VS_G retrospectively. Lastly, neither of the two SFs should be helpful for discovering new PPARA actives.

Figures S2 and S3 show the other two plots comparing SFs in terms of their retrieved actives. The SMINA *versus* RF-Score-VS_TS plot tells essentially the same story as that in Figure 5. The RF-Score-VS_G *versus* RF-Score-VS_TS plot shows a much higher proportion of common actives. Still, there are some targets where both ML-based SFs find a good proportion of different actives. For instance, a total of 17 HMGR actives have been retrieved by these SFs, of which only 6 are common to both SFs.

Conclusions

We have found that screening a larger compound library results in more potent actives being identified in all six targets (ACES, ADRB2, HMDH, PPARA, PPARG and RXRA). Even

though the difference was not significant in PPARA, the upward trend suggest that this would be significant too if we compare libraries that are more than four times larger (the limit in our experiments). This improvement in SBVS did not come at the cost of obtaining a smaller hit rate, which did not vary substantially with library (test set) size. Also, we observed that DEKOIS 2.0 is a far more demanding benchmark than DUD-E for SMINA.

We have also found that a way to improve the potency of the retrieved molecules further is to rank them with more accurate ML-based SFs (true for 4 of the 6 targets: ACES, HMDH, PPARG and RXRA). This is also hinted by prospective studies on relatively small libraries, where ML SFs for SBVS directly discovered relatively potent leads, e.g. 460 nM to 20 µM(Durrant *et al.*, 2015) or 359 nM to 18.2 µM(Sun *et al.*, 2016). The generic version of RF-Score-VS (RF-Score-VS_G) not only improved the potency of the retrieved actives, it also achieved an average hit rate that is three times higher than that of SMINA. Thus, RF-Score-VS should retrieve three times more actives with the same experimental effort. This strong improvement is not surprising given that RF-Score-VS was trained with about 910,000 docked complexes arising from docking 15,426 actives and 893,897 decoys into their respective DUD-E targets(Wójcikowski *et al.*, 2017). Given that RF-Score-VS_G is freely available and it can be easily used off-the-shelf, we recommend it to re-score SMINA poses. Because RF-Score-VS tends to find different actives than SMINA, we also recommend to use both SFs prospectively if both obtained reasonably good retrospective performance on the considered targets.

As each of these targets have many known ligands, we also built a target-specific version of RF-Score-VS (RF-Score-VS_TS) for each target by restricting training to ligands docked to that target. Training set sizes ranged from 6,390 (RXRA) to 25,581 (PPARG) docked molecules. This effort permitted a slight improvement over RF-Score-VS_G's performance in five of the six targets. The literature shows that devising the right features for each target, instead using the same v2 features as RF-Score-VS_G here in every target, would have led to further improvement of this target-specific SF. This requires considering a range of feature types and hence is out of the scope of this study.

While most of the experiments were carried at the commonly considered top 1% of the ranked library, the concordance of these performance metrics (EF1%, HR1%) with that considering all possible thresholds (AUC) across SFs means that conclusions would be unaffected by choosing another threshold.

Lastly, from a ML perspective, this study shows how big datasets are not only beneficial for training ML-based SFs, but also to boost the number and diversity of test set molecules with strong potency for any target.

References

Ain,Q.U. *et al.* (2015) Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *WIREs Comput. Mol. Sci.*, **5**, 405–424.

Ballester,P.J. *et al.* (2014) Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J. Chem. Inf. Model.*, **54**, 944–955.

Ballester,P.J. *et al.* (2012) Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. *J. R. Soc. Interface*, **9**, 3196–3207.

Ballester,P.J. and Mitchell,J.B.O. (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, **26**, 1169–1175.

Bauer,M.R. *et al.* (2013) Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.*, **53**, 1447–1462.

Bitencourt-Ferreira,G. *et al.* (2019) Application of Machine Learning Techniques to Predict Binding Affinity for Drug Targets. A Study of Cyclin-Dependent Kinase 2. *Curr. Med. Chem.*, **26**.

Boyles,F. *et al.* (2019) Learning from the ligand: Using ligand-based Features to Improve Binding Affinity Prediction. *Bioinformatics*, btz665.

Chaput,L. *et al.* (2016) Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J. Cheminform.*, **8**, 56.

Chen,P. *et al.* (2019) DLIGAND2: an improved knowledge-based energy function for protein–ligand interactions using the distance-scaled, finite, ideal-gas reference state. *J. Cheminform.*, **11**, 52.

Cheng,T. *et al.* (2009) Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.*, **49**, 1079–1093.

Coleman,R.G. *et al.* (2013) Ligand Pose and Orientational Sampling in Molecular Docking. *PLoS One*, **8**, e75992.

Doman,T.N. *et al.* (2002) Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. *J. Med. Chem.*, **45**, 2213–2221.

Durrant,J.D. *et al.* (2015) Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands. *J. Chem. Inf. Model.*, **55**, 1953–1961.

Elmessaoudi-Idrissi,M. *et al.* (2018) Virtual Screening in Hepatitis B Virus Drug Discovery: Current Stateof- the-Art and Future Perspectives. *Curr. Med. Chem.*, **25**, 2709–2721.

Gaulton,A. *et al.* (2017) The {ChEMBL} database in 2017. *Nucleic Acids Res.*, **45**, D945--D954.

Le Guilloux,V. *et al.* (2009) Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.

Hoffmann,T. and Gastreich,M. (2019) The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov. Today*, **24**, 1148–1156.

Koes,D.R. *et al.* (2013) Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.*, **53**, 1893–1904.

Kruger,F.A. and Overington,J.P. (2012) Global Analysis of Small Molecule Binding to Related Protein Targets. *PLoS Comput. Biol.*, **8**, e1002333.

Lagarde,N. *et al.* (2018) Online structure-based screening of purchasable approved drugs and natural compounds: Retrospective examples of drug repositioning on cancer targets. *Oncotarget*, **9**, 32346–32361.

Lätti,S. *et al.* (2016) Rocker: Open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization. *J. Cheminform.*, **8**, 45.

Li,H. *et al.* (2019) Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics*, **35**, 3989–3995.

Li,H., Leung,K.-S.K.-S., *et al.* (2016) Correcting the impact of docking pose generation error on binding affinity prediction. *BMC Bioinformatics*, **17**, 308.

Li,H. *et al.* (2015) Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inform.*, **34**, 115–126.

Li,H. *et al.* (2020) Machine-learning scoring functions for structure-based drug lead optimization. *WIREs Comput. Mol. Sci.*, e1465.

Li,H., Leung,K.-S., *et al.* (2016) USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic Acids Res.*, **44**, W436–W441.

Li,L. *et al.* (2011) Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.*, **51**, 2132–2138.

Lyu,J. *et al.* (2019) Ultra-large library docking for discovering new chemotypes. *Nature*, **566**, 224–229.

Mysinger,M. *et al.* (2012) Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.*, **55**, 6582–6594.

Park,H. *et al.* (2011) Structure-Based Virtual Screening Approach to the Discovery of Novel Inhibitors of Eyes Absent 2 Phosphatase with Various Metal Chelating Moieties. *Chem. Biol. Drug Des.*, **78**, 642–650.

Pettersen,E.F. *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–12.

dos Santos,R.N. *et al.* (2018) Practices in molecular docking and structure-based virtual screening. In, *Methods in Molecular Biology*. Humana Press Inc., pp. 31–50.

Schneider,G. (2010) Virtual screening: an endless staircase? *Nat Rev Drug Discov*, **9**, 273–276.

Shen,C. *et al.* (2019) From machine learning to deep learning: Advances in scoring functions for protein-ligand docking. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, e1429.

Sheridan,R.P. and Kearsley,S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today*, **7**, 903–11.

Singh,N. *et al.* (2020) Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. *Brief. Bioinform.*, (In Press).

Sun,H. *et al.* (2016) Constructing and Validating High-Performance MIEC-SVM Models in Virtual Screening for Kinases: A Better Way for Actives Discovery. *Sci. Rep.*, **6**, 24817.

Trott,O. and Olson,A.J. (2010) AutoDock Vina: Improving the speed and accuracy of

docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.

Wójcikowski,M. *et al.* (2019) Building Machine-Learning Scoring Functions for Structure-Based Prediction of Intermolecular Binding Affinity. In, *Methods in molecular biology (Clifton, N.J.)*., pp. 1–12.

Wójcikowski,M. *et al.* (2017) Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.*, **7**, 46710.

Yang,X. *et al.* (2019) Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.*, **119**, 10520–10594.

Yasuo,N. and Sekijima,M. (2019) An Improved Method of Structure-based Virtual Screening via Interaction-energy-based Learning. *J. Chem. Inf. Model.*, **59**, 1050–1061.