SCORCH: Improving virtual screening with a consensus of machine learning classifiers, data augmentation, and uncertainty estimation

Miles McGibbon[a,*], Sam Money-Kyrle[a,*], Vincent Blay[a,b**], Douglas R. Houston[a,**]

[a]Institute of Quantitative Biology, Biochemistry and Biotechnology, University of Edinburgh, Edinburgh, Scotland EH9 3BF, UK.

[b]Department of Microbiology and Environmental Toxicology, University of California at Santa Cruz, Santa Cruz, CA, 95064, USA.

Abstract

Introduction

The discovery of a new drug is a costly and lengthy endeavour. The computational prediction of which small molecules can bind to a protein target can accelerate this process if the predictions are fast and accurate enough. Recent machine-learning scoring functions re-evaluate the output of molecular docking to achieve more accurate predictions. However, previous scoring functions were trained on crystalised protein-ligand complexes and datasets of decoys. The limited availability of crystal structures and biases in the decoy datasets can lower the performance of scoring functions.

Objectives

To address key limitations of previous scoring functions and thus improve the predictive performance of structure-based virtual screening.

Methods

A novel machine-learning scoring function was created, named SCORCH (Scoring COnsensus for RMSD-based Classification of Hits). To develop SCORCH, training data is augmented by considering multiple ligand poses and stratifying pose classification by RMSD from the native pose. Decoy bias is addressed by generating property-matched decoys for each ligand and using the same methodology for preparing and docking decoys and ligands. A consensus of 3 different machine learning approaches is also used to improve performance.

Results

We find that multi-pose augmentation improves the docking power and screening power of machine-learning scoring functions on independent benchmark datasets. SCORCH outperforms an equivalent scoring function trained on single poses, with a 1% enrichment factor (EF) of 13.78 vs. 10.86 on 18 DEKOIS 2.0 targets and a mean native pose rank of 5.9 vs 30.4 on CSAR 2014. Additionally, SCORCH outperforms widely used scoring functions in virtual screening and pose prediction on independent benchmark datasets.

Conclusion

By rationally addressing key limitations of previous scoring functions, SCORCH improves the performance of virtual screening. SCORCH also provides an estimate of its uncertainty, which can help reduce the cost and time required for drug discovery.

Keywords: Docking; scoring; virtual screening; machine learning; drug discovery; neural networks

Introduction

Computational methods are playing an increasingly crucial role in the drug discovery process[1], aiming to reduce the time and cost of the discovery of novel therapeutics. In structure-based virtual screening (SBVS), libraries of compounds are docked against the crystal structure of a target biomolecule, generally a protein, in an attempt to produce low or minimal energy poses for each ligand in a specified binding site[2,3]. The interaction of different docked poses and the target can be evaluated by scoring functions (SFs), which seek to estimate the interaction strength between the compound and the target. There are three main use cases for SFs[4]: 1) distinguishing binder from non-binder compounds for a given target (screening power), 2) predicting relative binding affinities (ranking power), and 3) identifying a near-native pose for a known ligand (docking power). A high screening power is likely more useful in early stages of the drug discovery pipeline, providing a principled way to prioritise fewer compounds for experimental testing. Although the performance of SFs for virtual screening has improved over the last decade, novel strategies are needed to push the current capabilities and make computer-driven drug discovery widely preferred[5].

Classical SFs are generally linear functions with empirical, force field-based, or knowledge-based terms to estimate binding affinity[6]. While they are widely used, their performance is limited by not capturing complex non-linear relationships or taking advantage of the increasing amount of structural data available to improve performance[7]. To overcome the limitations of classical SFs, various machine learning-based scoring functions (MLSFs) have been developed over the last decade, leveraging convolutional neural networks[8], support vector machines[9], and random forests[10,11]. MLSFs have demonstrated notable improvements in performance over classical SFs in virtual screening benchmarks[10,12–15]. Most MLSFs are regressors, predicting the numerical pK_d of receptor-ligand interactions. However, a classification approach to affinity prediction is also a viable alternative[16] and could reduce false positive rates during virtual screening[17].

Neural networks underlie some of the best-performing MLSFs available. NNScore 1.0, a classification-based SF, outperformed Glide when it was first introduced[18]. Hassan et al. outperformed previous MLSFs on crystallographic protein-ligand complexes when introducing DLScore, a deep network for pK_d prediction[19]. Furthermore, a wide-and-deep network showed promise when implemented for binding affinity prediction using 1D amino acid sequences[20], although it has not been applied to 3D structure data. The application of some other ML algorithms, such as gradient boosted decision trees (GBDTs), has not received as much attention. In addition, while it is generally accepted that consensus strategies can improve performance in virtual screening and machine learning[21–24], no consensus scoring function has been proposed combining multiple machine learning algorithm architectures.

MLSFs are typically trained on co-crystalized structures and experimental affinity binding data for receptor-ligand complexes from datasets such as PDBbind[25] and BindingMOAD[26]. When training MLSFs on these datasets, providing the models with crystallographic ligand poses is the most common approach to provide examples of strong binders. However, redocking the co-crystallized ligands provides data closer to that which will be scored once the MLSF is deployed. This strategy was used in RFScoreVS[10] and significantly improved screening performance.

Despite the growing size of these datasets, the number of weak binders is often insufficient for SF training. Efforts have been made to produce datasets of experimentally verified active and inactive receptor-ligand pairs, but they are limited in their utility and size. For example, LIT-PCBA contains confirmed inactive ligands for only 15 target receptors[27]. To include more examples of non-binders, researchers have used datasets of decoy ligands, such as DUD-E[28]. These decoys are designed to mimic physicochemical properties of active ligands while differing in 2D topology, thereby minimising their likelihood of binding. However, some decoy datasets contain fundamental structural biases, allowing MLSFs to learn spurious differences between actives and decoys rather than actual patterns of ligand binding[29], limiting their performance and generalisability. Ideally, a well trained ML-based SF should have learned the importance of receptor-ligand atom interactions and their respective distances; high scores should only be assigned to well docked strong binders. Recently, DeepCoy has been introduced as a method to generate property-matched decoys[30], which could help alleviate systemic biases in training data. Additionally, use of multiple docked poses of a given compound as a data augmentation strategy has been previously explored as a method for improving docking power[31]. Stratifying pose labelling by both RMSD and binding affinity could potentially improve both docking and screening power.

In this work, we present 8 MLSFs for virtual screening that act as binary classifiers of docked receptor-ligand complexes. We explore the impact of different algorithms, as the SFs use either GBDTs, feedforward neural networks, wide and deep neural networks, or a consensus of all three methods. We aimed to address some of the limitations of existing MLSFs by employing unbiased decoy ligands generated by DeepCoy, identical pre-processing of actives and decoys from SMILES strings, and supplying multiple docked ligand poses labelled by RMSD from the crystallographic pose[30]. Each SF was trained on over 45,000 data points. The best performing scoring function, SCORCH (Machine Learning Scoring COnsensus for RMSD-based Classification of Hits), pushes the current capabilities of virtual screening and is openly available to the scientific community.

Materials and methods

The workflow in Figure 1 outlines the steps taken to develop SCORCH. Further details on the data preparation steps can be found in Figure S1.

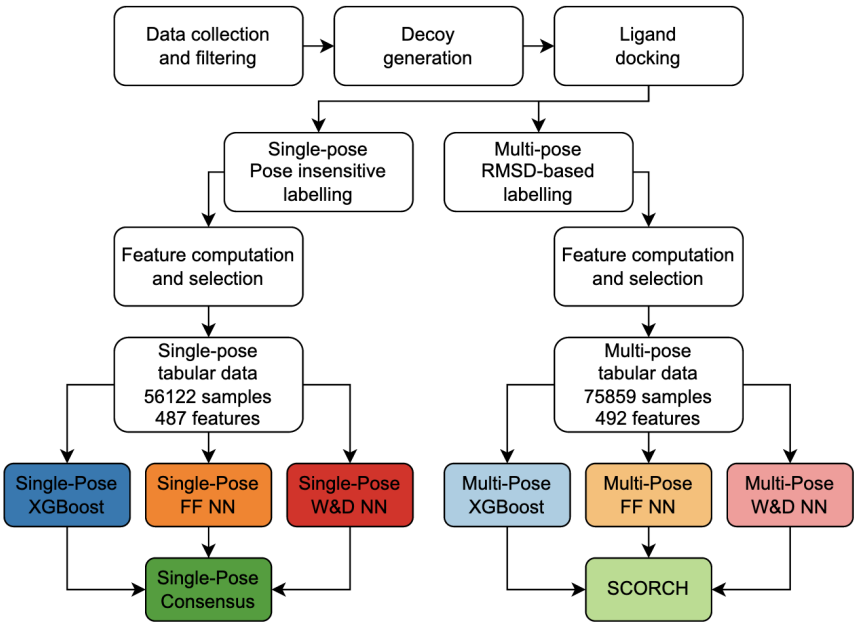

Figure 1. Production workflow of single-pose and multiple-pose machine learning models. Additional details on data preparation (white boxes) are indicated in Figure S1.

Datasets

The *Refined Set* (n=4,854 complexes), *Non-redundant Set* (n=3,187 complexes) and *Highly Trustworthy Set* (n=120 complexes) were downloaded from PDBBind, Binding MOAD, and Iridium, respectively[25,32,33]. These datasets were chosen to ensure that only high-resolution structures under 2.5 Å with reliable experimental binding data stated in primary literature were used for SF training.

Duplicate complexes (n=912) were removed. Complexes with peptide or amino acid-containing ligands, or complexes with multiple bound active ligands (n=757) were removed from the dataset using BioPython 1.78[34]. A 'receptor.pdb' file was created from each remaining complex protein (n=6,492) containing only non-water atoms within 14 Å of the ligand using BioPython 1.78. To ensure the accuracy of side chain positions close to bound ligands, any receptor files with alternative amino acid states within 8 Å of ligand atoms were excluded (n=590).

To maximise the likelihood of accurate docking results, ligands from the remaining complexes were required to have a molecular weight (MW) ≤ 650 Da and ≤ 20 rotatable bonds (nRot). MW and nRot for each ligand were calculated using OpenBabel 3.1.1 as part of Open Drug Discovery Toolkit (ODDT) 0.7 and MGLTools 1.5.6, respectively[35–37]. Ligands exceeding these cutoffs (n=583) or ligands incompatible with MGLTools 1.5.6 (n=54) were excluded from the dataset. Canonical SMILES were extracted for the ligands in the remaining complexes (n=5,265) from PDBBind for PDBBind and Iridium sourced structures, and from Binding MOAD for MOAD structures.

Decoys were used here to provide training data examples of poor or non-binder molecules. As commonly used decoy selection tools may introduce decoy bias, we employed a novel tool, DeepCoy, for decoy generation[29,30]. For each ligand SMILES, 1000 decoys were produced, and the best 10 decoys with nRot <20 were selected using DeepCoy's *select_and_evaluate_decoys.py* script according to previously described criteria[30]. Complexes with no available ligand canonical SMILES (n=52), a canonical SMILES incompatible with DeepCoy (n=51), incompatibilities with GWOVina (n=5) or incompatibilities with spyrmsd (n=55) were excluded, leaving a dataset of 5,102 protein-ligand complexes and 51,020 decoy ligands for a total of 56,122 individual protein-ligand instances.

Active protein-ligand complexes (n=5,102) were randomly split into training (n=4,131), test (n=511) and validation (n=460) sets, stratified by structure resolution and dissociation constant to ensure identical distributions across all three sets using scikit-learn 0.24.2 *StratifiedShuffleSplit* function[38]. PDB complex IDs were used to create splits. All decoy molecules and active ligand poses were assigned to the split of their source PDB complex ID to ensure no leakage between splits.

Data preparation and docking

Canonical active ligand smiles and DeepCoy generated decoy smiles were converted to pdb format using RDKit 2018.09.01's *MolFromSmiles*, *EmbedMolecule*, and *MolToPDBFile* functions[39]. MGLTools 1.5.6's *prepare_ligand4.py* script was used to add Gasteiger charges, add hydrogens, and merge non-polar hydrogens with their parent atoms for RDKit produced pdbs. MGLTools 1.5.6's *prepare_receptor4.py* was used to prepare receptor pdbs in the same way. Active and decoy ligands were then docked into their respective receptors using GWOVina 1.0[40], with up to 20 poses produced for active ligands, and up to 5 poses produced for decoy ligands. All docking was performed with the following settings: exhaustiveness=32, num_wolves=40, num_modes=20 (5 for decoys), energy_range=4. A padding of 12 Å in all directions was added to the native ligand pose position to define a box for docking.

Pose labelling

For labelling purposes, K_d and inhibition constant (K_i) were considered roughly equivalent. Strong binders were defined as complexes with a $K_d(K_i) \leq 25\mu M$ and were assigned a class of 1, and the remaining weak binder complexes with a $K_d(K_i) > 25\mu M$ were assigned a class of 0. A few complexes had IC_{50}. For these complexes, the class was estimated considering the Cheng-Prusoff equation[41]:

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}} \qquad (1)$$

Since IC_{50} values must be greater than the corresponding K_i values, complexes with an IC_{50} of $\leq 25\mu M$ were assigned a class of 1. Complexes with an $IC_{50} > 250\mu M$ were assigned a class of 0. Complexes with an IC_{50} value between 25μM and 250μM, and complexes with binding data stored in other formats (CC_{50} , AC_{50}) were previously excluded.

Additionally, for each active ligand pose, symmetry-corrected RMSD from the crystallographic ligand pose was calculated using spyrmsd 0.5.0[42] (RMSD could not be calculated for 5klp test set ligand poses). Two datasets were created, one considering a single docked pose for each ligand, and one considering multiple docked poses for each ligand.

Single-pose data consisted of the best pose for each active and decoy ligand as scored by GWOVina, resulting in a dataset of 56,122 protein-ligand complexes of which 45,441 were used for training.

For multi-pose data, poses of strong binders with an RMSD < 2.0 Å from the experimental pose were assigned a label of 1. All ligand poses with an RMSD >2.0 Å and < 4.5 Å were excluded, and poses with an RMSD >4.5 Å were assigned a label of 0. All poses of weak binders were assigned a label of 0 regardless of their RMSD. For each active, a maximum of three poses labelled 0 and five poses labelled 1 were included. Poses were prioritised according to their GWOVina affinity scores where the number of suitable poses exceeded these thresholds. For each decoy, a single pose was randomly selected for inclusion and assigned a label of 0, producing a dataset of 75,859 protein-ligand pose complexes of which 61,411 were used for training (Figure S1).

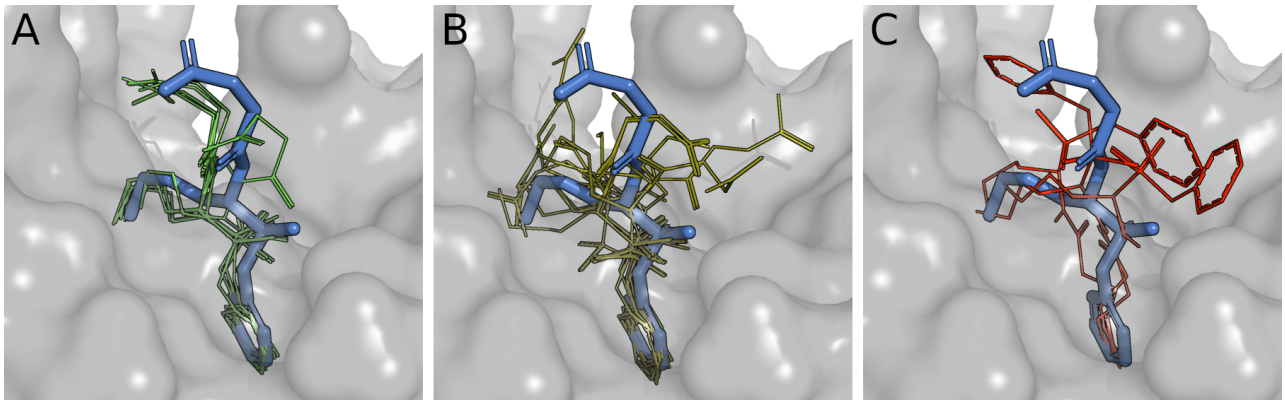

Figure 2. RMSD-based pose labelling example for PDB structure 1a0q. a) Docked poses (green) less than 2 Å from the crystal pose (thick blue lines) are given a label of 1. b) Docked poses (olive) between 2 Å and 4.5 Å from the crystal pose (thick blue lines) are excluded from the dataset. c) Docked poses (red) over 4.5 Å from the crystal pose (thick blue lines) are given a label of 0.

Feature computation

For every complex, BINANA 1.3 was used to calculate 531 features describing atom-type pairwise counts within 2.5 Å and 4.0 Å, ligand atom types and nRot, summed pairwise electrostatic energies, active-site flexibility, hydrogen bonds, hydrophobic contacts, stacking interactions, cation-pi interactions, and salt bridges[43].

Additionally, extended connectivity interaction features (ECIFs) were extracted for each protein-ligand complex as described previously[44]. ECIF atom types indicate atomic symbol, valency, count of bound heavy atoms and hydrogens, aromaticity state, and ring membership. Pairs of these detailed protein atom types (n=22) and ligand atom types (n=70) within 6.0 Å were tallied to produce 1,540 ECIFs.

Lastly, Kier flexibility for each compound was calculated as described previously[45]. Kier flexibility and nRot were the only features included pertaining only to the ligand, and therefore were unaffected by ligand pose.

Feature selection was performed on single pose and multiple pose training sets to identify the optimal combination of features for SF performance; validation and test set data were left unseen. Variances and correlations were calculated using SciPy 1.5.4[46]. Variance was calculated for all 2,072 features, and features with zero variance removed. Pairwise Pearson correlation was

calculated for the remaining features. Additionally, the Pearson correlation of each feature to the binding classification column was calculated. Features with a pairwise correlation >0.9 were identified and, for each pair, the feature with the lowest Pearson correlation to the binding classification was removed. Features were then scaled to between 0 and 1 using scikit-learn's *MaxAbsScaler* function. Recursive feature elimination with three-fold cross-validation was performed on the scaled features using scikit-learn's RFECV function using a random forest classifier. AUCPR (see section 2.9) of predictions was recorded, and the classifier was asked to rank the importance of each of the features in its predictions. The least important feature was iteratively removed until the subset of features that produced the highest average cross-validated AUCPR was identified (single pose n=487, multiple pose n=492 features).

Machine-learning models

Three separate machine learning classification approaches were employed - a gradient boosted decision tree (GBDT), a feedforward neural network, and a wide-and-deep neural network. Two models were produced using each approach, trained on either single pose or multiple pose datasets.

Gradient boosted decision trees

All GBDT models were trained using XGBoost 1.4.2[47]. GBDTs are a sequential ensemble of decision trees, where each tree is trained to predict the error of previous trees. Hyperparameter tuning was performed using five-fold cross-validation on both single and multiple pose training sets with scikit-optimize 0.8.1's *gp_minimize* function[48]. Single and multiple pose models were trained using the identified optimum hyperparameters with the native XGBoost API using previously prepared training and validation sets.

Neural network models

Feedforward and wide-and-deep neural networks were designed with a dropout layer followed by three dense layers (Figure S2). Wide-and-deep neural networks had a concatenation layer between the outputs of the final dense layer and those of the dropout layer. Hyperparameter tuning and model training were performed with Tensorflow 2.4.0, Keras 2.4.3, and scikit-learn[49,50]. Grid search hyperparameter tuning was performed using five-fold cross-validation on both single and multiple pose training sets (Table S1). Hyperparameter combinations were ranked by average cross-validated AUCPR. For each model type, the top 100 hyperparameter combinations were used to train 100 networks, which were ranked by validation set AUCPR. Consensus predictions were derived by averaging the network outputs. Consensus with varying numbers of networks were evaluated on the validation set using AUCPR (Figure S3). For each model type, a consensus of the top 15 models was subsequently selected for test set and independent benchmarking evaluation, and it is the default setting in SCORCH.

DEKOIS 2.0 virtual screening external dataset

All active ligands and decoys were obtained for a previously described diverse subset of 18 DEKOIS 2.0 protein targets[51,52]. Protein receptors were obtained from PDBBind where possible and the PDB where not possible. 14 Å receptor pdbqt files were produced using an in-house python script and MGLTools 1.5.6 as described above. Active and decoy ligands were converted to pdb format with RDKit 2018.09.01 and prepared with MGLTools 1.5.6 as described above. One decoy for ADRB2 (ZINC05822747) was discarded as it was incompatible with *prepare_receptor4.py*. 20 docked poses were produced for each active and decoy ligand using GWOVina 1.0.

CSAR 2014 benchmark pose prediction external dataset

CSAR benchmarking structures were obtained from the relevant publication site[53]. Pose prediction receptor files were converted to pdb format with openbabel 2.4.1 and prepared with MGLTools 1.5.6 as described above. As the dataset consists of docked poses to score, no docking was performed. Pose prediction ligand poses were converted to pdb format with RDKit 2018.09.01 and prepared with MGLTools 1.5.6 as described above.

Comparison with other scoring functions

Some third-party scoring functions were used to compare predictions on the test set, the DEKOIS 2.0 set, and the CSAR 2014 pose prediction benchmark dataset. DLSCORE was cloned from the publication GitHub and run in the supplied virtual environment[19]. NNScore 1.0 was obtained from the publication GitLab[18]. NNScore 2.0 scores were obtained simultaneously through the NNScore 2.0 version integrated with DLSCORE[54]. The RFScore-VS v2 binary was obtained from the publication GitHub[10].

Evaluation metrics

Model performance was evaluated using the area under the precision recall curve (AUCPR), Enrichment Factor (EF), and an in-house certainty metric.

AUCPR

AUCPR is a measure of binary classification performance across varying thresholds[55]. In this case, the threshold would be the value of the scoring function above which a compound is classified as a binder. The equations for recall and precision are given as:

$$Recall = \frac{TP}{TP + TN} \qquad (2)$$

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

Where *TP*, *FP*, and *TN* represent the number of true positive, false positive, and true negative predictions respectively. SF predictions are ranked from highest to lowest; at each ranking threshold, recall and precision were calculated using Rocker 0.1.4[56]. Minor modifications were made to Rocker 0.1.4 to enable calculation of precision. Recall and precision values are plotted as a precision-recall curve. The area under the curve was calculated using scikit-learn's *auc* function.

Enrichment factor

The enrichment factor (EF) is a useful indicator of screening power. It indicates how many times more active compounds would be discovered if experimentally testing the top x percent of compounds ranked by the scoring function, as opposed to randomly testing compounds in the library. The EF at a given % threshold x is calculated as follows[57].

$$FA_x = \frac{A_x}{N_x} \qquad (4)$$

$$FA_{total} = \frac{A_{total}}{N_{total}} \qquad (5)$$

$$EF_x = \frac{FA_x}{FA_{total}} \qquad (6)$$

Where A_x is the number of actives in the top $x\%$, N_x is the number of molecules in the top $x\%$, FA_x is the fraction of actives in the top $x\%$, A_{total} is the total number of actives, N_{total} is the total number of molecules, and FA_{total} is the total fraction of actives. EF was calculated using ODDT 0.7 at thresholds of 0.5%, 1.0%, 2.0%, 5.0%.

SCORCH certainty

We define a certainty metric for the consensus SFs by normalising the variation across the individual model predictions that contribute to the consensus.

$$\theta = 1 - \sigma \qquad (7)$$

Where σ is the population standard deviation of the contributing models' predictions. Since there are 3 models in each consensus with a prediction range of $(0, 1)$, the maximum and minimum θ values are:

$$\theta_{max} = 1 \qquad (8)$$

$$\theta_{min} = 0.529 \quad (3.s.f.) \qquad (9)$$

θ is then normalised between (0,1) as follows:

$$Certainty = \frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}} \qquad (10)$$

Thus, certainty values closer to 1 indicate that the model is more confident in the predicted score, whereas values closer to 0 indicate lower confidence.

Results and discussion

The aim of our work was to implement and study the impact of different strategies on the performance of machine learning-based scoring functions. Firstly, the mining and processing of three protein-ligand complex databases (see Methods) led to the generation of a dataset containing 5,102 high quality complexes with known experimental binding data. To our knowledge, this is the largest dataset of experimental high-quality protein-ligand complexes used to train an MLSF to date. Extensive feature engineering and selection were also performed to ensure a range of highly informative model inputs. A common approach to descriptor generation for MLSFs is to use a single program to produce pairwise atom tallies. In previous studies these descriptors have been passed directly to MLSFs without prior evaluation[10,19]. Here, we used multiple programs to generate over 2,000 descriptors, before performing feature selection to determine the optimal descriptors to use as features for training.

Several strategies were applied to mitigate decoy bias. Firstly, a novel ML-based method, DeepCoy[30], was employed to generate challenging unbiased decoys, avoiding biases present in existing decoy datasets. Additionally, an identical pre-processing pipeline was used for all ligands (see Methods). This removes any differences between active and decoy ligand 3D conformations arising from pre-processing steps. Thirdly, six models were produced using three different machine-learning algorithms (see Methods); three were trained on a single docked pose for each active ligand, and three were trained on multiple RMSD-labelled poses for each active ligand. In addition, single and multi-pose consensus models were produced by taking the mean of the three single-pose and multi-pose models predictions, respectively. The multi-pose consensus model will be referred to as SCORCH.

Lastly, for the multi-pose models, the RMSD-based labelling of active ligand poses ensured that only well docked strong binders were rewarded during training and that both docked decoys and poorly docked strong binders were punished. We reasoned that this should further reduce decoy bias, as examples of weak binders are not exclusively decoy molecules. This labelling strategy also serves to augment the size of the training data and produces scoring functions which should be better behaved in real virtual screening applications, where ligands are evaluated based on their docked poses.

Screening power evaluation on the test set

To assess the utility of our models in virtual screening, they were first evaluated on unseen test set complexes and their performance was compared to that of existing MLSFs. For all evaluated functions, all available poses for each ligand in the test set were scored and the maximum score out of all supplied poses was taken as the score for that protein-ligand complex. Hence, if a single pose out of many obtains a high score, the ligand is predicted to be a binder. When none of the supplied poses obtains a high score, the ligand is predicted as a non-binder. While there are other methods for aggregating scores from multiple poses, taking the maximum scoring pose showed the best performance by both AUCPR and EF across both our models and third-party scoring functions (Figure S4). Notably, this method may outperform the simpler approach of only scoring the lowest energy pose returned by molecular docking software.

AUCPR values were calculated for our models as well as third party scoring functions (Figure 3). All models trained on multiple poses achieved at least 0.814 AUCPR on the test set. The consensus scoring functions are the highest performing for both single-pose and multi-pose models, with 0.854 and 0.837 AUCPR respectively, evidencing the benefit of a consensus over different types of algorithms. It is notable that, for all single-pose models, AUCPR is greater than that of equivalent

multi-pose models, and all produced models outperform third-party scoring functions. The high AUCPR values on the test set do not indicate overfitting to the training data for any of our produced models, and show promise to improve the accuracy of virtual screening. However, validation against an independent benchmarking dataset is needed for fair comparison, as decoy bias and familiarity with the data pipeline could be contributing to the performance of our models.

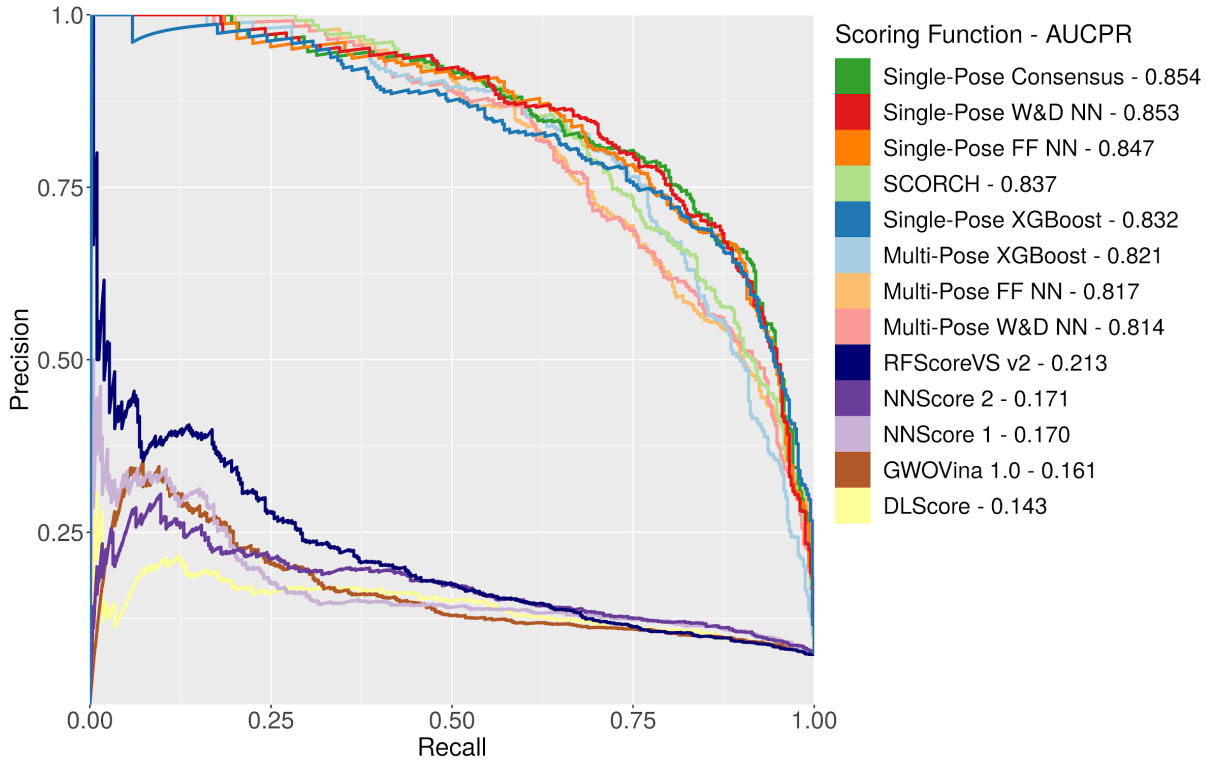

Figure 3. Precision-Recall curves for produced machine-learning models and third-party scoring functions on 5,621 test set complexes.

Additionally, while AUCPR is a useful global indicator of scoring function performance, it is insensitive to the importance of early recognition. Scoring functions are often tasked with identifying a small subset of top-ranked compounds below a predefined cutoff as “virtual hits” to be taken forward for experimental testing. Hence, we also evaluated our produced scoring functions using enrichment factor (EF) at four cutoffs (top 0.5%, 1%, 2%, and 5%). Maximum mean EF across all four cutoffs was achieved by SCORCH (Table S2). The multi-pose models all achieved higher mean ranks of actives in a top subset of the test set, favouring them for use in virtual screening.

Screening power on an independent benchmarking set

Since decoy bias and non-representative training data could produce misleading test-set performance, the models were further validated on a diverse subset of an independent benchmarking set, DEKOIS 2.0. This subset contains 18 diverse receptors, with 1200 decoys and 40 actives present for each target. Ligands were docked, scored and ranked by our models and third-party scoring functions for all 18 targets, and overall AUCPR values were calculated (Figure 4A). As 40 actives and 1200 decoys were present for each target receptor, an unskilled classifier should achieve an AUCPR of 40/1240, or ~0.03. All tested scoring functions performed better than this baseline score.

Except for the single-pose feedforward neural network model, which was outperformed by RFScoreVS v2, our models outperformed all third-party scoring functions. All multi-pose trained models outperformed the third-party scoring functions; the wide-and-deep neural network achieved the lowest multi-pose model AUCPR of 0.165, while the best performing third party scoring function RFScoreVS v2 achieved an AUCPR of 0.151. The results of the validation on this independent benchmark demonstrate a genuine increase in scoring function performance. The best performing model was the consensus of multi-pose models (SCORCH), with an AUCPR of 0.198. To our surprise, the third-party machine learning scoring functions NNScore 1, NNScore 2 and DLScore failed to outperform the GWOVina scoring on this benchmark set. These functions were trained with crystal poses as examples of active ligands, whereas the much better performing RFScoreVS v2 was trained on docked ligand poses; perhaps these models are overfitted to crystallographic active poses.

Notably, the multi-pose trained models outperformed the equivalent single pose-trained models. As the poses and labels supplied during training were the sole independent variable between these model types, this indicates that the employed methodology of RMSD-based pose labelling leads to better performing scoring functions.

While third-party models are still outperformed, there is a notable dropoff in AUCPR when comparing test set and DEKOIS 2.0 results. This could in part be due to single and multi-pose models overfitting to DeepCoy decoys despite the prior steps taken to mitigate bias. A possible solution to this would be to train on a heterogenous mix of decoys from multiple sources. Experimentally determined inactives from datasets such as LIT-PCBA would be excellent, however, these datasets are limited in size.

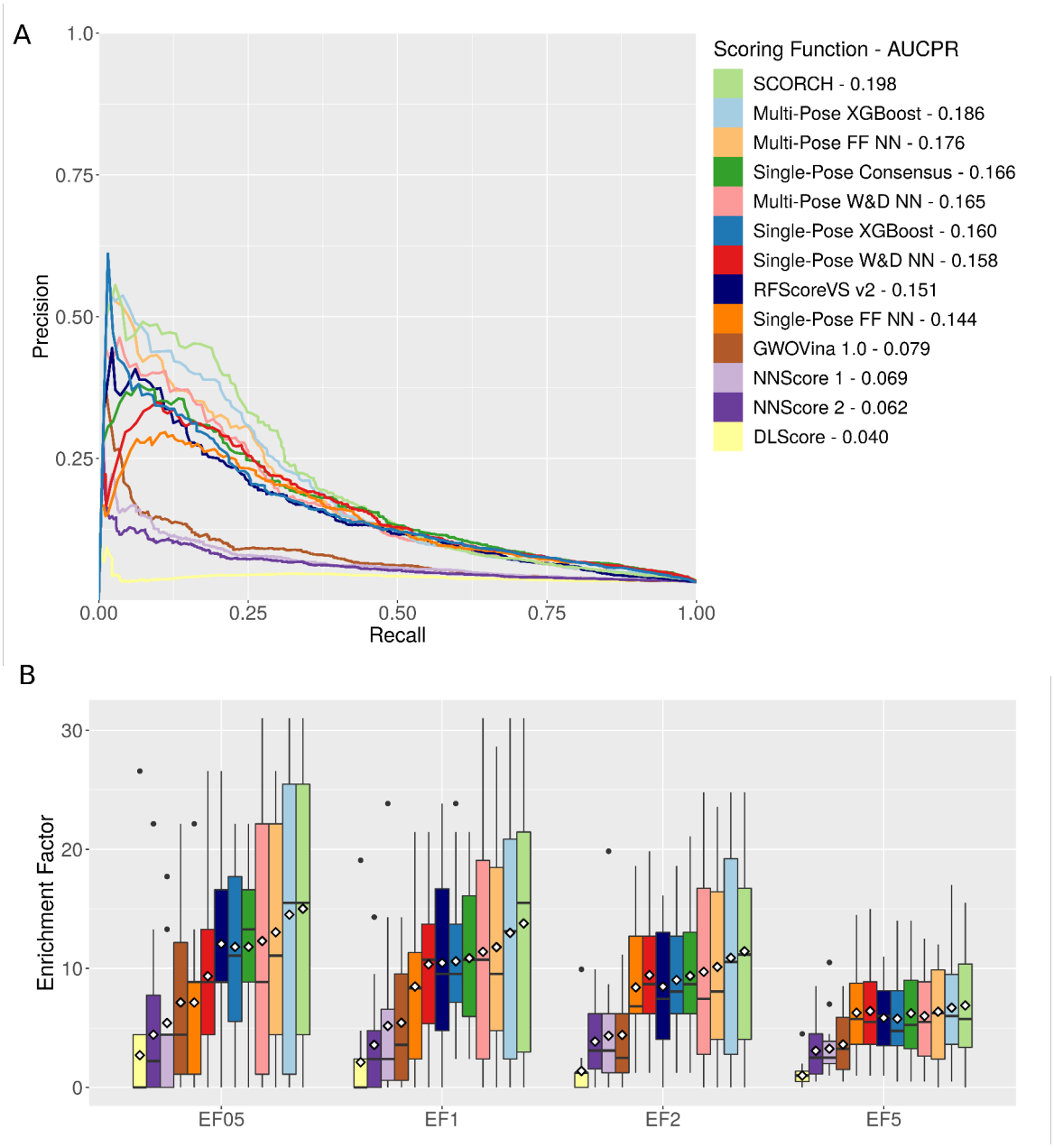

Figure 4. SCORCH is the best performing scoring function on a subset of the DEKOIS 2.0 independent benchmarking dataset. a) Precision-Recall curves for produced machine-learning models and third-party scoring functions on 18 DEKOIS 2.0 targets. b) Enrichment factors at 0.5%, 1%, 2% and 5% for produced machine-learning models and third-party scoring functions across 18 DEKOIS 2.0 targets. White diamonds indicate the mean EF.

The low proportion of actives and high proportion of inactives in DEKOIS 2.0 makes it well suited to assess virtual screening performance. For each receptor, EF at 0.5%, 1%, 2% and 5% was calculated (Figure 4B). As stated above, due to the aims of virtual screening campaigns, this enrichment factor is likely the best representation of real world scoring function utility.

At all chosen EF thresholds, SCORCH achieved the highest mean EF, with mean EF0.5=15.01, EF1=13.78, EF2=11.44 and EF5=6.89. As with AUCPR, this is superior to the consensus of the single-pose models, further demonstrating the performance boost provided by RMSD based pose labelling. Additionally, SCORCH outperformed all third-party scoring functions, with mean enrichment factors ~27% higher than the best performing third party scoring function, RFScoreVS

v2. Notably, RFScoreVS v2 achieves higher mean enrichment factors than both single-pose trained neural network models at EF05 and EF1. These results support previous observations that GBDTs and random forests are the best suited algorithms for structure based virtual screening[13]. Overall however, these EF results are in agreement with the AUCPR curves above. Additionally, by both EF and AUCPR, SCORCH's DEKOIS 2.0 performance increases with the number of poses it is shown for a given ligand (Figure S5). This indicates SCORCH is highly specific for well docked strong binders; perhaps supplying more than 20 poses per ligand could see further performance improvements.

As is common with machine learning based scoring functions, performance varies widely between receptors. Even with the best performing approach, SCORCH, EF0.5 was 0 for four DEKOIS 2.0 targets (ACHE, DHFR, CATL, MDM2), while the maximum possible EF0.5 of 31 was achieved for three targets (P38-alpha, ERBB2, ACE). To help navigate this variability in SCORCH, we investigated using a "certainty" metric. This metric is based on the standard deviation of SCORCH input models and can be regarded as a measure of confidence in the predicted classification. We observe that, across the whole DEKOIS 2.0 dataset, a higher SCORCH certainty is strongly and significantly correlated with smaller error in SCORCH predictions ($p=2e-16$, $Pr = -0.45$) (Figure 5A). The certainty value is thus a useful and robust metric which may allow *a priori* identification of poor enrichment performance. Moreover, we also observe that the average model certainty over the top 0.5% of ligands showed a significant positive correlation with enrichment factor ($p=0.013$, Pearson correlation $P_r=0.57$) (Figure 5B). Hence, the certainty values provided by SCORCH can help in assessing its overall utility for a given target as well as the likelihood of success of individual compounds in virtual screening results.

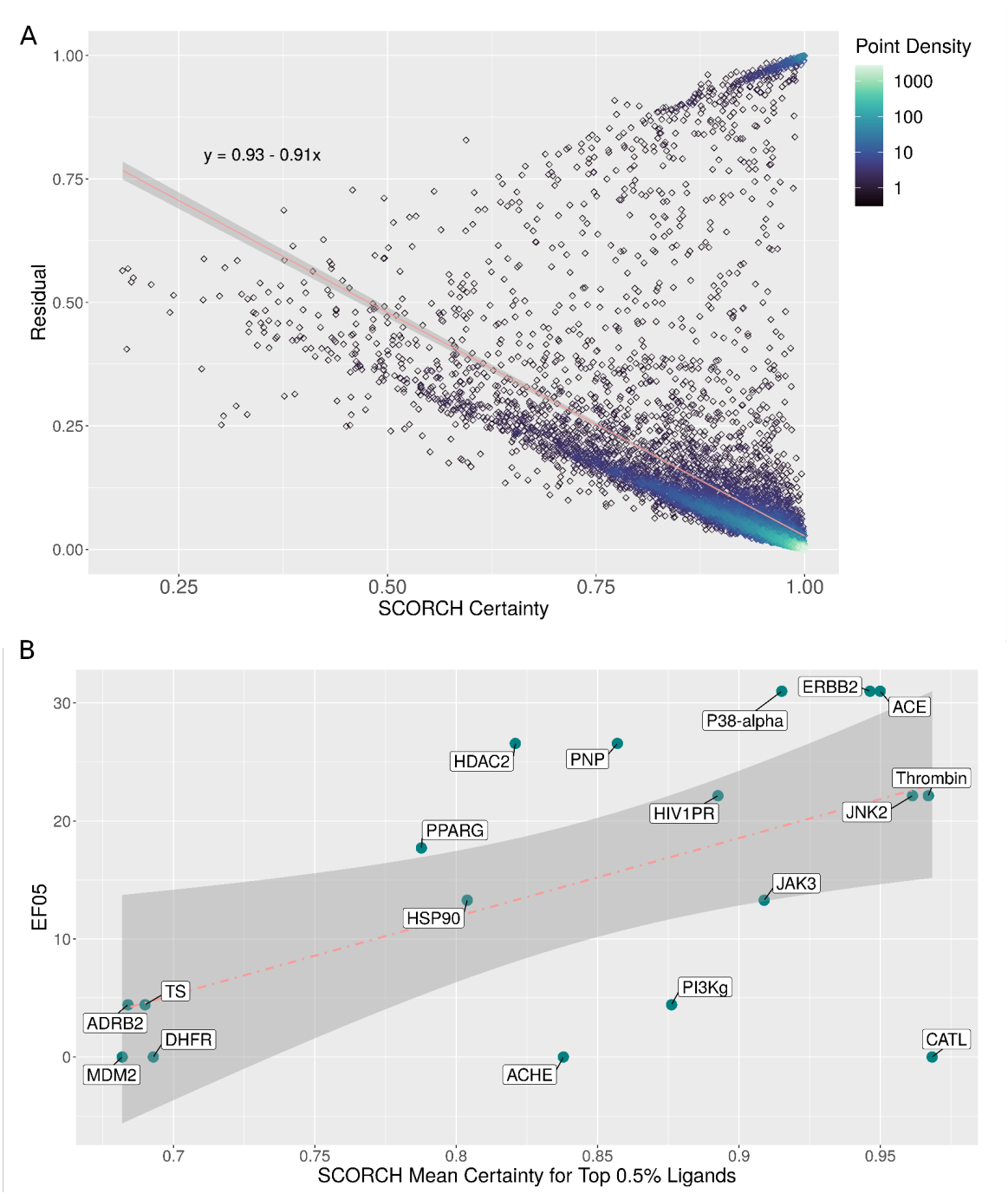

Figure 5. SCORCH certainty metric is a robust *a priori* indicator of enrichment success. a) Relationship between SCORCH certainty and residuals across all 22,319 DEKOIS 2.0 actives and decoys. Fitted regression line shown in red with the line equation displayed top right (p=2e-16, P_r=-0.45); 95% confidence intervals of the regression line are shown in grey. b) Relationship between enrichment factor and mean SCORCH certainty across top 0.5% of ligands for 18 DEKOIS 2.0 receptors. Fitted regression line shown in red (p=0.013, P_r=0.57); 95% confidence intervals of the regression line are shown in grey.

Docking power evaluation on the test set and CSAR 2014

Docking power is the ability of a scoring function to identify the docked pose that most closely resembles the native ligand pose. MLSFs trained on crystal structures frequently have poor docking power. It has been suggested that this problem is inherent to the use of pairwise atom tallies as features, allowing the ML models to learn ligand and receptor atom type counts rather than the impact of their pairwise distances[58]. If the strategy of training on multiple RMSD labelled docked poses is strengthening the performance of our scoring functions, then they should exhibit high docking power as well as screening power.

Docking power was assessed for 510 test set crystal ligands. For each complex, docked ligand poses were ranked by RMSD to the native ligand pose. An SF with good docking power would therefore rank the pose with the lowest RMSD highly. The mean rank of the lowest RMSD pose is shown in Figure 6A. Our single-pose models already had an improved mean rank over all evaluated third party ML-based SFs, showing a docking power comparable to GWOVina 1.0. Importantly, the models which assigned the highest rank to the lowest RMSD poses were all trained on multiple RMSD labelled docked poses, confirming that this training methodology greatly reduces pose insensitivity and increases docking power. Indeed, the mean RMSD of the highest scored pose for all multi-pose models was under 3 Å, and the median RMSD under 2 Å (Figure S6). This improvement in docking power is consistent across all poses, with multi-pose models averaging greater Spearman's rank and Kendall Tau between RMSD and model score (Table S3). Collectively, this demonstrates that these functions do in fact reward well docked strong binders[59].

To validate the docking power of our SFs, we performed an evaluation against the CSAR 2014 native pose identification benchmark. This dataset contains 22 active ligands across three receptors. Each ligand has one near native pose (<1Å RMSD) and 199 decoy poses (>2Å RMSD). For each ligand, all 200 poses were scored and the rank of the near native pose identified (Figure 6B). As with the test set, our SFs trained on multiple poses improve on preceding ML-based methods in terms of mean native pose ranking. Additionally, multi-pose trained models outperformed single-pose trained models in all cases. Interestingly, our multi-pose trained models show superior docking power to GWOVina on the test set, but are outperformed by GWOVina on this benchmarking dataset; perhaps our models are overfit to GWOVina produced poses. Docking poses with several docking programs could potentially mitigate this bias. Overall however, while GWOVina has high intra-ligand docking power, it demonstrates low inter-ligand screening power; SCORCH performs well by both metrics.

If the reason for improved docking and screening power is indeed that SCORCH is truly learning the key factors which drive receptor ligand binding, we hypothesised that this should be reflected in ranking power; ligands with higher affinities should be assigned higher scores. To investigate this, we analysed the relationship between pK_d and SCORCH score for all 511 test set complexes (Figure S7A). We observed a highly significant positive correlation between test set complex pK_d and SCORCH score ($p=2.2e-16$, $P_r=0.41$). When examined, a similar correlation was present for the 244 DEKOIS 2.0 actives where binding data was available as K_d or K_i, indicating that SCORCH's ranking power holds true on an independent benchmark set ($p=3.97e-05$, $P_r=0.26$) (Figure S7B). While this correlation is weak, it demonstrates that higher SCORCH scores are assigned to stronger binders, despite all training examples of ligands with a $K_d \leq 25\mu M$ being given a blanket

label of 1; no continuous pKd or Kd values were ever shown to the models during training. This significant relationship further indicates that SCORCH truly has learned which features drive strong receptor ligand binding.

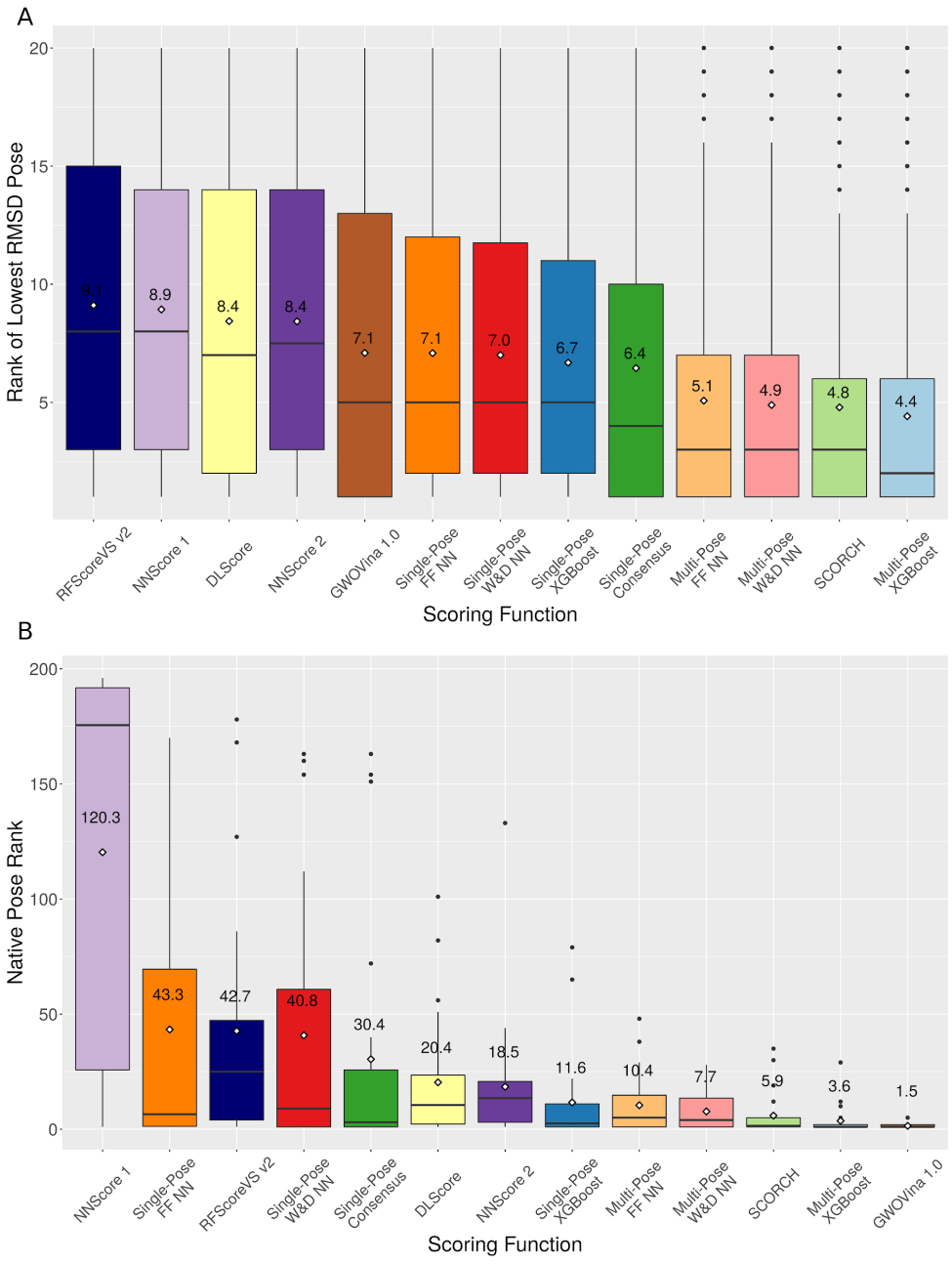

Figure 6. RMSD-based pose labelling used in SCORCH improves the docking power compared to other MLSFs. a) Ranks of the lowest RMSD docked ligand pose across 510 test set complexes for produced machine-learning models and third-party scoring functions; White diamonds indicate the mean rank with the value displayed above each point. b) Ranks of the near native pose across the CSAR 2014 pose prediction benchmark for produced machine-learning models and third-party scoring functions. White diamonds indicate the mean rank with the value displayed above each point.

Model Feature Importances

To gain insights on the origin of the models' performances, we examined feature importances for all our models by iteratively replacing each input feature with random noise and noting the performance decrease on the validation set (Figure 7). Interestingly, the neural network based models rely heavily on a single input, with multi-pose models experiencing a performance decrease of over 14% based on a single feature, 2.5 (A, HD). This BINANA feature is a count of hydrogen donor and acceptor atoms within 2.5 Å representing strong hydrogen bonds; these are certainly a key intermolecular interaction in many small molecule-protein complexes. The same was observed for single-pose neural network models, where randomising tallies of interactions between ligand phosphate groups (P54000) and receptor hydroxyl groups (O21100) lead to a ~15% performance decrease. Other similar features representing phosphate group interactions are highly important for all neural network models. Important XGBoost features, however, are limited to interactions involving ligand oxygen, carbon, and hydrogen atoms, in addition to overall ligand flexibility. These features are more globally applicable to all possible ligands, and suggest XGBoost-based models may be more robust across different types of receptor-ligand interactions. Indeed, randomising the most important feature led to a performance drop of less than 2.5%. This excellent robustness may explain the superior docking and screening power of this type of algorithm compared to neural network based models. Notably, feature importance differs minorly between neural network architectures, and majorly between neural network and XGBoost models. This heterogeneity in feature weightings explains the increased performance achieved when combining the models to form a consensus. Each model favours different features, which are then combined to yield a score and certainty with superior performance to previous scoring functions.

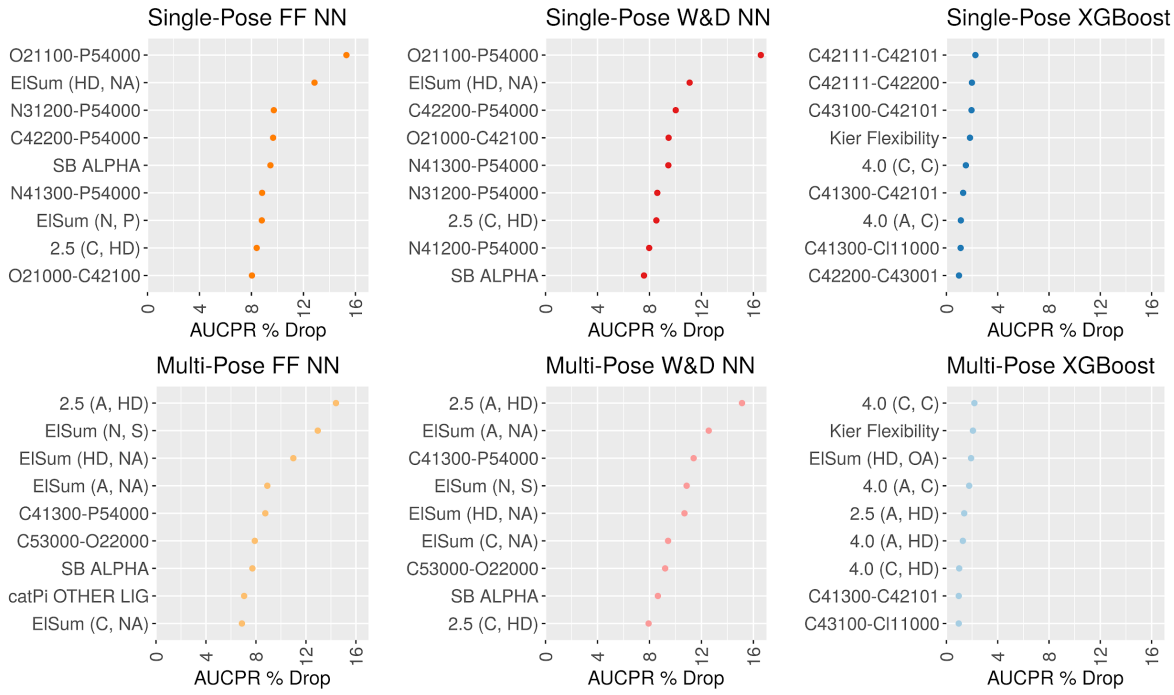

Figure 7. 10 features with highest influence on performance for six individual machine-learning models. Highly influential features for each model are listed on the y-axis. The x-axis represents the decrease in

model test set AUCPR when a given feature is replaced with random noise; a greater decrease indicates greater reliance on a given input feature for making predictions.

Conclusions

Classical structure-based virtual screening approaches suffered from too low or too variable predictive performance, which made many practitioners sceptical about their utility as a primary compound screening methodology. In recent years, machine-learning scoring functions offered superior scoring functions for SBVS, although their performance seemed to have reached a limit given the number of experimental structures of protein-ligand complexes available. In this work, we show that this needs not to be the case. It is possible to rationally identify biases and limitations of previous MLSFs, address them, and substantially improve their predictive performance as a result.

To achieve high-performance in SBVS, it is key that the training data includes the type of inputs that the model will use when deployed. Thus, training and validation in this study have been performed on docked poses; no crystallographic ligand pose was used in this publication. On the other hand, we have shown that both the docking and screening power of MLSFs can be improved by augmenting the training datasets with multiple RMSD-labelled docked poses. These additional poses can be readily labelled and provided to classifiers, which can outperform current regression-based scoring functions, especially in terms of early enrichment. We believe the pose-specific training data and property-matched decoys have forced our MLSFs to better learn the nature of receptor-ligand interactions. Remarkably, the MLSFs produced recapitulate to some extent the relative binding affinities of different compounds, although these affinities were never shown to the models. We have also demonstrated the potential of XGBoost models to build scoring functions and the benefits of a consensus of different algorithms. Furthermore, our model certainty metric allows the *a priori* identification of high or low enrichment in virtual screening, indicating how likely SBVS is to succeed on a specific target.

Taken together, these strategies have led to superior screening and docking power, as validated by two independent benchmarking datasets. Our multi-pose trained consensus model, SCORCH, thus advances the capabilities of current SBVS and positions it as a promising methodology for primary compound screening. SCORCH is freely available to the scientific community.

Data and Code availability

The scoring function code is available at https://github.com/miles-mcgibbon/SCORCH/

The repository will be made publicly available upon publication. Until then, reviewers may use the following temporary link: https://gitfront.io/r/mmcgibbon/cff5aca4804137fbf88d4c7357b32f0cd1c20ad1/SCORCH/

Conflict of interest

The authors declare no competing financial interest.

Acknowledgements

The authors thank Prof. Jie Dong (Central South University, China) for useful discussions.

References

[1] Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational Methods in Drug Discovery. Pharmacol Rev 2014;66:334–95. https://doi.org/10.1124/pr.112.007336.
[2] Tang YT, Marshall GR. Virtual screening for lead discovery. Methods Mol Biol Clifton NJ 2011;716:1–22. https://doi.org/10.1007/978-1-61779-012-6_1.
[3] Ma D-L, Chan DS-H, Leung C-H. Molecular docking for virtual screening of natural product databases. Chem Sci 2011;2:1656–65. https://doi.org/10.1039/C1SC00152C.
[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. https://doi.org/10.3389/fphar.2018.01089.
[5] Mehta S, Laghuvarapu S, Pathak Y, Sethi A, Alvala M, Priyakumar UD. MEMES: Machine learning framework for Enhanced MolEcular Screening. Chem Sci 2021;12:11710–21. https://doi.org/10.1039/D1SC02783B.
[6] Huang S-Y, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Phys Chem Chem Phys 2010;12:12899–908. https://doi.org/10.1039/C0CP00151A.
[7] Li H, Peng J, Sidorov P, Leung Y, Leung K-S, Wong M-H, et al. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. Bioinforma Oxf Engl 2019;35:3989–95. https://doi.org/10.1093/bioinformatics/btz183.
[8] Cang Z, Wei G-W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS Comput Biol 2017;13:e1005690. https://doi.org/10.1371/journal.pcbi.1005690.
[9] Kinnings SL, Liu N, Tonge PJ, Jackson RM, Xie L, Bourne PE. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J Chem Inf Model 2011;51:408–19. https://doi.org/10.1021/ci100369f.
[10] Wójcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep 2017;7:46710. https://doi.org/10.1038/srep46710.
[11] Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinforma Oxf Engl 2010;26:1169–75. https://doi.org/10.1093/bioinformatics/btq112.
[12] Li H, Sze K-H, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. WIREs Comput Mol Sci 2020;10:e1465. https://doi.org/10.1002/wcms.1465.
[13] Shen C, Hu Y, Wang Z, Zhang X, Zhong H, Wang G, 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 2021;22:497–514. https://doi.org/10.1093/bib/bbz173.
[14] Afifi K, Al-Sadek AF. Improving classical scoring functions using random forest: The non-additivity of free energy terms' contributions in binding. Chem Biol Drug Des 2018;92:1429–34. https://doi.org/10.1111/cbdd.13206.
[15] Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. Machine-learning methods for ligand–protein molecular docking. Drug Discov Today 2022;27:151–64. https://doi.org/10.1016/j.drudis.2021.09.007.
[16] Ballester PJ. Selecting machine-learning scoring functions for structure-based virtual

screening. Drug Discov Today Technol 2019;32–33:81–7. https://doi.org/10.1016/j.ddtec.2020.09.001.
[17] Adeshina YO, Deeds EJ, Karanicolas J. Machine learning classification can reduce false positives in structure-based virtual screening. Proc Natl Acad Sci U S A 2020;117:18477–88. https://doi.org/10.1073/pnas.2000585117.
[18] 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. https://doi.org/10.1021/ci100244v.
[19] Hassan M, Mogollon DC, Fuentes O, Sirimulla S. DLSCORE: A Deep Learning Model for Predicting Protein-Ligand Binding Affinities 2018. https://doi.org/10.26434/chemrxiv.6159143.v1.
[20] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 2018;34:i821–9. https://doi.org/10.1093/bioinformatics/bty593.
[21] Ericksen SS, Wu H, Zhang H, Michael LA, Newton MA, Hoffmann FM, et al. Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening. J Chem Inf Model 2017;57:1579–90. https://doi.org/10.1021/acs.jcim.7b00153.
[22] Lima AN, Philot EA, Trossini GHG, Scott LPB, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov 2016;11:225–39. https://doi.org/10.1517/17460441.2016.1146250.
[23] Wang D, Cui C, Ding X, Xiong Z, Zheng M, Luo X, et al. Improving the Virtual Screening Ability of Target-Specific Scoring Functions Using Deep Learning Methods. Front Pharmacol 2019;10.
[24] Houston DR, Walkinshaw MD. Consensus docking: improving the reliability of docking in a virtual screening context. J Chem Inf Model 2013;53:384–90. https://doi.org/10.1021/ci300399w.
[25] Wang R, Fang X, Lu Y, Wang S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. J Med Chem 2004;47:2977–80. https://doi.org/10.1021/jm030580l.
[26] Wójcikowski M, Siedlecki P, Ballester PJ. Building Machine-Learning Scoring Functions for Structure-Based Prediction of Intermolecular Binding Affinity. In: de Azevedo Jr. WF, editor. Docking Screens Drug Discov., New York, NY: Springer; 2019, p. 1–12. https://doi.org/10.1007/978-1-4939-9752-7_1.
[27] Tran-Nguyen V-K, Jacquemard C, Rognan D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. J Chem Inf Model 2020;60:4263–73. https://doi.org/10.1021/acs.jcim.0c00155.
[28] Mysinger MM, Carchia M, Irwin JohnJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J Med Chem 2012;55:6582–94. https://doi.org/10.1021/jm300687e.
[29] Chen L, Cruz A, Ramsey S, Dickson CJ, Duca JS, Hornak V, 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. https://doi.org/10.1371/journal.pone.0220113.
[30] Imrie F, Bradley AR, Deane CM. Generating property-matched decoy molecules using deep learning. Bioinformatics 2021;37:2134–41. https://doi.org/10.1093/bioinformatics/btab080.
[31] Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model 2017;57:942–57. https://doi.org/10.1021/acs.jcim.6b00740.
[32] Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All Databases). Proteins Struct Funct Bioinforma 2005;60:333–40. https://doi.org/10.1002/prot.20512.
[33] Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD. Essential considerations for using protein–ligand structures in drug discovery. Drug Discov Today 2012;17:1270–81. https://doi.org/10.1016/j.drudis.2012.06.011.

[34] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinforma Oxf Engl 2009;25:1422–3. https://doi.org/10.1093/bioinformatics/btp163.
[35] Wójcikowski 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. https://doi.org/10.1186/s13321-015-0078-2.
[36] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminformatics 2011;3:33. https://doi.org/10.1186/1758-2946-3-33.
[37] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. J Comput Chem 2009;30:2785–91. https://doi.org/10.1002/jcc.21256.
[38] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.
[39] Landrum G, Kelley B, Tosco P, Sriniker, Gedeck, NadineSchneider, et al. Rdkit/Rdkit: 2018_09_1 (Q3 2018) Release. Zenodo; 2018. https://doi.org/10.5281/ZENODO.1468109.
[40] Wong KM, Tai HK, Siu SWI. GWOVina: A grey wolf optimization approach to rigid and flexible receptor docking. Chem Biol Drug Des 2021;97:97–110. https://doi.org/10.1111/cbdd.13764.
[41] Cheng Y, Prusoff WH. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem Pharmacol 1973;22:3099–108. https://doi.org/10.1016/0006-2952(73)90196-2.
[42] Meli R, Biggin PC. spyrmsd: symmetry-corrected RMSD calculations in Python. J Cheminformatics 2020;12:49. https://doi.org/10.1186/s13321-020-00455-2.
[43] Durrant JD, McCammon JA. BINANA: a novel algorithm for ligand-binding characterization. J Mol Graph Model 2011;29:888–93. https://doi.org/10.1016/j.jmgm.2011.01.004.
[44] Sánchez-Cruz N, Medina-Franco JL, Mestres J, Barril X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. Bioinformatics 2021;37:1376–82. https://doi.org/10.1093/bioinformatics/btaa982.
[45] Kier LB. An index of flexibility from molecular shape descriptors. Prog Clin Biol Res 1989;291:105–9.
[46] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
[47] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., San Francisco California USA: ACM; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.
[48] Head T, Kumar M, Nahrstaedt H, Louppe G, Shcherbatyi I. Scikit-Optimise. 2020.
[49] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. Proc. 12th USENIX Conf. Oper. Syst. Des. Implement., USA: USENIX Association; 2016, p. 265–83.
[50] Keras: Deep Learning for humans. Keras; 2022.
[51] Ibrahim TM, Bauer MR, Boeckler FM. Applying DEKOIS 2.0 in structure-based virtual screening to probe the impact of preparation procedures and score normalization. J Cheminformatics 2015;7:21. https://doi.org/10.1186/s13321-015-0074-6.
[52] Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM. 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. https://doi.org/10.1021/ci400115b.
[53] Carlson HA, Smith RD, Damm-Ganamet KL, Stuckey JA, Ahmed A, Convery MA, et al. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. J Chem Inf Model 2016;56:1063–77. https://doi.org/10.1021/acs.jcim.5b00523.
[54] Durrant JD, McCammon JA. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring

Function. J Chem Inf Model 2011;51:2897–903. https://doi.org/10.1021/ci2003889.
[55] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432.
[56] Lätti S, Niinivehmas S, Pentikäinen OT. Rocker: Open source, easy-to-use tool for AUC and enrichment calculations and ROC visualization. J Cheminformatics 2016;8:45. https://doi.org/10.1186/s13321-016-0158-y.
[57] Bender A, Glen RC. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. J Chem Inf Model 2005;45:1369–75. https://doi.org/10.1021/ci0500177.
[58] 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. https://doi.org/10.1021/ci500406k.
[59] Ramírez D, Caballero J. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? Molecules 2018;23:1038. https://doi.org/10.3390/molecules23051038.

Supporting Information

SCORCH: Improving virtual screening with a consensus of machine learning classifiers, data augmentation, and uncertainty estimation

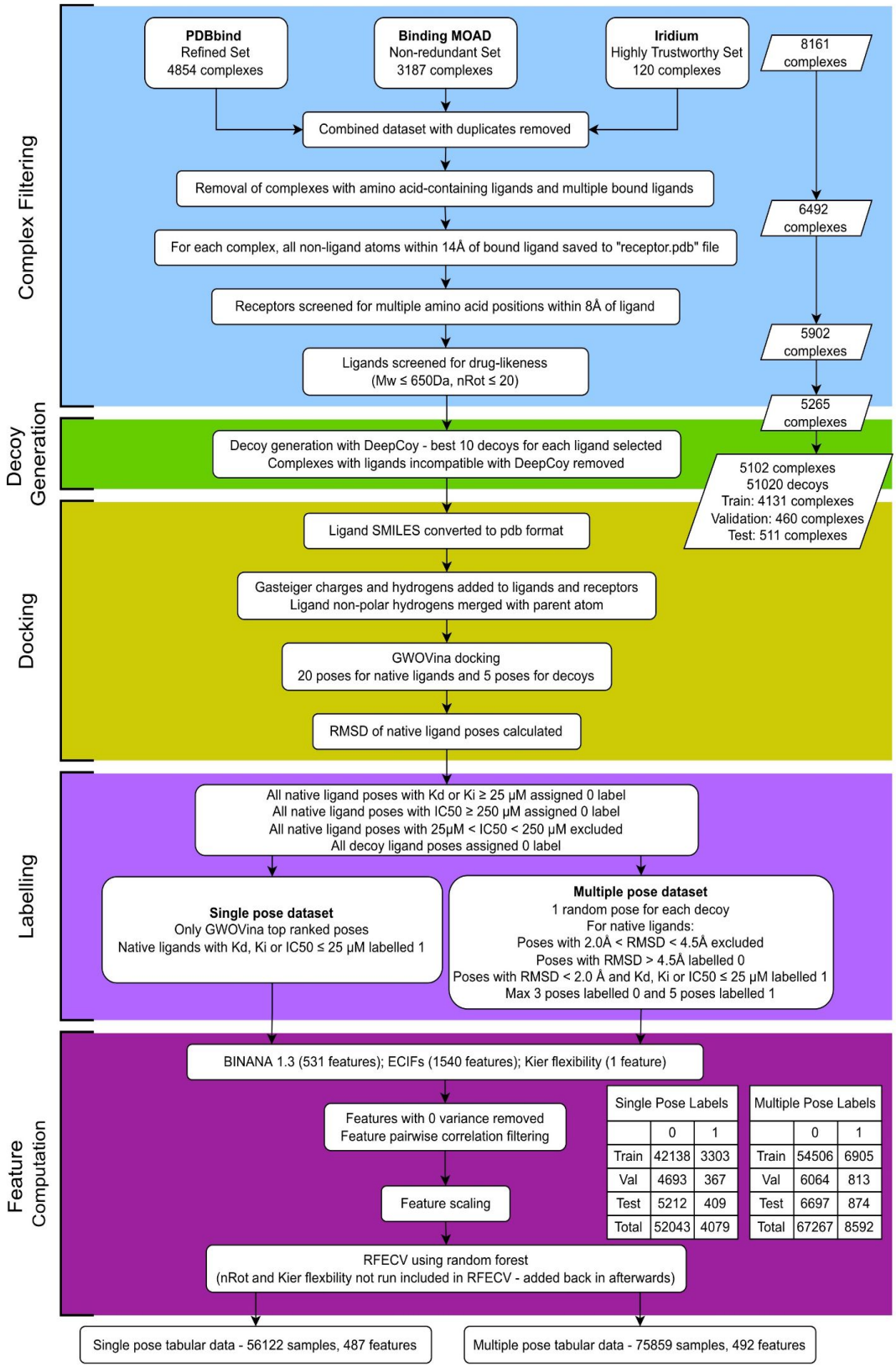

Figure S1. Full data preparation workflow. The steps prior to model development are outlined here, detailing initial data collection and filtering (blue), decoy generation (green), docking (yellow), labelling (light purple), and feature engineering (dark purple).

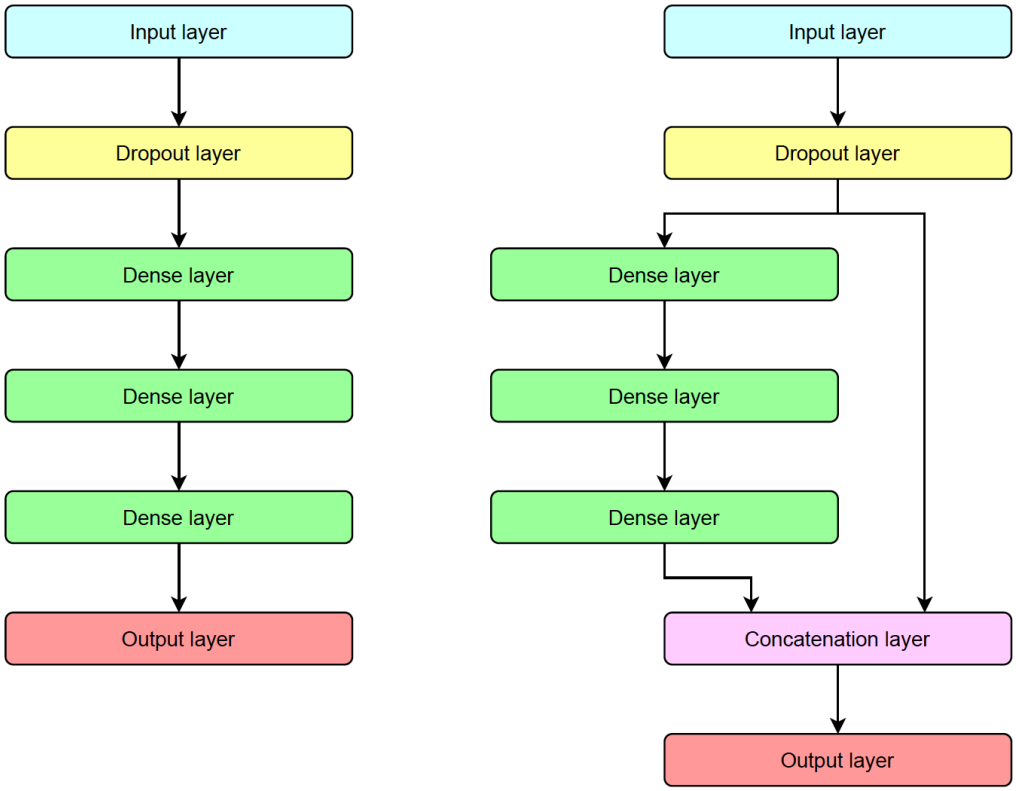

Figure S2. Overall model architecture for feedforward (left) and wide and deep (right) neural networks. Specific parameters for architectures were determined using grid search. Final models used a consensus of 15 networks with varying parameters.

Table S1. Hyperparameters tested for neural network models during the grid search process.

Parameter	Grid Space
Hidden Activation	ReLU; LeakyReLU; ELU
Optimizer	adam; nadam
Learning Rate	0.0005; 0.001; 0.005; 0.01
Input Layer Dropout	0.2; 0.4; 0.6
Kernel Initialiser	Glorot Normal; Glorot Uniform; He Normal
First Layer Nodes	100; 250; 400
FF NN Second Layer Nodes	100; 250; 400
W&D NN Second Layer Nodes	50; 150; 250
FF NN Third Layer Nodes	100
W&D NN Third Layer Nodes	50
L2 Kernel Regularizer	None; 0.02; 0.04
L2 Activity Regularizer	None; 0.02; 0.04
Loss	Binary Crossentropy
Output Activation	Sigmoid
Output L1 Kernel Regularizer	0.02
Output Kernel Constraint Max Value	0.5
Output Kernel Constraint Axis	0

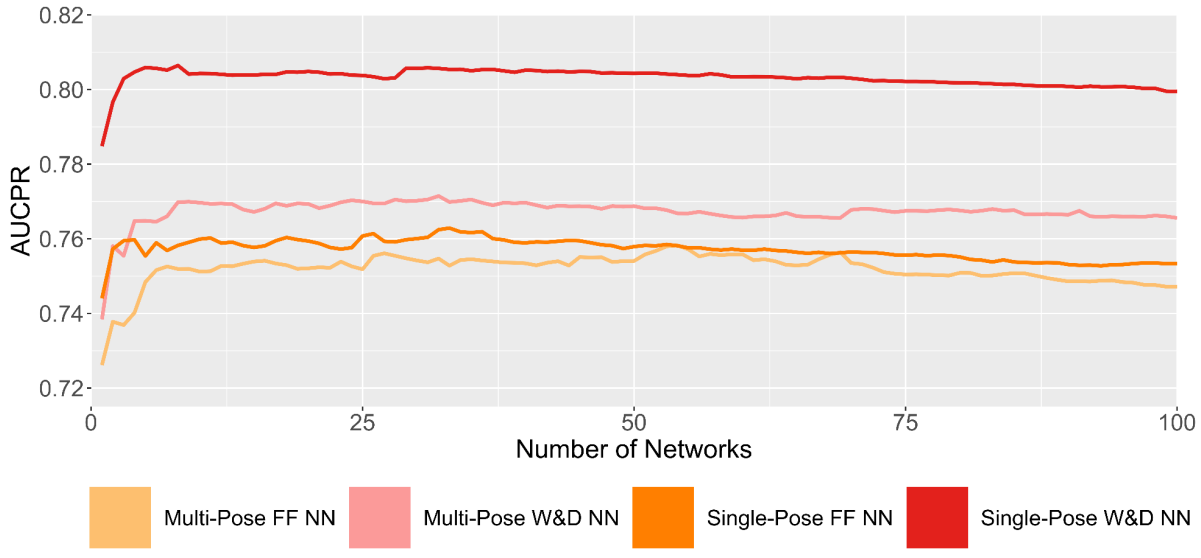

Figure S3. AUCPR on the validation set for varying numbers of networks. Network consensus improves performance, however AUCPR reaches saturation within 15 networks.

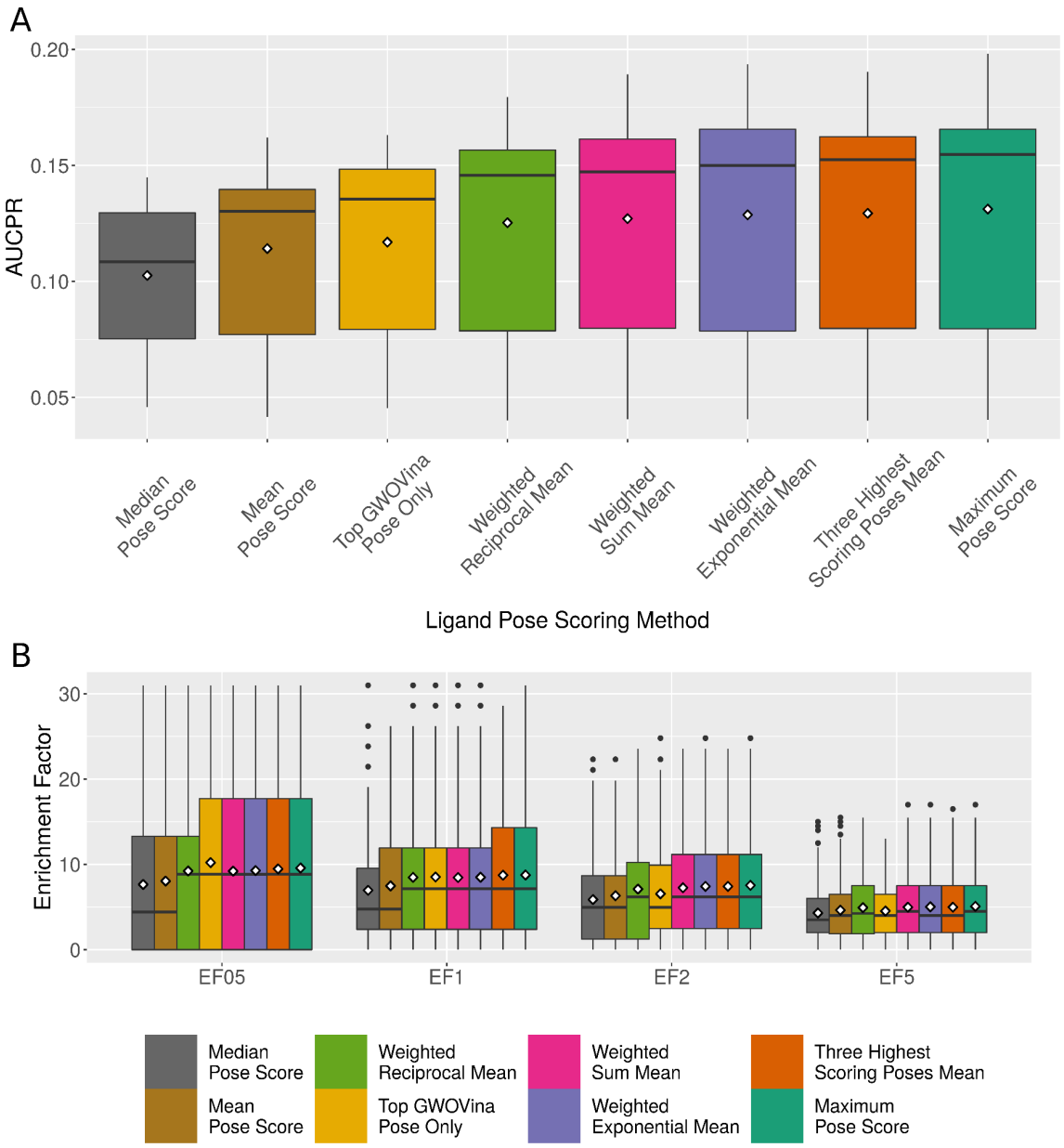

Figure S4. Taking the maximum score of all supplied poses is the best performing ligand scoring method. a) AUCPR for all scoring functions by method of obtaining a single score from 20 docked poses of DEKOIS 2.0 ligands. Weighted means were calculated using five highest scoring poses only. Mean values indicated by white diamonds. b) EF for all scoring functions at four commonly used thresholds based on the method for obtaining a single score from 20 docked poses of DEKOIS 2.0 ligands. Mean values indicated by white diamonds.

Table S2. Enrichment factors at 0.5%, 1%, 2% and 5% for produced machine-learning models and third-party scoring functions on 5,621 test set complexes.

Model	EF0.5	EF1	EF2	EF5	Mean
Multi-Pose XGBoost	13.74	13.74	13.50	11.99	13.24
Multi-Pose W&D NN	13.74	13.74	13.62	11.89	13.25
Multi-Pose FF NN	13.74	13.74	13.62	11.70	13.20
SCORCH	13.74	13.74	13.74	11.99	13.30
Single-Pose XGBoost	13.27	13.50	13.26	11.45	12.87
Single-Pose W&D NN	13.74	13.74	13.38	11.99	13.21
Single-Pose FF NN	13.74	13.74	13.14	12.04	13.16
Single-Pose Consensus	13.74	13.74	13.26	11.99	13.18
DLScore	3.32	1.69	1.70	2.63	2.34
NNScore 1	4.74	4.58	4.50	3.51	4.33
NNScore 2	2.84	3.13	3.77	3.27	3.25
GWOVina	0.47	3.86	4.74	3.61	3.17
RFScoreVS v2	6.63	6.03	5.35	4.24	5.56

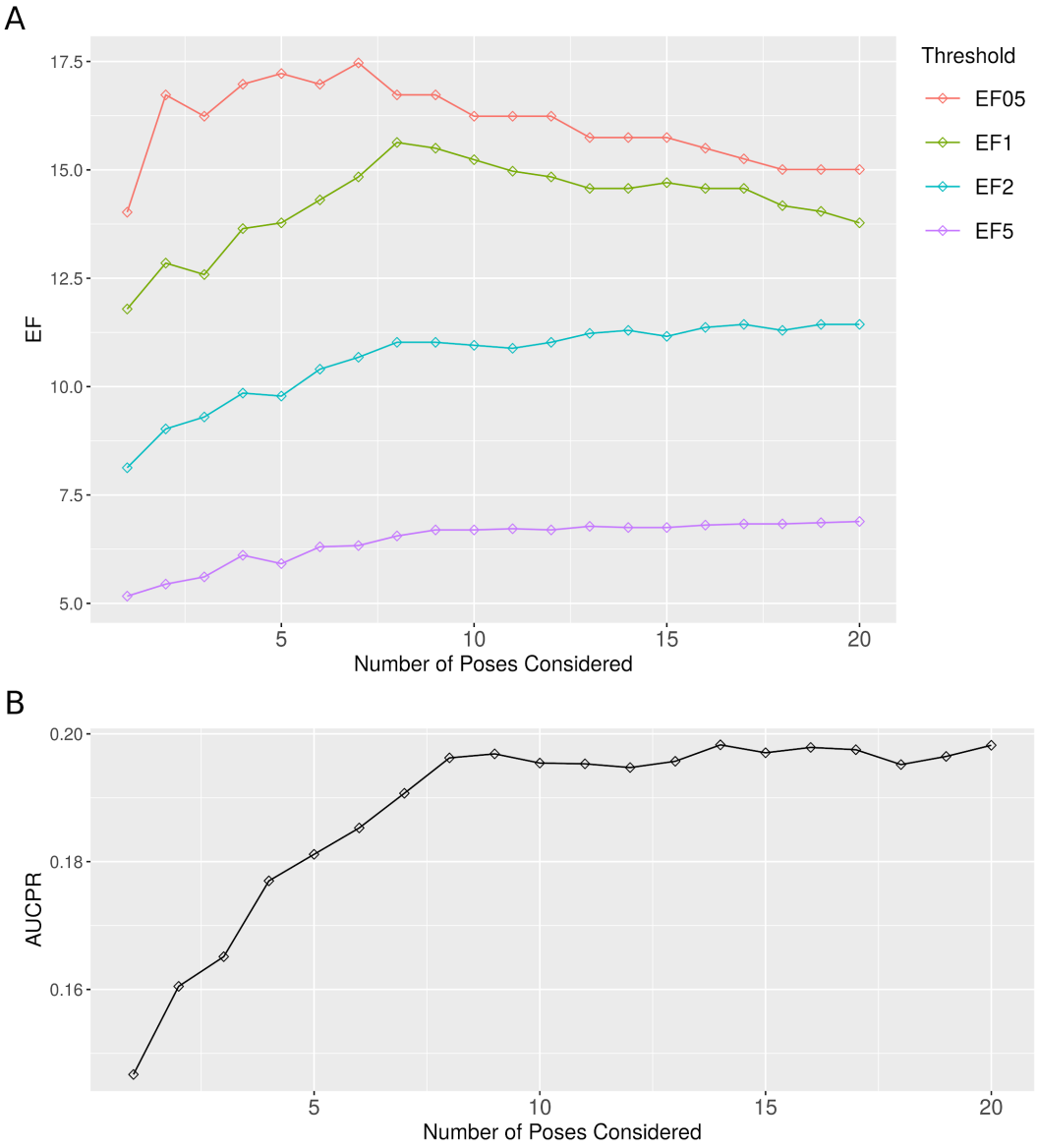

Figure S5. SCORCH performance improves with the number of ligand poses supplied for scoring. a) DEKOIS 2.0 subset EFs with varying numbers of GWOVina produced poses. b) DEKOIS 2.0 subset AUCPR with varying numbers of GWOVina produced poses. Poses are ordered according to GWOVina predicted affinities.

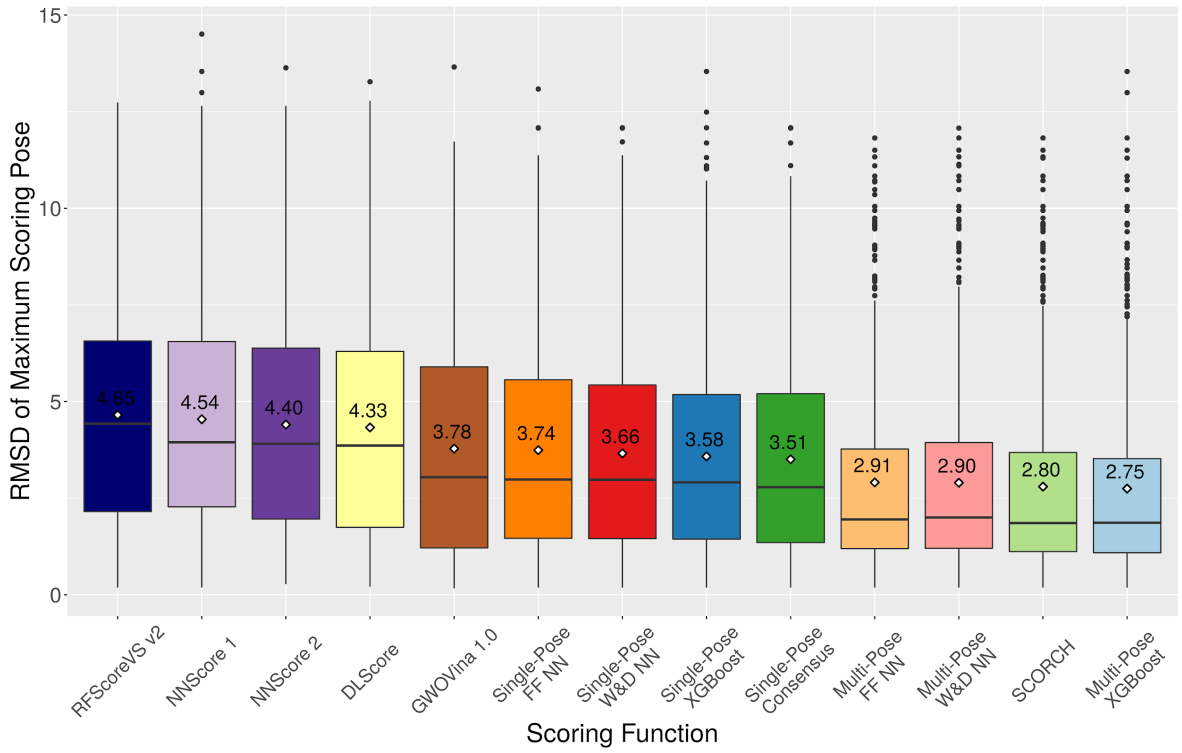

Figure S6. Mean RMSD of highest scoring pose for 510 test set complexes across all evaluated scoring functions. White diamonds indicate the mean rank with the value displayed above each point. Multi-pose trained models assigned higher scores to lower RMSD poses compared to all other tested scoring functions.

Table S3. Mean Spearman's Rank Correlation (ρ) and Kendall Tau (τ) between RMSD and model score for test set pose prediction; RMSDs were ranked from lowest (1) to highest (20). SCORCH predictions have the highest correlation with RMSD.

Model	Mean ρ	Mean τ
SCORCH	0.496	0.376
Multi-Pose XGBoost	0.486	0.368
Multi-Pose W&D NN	0.481	0.363
Multi-Pose FF NN	0.480	0.364
Single-Pose Consensus	0.372	0.279
Single-Pose XGBoost	0.372	0.277
Single-Pose W&D NN	0.348	0.262
Single-Pose FF NN	0.345	0.259
GWOVina	0.260	0.188
NNScore 2	0.221	0.158
DLScore	0.192	0.138
RFScoreVS v2	0.169	0.126
NNScore 1	0.147	0.106

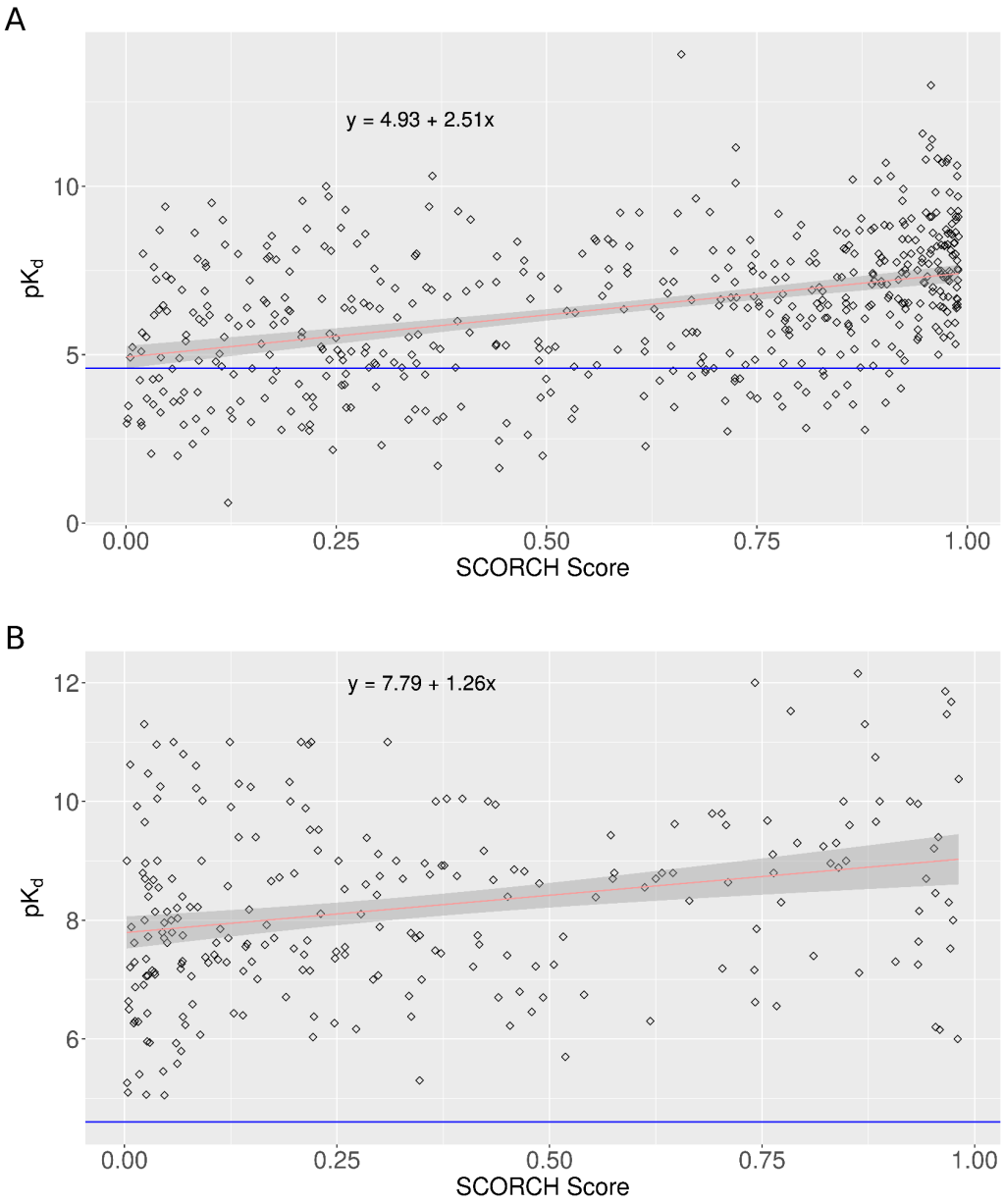

Figure S7. Relationship between SCORCH maximum pose score and pK_d. Fitted regressions line shown in red with the line equation displayed top left; 95% confidence intervals of regression lines are shown in grey. pK_d equivalent to 25μM threshold for active ligands used in SCORCH training is shown with a blue horizontal line. a) Positive relationship between SCORCH score and pK_d for 511 test set complexes (p=2.2e-16 , P_r=0.41). b) Positive relationship between SCORCH score and pK_d for 244 DEKOIS 2.0 test set of active ligands (p=3.97e-05, P_r=0.26).