Briefings in Bioinformatics, 00(00), 2020, 1–22

doi: 10.1093/bib/bbaa070
Problem Solving Protocol

Beware of the generic machine learning-based scoring functions in structure-based virtual screening

Chao Shen, Ye Hu, Zhe Wang, Xujun Zhang, Jinping Pang, Gaoang Wang, Haiyang Zhong, Lei Xu, Dongsheng Cao and Tingjun Hou

Corresponding authors: Tingjun Hou, College of Pharmaceutical Sciences, Hangzhou Institute of Innovative Medicine, Zhejiang University, Hangzhou 310058, Zhejiang, P. R. China. E-mail: tingjunhou@zju.edu.cn; Dongsheng Cao, Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, Hunan, P. R. China. E-mail: oriental-cds@163.com

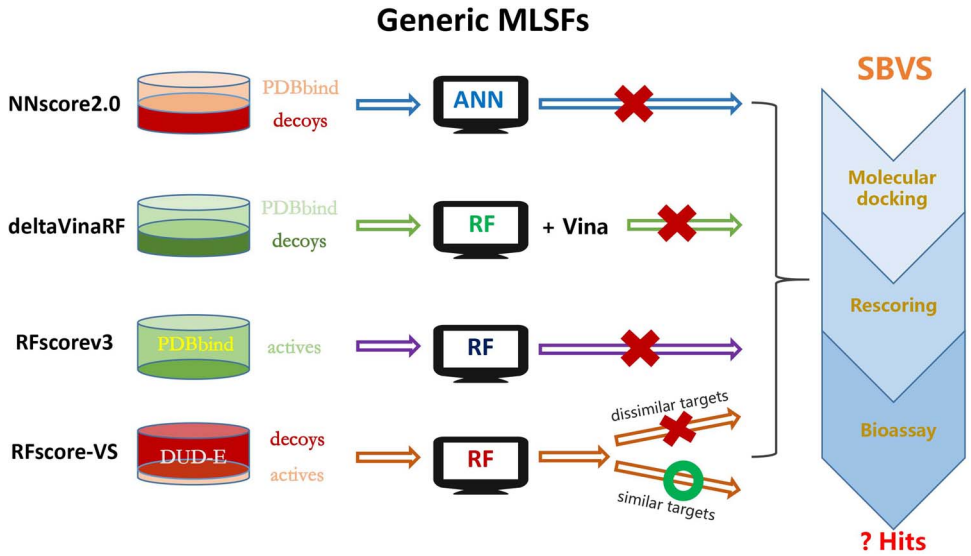
Abstract

Machine learning-based scoring functions (MLSFs) have attracted extensive attention recently and are expected to be potential rescoring tools for structure-based virtual screening (SBVS). However, a major concern nowadays is whether MLSFs trained for generic uses rather than a given target can consistently be applicable for VS. In this study, a systematic assessment was carried out to re-evaluate the effectiveness of 14 reported MLSFs in VS. Overall, most of these MLSFs could hardly achieve satisfactory results for any dataset, and they could even not outperform the baseline of classical SFs such as Glide SP. An exception was observed for RFscore-VS trained on the Directory of Useful Decoys-Enhanced dataset, which showed its superiority for most targets. However, in most cases, it clearly illustrated rather limited performance on the targets that were dissimilar to the proteins in the corresponding training sets. We also used the top three docking poses rather than the top one for rescoring and retrained the models with the updated versions of the training set, but only minor improvements were observed. Taken together, generic MLSFs may have poor generalization capabilities to be applicable for the real VS campaigns. Therefore, it should be quite cautious to use this type of methods for VS.

Professor Dongsheng Cao received his PhD degree in 2013 from Central South University, China. He is currently a professor in the Xiangya School of Pharmaceutical Sciences, Central South University, China. His research interests include (i) artificial intelligent systems for drug discovery and disease diagnosis, (ii) the development of software, web service and database in systems biology and drug discovery and (iii) design and discovery of small molecular inhibitors of important protein targets. More information can be found at the website of his group: http://www.scbdd.com.
Professor Tingjun Hou received his PhD degree in 2002 from Peking University, China. He is currently a professor in the College of Pharmaceutical Sciences, Zhejiang University, China. His research interests include (i) development of structure-based virtual screening methodologies, (ii) prediction of ADMET and drug-likeness and (iii) design and discovery of small molecular inhibitors of important protein targets. More information can be found at the website of his group: http://cadd.zju.edu.cn.

Submitted: 28 January 2020; **Received (in revised form):** 17 April 2020

Graphical Abstract



A comparative assessment in terms of screening power was conducted towards 14 existing MLSFs.

Key words: scoring function; machine learning; virtual screening; machine learning-based scoring function

Introduction

During the past two decades, we have witnessed the emergence of structure-based virtual screening (SBVS) as a routine technique in the field of computer-aided drug design (CADD), and the involvement of SBVS has greatly contributed to the discovery and development of a wide range of new drugs [1–4]. During a representative SBVS campaign, large compound libraries are computationally screened by molecular docking, and the top-ranked compounds are chosen for further experimental verification. On the one hand, SBVS owns its advantages of lower costs and higher efficiency than experimental high-throughput screening; on the other hand, compared with ligand-based VS, SBVS is also more likely to identify lead compounds with novel scaffolds [5, 6].

A wide variety of scoring schemes have been developed for SBVS, among which scoring functions (SFs) for molecular docking based on force field, empirical or knowledge are most widely utilized (Figure 1) [7]. A common feature of these classical SFs is that they all assume a theory-driven or expert-judged additive function form to represent the relationship between the experimentally determined binding affinities and the features that characterize protein–ligand interactions. However, it has been known that this assumption may not always exist in the real scenario [8]. Recently, to further improve the accuracy of classical SFs, machine learning-based SFs (MLSFs) have gained increasing attention, and hence numerous methods employing different ML technologies, such as random forest (RF) [9–12], support vector machine (SVM) [13, 14], artificial neural network (ANN) [15, 16], gradient boosting decision tree [17–19] and convolutional neural network (CNN) [20–23], have been developed. Unlike classical SFs, MLSFs can implicitly learn the function form from the training data and use adjustable parameters to further improve the performance, thus showing higher flexibility and convenience [24]. In general, MLSFs can be classified into generic and target-specific MLSFs. The former is always trained on the datasets composed of multiple targets and designed for generic uses just as classical SFs, while the latter is specifically designed for a certain target. Despite the emergence of MLSFs, until now, only a few practical applications of these approaches have been reported [25–29], and most related publications mainly focus on the development of methodology, so how such methods actually perform remains unclear. Thus, a systematic assessment of these reported MLSFs from third party is urgently needed.

Apart from the improved flexibility and accuracy, existing MLSFs also display limitations. One major issue is how to balance three aspects for a given SF, i.e. the scoring power (the ability to predict binding affinity), the docking power (the ability to distinguish true poses from decoy poses) and the screening power (the ability to distinguish active compounds from decoy compounds). The concern was first raised in 2014 by Gabel *et al.* [30], who utilized simple protein–ligand element–element distance counts as the features in RFscore [9] and constructed two MLSFs using RF and SVM, respectively. These MLSFs could reproduce the claimed superiority of RFscore in terms of scoring power, but their docking power and screening power were rather poor and even significantly worse than classical Surflex-Dock. In the following years, extensive efforts were dedicated to solving this problem. Unfortunately, no substantial progress was reported until the development of deltaVinaRF [14] and deltaVinaXGB [17] reported by Zhang's group. The incorporation of special decoys into the training set and the introduction of a novel parameterization strategy to fit a correction term rather than the final score enabled the ranking of the top three in all four metrics in the CASF benchmark [31]. However, each target in the screening power test set of the CASF benchmark only contains 5 active and 280 inactive compounds, and therefore, the screening power

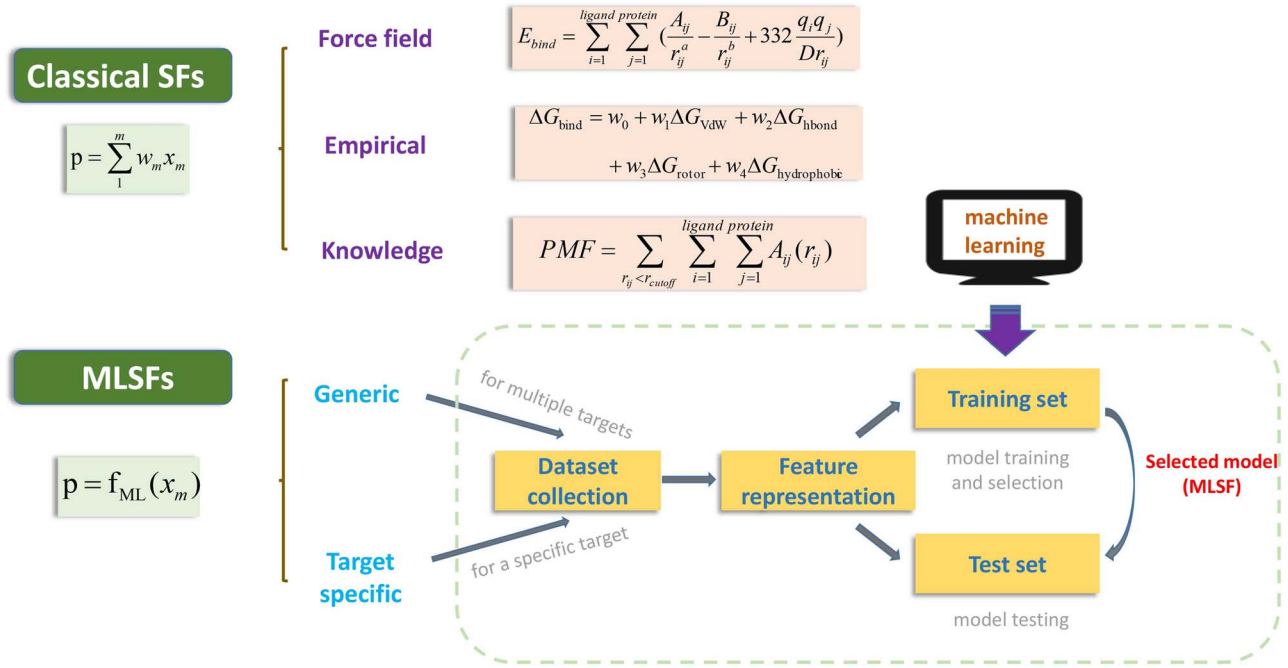


Figure 1. The general workflow to construct MLSFs. Unlike the classical SFs that use a linear function form, MLSFs can implicitly learn the function form from the training data. Generic or target-specific data should firstly be collected and prepared and then represented by some types of feature vectors and trained by some advance machine learning algorithms, thus succeeding in the construction of a MLSF.

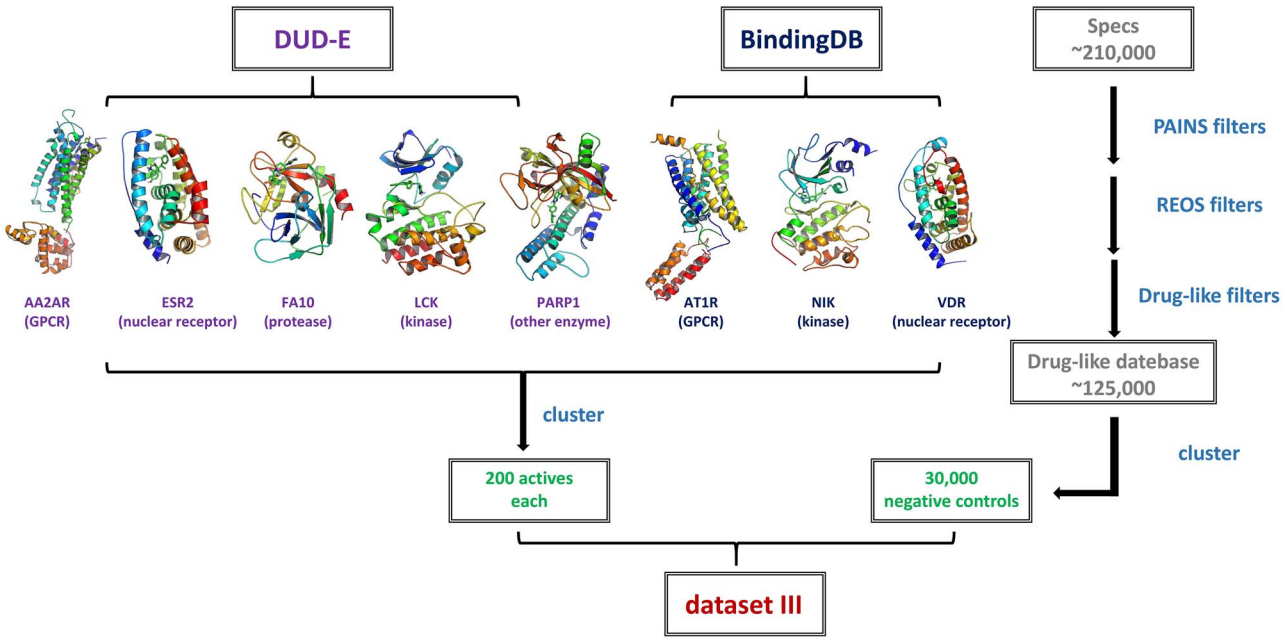


Figure 2. Workflow of the generation for dataset III.

of the above two methods needs further exploration. Other attempts of designing MLSFs with balanced powers include AGL-Score [18] and MT-Net [32], of which their applicability might also be limited for practical uses. In general, MLSFs focused on scoring power are mainly trained on some datasets such as PDBbind [33], where all the protein–ligand complexes and their corresponding binding affinities are experimentally determined, and then a regression model was constructed for the prediction of binding affinity [34]. However, in terms of docking power or screening power, additional decoy binding poses (or decoy compounds) are also included in the training set, so that the incorrect poses (or inactive compounds) in the test set can also be differentiated. It seems that a consensus has been reached that the MLSFs designed for different tasks may require different training sets due to the inherent attributes of individual ML models. However, how the MLSFs designed for binding affinity

prediction would perform in terms of docking power or screening power, and how the MLSF trained for VS would perform in terms of scoring power, still need more thorough investigation, which would lead to the further development of MLSFs with higher precision and wider applicability.

In this study, to figure out whether MLSFs could be generally applicable to VS, a comparable assessment of the screening power was carried out using 14 existing MLSFs. Most of these commonly used SFs that were trained on the PDBbind dataset mainly focused on scoring power. In addition, deltaVinaRF and deltaVinaXGB were also involved to analyze if they could still perform well in a much larger dataset. Moreover, RFscore-VS [12], specially designed for VS, was also re-evaluated here. We intend to not only compare their performance but also to solve the following puzzles: (1) Generic MLSFs have been reported in many cases with higher scoring power than classical SFs, but can they also perform better in terms of other capabilities? (2) The performance of MLSFs may vary dramatically on different datasets. Therefore, we investigated how MLSFs would perform on the data sets that have been widely used for the evaluation of classical SFs by means of their screening power. (3) Docking poses may exert significant influence on the performance of the rescoring methods. Hence, we explored how performance would change when different docking programs are utilized to generate binding poses. (4) MLSFs may be greatly affected by several determining factors such as training sets, feature representation methods and ML algorithms. Thus, how can these factors impact the performance?

Materials and methods

Data sets

A total of three datasets, i.e. Directory of Useful Decoys-Enhanced (DUD-E) [35], demanding evaluation kits for objective *in silico* screening (DEKOIS) 2.0 [36] and another dataset constructed by ourselves (dataset III), were used in this study. DUD-E is often considered as the most widely used dataset for benchmarking VS protocols, while DEKOIS2.0 as a complementary benchmark was utilized for further validation. DUD-E and DEKOIS2.0 have different actives versus decoys ratio and utilize different methodologies to construct the decoys. As both of these two datasets utilized the artificially constructed decoys as the negative controls, we further proposed the dataset III that directly collected compounds from drug-like compound library as the negative controls to better mimic the real VS campaigns. Accordingly, dataset III shows lower actives versus decoys ratio of 1:150, indicating a more complex VS scenario.

Directory of Useful Decoys-Enhanced

DUD-E contains a total of 22 886 active compounds against 102 targets from 8 diverse protein families, including 5 G-protein-coupled receptors (GPCRs), 26 kinases, 11 nuclear receptors, 15 proteases, 2 ion channels, 2 cytochrome P450s, 36 other enzymes and 5 miscellaneous proteins, as reported in Supplementary Table S1. These active compounds were originally retrieved from the ChEMBL09 database [37]. For each active compound, 50 decoys with similar physicochemical properties but dissimilar 2D topology were generated from ZINC [38] as the negative controls. The ligands were preprocessed by OpenEye's Omega [39] and directly employed for the following docking calculations. Furthermore, the proteins were retrieved from RCSB Protein Data Bank (PDB) [40] and then prepared using the *Protein Preparation Wizard* [41] module implemented in Schrödinger (version 2019), including removing waters and redundant chains, assigning bond orders, adding hydrogen atoms, filling in missing side chains, optimizing H-bond network and minimizing the system with the OPLS2005 force field [42] until the root-mean-square deviation of heavy atoms reached to 0.30 Å. PROPKA [43] was utilized to determine the protonation states of residues at pH = 7.0.

DEKOIS2.0

DEKOIS2.0 contains 81 structurally diverse targets, as listed in Supplementary Table S2. Each target has 40 active compounds extracted from BindingDB [44] and 1200 decoys generated from ZINC. The proteins and ligands from DEKOIS2.0 were directly used in this assessment because they had been prepared with the *Protein Preparation Wizard* and *LigPrep* [45] modules in Schrödinger, respectively.

Dataset III

The preparation of dataset III is outlined in Figure 2. First, a library of approximately 210 000 commercially available compounds was filtered with the Pan-Assay Interference Compounds (PAINS) [46] and Rapid Elimination Of Swill (REOS) [47] rules in Canvas [48] and several in-house drug-like filters. The remaining molecules must satisfy the criteria that the violation counts of the Oprea's rule [49] and Lipinski's rule [50] were less than 3 and 2, respectively, thus resulting in a drug-like database with nearly 125 000 compounds. Next, 30 000 compounds were selected from this database by using the *Find Diverse Molecules* module in Discovery Studio 2.5 [51] with the FCFP_6 fingerprints and considered as the negative controls. Furthermore, eight representative targets, i.e. five targets from DUD-E including AA2AR (GPCR), ESR2 (nuclear receptor), FA10 (protease), LCK (kinase) and PARP1 (other enzyme) and 3 targets absent in DUD-E including AT1R (GPCR), NIK (kinase) and VDR (nuclear receptor) were selected for testing. For each of the former five targets, all active compounds were collected from DUD-E and clustered from the original dataset with the *Find Diverse Molecules* module. For each of the latter three targets, all active compounds were extracted from BindingDB with their K_i (or K_d or IC_{50}) values less than 10 μM. Then, the *LigPrep* module was employed for ligand preparation with all the default settings and the *Find Diverse Molecules* module was utilized to further reduce the number of compounds to 200 for each target. Proteins with the PDB entries 4YAY [52], 5T8O [53] and 1S19 [54] were assigned to AT1R, NIK and VDR, respectively, and were prepared with the *Protein Preparation Wizard* module.

Docking programs

In our previous study [55], three docking programs, i.e. Glide SP (version 8.2) [56], GOLD (version 2019) [57] with Piecewise Linear Potential (CHEMPLP) SF and LeDock (version 1.0) [58] showed high docking power and computational efficiency and thus were chosen to generate appropriate binding poses for each ligand. In each docking calculation, the binding site was defined by the co-crystalized ligand, and only the poses with best scores were remained. All the other parameters were set without tuning of the optional parameters, unless otherwise noted as followed.

Glide SP

The grids were firstly generated by using the *Receptor Grid Generation* utility with the size of binding box set to 10 Å × 10 Å × 10 Å

centered on the co-crystallized ligand. Then, the Glide docking calculations with the SP scoring mode were carried out.

GOLD CHEMPLP

Proteins were firstly prepared using the built-in protein preparation plugin including adding hydrogens and deleting unnecessary waters. The binding site was determined as the residues within 10 Å around the co-crystalized ligand. The genetic algorithm search efficiency was set to 'slow', and CHEMPLP was utilized for scoring.

LeDock

A combination of simulated annealing and evolutionary optimization algorithm was used to sample the conformations for each ligand. The given protein was firstly processed by the *lepro* utility, and then docking calculation was performed with the default settings.

MLSFs

A total of 14 MLSFs with source codes or executable scripts were systematically assessed (Table 1). It should be noted that all these methods are generic, as target-specific approaches are hard to be tested here. In this study, each compound was first docked into the binding site by three docking programs and then rescored by each tested MLSF. For most MLSFs, each protein–ligand complex was fed into the MLSF model at a time, and then a final predicted score was computed. All the other parameters were set without tuning the optional parameters, unless otherwise noted as followed.

deltaVinaRF && deltaVinaXGB

deltaVinaRF and deltaVinaXGB were executed directly using the *dvrf20.py* and *run_DXGB.py* scripts, respectively. deltaVinaXGB was originally designed to consider the impacts of explicit water molecules. However, in this study, water molecules were removed in advance and thus were not considered. Besides, as the computation of deltaVinaXGB was too time-consuming, it was not assessed on DUD-E.

RFscorev3

The model was trained on pre-calculated descriptors of the PDBbind v2007 dataset by the *rf-train* utility, and then used to rescore our test sets with the *rf-score* utility.

RFscorev4

Unlike RFscorev3, RFscorev4 used the docking poses generated from Vina instead of the experimentally determined poses to train the model. With the use of the *rf -score* utility, the input complex could directly obtain its final score. Two versions of training sets including the PDBbind v2013 and v2014 refined sets were available. However, only the results of v2014 were reported here as it showed better performance.

RFscore-VS

RFscore-VS rescoring was carried out by using the *rf-score-vs* utility for rescoring.

NNscore1.0 && NNscore2.0

Proteins and ligands were firstly converted to the *pdbqt* format using the *prepare_receptor4.py* and *prepare_ligand4.py* scripts implemented in AutoDockTools 1.5.6 [61], respectively, along with the addition of hydrogen atoms, assignment of Gasteiger charges and cleanup of unwanted elements. Then, the *NNScore.py* and *NNScore2.01.py* scripts were used to generate the predicted scores of NNscore1.0 and NNscore2.0, respectively. The average score of 24 networks and the average score of 20 networks were utilized as the final scores of NNscore1.0 and NNscore2.0, respectively. It was worth noting that NNscore1 .0 would give a complex score of −999999.9 if this complex was regarded as an extremely bad binder.

pafnucy

The *prepare.py* script was used to prepare the inputs for the neural network and *predict.py* was utilized to predict the binding affinities.

OnionNet

The *generate_features.py* and *predict_pKa.py* scripts were employed for the generation of input features and the prediction of binding affinities, respectively.

Open Drug Discovery Toolkit

Open Drug Discovery Toolkit (ODDT) was developed as an open source tool for CADD developers and researchers, and it could reproduce the execution of several state-of-the-art MLSFs including NNscore2.0, RFscorev1, RFscorev2 and RFscorev3. Therefore, these methods were also tested in this study. Six versions of the PDBbind datasets (v2007, v2012, v2013, v2014, v2015 and v2016) were available to serve as the training set, and all of them were tested for training but only the results of the latest v2016 were finally utilized for comparison. Protein–ligand extended connectivity (PLEC) fingerprint could implicitly encode protein–ligand interactions by pairing the extended connectivity fingerprint environments from the ligand and the protein. Based on this feature representation method, some MLSFs were correspondingly developed [63]. Three ML approaches, namely linear model fitted by minimizing a regularized empirical loss with Stochastic Gradient Descent, ANN and RF, were used to train the models. In this study, PLECRF tended to show the best early recognition performance; thus, it was remained for further comparison. All the above rescoring calculations were conducted by the *oddt_cli* utility.

Evaluation metrics

Four widely used metrics were employed in this study to comprehensively evaluate the screening power of all 14 MLSFs, including the area under the receiver operating characteristic curve (AUROC) [64], area under the semilog ROC curve (LogAUC) [65], enrichment factors (EFs) [66] and Boltzmann enhanced discrimination of receiver operating characteristic (BEDROC) [67] scores. The ROC curve reflects the relationship between the true positive rate and false positive rate. The corresponding AUC value ranges from 0 for a complete failure to 1 for a perfect enrichment, and a value of 0.5 suggests a random prediction. When plotting the *x*-axis as the form of $\log_{10}$, ROC curve can be transformed to semilog ROC curve, and the area under the curve is defined as LogAUC_{λ}, where the log area computations run from λ to 1, and here $\text{LogAUC}_{0.001}$ is simply referred to as LogAUC, and the random prediction of LogAUC is about 0.1446. $\text{EF}_{x\%}$ is an intuitive parameter that is defined as the percentage of true positives found among all the active compounds for a given percentile of

Table 1. Basic information of MLSFs tested here

MLSFs	Training sets	Features	ML methods	Website
deltaVinaRF [11]	Experimental subset (PDBbind v2014 refined set, native poses in CSAR decoy dataset and weak-binding structures in PDBbind v2014 general set) and decoy subset (computationally estimated binding affinities)	10 terms from Smina and 10 terms related to bSASA	RF	http://www.nyu.edu/projects/yzhang/DeltaVina/
deltaVinaXGB [17]	PDBbind v2016 refined set; CSAR decoys	58 Smina terms, 10 bSASA, terms related to water effects, ligand conformation stability and the number of binding site ions	XGB	http://www.nyu.edu/projects/yzhang/DeltaVina/
RFscorev3 [59]	PDBbind v2007	36 atom type pair counts and 6 Vina terms	RF	http://istar.cse.cuhk.edu.hk/rf-score-3.tgz
RFscorev4 [60]	PDBbind v2014 refined set (Vina generated poses)	36 atom type pair counts and 11 Vina terms	RF	http://ballester.marseille.inserm.fr/rf-score-4.tgz
RFscore-VS [12]	DUD-E (Vina generated poses)	Terms from rfscorev2	RF	http://wojcikowski.pl/travis/rf-score-vs_v1.0_linux_2.7.zip
NNscore1.0 [15]	2710 complexes from PDB (good and poor binders); 1431 docking poses	194 terms (mainly related with atom pair counts)	ANN	http://nbcr.ucsd.edu/software-nnscore
NNscore2.0 [16]	Experimentally measured K_d complexes from MOAD and PDBbind; Vina or AutoDock docking poses from DUD	384 terms (mainly related with atom pair counts)	ANN	http://nbcr.ucsd.edu/software-nnscore
pafnucy [22]	PDBbind v2016 refined set	A 4D tensor (Cartesian coordinates and 19 features)	CNN	https://gitlab.com/cheminfIBB/pafnucy
OnionNet [23]	PDBbind v2016 general set	element-pair specific contacts	CNN	https://github.com/zhenglz/onionnet
NNscore(ODDT) [16, 62]	PDBbind v2016 refined set	NNscore2.0 excluding terms from Vina	ANN	https://github.com/oddt/oddt
RFscorev1(ODDT) [9, 62]	PDBbind v2016 refined set	rfscorev1	RF	
RFscorev2(ODDT) [10, 62]	PDBbind v2016 refined set	rfscorev2	RF	
RFscorev3(ODDT) [59, 62]	PDBbind v2016 refined set	rfscorev3	RF	
PLECRF(ODDT) [62, 63]	PDBbind v2016 general set	PLEC fingerprint	RF	

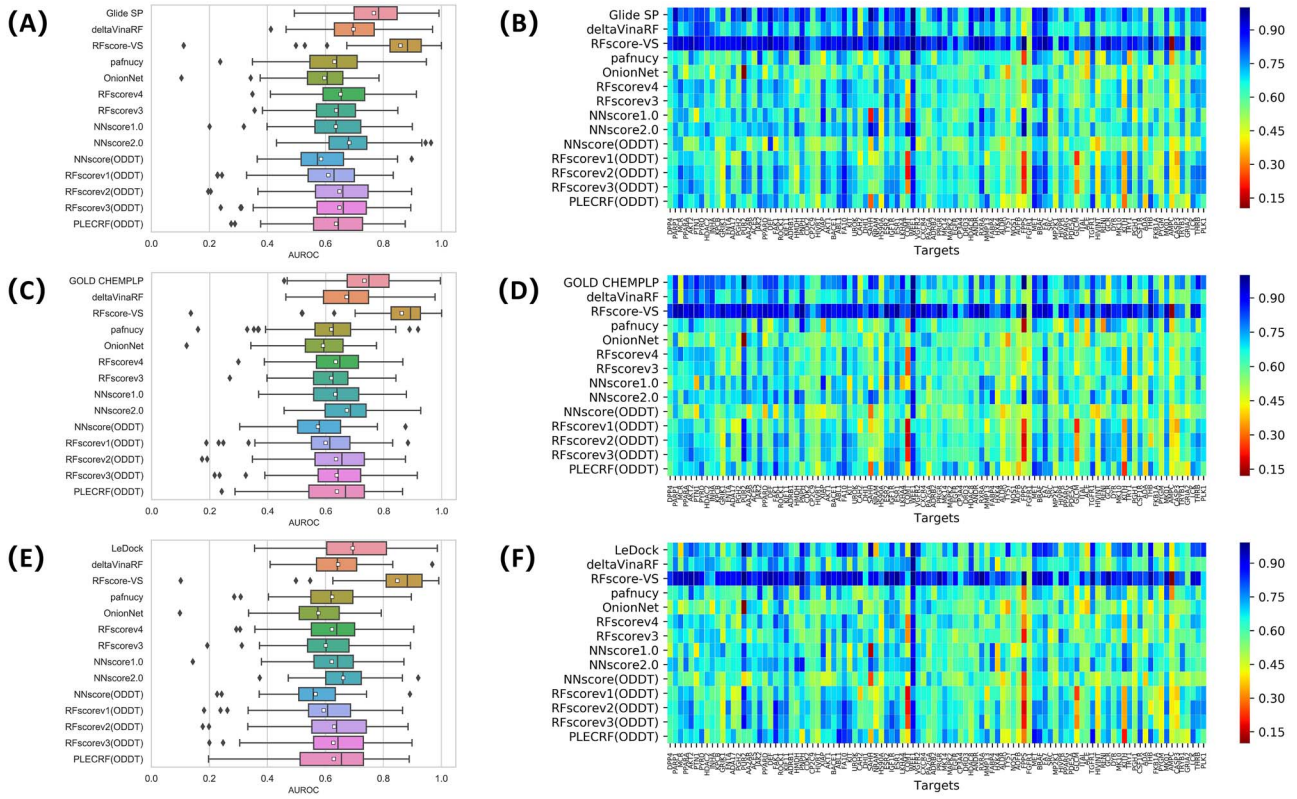


Figure 3. The AUROC values of 13 MLSFs on DUD-E based on the docking poses predicted by (A, B) Glide SP, (C, D) GOLD CHEMPLP and (E, F) LeDock, respectively. (A), (C) and (E) show the overall performance towards 102 targets in box plots, while (B), (D) and (E) show the performance of individual targets in heat maps. The white squares in boxplots represent the mean AUROC values for each MLSF.

the top-ranked compounds (x%) of a chemical library. In particular, here the percentiles were defined as 0.1%, 0.5%, 1% and 5%. Unlike EFs, BEDROC score takes all the compounds into consideration rather than proportion of the chemical library. This score can be modulated by parameter α to adjust the weight given to the top-ranked compounds, and here the α-value was set to 80.5, meaning that the 2% top-ranked molecules were utilized to account for 80% of the BEDROC score. All the above metrics were calculated by using the in-house python scripts mainly based on the *scikit-learn* [68] and *numpy* [69] modules. The post hoc Nemernyi test [70] implemented in the *scikit-posthocs* [71] package was selected to determine whether the difference between any two of compared methods was statistically significant. If the computed *P*-value was higher than 0.10, the performance of two tested methods was roughly considered to have no significant difference. In addition, nonparametric Mann–Whitney test and Wilcoxon signed-rank test were also used somewhere to handle the two independent samples and paired samples, respectively, with the *P*-value cutoff set to 0.05.

Results and discussion

Assessment of screening power on DUD-E

A total of 13 MLSFs (except deltaVinaXGB) were first tested on the DUD-E dataset using three different docking programs (i.e. Glide SP, GOLD CHEMPLP and LeDock) for pose generation. The performance reported by the AUROC values is shown in Figure 3, with the corresponding *P*-values between any two of the compared SFs shown in Supplementary Figure S1. Among all the MLSFs, RFscore-VS trained on DUD-E performs unsurprisingly better than most of the other methods based on the binding poses generated from all docking programs ($P < 0.10$ except for Glide SP), indicating that RFscore-VS may be a versatile rescoring tool but not limited to rescoring poses from Vina. In addition, least variations are found over different targets. However, for the other MLSFs, the mean AUROC values of all these SFs range from 0.55 to 0.70, which is only slightly higher than random prediction but worse than the baseline of classical SFs (i.e. the first row of individual plots in Figure 3). Furthermore, deltaVinaRF that was originally reported to perform well on the CASF benchmark results in rather low AUROC values on this dataset. This implies that the CASF benchmark may be not so applicable for the comprehensive assessment of the performance of MLSFs in this context. One reason of generally poor performance might be due to the fact that their training sets are not large and diverse enough. For the development of some MLSFs such as deltaVinaRF, NNscore1.0 and NNscore2.0, additional decoys were incorporated into their training sets, but they may be still far from enough to cover sufficient chemical space. Another important reason may be the intrinsic nonlinear essences of these advanced ML models. Despite those MLSFs trained on PDBbind show extremely superior scoring power, overfitting effects may also lead to overestimation of certain features, thus resulting in relatively poor screening power. As can be expected, different feature representation methods and ML algorithms could also affect final performance. For example, it seems that RFscorev3 (ODDT) and RFscorev2 (ODDT) perform slightly better than RFscorev1 (ODDT) (despite $P > 0.10$). Recently, CNN-based SFs have emerged and shown to be more convenient

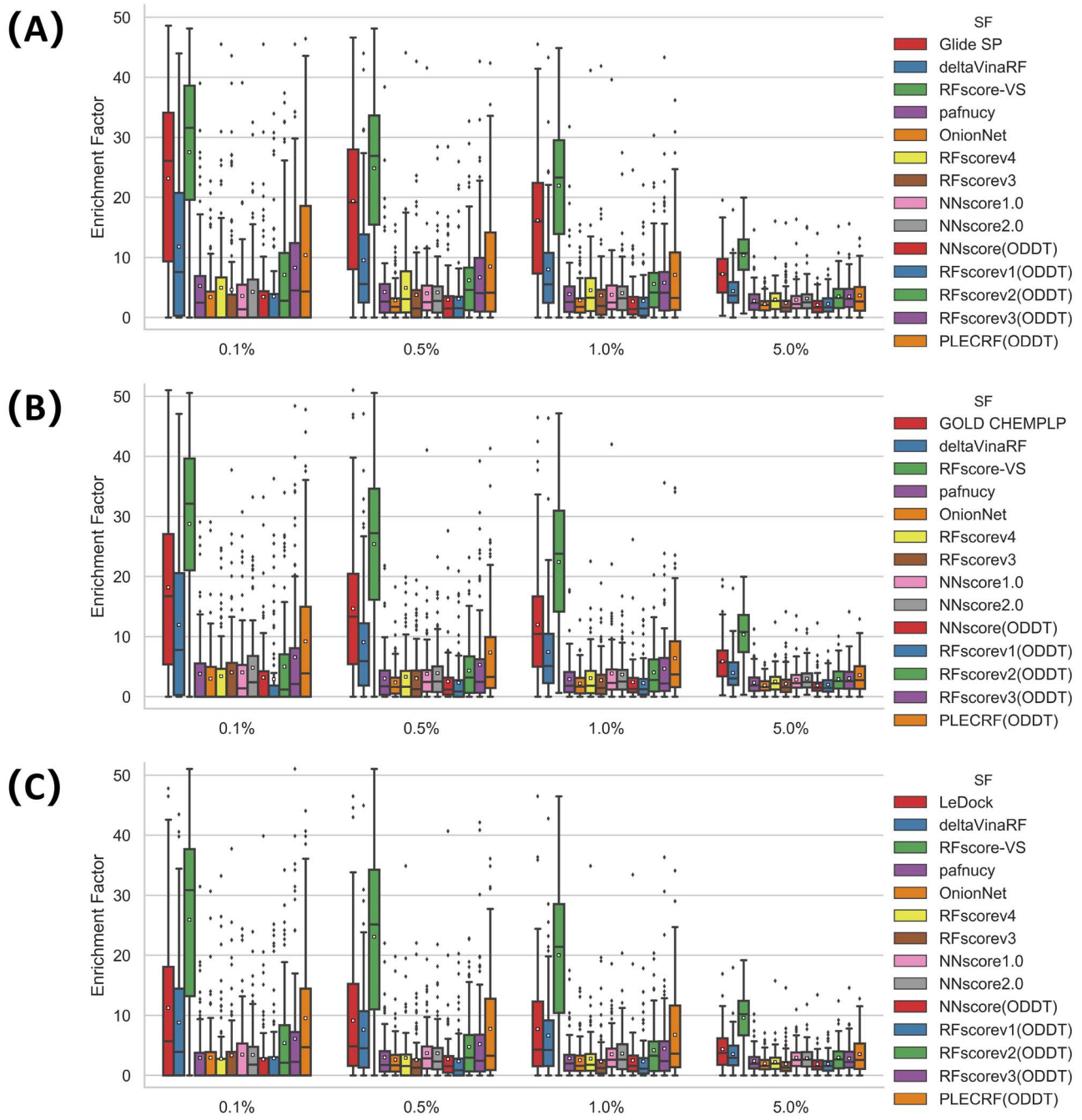


Figure 4. The EFs of 13 MLSFs on DUD-E based on the docking poses predicted by (A) Glide SP, (B) GOLD CHEMPLP and (C) LeDock, respectively. The white squares in boxplots represent the mean values for each MLSF.

than traditional ML-based SFs [24]. However, the application of CNN (pafnucy and OnionNet) here does not show distinct effect (Figure 3). These phenomena further indicate that feature representation methods and ML algorithms may be important to their performance here but do not play the decisive roles.

Then, the enrichment performance is evaluated by the values of EFs, LogAUC and BEDROC scores as shown in Figure 4 and Supplementary Figures S2 and S3, respectively, with the corresponding *P*-values shown in Supplementary Figures S4–S6. Unsurprisingly, a similar trend is found for different docking programs. RFscore-VS again performs the best, whereas the remaining MLSFs perform worse than the baseline of classical SFs ($P < 0.10$ for most MLSFs). Unlike classical Glide SP that can achieve the average $EF_{0.1\%}$, $EF_{0.5\%}$, $EF_{1\%}$ and $EF_{5\%}$ of 23.13, 19.39, 16.18 and 7.23, respectively, the corresponding value ranges of most MLSFs are only 3–12, 2–10, 2–8 and 2–5, respectively (Figure 4A). More importantly, the median values (thick lines in box plots) are much higher than the mean values (white squares), suggesting that the EFs distribution is dominated by large values resulted from certain small numbers of targets. Hence, most MLSFs tested here cannot show reasonable early enrichment on the DUD-E dataset.

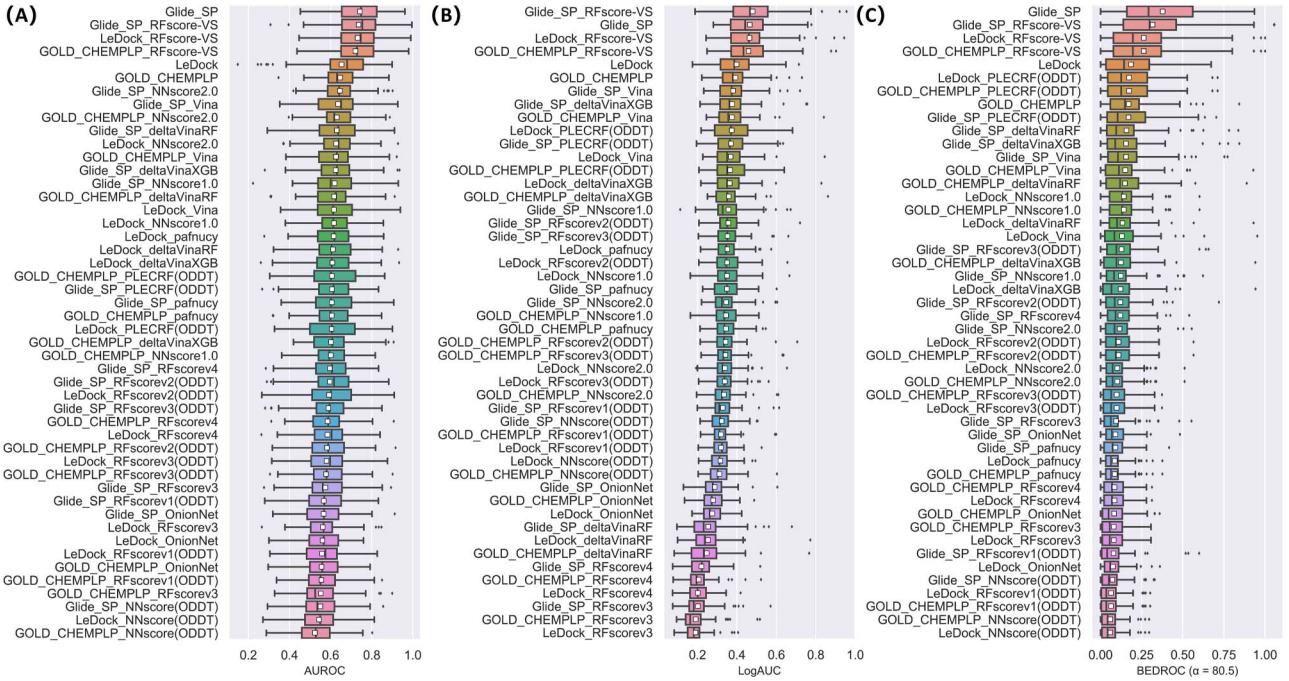


Figure 5. Evaluation of 14 MLSFs on DEKOIS2.0 based on three docking programs using (A) AUROC, (B) LogAUC, and (C) BEDROC (α = 80.5). The white squares in boxplots represent the mean values for each MLSF. The SFs are sorted by their mean values. Vina means using Vina SF to rescore the docking poses, and this score is generated by deltaVinaXGB.

Furthermore, the average AUROC and BEDROC values calculated from three docking programs based on all MLSFs are compared among different protein target families, as reported in Table 2 and Supplementary Figure S7. Among all eight families, cytochrome P450 is generally the most difficult case, for which the best performance is found by RFscore-VS with an average AUROC of 0.731 (Glide SP) and an average BEDROC score of 0.187 (GOLD CHEMPLP), which can be expected from the intrinsic characteristics of this protein family. Cytochrome P450 enzymes are responsible for drug metabolism and their binding sites are often large enough to accommodate diverse substrates and inhibitors [72], thus leading to their poor performance to distinguish actives from decoys. Except the cytochrome P450 family, Glide SP and GOLD CHEMPLP can obtain relatively comparable results among the other seven families. However, larger variations among different target families are found for LeDock, for which the best performance is observed for kinases, indicating that LeDock might be more reliable for kinases. Among all MLSFs, RFscore-VS still yields the best average performance for all eight families, with the highest AUROC score of 0.955 and the highest BEDROC score of 0.780 for nuclear receptors. For the other MLSFs, the performance is also generally poor and often worse than the baseline of classical SFs. However, some MLSFs perform not so badly on some categories such as kinases and proteases. Two types of biases within certain target sets, i.e. analogue bias and decoy bias, may mainly account for these phenomena, which have been reported previously [73, 74]. On the one hand, the actives for some targets may be analogous despite they have been clustered, thus making them extremely easy to be discriminated from the decoys. On the other hand, although the decoys are reported to be dedicatedly constructed to mimic the physicochemical properties but dissimilar 2D topology, some molecular properties may still be different.

Assessment of screening power on DEKOIS2.0

We further explored the screening power of all MLSFs on DEKOIS2.0. The average performance based on different combinations of docking programs and MLSFs is evaluated by AUROC, LogAUC, BEDROC (Figure 5) and EFs (Figure 6). The corresponding *P*-values and the corresponding target-based results for individual docking programs are provided in Supplementary Figures S8–S14 and S15–S17, respectively. It is unfortunately to see that the differences between most SF pairs are not statistically significant as the results vary so much over different targets, but these comparisons may still provide some valuable guidance. As shown in Figure 5, RFscore-VS performs similarly to the best-performed classical SF Glide SP. In addition, GOLD CHEMPLP and LeDock also display comparable performance. The remaining MLSFs still show worse screening power than classical SFs. As for the EFs, a similar phenomenon can be observed, and it is striking to observe that the enrichment of the active compounds in the top 0.1% for most targets completely fails for 10 MLSFs (Figure 6). DeltaVinaXGB is newly added here so it should be specifically noted. To better illustrate the essence of this deltavina parameterization strategy, Vina SF directly generated from deltaVinaXGB is also included to make a comparison. Surprisingly, as shown in Figure 6, deltaVinaXGB and deltaVinaRF only yield comparable mean performance to Vina, suggesting that this deltavina parameterization strategy actually does not exert obvious effects. Comparison of the differences between their scores illustrates that the Vina scores still account for the largest proportion of the final scores while the contribution of the correction terms fitted by ML methods is relatively low, which may partially explain why these methods yield similar results. Another crucial reason may be attributed to the intrinsic nonlinear essences of these constructed ML models, which has been mentioned in the previous section, thus

Table 2. The mean AUROC and BEDROC of each MLSF towards different target protein families of DUD-E

Docking program	Rescoring MLSF	Kinase		Protease		Nuclear receptor		GPCR		Cytochrome P450		Other enzymes		Miscellaneous	
		AUROC	BEDROC	AUROC	BEDROC	AUROC	BEDROC	AUROC	BEDROC	AUROC	BEDROC	AUROC	BEDROC	AUROC	BEDROC
Glide SP	/	0.788	0.436	0.801	0.490	0.832	0.545	0.707	0.226	0.593	0.078	0.732	0.357	0.792	0.471
	deltaVinaRF	0.722	0.248	0.720	0.211	0.716	0.297	0.612	0.034	0.552	0.060	0.692	0.195	0.662	0.289
	RFscore-VS	0.861	0.539	0.885	0.635	0.944	0.767	0.856	0.435	0.731	0.173	0.839	0.528	0.836	0.591
	pafnucy	0.734	0.252	0.570	0.108	0.679	0.193	0.579	0.084	0.601	0.157	0.579	0.126	0.596	0.153
	OnionNet	0.632	0.094	0.623	0.095	0.575	0.083	0.668	0.096	0.547	0.034	0.567	0.082	0.559	0.073
	RFscorev4	0.705	0.132	0.682	0.187	0.699	0.187	0.630	0.073	0.596	0.107	0.601	0.084	0.638	0.130
	RFscorev3	0.668	0.121	0.667	0.186	0.683	0.063	0.608	0.040	0.569	0.129	0.600	0.071	0.641	0.134
	NNscore1.0	0.599	0.089	0.711	0.143	0.729	0.138	0.631	0.058	0.585	0.072	0.602	0.127	0.681	0.118
	NNscore2.0	0.684	0.141	0.699	0.096	0.710	0.141	0.640	0.094	0.611	0.071	0.687	0.128	0.655	0.095
	NNscore(ODDT)	0.571	0.086	0.620	0.101	0.626	0.110	0.532	0.025	0.498	0.046	0.586	0.072	0.596	0.047
	RFscorev1(ODDT)	0.661	0.066	0.647	0.115	0.670	0.225	0.590	0.074	0.569	0.059	0.544	0.048	0.632	0.142
	RFscorev2(ODDT)	0.720	0.197	0.690	0.219	0.705	0.219	0.623	0.112	0.611	0.138	0.572	0.081	0.649	0.186
	RFscorev3(ODDT)	0.706	0.193	0.689	0.209	0.726	0.257	0.629	0.129	0.571	0.126	0.576	0.089	0.653	0.163
	PLECRF(ODDT)	0.692	0.285	0.665	0.200	0.736	0.299	0.624	0.071	0.576	0.065	0.573	0.092	0.605	0.166
GOLD CHEMPLP	/	0.747	0.293	0.793	0.451	0.697	0.244	0.718	0.201	0.587	0.095	0.726	0.310	0.703	0.361
	deltaVinaRF	0.691	0.209	0.676	0.178	0.670	0.244	0.583	0.030	0.538	0.081	0.678	0.183	0.684	0.255
	RFscore-VS	0.867	0.520	0.872	0.611	0.955	0.780	0.860	0.401	0.709	0.187	0.841	0.518	0.859	0.600
	pafnucy	0.705	0.118	0.564	0.055	0.664	0.103	0.583	0.078	0.604	0.112	0.571	0.071	0.586	0.066
	OnionNet	0.627	0.073	0.611	0.083	0.560	0.060	0.667	0.072	0.549	0.036	0.560	0.076	0.550	0.015
	RFscorev4	0.651	0.066	0.688	0.170	0.677	0.161	0.614	0.063	0.586	0.099	0.597	0.064	0.618	0.075
	RFscorev3	0.631	0.082	0.654	0.162	0.641	0.046	0.594	0.054	0.567	0.115	0.599	0.067	0.620	0.111
	NNscore1.0	0.607	0.077	0.684	0.129	0.709	0.163	0.657	0.056	0.581	0.067	0.598	0.124	0.680	0.097
	NNscore2.0	0.664	0.132	0.695	0.100	0.687	0.149	0.636	0.063	0.599	0.052	0.674	0.108	0.695	0.102
	NNscore(ODDT)	0.571	0.099	0.612	0.081	0.589	0.074	0.521	0.028	0.492	0.060	0.569	0.062	0.565	0.033
	RFscorev1(ODDT)	0.622	0.053	0.645	0.102	0.652	0.165	0.594	0.075	0.568	0.064	0.547	0.061	0.625	0.099
	RFscorev2(ODDT)	0.691	0.121	0.686	0.207	0.666	0.181	0.607	0.112	0.611	0.144	0.573	0.070	0.636	0.124
	RFscorev3(ODDT)	0.676	0.124	0.681	0.188	0.702	0.217	0.620	0.149	0.574	0.098	0.582	0.082	0.636	0.159
	PLECRF(ODDT)	0.722	0.256	0.667	0.198	0.713	0.198	0.625	0.070	0.592	0.094	0.562	0.099	0.586	0.162
LeDock	/	0.799	0.344	0.681	0.165	0.633	0.162	0.663	0.082	0.560	0.073	0.663	0.168	0.667	0.190
	deltaVinaRF	0.663	0.200	0.634	0.149	0.646	0.234	0.593	0.039	0.521	0.031	0.644	0.150	0.691	0.303
	RFscore-VS	0.866	0.515	0.872	0.494	0.952	0.757	0.856	0.402	0.689	0.163	0.810	0.422	0.843	0.619
	pafnucy	0.724	0.137	0.568	0.054	0.672	0.105	0.568	0.052	0.604	0.098	0.562	0.057	0.627	0.061
	OnionNet	0.634	0.090	0.584	0.075	0.550	0.071	0.656	0.091	0.522	0.039	0.539	0.075	0.512	0.060
	RFscorev4	0.674	0.096	0.643	0.111	0.668	0.142	0.611	0.050	0.590	0.068	0.569	0.046	0.616	0.093
	RFscorev3	0.589	0.091	0.673	0.127	0.650	0.041	0.585	0.037	0.554	0.085	0.570	0.042	0.625	0.116
	NNscore1.0	0.613	0.109	0.652	0.125	0.670	0.145	0.651	0.057	0.573	0.047	0.588	0.112	0.688	0.130
	NNscore2.0	0.643	0.127	0.679	0.099	0.673	0.158	0.645	0.068	0.594	0.063	0.668	0.102	0.684	0.094
	NNscore(ODDT)	0.585	0.104	0.576	0.065	0.597	0.091	0.539	0.033	0.486	0.049	0.551	0.058	0.569	0.042
	RFscorev1(ODDT)	0.650	0.059	0.623	0.091	0.653	0.173	0.598	0.078	0.563	0.062	0.519	0.043	0.631	0.127
	RFscorev2(ODDT)	0.717	0.158	0.658	0.175	0.672	0.194	0.611	0.124	0.605	0.110	0.548	0.048	0.647	0.155
	RFscorev3(ODDT)	0.702	0.145	0.639	0.149	0.701	0.211	0.625	0.155	0.575	0.097	0.549	0.052	0.647	0.199
	PLECRF(ODDT)	0.719	0.295	0.640	0.165	0.717	0.218	0.604	0.072	0.532	0.074	0.549	0.089	0.605	0.181

Table 3. The mean AUROC and BEDROC of each MLSF towards different protein families of DEKOIS2.0

Docking program	Rescoring MLSF	Kinase		Protease		Nuclear receptor		GPCR		Hydrolase		Isomerase		Ligase		Oxidored-uctase		Transferase		Other	
		AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC
Glide SP	/	0.756	0.352	0.770	0.445	0.770	0.422	0.806	0.356	0.670	0.275	0.620	0.589	0.587	0.221	0.743	0.404	0.735	0.474	0.758	0.249
	deltaVinaRF	0.652	0.139	0.596	0.159	0.634	0.184	0.579	0.045	0.637	0.187	0.515	0.053	0.292	0.018	0.615	0.202	0.681	0.186	0.629	0.082
	deltaVinaXGB	0.623	0.106	0.642	0.175	0.654	0.236	0.641	0.064	0.616	0.157	0.580	0.346	0.280	0.000	0.632	0.235	0.635	0.193	0.545	0.002
	RFscore-VS	0.752	0.308	0.742	0.332	0.879	0.589	0.790	0.286	0.703	0.235	0.774	0.403	0.585	0.029	0.731	0.372	0.600	0.198	0.569	0.028
	pafnucy	0.709	0.144	0.542	0.088	0.575	0.040	0.605	0.016	0.534	0.069	0.496	0.000	0.396	0.000	0.516	0.061	0.562	0.032	0.749	0.155
	OnionNet	0.594	0.097	0.581	0.094	0.561	0.095	0.659	0.136	0.538	0.107	0.705	0.161	0.506	0.040	0.503	0.056	0.504	0.069	0.627	0.052
	RFscorev4	0.652	0.121	0.619	0.167	0.593	0.113	0.642	0.167	0.577	0.131	0.518	0.006	0.371	0.068	0.463	0.058	0.541	0.105	0.689	0.246
	RFscorev3	0.557	0.065	0.648	0.219	0.605	0.054	0.606	0.048	0.576	0.091	0.658	0.014	0.476	0.008	0.522	0.055	0.528	0.100	0.724	0.350
	NNscore1.0	0.577	0.064	0.693	0.166	0.724	0.251	0.611	0.086	0.612	0.106	0.774	0.121	0.754	0.507	0.644	0.185	0.531	0.100	0.502	0.001
	NNscore2.0	0.644	0.135	0.644	0.059	0.717	0.214	0.645	0.111	0.602	0.070	0.648	0.003	0.770	0.172	0.681	0.177	0.604	0.051	0.546	0.047
	NNscore(ODDT)	0.538	0.066	0.580	0.103	0.525	0.048	0.548	0.078	0.607	0.048	0.500	0.193	0.293	0.000	0.514	0.086	0.558	0.069	0.729	0.206
	RFscorev1(ODDT)	0.615	0.048	0.594	0.119	0.562	0.225	0.580	0.037	0.546	0.077	0.415	0.000	0.350	0.003	0.450	0.009	0.514	0.055	0.791	0.600
	RFscorev2(ODDT)	0.670	0.120	0.609	0.216	0.569	0.178	0.648	0.109	0.564	0.080	0.368	0.000	0.321	0.019	0.438	0.034	0.566	0.094	0.763	0.317
	RFscorev3(ODDT)	0.648	0.113	0.600	0.192	0.590	0.232	0.645	0.184	0.552	0.082	0.409	0.000	0.310	0.031	0.478	0.043	0.545	0.108	0.766	0.601
	PLECRF(ODDT)	0.666	0.219	0.611	0.193	0.603	0.131	0.578	0.165	0.602	0.222	0.775	0.182	0.374	0.109	0.510	0.079	0.494	0.028	0.614	0.009
GOLD CHEMPLP	/	0.650	0.176	0.706	0.275	0.607	0.109	0.651	0.215	0.631	0.133	0.599	0.017	0.481	0.262	0.631	0.169	0.621	0.093	0.621	0.163
	deltaVinaRF	0.645	0.144	0.552	0.103	0.592	0.155	0.552	0.004	0.653	0.187	0.581	0.056	0.498	0.041	0.601	0.237	0.611	0.118	0.611	0.254
	deltaVinaXGB	0.594	0.102	0.607	0.093	0.600	0.140	0.625	0.041	0.642	0.145	0.745	0.105	0.450	0.000	0.600	0.261	0.549	0.099	0.549	0.071
	RFscore-VS	0.727	0.244	0.737	0.223	0.855	0.508	0.767	0.261	0.706	0.172	0.860	0.160	0.519	0.007	0.711	0.359	0.591	0.274	0.591	0.000
	pafnucy	0.679	0.138	0.570	0.074	0.592	0.074	0.597	0.011	0.563	0.055	0.563	0.106	0.510	0.004	0.531	0.041	0.538	0.032	0.538	0.188
	OnionNet	0.575	0.089	0.578	0.089	0.554	0.096	0.653	0.092	0.519	0.056	0.729	0.208	0.510	0.002	0.516	0.063	0.485	0.049	0.485	0.329
	RFscorev4	0.598	0.095	0.641	0.103	0.580	0.118	0.613	0.036	0.591	0.077	0.613	0.001	0.617	0.001	0.489	0.015	0.508	0.105	0.508	0.189
	RFscorev3	0.496	0.067	0.638	0.139	0.568	0.051	0.547	0.039	0.588	0.093	0.688	0.059	0.770	0.010	0.514	0.044	0.541	0.102	0.541	0.231
	NNscore1.0	0.583	0.127	0.604	0.211	0.714	0.235	0.654	0.113	0.574	0.090	0.763	0.106	0.713	0.281	0.619	0.158	0.519	0.064	0.519	0.002
	NNscore2.0	0.634	0.086	0.602	0.091	0.701	0.205	0.663	0.071	0.590	0.103	0.689	0.021	0.690	0.130	0.649	0.142	0.598	0.058	0.598	0.044
	NNscore(ODDT)	0.513	0.053	0.593	0.090	0.492	0.050	0.509	0.025	0.577	0.056	0.501	0.001	0.348	0.009	0.474	0.055	0.478	0.067	0.478	0.230
	RFscorev1(ODDT)	0.582	0.047	0.575	0.094	0.563	0.189	0.570	0.020	0.551	0.044	0.522	0.000	0.459	0.000	0.449	0.010	0.520	0.074	0.520	0.170
	RFscorev2(ODDT)	0.625	0.129	0.613	0.163	0.554	0.167	0.627	0.071	0.569	0.076	0.533	0.001	0.507	0.001	0.432	0.035	0.555	0.067	0.555	0.143
	RFscorev3(ODDT)	0.610	0.101	0.596	0.124	0.563	0.217	0.638	0.086	0.568	0.076	0.540	0.000	0.550	0.000	0.478	0.022	0.529	0.061	0.529	0.225
	PLECRF(ODDT)	0.675	0.261	0.631	0.190	0.582	0.131	0.528	0.095	0.585	0.147	0.763	0.178	0.470	0.064	0.492	0.057	0.486	0.059	0.486	0.053
LeDock	/	0.748	0.311	0.566	0.130	0.591	0.054	0.656	0.111	0.617	0.086	0.262	0.000	0.609	0.072	0.572	0.120	0.673	0.180	0.728	0.246
	deltaVinaRF	0.618	0.123	0.597	0.095	0.606	0.129	0.588	0.045	0.647	0.180	0.347	0.002	0.449	0.076	0.600	0.188	0.634	0.145	0.672	0.194
	deltaVinaXGB	0.609	0.097	0.599	0.081	0.606	0.157	0.589	0.032	0.644	0.167	0.642	0.033	0.317	0.002	0.618	0.224	0.589	0.114	0.664	0.097
	RFscore-VS	0.758	0.244	0.733	0.223	0.872	0.508	0.755	0.261	0.693	0.172	0.746	0.160	0.566	0.007	0.703	0.359	0.626	0.274	0.458	0.000
	pafnucy	0.719	0.138	0.571	0.074	0.605	0.074	0.579	0.011	0.551	0.055	0.584	0.106	0.552	0.004	0.496	0.041	0.513	0.032	0.740	0.188
	OnionNet	0.607	0.103	0.566	0.061	0.554	0.122	0.608	0.102	0.526	0.038	0.751	0.160	0.557	0.027	0.467	0.026	0.471	0.037	0.726	0.184
	RFscorev4	0.646	0.095	0.586	0.103	0.600	0.118	0.590	0.036	0.561	0.077	0.474	0.001	0.341	0.001	0.438	0.015	0.561	0.105	0.681	0.189
	RFscorev3	0.550	0.067	0.605	0.139	0.597	0.051	0.550	0.039	0.549	0.093	0.603	0.059	0.577	0.010	0.500	0.044	0.561	0.102	0.820	0.231
	NNscore1.0	0.616	0.127	0.610	0.211	0.671	0.235	0.617	0.113	0.619	0.090	0.679	0.106	0.721	0.281	0.625	0.158	0.538	0.064	0.556	0.002

(Continued)

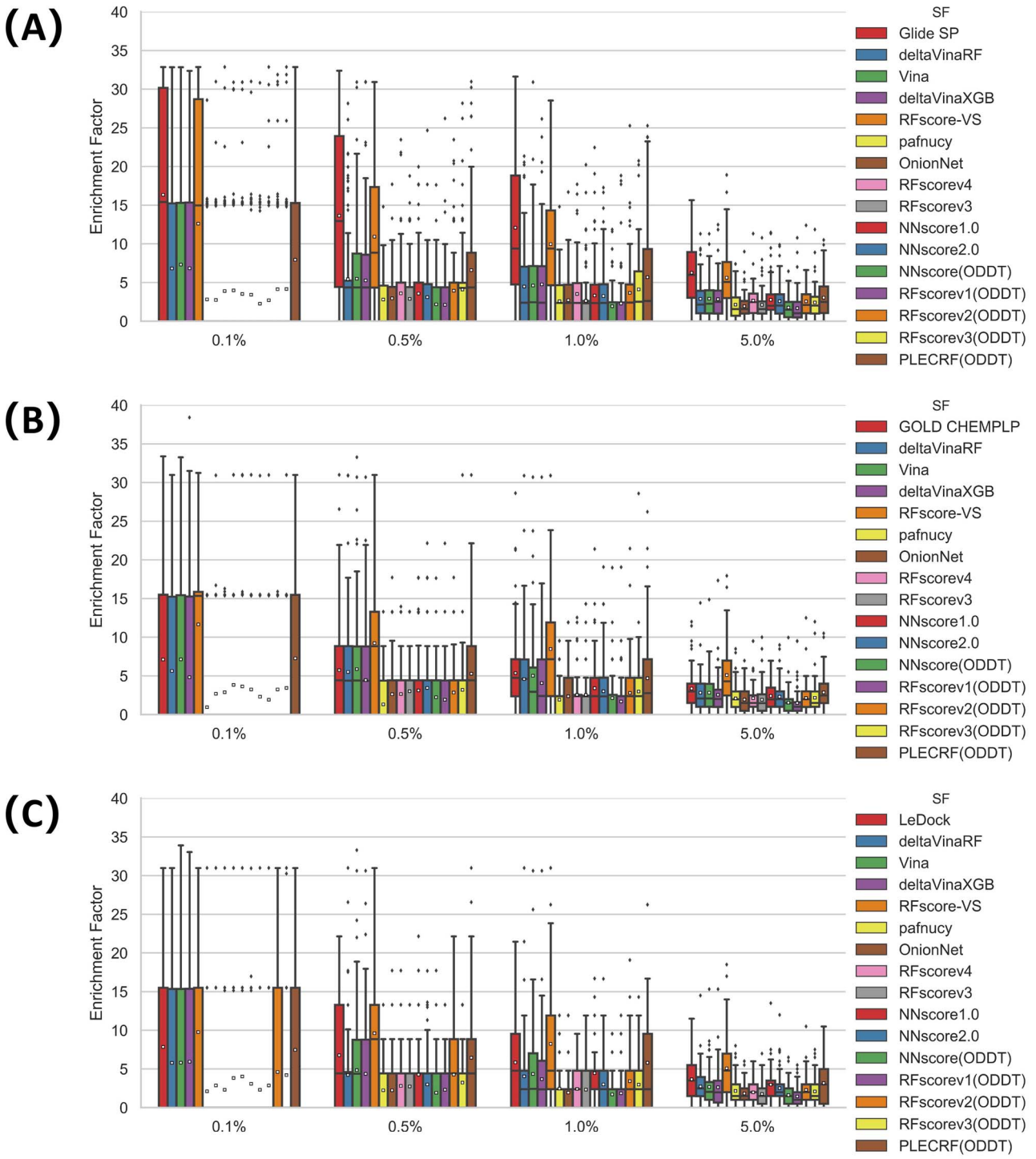


Figure 6. The EFs of 14 MLSFs on DEKOIS2.0 based on the docking poses predicted by (A) Glide SP, (B) GOLD CHEMPLP and (C) LeDock, respectively. The white squares in boxplots represent the mean values for each MLSF. Vina means using Vina SF to rescore the docking poses, and this score is generated by deltaVinaXGB.

leading to their excellent performance on the CASF benchmark but significantly worse performance here.

To further investigate whether RFscore-VS can surely improve the screening power, the sequence identities and structural similarities between the protein targets in DUD-E and DEKOIS2.0 were calculated with NW-align [75] and TM-Score [76], respectively. It should be noted that a same protein from different PDB entries may result in a low similarity score such as SRC existing in both two datasets. The calculation here is based on the whole protein rather than the binding pocket. For each protein in DEKOIS2.0, the highest score with this protein in DUD-E is assigned, and the results are depicted in Figure 7 and Supplementary Figure S18. Taking 60% as the cutoff (as most sequence identity or structural similarity scores are close to 100%), it is obvious that the targets with high similarity perform better than those with low similarity when using

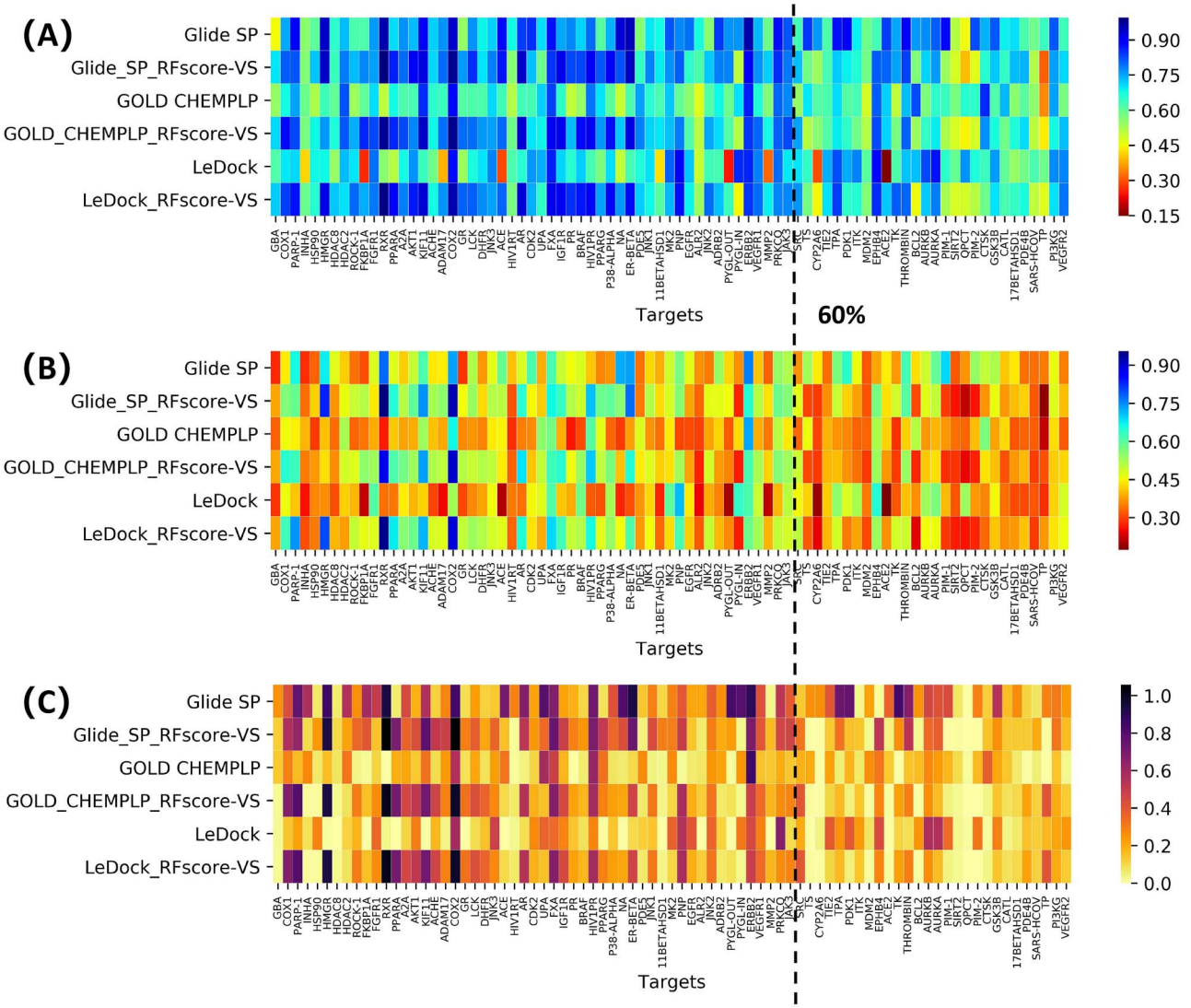


Figure 7. (A) AUROC, (B) LogAUC and (C) BEDROC of each target on DEKOIS2.0. The targets are sorted by their sequence identity with proteins in DUD-E with the descending order. The proteins in the left of the dotted line have a sequence identity over 60%, while those on the right have their scores below 60%.

RFscore-VS. However, the effect of sequence similarity does not influence classical SFs. To intuitively compare the difference, the quantitative comparison of the performance is also shown as boxplots in Figure 8 and Supplementary Figure S19. In terms of whichever metrics, there is almost no difference for classical SFs ($P > 0.05$ except the Glide SP split by 60% structural similarity) while significant difference for RFscore-VS. Hence, it can be concluded that RFscore-VS may perform well to the targets that are either present in the training set or similar to some proteins in the training set, but for novel or dissimilar protein targets, the performance might be unsatisfactory. The construction of RFscore-VS was trained directly by linearly fitting the data of the whole DUD-E dataset, in which all the decoys were assigned an identical inactive value of $pK_{d/i} = 5.95$. The involvement of large numbers of decoys in the training set could better differentiate inactive compounds and thereby obtain acceptable screening power on most similar targets. However, unlike PDBbind that collects various types of targets, only 102 targets are assembled in DUD-E, implying its applicability domain may be still limited towards those significantly novel targets. Hopefully, such limitation can be alleviated in the future with more available data.

The mean AUROC and BEDROC values on the individual target families are reported in Table 3 and Supplementary Figure S20. Unlike DUD-E, the targets in DEKOIS2.0 can be classified into 10 families, and the other enzymes in DUD-E are further split into hydrolase, isomerase, ligase, oxidoreductase and transferase here. There is only one target belonging to isomerase, ligase and other classes each, so their results may be not convincing enough to be included in our further analysis. Among the three classical SFs, Glide SP still yields the best performance with the AUROC and BEDROC values higher than 0.73 and 0.35, respectively, for almost all target families. By contrast, GOLD CHEMPLP produces rather bad scores with the best AUROC of 0.706 for proteases, and LeDock only shows acceptable results for kinases. The best-performed MLSF RFscore-VS only shows higher scores than Glide SP for nuclear receptors with the AUROC and BEDROC values of exceeding 0.85 and 0.50, respectively, which is similar to the excellent performance in DUD-E for the same family. However, for the

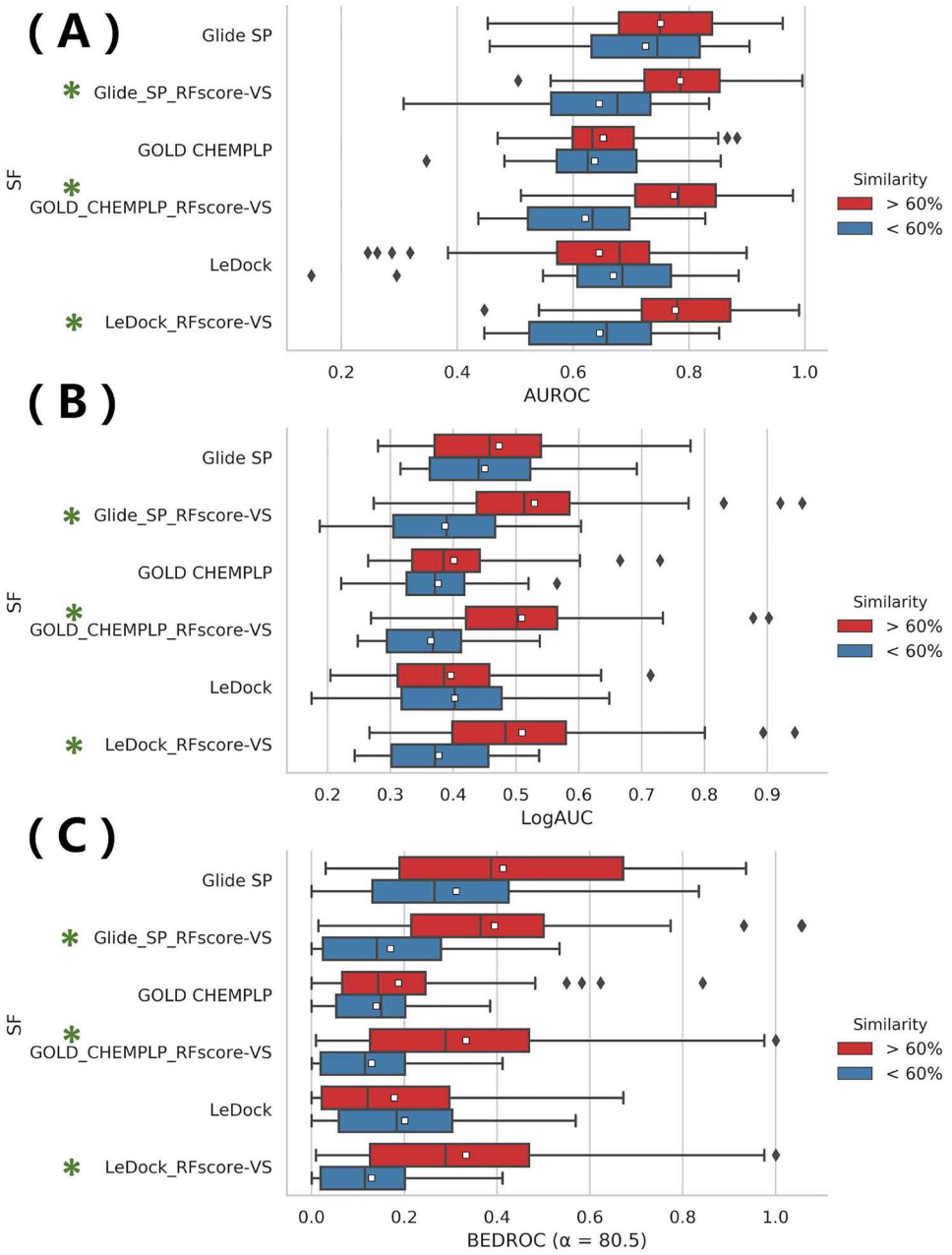


Figure 8. The average performance in terms of (A) AUROC, (B) LogAUC and (C) BEDROC on DEKOIS2.0. Red and blue boxes represent the targets with their sequence identity scores above 60% and below 60%, respectively. The white squares in boxplots represent the mean values for each MLSF. ∗Statistical difference is observed according to Mann–Whitney test.

other families, RFscore-VS exhibits worse performance. Such poor performance might be contributed by the introduction of some targets dissimilar to the targets in DUD-E. For other MLSFs, despite their generally poor performance, decent results can also be found for certain families, such as NNscore1.0 and NNscore2.0 on nuclear receptors and PLECRF(ODDT) on kinases.

Impacts of docking poses on the performance of a certain MLSF

It has been shown previously that the screening power of the same MLSF may vary dramatically among different docking programs. We wonder how the docking poses would influence the performance of a certain MLSF. Accordingly, the performances are compared among the three docking programs in a pairwise manner, leading to three pairs for each MLSF. Then, the Pearson's correlation coefficient (R_p) of each pair is calculated, where both the actives and decoys are included to obtain an overall result. The absolute values of R_p towards the 102 targets in DUD-E and 81 targets in DEKOIS2.0 are reported as the heat maps in Figure 9 and Supplementary Figure S21, respectively. On the basis of the first three rows, it can be found that the correlations between any two of the docking scores of Glide SP, GOLD CHEMPLP and LeDock are mostly intermediate, suggesting that SFs from different docking programs vary moderately. However, much higher correlations are observed between

Table 3. Continued

Docking program	Rescoring MLSF	Kinase		Protease		Nuclear receptor		GPCR		Hydrolase		Isomerase		Ligase		Oxidored-uctase		Transferase		Other	
		AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC	AU ROC	BED ROC
NNscore2.0	0.629	0.086	0.610	0.091	0.656	0.205	0.660	0.071	0.605	0.103	0.586	0.021	0.678	0.130	0.657	0.142	0.598	0.058	0.641	0.044	
NNscore(ODDT)	0.553	0.053	0.539	0.090	0.556	0.050	0.513	0.025	0.603	0.056	0.485	0.001	0.271	0.009	0.480	0.055	0.509	0.067	0.795	0.230	
RFscorev1(ODDT)	0.604	0.047	0.583	0.094	0.586	0.189	0.573	0.020	0.530	0.044	0.411	0.000	0.304	0.000	0.420	0.010	0.552	0.074	0.678	0.170	
RFscorev2(ODDT)	0.669	0.129	0.610	0.163	0.584	0.167	0.623	0.071	0.546	0.076	0.460	0.001	0.265	0.001	0.408	0.035	0.590	0.067	0.748	0.143	
RFscorev3(ODDT)	0.643	0.101	0.578	0.124	0.600	0.217	0.607	0.086	0.536	0.076	0.367	0.000	0.313	0.000	0.448	0.022	0.572	0.061	0.719	0.225	
PLECRF(ODDT)	0.700	0.261	0.580	0.190	0.599	0.131	0.528	0.095	0.557	0.147	0.656	0.178	0.487	0.064	0.474	0.057	0.503	0.059	0.720	0.053	

docking programs for most RFscore-related MLSFs, including RFscorev4, RFscorev3, RFscorev1(ODDT), RFscorev2(ODDT) and RFscorev3(ODDT), as shown in Figure 9. It indicates that these approaches may be not so sensitive to the differences between different binding poses. We assume it may be explained by the fact that they mainly utilize the frequency occurrence of atomic pairs as the features to train the models. With slight changes of the binding poses, the atomic pair counts within a given distance hardly change, thus resulting in an almost unchanged score of a certain ligand as long as the poses generated by different docking programs are not significantly different. Besides, these MLSFs are constructed by training PDBbind dataset as regression models and show excellent scoring power with their R_ps close to 0.80. RFscore-VS also utilizes the features from RFscore to characterize protein–ligand interactions, and the number of the targets with relatively high correlations for RFscore-VS are less than those for the other RFscore-related MLSFs but still much more than those for most MLSFs. The most prominent advantage of these highly correlated approaches is that they may be hardly affected by the quality of docking programs or the relatively low structural flexibility of protein-binding pockets. However, a drawback is that they tend to easily overfit due to their too simple feature representation. Therefore, how to balance these two aspects may be a promising direction in the future. Among the other MLSFs, NNscore1.0 yields significantly low correlations. It is highly likely that some extremely bad poses are given a score of −999999.9. The R_p values of NNscore2.0, deltaVinaRF and deltaVinaXGB are also always low, and similarly, it may be caused by the introduction of decoys into the training sets. For the latter two MLSFs, another reason may be their special parameterization method to fit a correct term rather than the final score. In summary, different docking poses can indeed exert large influences on most MLSFs studied here.

Next, we further explore whether using the top three docking poses rather than the top one could improve the screening power of MLSFs. The results of the seven selected MLSFs based on the binding poses generated by Glide SP and GOLD CHEMPLP are reported in Figure 10. However, performances are quite comparable with rather limited improvements for several MLSFs. Although more appropriate binding poses are more likely to be predicted for some active compounds, the improvement of the scores of decoys cannot be neglected also. Therefore, it seems that using the top three docking poses may to some extent increase the prediction quality of binding poses, but it is not a reliable strategy to improve the screening power.

Impacts of PDBbind versions of the training set on the performance of MLSFs

Some doubts may have been raised towards the reliability to assess the performance of MLSFs with different versions of training sets, so here we also conducted a simple experiment to explore the impacts of different PDBbind versions. Six different versions of PDBbind available in ODDT were utilized as the training sets, and the performance for four representative MLSFs based on the Glide and GOLD docking poses is shown in Figure 11. With the update of PDBbind, the performance improves slightly and then tends to be stable [especially for RFscorev2(ODDT) and RFscorev3(ODDT)]. However, the improvement is so small, and obvious statistical significance cannot be found among any two versions according to the *post hoc* Nemernyi test. Therefore, it can be concluded that the

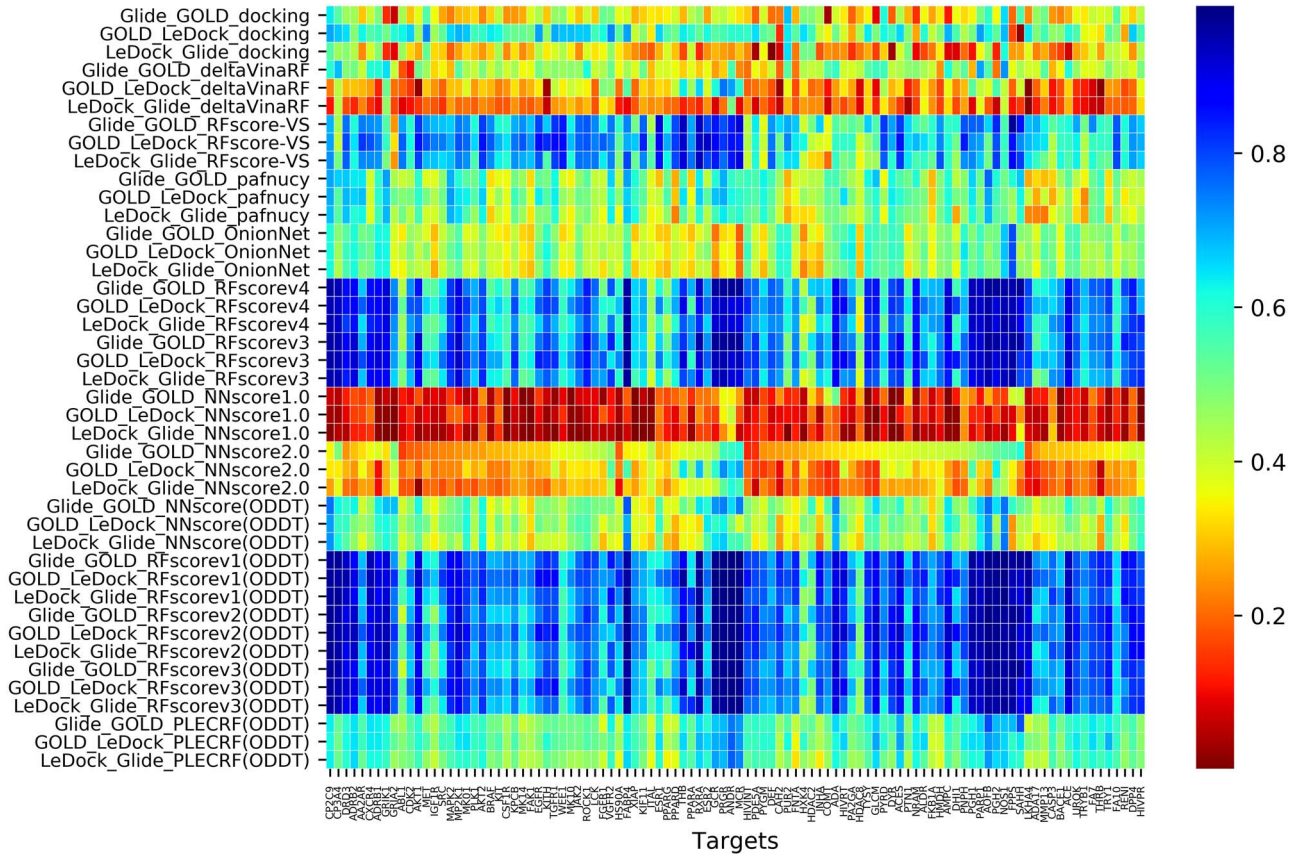


Figure 9. Correlations of the scores based on the binding poses generated from different docking programs over different targets in DUD-E. The correlation is represented by the absolute value of Pearson's correlation coefficient.

versions of the training set may indeed affect the performance but these contributions are still not decisive.

Assessment of screening power on dataset III

For all MLSFs, a simple simulation to mimic the real VS campaigns is also carried out. The performance assessed by AUROC, BEDROC and $EF_{0.5\%}$ is listed for the eight representative targets in Table 4 and the other metrics in Supplementary Table S3. In addition, the ROC and semilog ROC curves are shown in Supplementary Figures S22 and S23, respectively. In fact, dataset III is expected to be more complicated than DUD-E and DEKOIS2.0, because the decoys in the latter two sets possess similar physicochemical properties with the active compounds, whereas the compounds from dataset III are more structurally diverse. Among all the eight targets, the highest screening power is observed for almost all MLSFs against FA10 and AT1R. Intrinsically, for some targets, the corresponding actives are easily differentiated from the negative controls. However, whether the binding poses of these ligands are correctly predicted still remains unclear. In general, RFscore-VS shows better performance for five targets present in DUD-E than other two targets (NIK and VDR) that are unique to dataset III, indicating possible limitation of RFscore-VS for novel targets. Furthermore, distinct performances of varying magnitude are observed for targets from the same family over different MLSFs, such as AA2AR versus AT1R (GPCR) and ESR2 versus VDR (nuclear receptors). It should be noted that NIK generally yields the worst performance among all targets, which might be related to the fact that most NIK inhibitors belong to type $I^{1/2}$ inhibitors that often rarely occur in other kinases [77]. As a consequence, MLSFs cannot learn sufficient information from the training sets. On the contrary, Glide SP can well represent the interactions by empirical analysis. Thus, another concern towards the MLSFs arises: how to deal with ligands bound to the unconventional binding pockets. The most direct solution might be to include as many representative cases in the training sets as possible. Instead of individual targets, the best average performance over all targets is still observed for Glide SP, irrespective of the calculated metrics (except LogAUC; Table 4 and Supplementary Table S3, despite $P > 0.10$). Therefore, RFscore-VS seems to own little superiority over the baseline of Glide SP. Taken together, despite that MLSFs have been widely reported to be able to significantly improve the performance of classical SFs, their poor generalization capability and applicability in real practice represent additional challenges and require further understanding and development.

Conclusion

Herein, a comprehensive and extensive assessment of screening power has been conducted for 14 existing MLSFs on the basis of three data sets. For DUD-E and DEKOIS2.0 data sets, most of these MLSFs can hardly achieve a satisfactory result according to their average performance, despite that marginally better performance than the baseline of classical SFs is observed for certain targets. RFscore-VS that is trained on DUD-E often

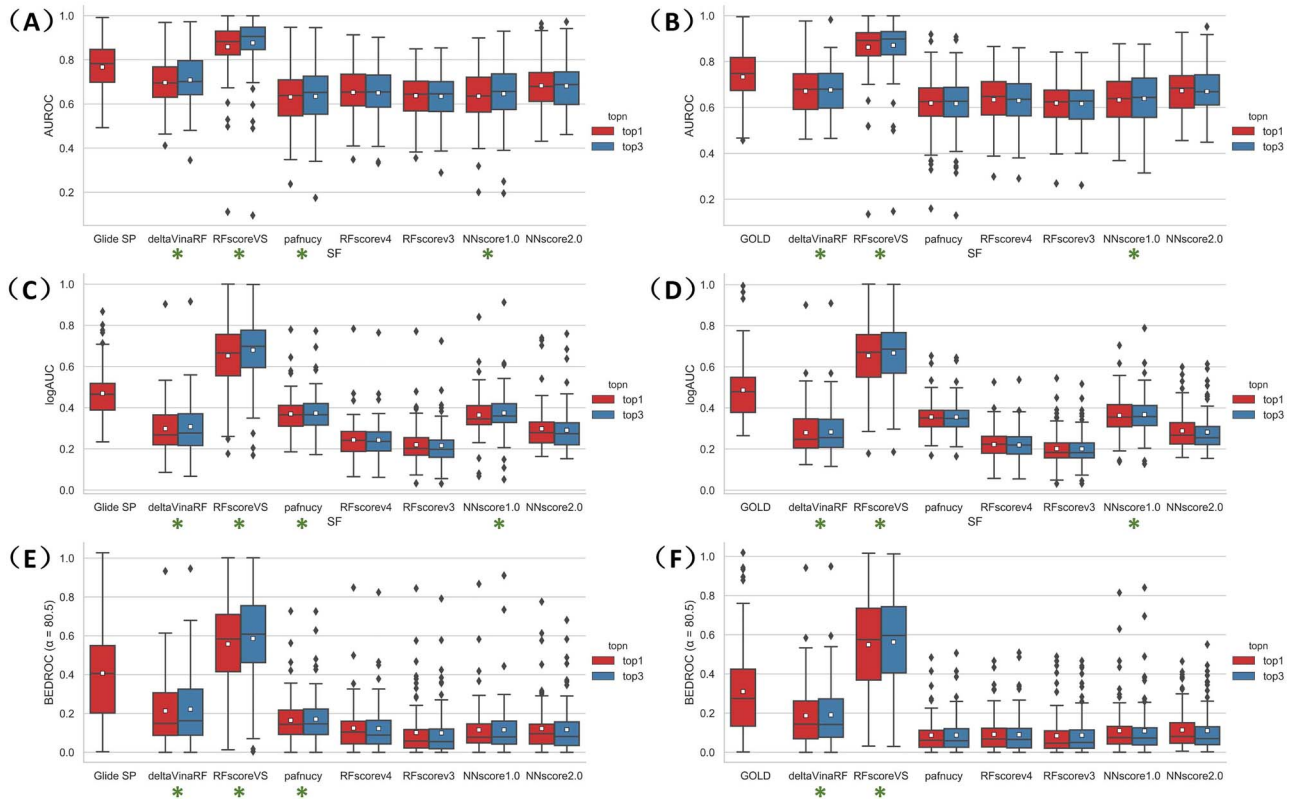


Figure 10. Comparison of the screening power of seven MLSFs on DUD-E in terms of (A, B) AUROC, (C, D) LogAUC and (E, F) BEDROC (α = 80.5) when using the top one and the top three docking poses for rescoring. (A), (C) and (E) are based on the docking poses of Glide SP and (B), (D) and (F) are based on GOLD CHEMPLP. ∗Statistical difference is observed according to Wilcoxon signed-rank test.

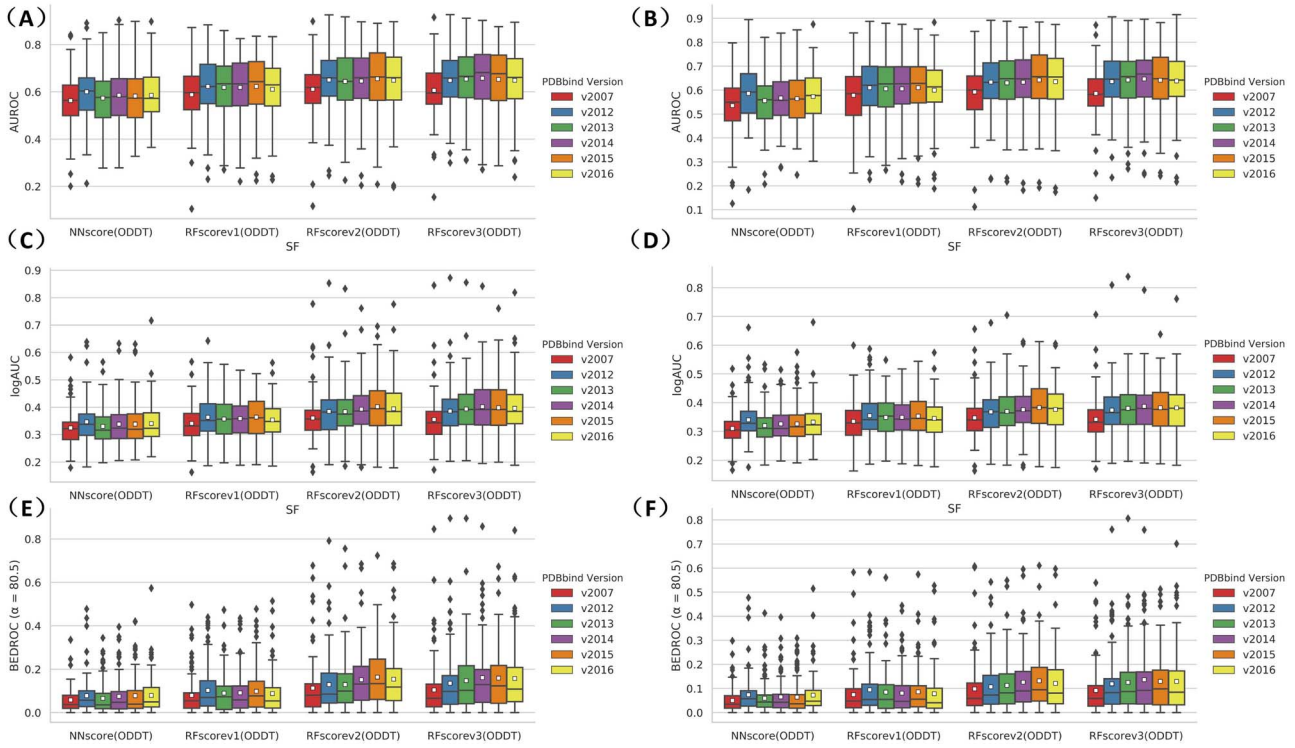


Figure 11. Comparison of the screening power of four MLSFs implemented in ODDT on DUD-E in terms of (A, B) AUROC, (C, D) LogAUC and (E, F) BEDROC (α = 80.5) when using different versions of PDBbind as the training sets. (A), (C) and (E) are based on the docking poses of Glide SP and (B), (D) and (F) are based on GOLD CHEMPLP.

Table 4. The AUROC, BEDROC (α = 80.5) and $EF_{0.5\%}$ of the assessment of several MLSFs towards eight representative targets in a real VS campaign

SFs	AA2AR	ESR2	FA10	LCK	PARP1	AT1R	NIK	VDR	average
AUROC									
Glide SP	0.830	0.933	0.948	0.894	0.964	0.901	0.852	0.825	0.893
deltaVinaRF	0.646	0.840	0.901	0.823	0.831	0.944	0.815	0.663	0.808
RFscore-VS	0.848	0.928	0.969	0.937	0.900	0.966	0.773	0.798	0.890
pafnucy	0.711	0.729	0.789	0.833	0.670	0.978	0.774	0.563	0.756
OnionNet	0.673	0.789	0.821	0.735	0.691	0.882	0.630	0.732	0.744
RFscorev4	0.734	0.813	0.944	0.865	0.674	0.973	0.742	0.916	0.833
RFscorev3	0.589	0.797	0.941	0.762	0.594	0.951	0.694	0.868	0.774
NNscore1.0	0.681	0.806	0.832	0.765	0.740	0.818	0.747	0.689	0.760
NNscore2.0	0.733	0.812	0.880	0.765	0.703	0.895	0.710	0.511	0.751
NNscore(ODDT)	0.585	0.687	0.906	0.715	0.685	0.879	0.592	0.889	0.742
RFscorev1(ODDT)	0.797	0.702	0.946	0.805	0.587	0.932	0.773	0.908	0.806
RFscorev2(ODDT)	0.833	0.726	0.962	0.871	0.628	0.991	0.779	0.945	0.842
RFscorev3(ODDT)	0.830	0.764	0.964	0.852	0.616	0.982	0.801	0.940	0.844
PLECRF(ODDT)	0.613	0.846	0.883	0.811	0.655	0.924	0.726	0.930	0.799
BEDROC									
Glide SP	0.229	0.644	0.767	0.458	0.751	0.587	0.525	0.452	0.552
deltaVinaRF	0.013	0.264	0.480	0.244	0.221	0.566	0.223	0.120	0.266
RFscore-VS	0.398	0.544	0.827	0.540	0.546	0.754	0.129	0.184	0.490
pafnucy	0.053	0.112	0.111	0.285	0.011	0.645	0.167	0.031	0.177
OnionNet	0.053	0.112	0.111	0.285	0.011	0.645	0.167	0.084	0.184
RFscorev4	0.052	0.182	0.544	0.438	0.087	0.670	0.132	0.448	0.319
RFscorev3	0.029	0.104	0.560	0.092	0.064	0.599	0.066	0.094	0.201
NNscore1.0	0.085	0.168	0.204	0.150	0.103	0.185	0.110	0.315	0.165
NNscore2.0	0.054	0.183	0.254	0.097	0.112	0.262	0.113	0.072	0.143
NNscore(ODDT)	0.012	0.188	0.309	0.111	0.180	0.432	0.065	0.196	0.187
RFscorev1(ODDT)	0.060	0.179	0.314	0.095	0.012	0.229	0.038	0.405	0.167
RFscorev2(ODDT)	0.159	0.297	0.686	0.437	0.054	0.875	0.159	0.513	0.397
RFscorev3(ODDT)	0.188	0.335	0.634	0.352	0.073	0.851	0.243	0.499	0.397
PLECRF(ODDT)	0.061	0.369	0.468	0.356	0.083	0.393	0.174	0.462	0.296

(Continued)

performs better than other MLSFs. However, for novel targets that are absent or dissimilar to the proteins in the training sets, the resulting performance might be significantly worsened. Furthermore, several RFscore-related MLSFs are found to be relatively insensitive to docking programs that generate binding poses, whereas other methods may be largely influenced. While using the top three docking poses rather than the top one for rescoring, or employing an updated version of the training set, no obvious improvement of screening power is observed. We also explored the performance of these MLSFs in real VS campaigns. Unfortunately, they still do not show better performance than classical Glide SP.

Using ML technologies to predict protein–ligand binding affinities or conduct SBVS campaigns has become increasingly

Table 4. Continued

$EF_{0.5\%}$									
Glide SP	22.079	86.037	113.951	59.569	91.990	85.321	76.145	68.721	75.477
deltaVinaRF	0.000	26.256	61.868	24.364	24.989	75.971	27.392	14.526	31.921
RFscore-VS	51.184	76.928	115.950	65.627	74.992	112.136	8.122	10.573	64.439
pafnucy	5.018	8.098	8.996	32.309	0.000	79.227	16.244	3.965	19.232
OnionNet	6.022	21.256	9.996	3.029	3.000	28.034	3.046	9.251	10.454
RFscorev4	4.014	21.256	62.973	48.463	8.999	101.166	13.198	71.365	41.429
RFscorev3	2.999	3.037	69.970	6.058	5.999	95.072	6.092	5.286	24.314
NNscore1.0	4.014	11.134	9.996	13.125	9.999	19.502	9.137	47.576	15.561
NNscore2.0	3.011	16.195	23.990	5.048	7.999	25.596	9.137	0.000	11.372
NNscore(ODDT)	0.000	22.269	20.991	8.077	19.998	56.068	5.076	22.467	19.368
RFscorev1(ODDT)	1.004	25.305	11.995	5.048	0.000	10.970	1.015	46.255	12.699
RFscorev2(ODDT)	9.032	38.464	87.962	49.472	5.999	131.638	15.229	79.294	52.136
RFscorev3(ODDT)	17.061	41.500	86.962	41.395	9.999	142.608	28.428	70.043	54.750
PLECRF(ODDT)	8.029	44.537	64.972	42.405	9.999	39.004	21.321	63.435	36.713

popular and been applied routinely. However, from this study, we would propose that one might need to be careful of directly using the generic MLSFs for VS. A major concern may relate to their poor generalization capability, which is reflected by how to choose an appropriate training set to construct them. MLSFs trained on PDBbind have been proven to be inapplicable to VS, and RFscore-VS shows poor screening power on targets that are novel or dissimilar to proteins within the training sets, as well as targets with unconventional binding pockets. Thus, the development of MLSFs seems to reach the bottleneck. Target-specific MLSFs may provide potential solutions. However, whether this type of methods is consistently better than ligand-based strategies remains unclear [78, 79], as the introduction of molecular docking itself may bring in uncertainty. Besides, how to choose an appropriate dataset for target-specific SFs is also faced with a great challenge, as most widely used benchmark datasets for classical SFs such as DUD-E and DUD have shown their inherent biases, and thereby are hard to be trained by ML methods [74, 80]. Taken together, it seems there is still space to further investigate and develop MLSFs with high precision and balanced applicability. We expect this paper may provide some valuable guidance and drive some caution while using these MLSFs in real VS campaigns.

Key Points

- A systematic assessment was carried out to re-evaluate the effectiveness of 14 reported machine learning-based scoring functions in virtual screening.
- According to the predictions to the dataset DUD-E, DEKOIS2.0 or dataset III constructed by ourselves, most of the tested machine learning-based scoring functions could hardly achieve satisfactory results, and they could even not outperform the baseline of classical scoring functions such as Glide SP.
- RFscore-VS trained on the DUD-E dataset could show its superiority for most targets, but it clearly illustrated rather limited performance on the targets that were dissimilar to the proteins in the corresponding training sets.
- Using the top three docking poses rather than the top one for rescoring and the re-trained models with the updated versions of the training set could not gain significant improvements.
- The use of the reported machine learning-based scoring functions for virtual screening should be cautious.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Dataset availability

The docking poses, docking/rescoring scores and some representative scripts used in this study can be found at https://112.17.133.17:8888/downloads/BIB_dataset.tar.gz.

Funding

National Key R&D Program of China (2016YFA0501701); Key R&D Program of Zhejiang Province (2020C03010); National Natural Science Foundation of China (81773632).

Conflict of interest

There are no conflicts to declare.

References

1. da Silva Rocha SFL, Olanda CG, Fokoue HH, *et al*. Virtual screening techniques in drug discovery: review and recent applications. *Curr Top Med Chem* 2019;**19**:1751–67.
2. Wang Z, Sun H, Shen C, *et al*. Combined strategies in structure-based virtual screening. *Phys Chem Chem Phys* 2020;**22**:3149–59.
3. Rifaioglu AS, Atas H, Martin MJ, *et al*. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform* 2019;**20**:1878–912.
4. Guedes IA, Pereira FSS, Dardenne LE. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front Pharmacol* 2018;**9**: 1089.
5. Hou TJ, Xu XJ. Recent development and application of virtual screening in drug discovery: an overview. *Curr Pharm Des* 2004;**10**:1011–33.
6. Kitchen DB, Decornez H, Furr JR, *et al*. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 2004;**3**:935–49.
7. Cheng T, Li Q, Zhou Z, *et al*. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J* 2012;**14**:133–41.
8. Ain QU, Aleksandrova A, Roessler FD, *et al*. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* 2015;**5**:405–24.
9. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* 2010;**26**:1169–75.
10. Ballester PJ, Schreyer A, Blundell TL. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J Chem Inf Model* 2014;**54**:944–55.
11. Wang C, Zhang Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *J Comput Chem* 2017;**38**:169–77.
12. Wojcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci Rep* 2017;**7**:46710.
13. Ding B, Wang J, Li N, *et al*. Characterization of small molecule binding. I. Accurate identification of strong inhibitors in virtual screening. *J Chem Inf Model* 2013;**53**:114–22.
14. Yan Y, Wang W, Sun Z, *et al*. Protein-ligand empirical interaction components for virtual screening. *J Chem Inf Model* 2017;**57**:1793–806.
15. Durrant JD, McCammon JA. NNScore: a neural-network-based scoring function for the characterization of protein-ligand complexes. *J Chem Inf Model* 2010;**50**:1865–71.
16. Durrant JD, McCammon JA. NNScore 2.0: a neural-network receptor-ligand scoring function. *J Chem Inf Model* 2011;**51**:2897–903.
17. Lu J, Hou X, Wang C, *et al*. Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. *J Chem Inf Model* 2019;**59**:4540–9.
18. Nguyen DD. Wei G-W. AGL-score: algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *J Chem Inf Model* 2019;**59**:3291–304.
19. Cang Z, Mu L, Wei G-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput Biol* 2018;**14**: e1005929.
20. Pereira JC, Caffarena ER, dos Santos CN. Boosting docking-based virtual screening with deep learning. *J Chem Inf Model* 2016;**56**:2495–506.
21. Ragoza M, Hochuli J, Idrobo E, *et al*. Protein-ligand scoring with convolutional neural networks. *J Chem Inf Model* 2017;**57**:942–57.
22. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics (Oxford, England)* 2018;**34**:3666–74.
23. Zheng L, Fan J, Mu Y. OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction. *Acs Omega* 2019;**4**:15956–65.
24. Shen C, Ding J, Wang Z, *et al*. From machine learning to deep learning: advances in scoring functions for protein-ligand docking. *Wiley Interdiscip Rev Comput Mol Sci* 2019;e1429.
25. Sun H, Pan P, Tian S, *et al*. Constructing and validating high-performance MIEC-SVM models in virtual screening for kinases: a better way for actives discovery. *Sci Rep* 2016; **6**:24817.
26. Durrant JD, Amaro RE. Machine-learning techniques applied to antibacterial drug discovery. *Chem Biol Drug Des* 2015;**85**:14–21.
27. Durrant JD, Carlson KE, Martin TA, *et al*. Neural-network scoring functions identify structurally novel estrogen-receptor ligands. *J Chem Inf Model* 2015;**55**:1953–61.
28. Hsieh C-H, Li L, Vanhauwaert R, *et al*. Miro1 marks Parkinson's disease subset and Miro1 reducer rescues neuron loss in Parkinson's models. *Cell Metab* 2019;**30**:1131–1140.e7.
29. Zhang L, Ai H-X, Li S-M, *et al*. Virtual screening approach to identifying influenza virus neuraminidase inhibitors using molecular docking combined with machine-learning-based scoring function. *Oncotarget* 2017;**8**:83142–54.
30. 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**:2807–15.
31. Li Y, Liu Z, Li J, *et al*. Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. *J Chem Inf Model* 2014;**54**:1700–16.
32. Ashtawy HM, Mahapatra NR. Task-specific scoring functions for predicting ligand binding poses and affinity and for screening enrichment. *J Chem Inf Model* 2018;**58**:119–33.
33. Wang RX, Fang XL, Lu YP, *et al*. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* 2004;**47**:2977–80.
34. Shen C, Hu Y, Wang Z, *et al*. Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions. *Brief Bioinform* 2020.
35. Mysinger MM, Carchia M, Irwin JJ, *et al*. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* 2012;**55**:6582–94.
36. Bauer MR, Ibrahim TM, Vogel SM, *et al*. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0-a public library of challenging docking benchmark sets. *J Chem Inf Model* 2013;**53**:1447–62.
37. Mendez D, Gaulton A, Bento AP, *et al*. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**: D930–40.
38. Irwin JJ, Shoichet BK. ZINC - a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;**45**:177–82.

39. Hawkins PCD, Skillman AG, Warren GL, *et al*. Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. *J Chem Inf Model* 2010;**50**:572–84.
40. Berman HM, Westbrook J, Feng Z, *et al*. The protein data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
41. Sastry GM, Adzhigirey M, Day T, *et al*. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* 2013;**27**:221–34.
42. Kaminski GA, Friesner RA, Tirado-Rives J, *et al*. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 2001;**105**:6474–87.
43. Olsson MHM, Sondergaard CR, Rostkowski M, *et al*. PROPKA3: consistent treatment of internal and surface residues in empirical pK(a) predictions. *J Chem Theory Comput* 2011;**7**:525–37.
44. Liu T, Lin Y, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2007;**35**:D198–201.
45. *Schrödinger Release 2019-1: LigPrep*. New York, NY: Schrödinger, LLC, 2019.
46. Baell JB, Holloway GA. New substructure filters for removal of Pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 2010;**53**:2719–40.
47. Walters WP, Murcko A, Murcko MA. Recognizing molecules with drug-like properties. *Curr Opin Chem Biol* 1999;**3**:384–7.
48. Duan J, Dixon SL, Lowrie JF, *et al*. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J Mol Graph Model* 2010;**29**:157–70.
49. Oprea TI. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* 2000;**14**:251–64.
50. Lipinski CA, Lombardo F, Dominy BW, *et al*. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 2012;**64**:4–17.
51. *Discovery Studio 2.5 Guide*. San Diego: Accelrys Inc., 2009, http://www.accelrys.com.
52. Zhang H, Unal H, Gati C, *et al*. Structure of the angiotensin receptor revealed by serial femtosecond crystallography. *Cell* 2015;**161**:833–44.
53. Castanedo GM, Blaquiere N, Beresini M, *et al*. Structure-based Design of Tricyclic NF-kappa B inducing kinase (NIK) inhibitors that have high selectivity over Phosphoinositide-3-kinase (PI3K). *J Med Chem* 2017;**60**:627–40.
54. Tocchini-Valentini G, Rochel N, Wurtz JM, *et al*. Crystal structures of the vitamin D nuclear receptor liganded with the vitamin D side chain analogues calcipotriol and seocalcitol, receptor agonists of clinical importance. Insights into a structural basis for the switching of calcipotriol to a receptor antagonist by further side chain modification. *J Med Chem* 2004;**47**:1956–61.
55. Wang Z, Sun H, Yao X, *et al*. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys Chem Chem Phys* 2016;**18**:12964–75.
56. Friesner RA, Banks JL, Murphy RB, *et al*. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* 2004;**47**:1739–49.
57. Jones G, Willett P, Glen RC, *et al*. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 1997;**267**:727–48.
58. Zhang N, Zhao H. Enriching screening libraries with bioactive fragment space. *Bioorg Med Chem Lett* 2016;**26**:3594–7.
59. Li H, Leung K-S, Wong M-H, *et al*. Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Mol Inform* 2015;**34**:115–26.
60. Li H, Leung K-S, Wong M-H, *et al*. Correcting the impact of docking pose generation error on binding affinity prediction. *BMC Bioinform* 2016;**17**:308.
61. Morris GM, Huey R, Lindstrom W, *et al*. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 2009;**30**:2785–91.
62. Wojcikowski M, Zielenkiewicz P, Siedlecki P. Open drug discovery toolkit (ODDT): a new open-source player in the drug discovery field. *J Cheminformatics* 2015;**7**:26.
63. Wojcikowski M, Kukielka M, Stepniewska-Dziubinska MM, *et al*. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* 2019;**35**:1334–41.
64. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.
65. Mysinger MM, Shoichet BK. Rapid context-dependent ligand desolvation in molecular docking. *J Chem Inf Model* 2010;**50**:1561–73.
66. Jain AN, Nicholls A. Recommendations for evaluation of computational methods. *J Comput Aided Mol Des* 2008;**22**:133–9.
67. Truchon J-F, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model* 2007;**47**:488–508.
68. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
69. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: a structure for efficient numerical computation. *Comput Sci Eng* 2011;**13**:22–30.
70. Nemenyi P. Distribution-free multiple comparisons. *Biometrics* 1962;**18**:263.
71. Terpilowski M. Scikit-posthocs: pairwise multiple comparison tests in python. *J Open Source Softw* 2019;**4**:1169.
72. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther* 2013;**138**:103–41.
73. Chaput L, Martinez-Sanz J, Saettel N, *et al*. Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J Cheminformatics* 2016;**8**:56.
74. Chen L, Cruz A, Ramsey S, *et al*. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLos One* 2019;**14**:e0220113.
75. Zhang Y. NW-align. http://zhanglab.ccmb.med.umich.edu/NW-align/. 2019.
76. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;**57**:702–10.
77. Shen C, Liu H, Wang X, *et al*. Importance of incorporating protein flexibility in molecule modeling: a theoretical

study on type I-1/2 NIK inhibitors. *Front Pharmacol* 2019; **10**:345.
78. Gonczarek A, Tomczak JM, Zareba S, *et al*. Interaction prediction in structure-based virtual screening using deep learning. *Comput Biol Med* 2018;**100**:253–8.
79. Morrone JA, Weber JK, Huynh T, *et al*. Combining docking pose rank and structure with deep learning improves protein-ligand binding mode prediction over a baseline docking approach. *J Chem Inf Model* 2020. https://doi.org/10.1021/acs.jcim.9b00927.
80. 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**: 947–61.