Rapid Traversal of Ultralarge Chemical Space using Machine Learning Guided Docking Screens

Andreas Luttens[1], Israel Cabeza de Vaca[1], Leonard Sparring[1], Ulf Norinder[2,3,4,*], Jens Carlsson[1,*]

[1]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, BMC, Box 596, SE-75124 Uppsala, Sweden

[2]Department of Pharmaceutical Biosciences, Uppsala University, Box 591, SE-75124, Uppsala, Sweden

[3]Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-16407, Kista, Sweden

[4]MTM Research Centre, School of Science and Technology, Örebro University, SE-70182, Örebro, Sweden

[*]To whom correspondence should be addressed: jens.carlsson@icm.uu.se

Keywords: Molecular docking, chemical space, conformal prediction, virtual screening

Abstract

The accelerating growth of make-on-demand chemical libraries provides novel opportunities to identify starting points for drug discovery with virtual screening. However, the recently released multi-billion-scale libraries are too challenging to screen even for the fastest structure-based docking methods. Here, we introduce a strategy that combines machine learning and molecular docking to enable rapid virtual screening of databases containing billions of compounds. In our workflow, a classification algorithm is first trained to identify top-scoring compounds based on molecular docking of one million compounds to the target protein. The conformal prediction framework is then used to make selections from the multi-billion-scale library, drastically reducing the number of compounds to be scored by the docking algorithm. The performance of the approach was benchmarked on a set of eight different target proteins, and classifiers based on gradient boosting, deep neural network, and transformer architectures were evaluated. The CatBoost classifier exhibited the optimal balance between speed and accuracy and was used to adapt the workflow for screens of ultralarge libraries. The optimized workflow was demonstrated to identify >90% of the very top-scoring molecules in a library with 0.2 billion compounds, which only required docking of 3-5% of this set. Application to a library with >3.5 billion compounds showed that molecules with substantially improved docking scores can be identified by machine learning, enabling efficient virtual screening of the largest commercial chemical libraries available. The accelerated virtual screening workflow has been made publicly available to facilitate exploration of vast chemical libraries for drug discovery.

Introduction

The number of possible drug-like molecules has been estimated to be more than 10^{60}, which exceeds the size of chemical libraries evaluated in early drug discovery by many orders of magnitude[1–3]. In fact, only ~13 million compounds are currently available in-stock from chemical suppliers, which clearly illustrates the limited chemical space coverage[4]. Advances in synthetic organic chemistry have provided access to increasingly larger compound collections and make-on-demand libraries currently contain >30 billion readily available molecules[5]. The diverse scaffolds available in these libraries represent a major opportunity for drug discovery, but identifying the subset of compounds relevant for a specific therapeutic target remains a major challenge.

Recently, structure-based virtual screens of ultralarge libraries have identified ligands of important therapeutic targets, demonstrating that expanding the coverage of chemical space can accelerate hit discovery.[6–10] The most recently published docking screens have reached billions of compounds[11–13], but these massive libraries are demanding to evaluate due to the substantial computational resources required. The make-on-demand databases will also continue to grow and likely reach several hundred billion compounds in the near future, which will be unfeasible to screen even with the fastest structure-based docking algorithms.[14] Therefore, there is an urgent need for more efficient virtual screening approaches able to evaluate multi-billion-scale libraries.

Recent breakthroughs in artificial intelligence have revived interest in using Quantitative Structure-Activity Relationship (QSAR) models in drug discovery. QSAR has been widely used by the pharmaceutical industry to predict on- and off-target activities, physicochemical and pharmacokinetic properties.[15–18] By representing compounds using molecular descriptors (*e.g.*, fingerprints), machine learning methods can rapidly evaluate large compound databases. Traditionally, QSAR models have been trained on experimental data, but there is an increasing interest to predict which compounds in make-on-demand libraries are likely to receive favorable scores from computationally expensive virtual screening methods.[19–23] This combination of machine learning and molecular docking screening has the potential to enable virtual screens of multi-billion-scale compound libraries at a modest computational cost.

In this work, we developed an ultra-fast workflow based on conformal prediction (CP) for screening of vast chemical libraries. The CP framework can be applied to any machine learning classifier and allows the user to control the error rate of the predictions.[24–26] CP also performs well on imbalanced datasets, which is the case in virtual screening applications because only the very top-scoring compounds in the library ("virtual actives") are of interest.[27] This framework has previously been applied successfully to predict pharmacokinetic properties and bioactivity[28–30]. Recently, strategies to improve the virtual screening efficiency using the CP framework have been explored, but these workflows were not suitable for multi-billion-scale libraries and focused on traditional classifiers.[31,32] Applications of more recently developed techniques such as gradient boosting, deep neural networks, and transformers to early-phase drug discovery have been very successful, including applications to molecular docking.[33–36] Here, we combined the CP framework with several state-of-

the-art classification algorithms to develop a workflow for accelerated structure-based virtual screening. We demonstrate that our most efficient workflow identifies the top-scoring compounds in ultralarge compound libraries and reduces the number of molecules to be explicitly docked by three orders of magnitude.

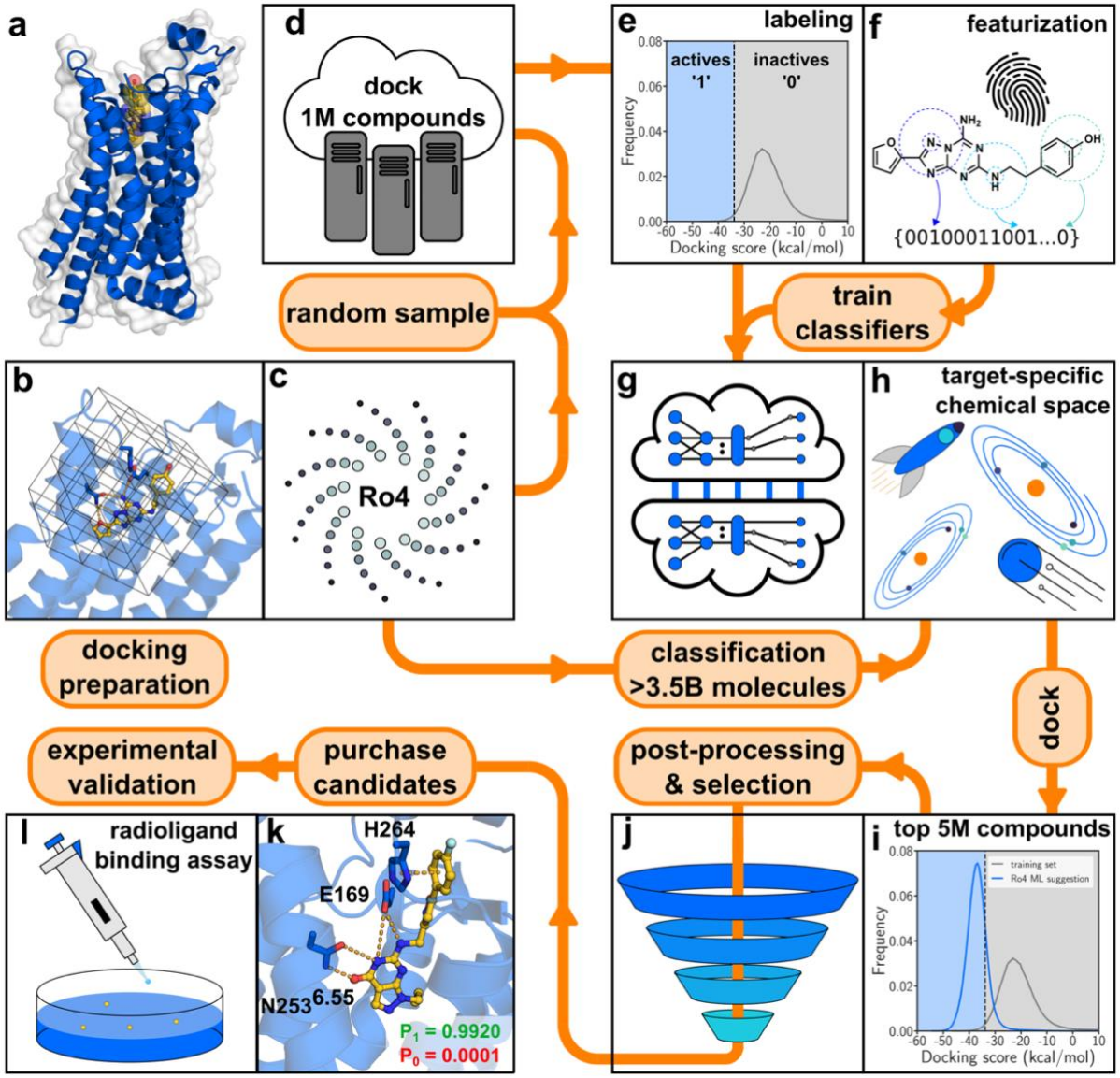


Figure 1. Machine learning accelerated virtual screening workflow. (a-b) Selection and preparation of a target protein for molecular docking calculations. (c) A subset from an ultralarge chemical library is extracted and prepared for docking screens. (d) Docking scores for compounds in the training set are generated. (e) A docking score threshold splits the training set into virtual actives (1-class) and inactives (0-class). (f-g) Molecules in the training set are represented by molecular descriptors (*e.g.,* fingerprints) and a classifier is trained to distinguish virtual actives from inactives. (h) The trained classifier is used to identify a subset of predicted virtual actives in the ultralarge library. (i) A set of compounds is selected for docking to the target. (j) Post-processing of docking results and selection of compounds. (k-l) Selected compounds are synthesized and experimentally evaluated.

Results and discussion

In the development of the virtual screening workflow, the use of classifiers to enable the evaluation of ultralarge compound libraries was explored. Our approach was first evaluated by conducting docking screens of 10 million compounds against eight different protein targets, and this benchmarking set guided the selection of classifiers and molecular descriptors. In the second step, the method was optimized to perform virtual screens of multi-billion-scale libraries.

Machine learning accelerated virtual screening pipeline. Our workflow for combining machine learning and molecular docking (Figure 1) is freely distributed and consists of the following consecutive steps, which are described in detail in the methods section and in Supplementary Figure S1:

Step 1. *Preparation and docking of the training set.* A set of randomly selected molecules from an ultralarge chemical library is docked to the target protein structure (Figure 1a-d). We recommend a training set of one million molecules in virtual screens of multi-billion-scale libraries.

Step 2. *Generation and labeling of the training set.* A docking score threshold (Figure 1e) is selected to label each compound in the training set as either virtual active (better score than the selected threshold) or inactive (equal or worse score than the selected threshold). As our CP approach is based on aggregating predictions made by several classifiers, multiple independent training sets are generated. Our recommendation is to label the top-scoring 1% of the training set as virtual active and generate five independent training sets.

Step 3. *Molecule featurization and training of the classifier.* Molecular descriptors of each molecule in the training set are generated as input for the classifier. Each of the training sets is used to train an independent classification model to distinguish virtual actives from inactives (Figure 1f-g).

Step 4. *Conformal prediction for the ultra-large library.* The trained classification models are used to evaluate compounds from the ultralarge chemical library (Figure 1h). The Mondrian CP framework is then used to categorize the compounds into one of the following four sets based on a selected significance level (ε): virtual active, virtual inactive, both = virtual active or inactive, and null = no class assignment. The significance level can be tuned to control the size of the virtual active set, which is predicted to contain compounds with a docking score better than the selected threshold.

Step 5. *Post-processing and compound selection.* The database pruning level by the workflow is target-dependent, and additional post-processing steps can be applied to identify the most promising compounds. The compounds assigned to the virtual active set are rank-ordered by sorting them based on the quality of information (*i.e.,* prioritizing the predictions in which the classification model has the highest confidence) and a subset of these are docked to the target. Top-scoring molecules are clustered by chemical similarity and representative compounds are visually inspected (Figure 1i-j), followed by synthesis and experimental evaluation of selected compounds (Figure 1k-l). We recommend docking a set of 1-5 million molecules selected based on the quality of information.

Benchmarking set and training of classifiers. Docking screens against eight therapeutically relevant proteins were carried out to optimize the performance of the workflow. The benchmarking set represented different types of protein folds, binding sites, protein-ligand interactions, and ligand chemotypes. G protein-coupled receptors (GPCRs) were represented by the A_{2A} adenosine receptor ($A_{2A}R$) and the D_2 dopamine receptor (D_2R)[37–40]. The SARS-CoV-2 main protease (M^{pro}), 8-oxoguanine glycosylase 1 (OGG1), ecto-5'-nucleotidase (5'-NT), and AmpC β-lactamase (AmpC) exemplified different types of soluble enzymes.[6,9,41–43] Finally, the Kelch-like ECH-associated protein 1 (KEAP1) and Sortilin (SORT1) represented protein-protein interaction interfaces[44–46]. A set of 11 million randomly sampled molecules from the Rule-of-Four space (molecular weight < 400 Da and LogP < 4) of the largest available make-on-demand library (Enamine Real-Space) was docked to each target, resulting in a final benchmarking set of 120 million complexes and their corresponding docking scores. For each target, the docking scores and chemical structures of the compounds were used to create training (10^6 compounds) and test (10^7 compounds) sets suitable for the CP framework. Unless noted otherwise, the energy threshold for the active class was determined based on the top-scoring 1% of each screen.

Three different ML classifiers were trained for each target in the benchmarking set: CatBoost[47], Deep Neural Networks (DNNs)[48], and Bidirectional Encoder Representations from Transformers (BERT)[49]. CatBoost and DNN classifiers were evaluated using two types of molecular descriptors: Morgan2 fingerprints and continuous data-driven descriptors (CDDD)[50,51]. The BERT classifier is based on a pre-trained encoder (RoBERTa) that uses SMILES as input.[49] Detailed descriptions of

the hyperparameters used in the training of each classifier are provided in Supplementary Table S1 and Figures S2-S4. Unless explicitly stated otherwise, five independent models were generated based on the training set. The compounds in the test set (10 million compounds) were assigned normalized p-values (p_1 and p_0) by each individual classification model and its corresponding calibration set. The resulting sets of five p_1 and p_0 values were aggregated into a single pair of p-values by taking the medians[52]. Based on the aggregated p-values and the selected significance level, the Mondrian CP framework was used to divide the compounds into virtual active, virtual inactive, both (*i.e.* either virtual active or inactive), or null (no class assignment) sets (Figure 2a). Performance for the benchmarking set was assessed using the significance level at which the CP framework resulted in the maximal number of useful predictions, *i.e.* single-label predictions for compounds (ε_{opt}) (Figure 2b)[53]. If the training and test data are exchangeable, the CP framework leads to agreement between the prediction error rate and the selected significance level (Figure 2c).[25] The performance of each configuration of classifier and molecular representation was assessed based on analysis of the resulting sensitivity, precision, efficiency, and prediction error rate (for definitions, see methods).

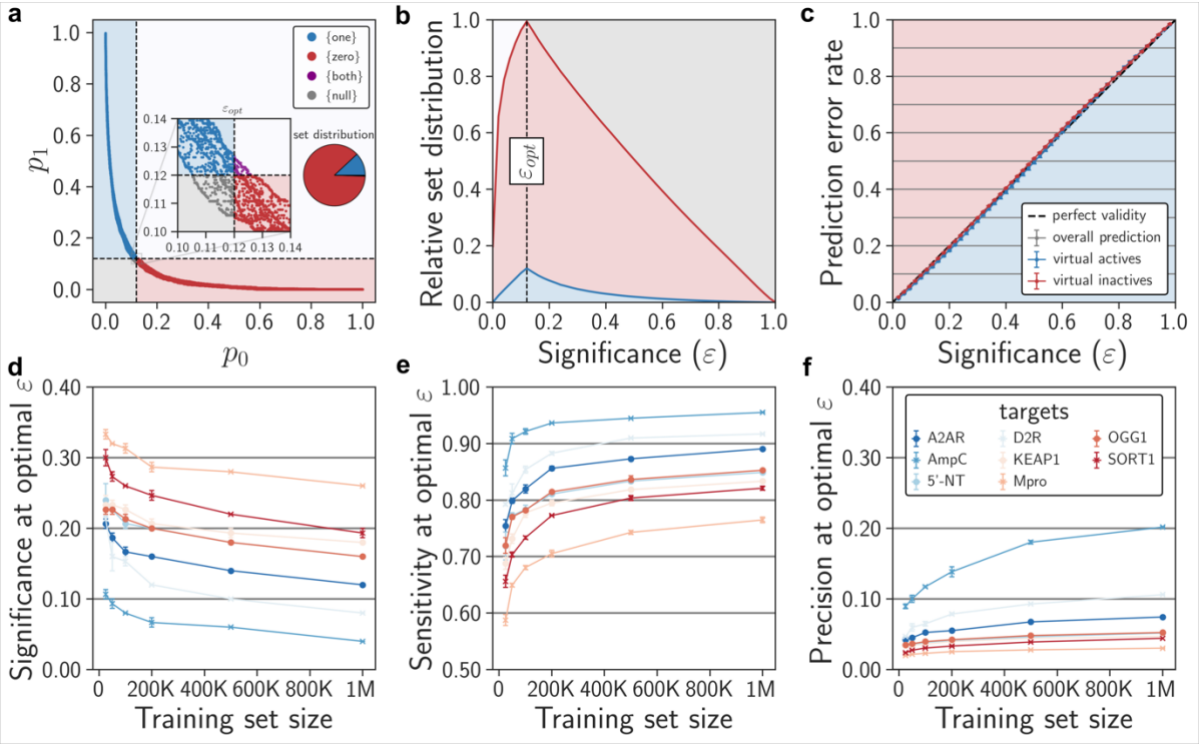


Figure 2. Benchmarking of classifiers and molecular descriptors. (a-c) Summary of application of the Mondrian CP framework to one of the targets in the benchmarking set ($A_{2A}R$). (a) Molecules were classified into four distinct sets based on their p-values and a selected significance threshold (ε): virtual actives (blue, 1-class), virtual inactives (red, 0-class), both (purple, 1- or 0-class), null (grey, no class assignment). (b) The $A_{2A}R$ test set molecules were divided into four prediction sets depending on the significance level. The optimal significance (ε_{opt}) corresponds to the value at which the maximal number of compounds have been assigned to single-label set (i.e., either virtual actives or inactives), *i.e.*, at maximal efficiency. (c) The error rate obtained for predictions of the $A_{2A}R$ benchmarking set compounds with respect to the significance threshold (calibration plot). There was a close agreement between the significance value and the prediction error rate. (d) The optimal significance level improved if the classification models are trained on larger datasets. (e) At optimal efficiency, the sensitivity values improved with increasing size of the training set. (f) At optimal efficiency, the precision values improved with increasing training set size. In (d-e), three independent calculations (training and prediction) were performed for the eight targets and error bars correspond to the standard error of the mean.

Evaluation of classifiers and molecular descriptors. An optimal size of the training set is crucial to minimize the number of compounds to dock and maximize the performance of the classifier. The effect of the training set size on the test set sensitivity and precision values was evaluated using sets ranging from 25000 to 1000000 compounds. The average sensitivity and precision values improved with increased training set sizes for all combinations of classifiers and molecular descriptors (Figure 2 and Supplementary Table S2). A sharp improvement of

sensitivity and precision was obtained by increasing the training set from 25000 to 200000 compounds. Training sets of 500000 compounds further increased performance, but only incremental improvements were obtained for more than one million compounds. Based on these results, molecular descriptors were evaluated using a training set with one million compounds (Table 1).

Table 1. Performance of classifiers and molecular representations using a training set of one million compounds.

Classifier[a]	Descriptor	Performance for benchmarking set[b]		
		Sensitivity	Precision	Significance
CatBoost	Morgan2	0.86 ± 0.01	0.08 ± 0.01	0.15 ± 0.01
	CDDD	0.84 ± 0.01	0.06 ± 0.01	0.17 ± 0.01
DNN	Morgan2	0.82 ± 0.01	0.05 ± 0.01	0.18 ± 0.01
	CDDD	0.84 ± 0.01	0.06 ± 0.01	0.15 ± 0.01
RoBERTa		0.85 ± 0.01	0.07 ± 0.01	0.15 ± 0.01

[a] Five independent models were trained on one million compounds. Detailed descriptions of the hyperparameters used in the training of each classifier are provided in Supplementary Table 1. [b] Values represent mean ± SEM of 24 test set predictions of 10 million compounds (three replicates of eight individual target datasets).

Morgan2, CDDD, and RoBERTa consistently resulted in high average sensitivity values (0.82-0.86) and the three classifiers showed similar performance. The main differences between the classifiers were instead in the precision, significance, and computational cost. On average, the significance values ranged from 0.15 to 0.18 with prediction efficiencies exceeding 0.99. In other words, the CP framework was able to classify nearly all evaluated compounds as either virtual active or virtual inactive with an average error rate of 15-18%. Whereas deviations in *validity*, i.e., the agreement between the selected significance and resulting error rate, is often observed in applications where insufficient data is available[54], the performance of the CP on molecular docking data yielded the expected error rate for all targets in the benchmarking set (Figure 3c).

The CatBoost classifier together with Morgan2 fingerprints resulted in the best average precision and had comparable or slightly better significance and sensitivity values compared to the other combinations. In addition, CatBoost/Morgan2 required the least computational resources, both in the training of the classifier, predictions for the test set, and storage of molecular descriptors. Based on these results, further assessments of the parameters used by the classifier (*e.g.*, the number of models and class imbalance) were focused on the CatBoost/Morgan2 combination. Increasing the number of classification models from five to ten did not significantly increase the performance of CatBoost/Morgan2, and the results were also robust if the size of the minority class (virtual actives) was decreased from 1% to 0.1% or 0.01% (Supplementary Figure S5).

Analysis of the results for each protein target in the benchmarking set demonstrated that the performance varied depending on the target, with sensitivity values ranging from 0.76 to 0.96 for CatBoost/Morgan2. As 100000 compounds in the test set belonged to the actives class, a maximal reduction of 100-fold could be achieved if all compounds were correctly classified. The largest database reduction was obtained for AmpC, a beta-lactamase targeted for the development of antibiotics[6,55]. For AmpC, 474646 out of the 10 million compounds in the test set were assigned to the virtual active class, corresponding to a 21-fold database reduction, and 96% of the true virtual actives were among these. The worst performance was obtained for the target M^{pro}, which is a viral protease relevant for development of drugs for treatment of COVID-19[9,56]. In this case, the database was reduced by four-fold and 76% of the true virtual actives were identified. The target dependent results of machine learning accelerated protocols have been observed previously[20,31], and analysis of our docking results

indicate that the performance is influenced by the nature of the binding site, the diversity of the top-ranked compounds, and the docking score distribution. For example, the top-scoring compounds of open and solvent-exposed binding sites tend to be more structurally diverse, which affects the ability of the classifier to recognize patterns in the docking data.

The results from the benchmarking set clearly demonstrated that top-scoring compounds in a chemical library were identified by the machine learning protocol. However, the potential activity of these compounds in experimental assays was unknown and, considering the high false positive rate of molecular docking screens, a majority of these likely do not bind to the target. To assess if the workflow was able to identify experimentally confirmed actives, we extracted known ligands from the ChEMBL database for two targets in the benchmarking set.[57] For the $A_{2A}R$ and D_2R, there were thousands of ligands with activity values better than 10 μM, and predictions were made for these sets using the CatBoost/Morgan2 classifier trained only on docking results. Encouragingly, the CP framework correctly assigned 92% and 86%, respectively, of the experimentally confirmed $A_{2A}R$ and D_2R ligands to the virtual active class at optimal efficiency. Assessment of the classification of known actives can serve as an important control in preparation of a prospective virtual screen. In this context, it should be noted that the machine learning step of the workflow relies on careful selection of the protein structure and docking parameters, *e.g.*, by performing enrichment calculations with small sets of known actives and decoys.[10] If the docking scoring function performs poorly in these assessments, the final results tend to be even worse in screens of large libraries.[6,14]

Optimization of performance for ultralarge chemical libraries. A primary goal in the development of the workflow was that the machine learning step must be able to reduce a multi-billion-scale database to a few million compounds. To optimize the performance for ultralarge databases, we docked 235 million compounds to two proteins from the benchmarking set ($A_{2A}R$ and D_2R). A CatBoost classifier in combination with Morgan2 fingerprints was then trained on one million compounds for each target, followed by predictions for the remaining part of the library. As the docking scores were available for all the compounds in the library, efficient strategies to identify the top-scoring molecules could be identified.

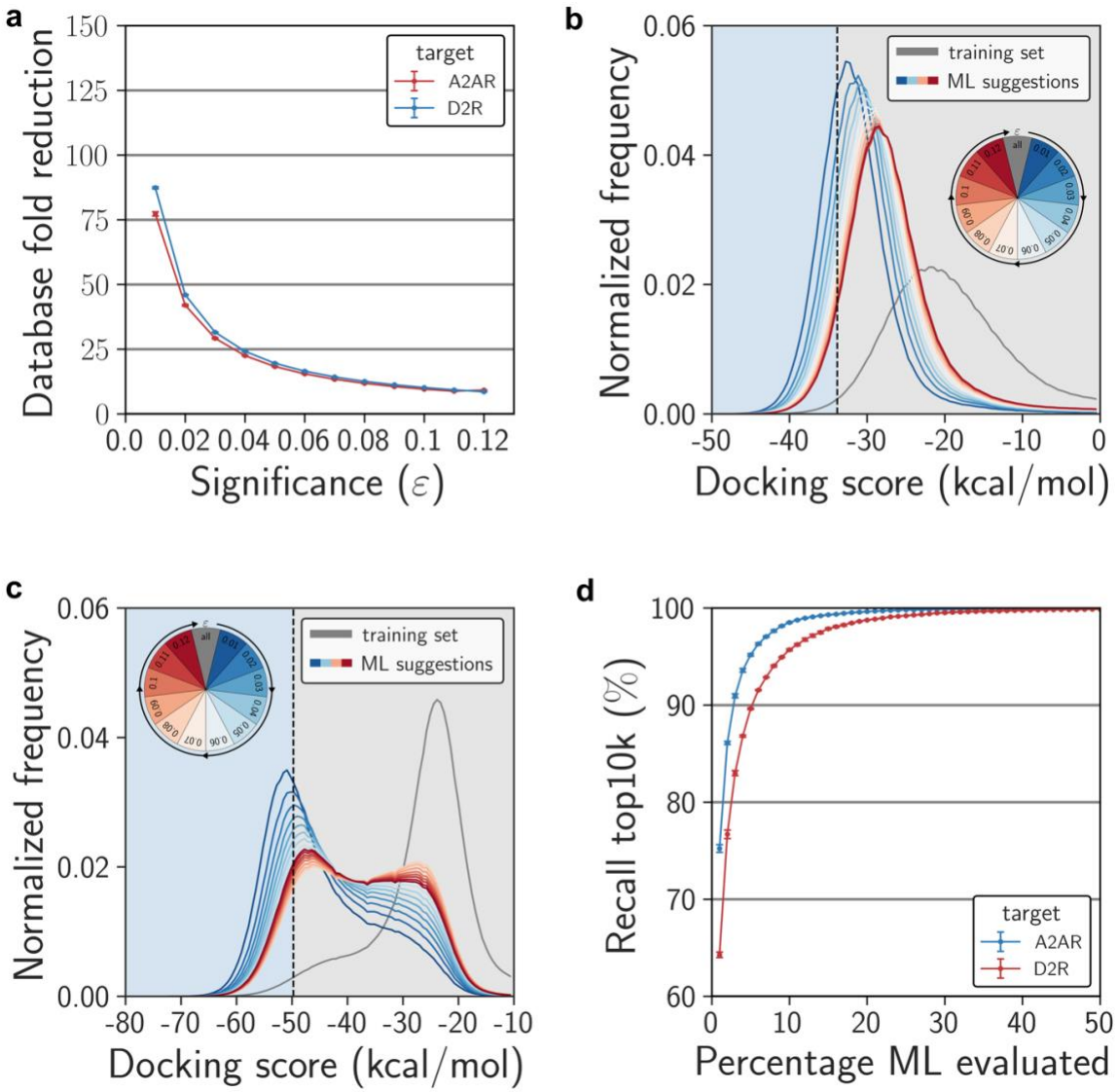


Figure 3. Machine learning performance for ultralarge docking screening data. Five independent Catboost classifiers were trained on one million molecules from a docking screen of 235 million molecules against the $A_{2A}R$ and D_2R. (a) The size of the predicted active class decreases with more stringent significance values. (b) Normalized frequency distributions of DOCK scores present in the ultralarge docking screen. The score distribution of the training set is shown in gray. In color (red to blue), score distributions of molecules predicted to be active at a given (increasing stringency) significance threshold (ε). (d) Molecules in the test set were sorted based on the quality of information. The percentage recall of the 10000 best-scoring molecules is shown as a function of the percentage evaluated compounds in the test set.

In the CP framework, the selected significance level determines the size of the predicted virtual active set, which is the set that will be docked to the target. The significance level was set to achieve the maximal efficiency, and close to all compounds received a single label (>98% for both targets). CP reduced the ultralarge library from 234 to 25 and 19 million compounds for the $A_{2A}R$ and D_2R, respectively,

with high sensitivity values (0.87 and 0.88, respectively). The workflow would hence be able to identify close to 90% of the virtual actives by docking only ~10% of the ultralarge library and the CP framework guaranteed that the percentage of incorrectly classified compounds less than 12%. For libraries of this size, molecular docking screens of the predicted virtual active set in order to select compounds for experimental evaluation would be viable. However, further reduction of the database would be required to apply the workflow to multi-billion-scale libraries, as in these cases docking calculations for even a small percentage of the library would be computationally too demanding. In theory, decreasing the significance level should lead to a reduction of the virtual active set and enrich predictions in which the classifier has high confidence. This approach was evaluated by gradually reducing the significance level and assessing how the distribution of docking scores in the virtual active set was influenced. As anticipated, lowering the significance level did reduce the virtual active set size (Figure 3a) and also led to marked shifts of the docking score distribution towards better energies for both protein targets (Figure 3b-c). At the lowest evaluated significance level (0.01), the database was reduced to 3.0 and 2.6 million molecules for the $A_{2A}R$ and D_2R, respectively, and the largest shifts in docking score distributions were obtained. For example, the most populated bin in the docking score distribution for the training set was −21.7 kcal/mol for the $A_{2A}R$, which was improved to −28.5 and −32.8 kcal/mol for significance levels of 0.12 (ε_{opt}) and 0.01, respectively. At the strictest significance level (0.01), 80% of the 10000 top-scoring molecules, corresponding to 0.004% of the chemical library, were identified. These results showed that the significance level can be tuned to achieve substantial database reduction, whilst retaining most of the very top-scoring candidates for the subsequent docking step.

An alternative approach to reduce the size of the set to evaluate by molecular docking is to sort the compounds based on the difference between the p_1 and p_0 values (the quality of information, $p_1 - p_0$), which gives priority to subsets in which the predictor has the highest confidence. The enrichment of the top-scoring 10000 molecules from the $A_{2A}R$ and D_2R screens were assessed based on prioritizing the compounds using the quality of information. Remarkably, the single-iteration workflow identified more than 90% of the very top-scoring molecules after only 3% ($A_{2A}R$) and 5% (D_2R) of the 234 million compounds had been evaluated (Figure 3d). Using the quality of information to reduce the docked set of compounds hence had a similar effect as decreasing the significance level, and these two techniques can be combined in prospective screens of multi-billion-scale libraries.

Virtual screening of a multi-billion-scale library. The goal of this study was to develop a method suitable for the virtual screening of multi-billion-scale libraries, which would be computationally too demanding to evaluate explicitly by molecular docking. To assess the efficiency of our combined machine learning approach, we performed predictions for the entire Rule-of-Four chemical space of the Enamine REAL space database. This database contained more than 3.5 billion molecules, which we estimate would require 7.3 million core-hours/target (*i.e.,* 833 years on a single core) to screen by molecular docking using the same parameters as for the benchmarking set. In contrast to the benchmarking sets, the docking scores of the compounds in this library were unknown and this screen hence represented the ultimate test case for the method. The predictions were performed for two targets ($A_{2A}R$ and D_2R), and the

ability of our workflow to reduce the size of the database and the docking scores of the predicted compounds were evaluated.

Docking of the training set, training of the classifier, and predictions for 3.5 billion compounds for one target were performed in approximately 2500 core-hours. The significance level was set to 0.005, resulting in 25 and 24 million predicted virtual actives for the $A_{2A}R$ and D_2R, respectively. Of these, five million compounds per target were prioritized for docking calculations based on the quality of information, corresponding to a 700-fold reduction of the library and was performed in 10344 core hours per target. Compared to explicit docking of the 3.5 billion compounds, the workflow hence achieved a 568-fold reduction of compute cost. The docking score distribution of the five million compounds was substantially shifted towards better energies for both targets (Figure 4a-b). For example, the most populated bin in the docking score distribution of the training set was −23.2 kcal/mol for the $A_{2A}R$, which was shifted to −35.3 kcal/mol for the predicted virtual actives. A majority of the predicted compounds (58%) had a docking score better than the energy threshold (−33.9 kcal/mol) used for labeling of the training set, corresponding to a 58-fold enrichment of virtual actives. Similar distributions of docking energy were obtained if only one million predicted virtual actives were selected for molecular docking (Figure 4a-b), demonstrating that the user can select the degree database reduction and achieve reductions of 3500-fold for this library.

To increase the molecular diversity, the 100000 top-scoring compounds, corresponding to 0.003% of the library, were clustered by topological similarity. The best-scoring molecule from each cluster was visually inspected for their

complementarity with the binding site. Encouragingly, the docked molecules formed interactions with residues that have been observed to be important for ligand binding to the $A_{2A}R$ (Asn253) and D_2R (Asp114). These observations indicate that the machine learning accelerated virtual screening workflow can identify molecules that have excellent docking scores and similar interactions as known ligands, which would be relevant for experimental evaluation in a prospective screen.

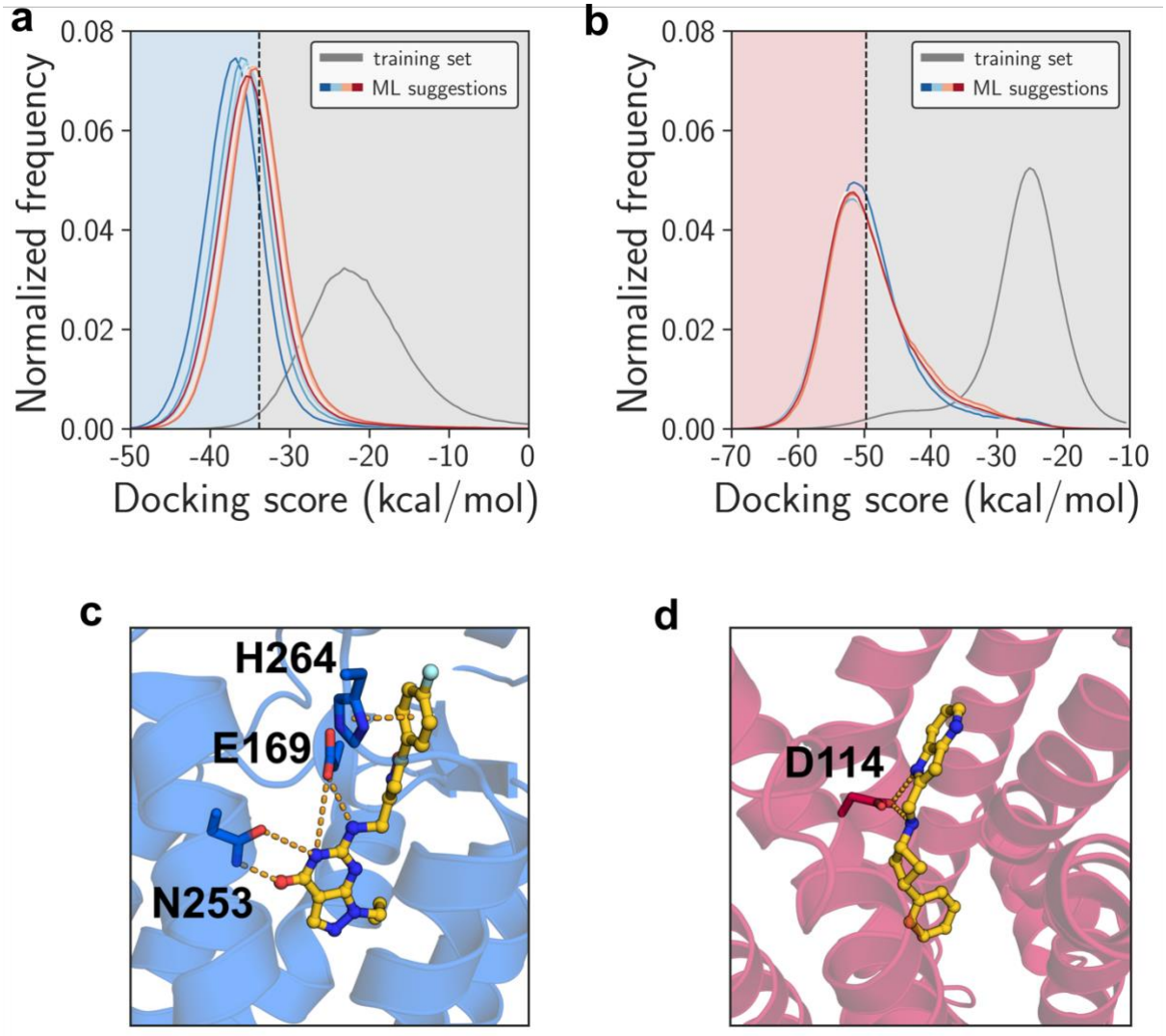


Figure 4. Virtual screen of multi-billion-scale libraries. The machine learning accelerated workflow was used to predict compounds for the $A_{2A}R$ and D_2R targets in a database with 3.5 billion compounds. (a) Normalized frequency distributions of $A_{2A}R$ docking scores. The docking score distribution of the training set is shown in gray. In color (red to blue), docking score distributions of top-ranked molecules according to their difference in p-values. Different distributions are based on one million molecules. (b) Normalized frequency distributions of D_2R docking scores. The docking score distribution of the training set is shown in gray. In color (red to blue), score distributions of top-ranked molecules according to their difference in p-values. Different distributions represent one million molecules. (c, d) Examples of top-scoring molecules that form key interactions with the binding site for the $A_{2A}R$ (c) and D_2R (d).

Conclusions

Make-on-demand chemical libraries can be expected to continue to grow and will likely reach one trillion compounds within a few years. However, navigating in this chemical space has become increasingly difficult due to the tremendous computational costs of screening these libraries. In this work, we demonstrate that classifiers can be trained to prioritize relevant chemical space for a protein target based on molecular docking screens of small sets of compounds. By using a conformal predictor to guide structure-based virtual screening, top-scoring compounds in ultralarge libraries can be rapidly identified after docking only a few million molecules, which will facilitate discovery of starting points for development of novel therapeutics. The workflow is freely distributed and allows exploration of vast chemical space with modest computational resources.

Materials and methods

Docking library preparation. Enamine's November 2019 REAL space library (12.3 billion compounds) was reduced to a Rule-of-Four chemical subspace by excluding compounds with a molecular weight over 400 MW and cLogP over 4 as calculated by the RDKit.[58] The total size of the Rule-of-Four subspace was 3,541,746,925 compounds. A representative random sample containing 15 million compounds (0.4%) from this library was obtained after shuffling the SMILES with Terashuf.[59] Molecules were prepared for docking using DOCK3.7 standard protocols.[60] ChemAxon's CXCalc (from ChemAxon's Marvin package Marvin 18.10.0) was used for calculating predominant protomers at relevant pH levels (6.9, 7.4, 7.9). Conformational ensembles were generated with OMEGA (OpenEye, version 2020.2) and were capped at 400 conformations per rigid segment and an inter-conformer RMSD diversity threshold of 0.25 Å.

Molecular descriptors. Canonical SMILES were used to generate three different molecular descriptors as input data for the machine learning classifiers. Extended-Connectivity fingerprints (ECFP4 with 1024 bits and radius 2) were generated using the RDKit.[50,58] Continuous data-driven descriptors were generated using the CDDD Python library.[51] The RoBERTa model generates its own descriptors directly from the SMILES. We used a pretrained RoBERTa model[61] to generate the internal encoded representation of each molecule during runtime using the Python library simpletransformers[62].

Preparation of proteins for docking. Crystal structures of M^{pro}, SORT1, 5'-NT, $A_{2A}R$, D_2R, OGG1, AmpC, and KEAP1 were extracted from the PDB.[6,38,39,41,42,44,46,63] Details

regarding preparation of crystal structures for molecular docking are provided in Supplementary Table S3. Unless stated otherwise, water molecules and other solutes were removed from the crystal structure. The N-termini and C-termini were capped with acetyl and methyl groups respectively using PyMOL.[64] The atoms of the co-crystallized ligands were used to generate matching spheres in the binding site. DOCK3.7 uses a flexible ligand algorithm that superimposes rigid segments of a molecule's pre-calculated conformational ensemble on top of the matching spheres.[60] Histidine protonation states were assigned manually after visual inspection of the hydrogen bonding network. The remainder of the protein structure was protonated by REDUCE[65] and assigned AMBER[66] united atom charges. The dipole moments of keys residues involved in recognition of the co-crystallized ligands were increased to favor interactions with these. This technique is common practice for users of DOCK3.7 to improve docking performance and has been used in previous virtual screens.[10] The atoms of the co-crystallized ligands were used to create two sets of sphere layers on the protein surface (referred to as thin spheres). One set of thin spheres described the low protein dielectric and defines the boundary between solute and solvent. A second set of thin spheres was used to calibrate ligand desolvation penalties. Scoring grids were pre-calculated using QNIFFT[67] for Poisson-Boltzmann electrostatic potentials, SOLVMAP[68] for ligand desolvation potentials, and CHEMGRID[69] for AMBER van der Waals potentials. Tailored control sets, as described above, were used to evaluate the performance of the docking grids by means of ligands-over-decoys enrichments. Finally, ligands-over-decoys enrichments and predicted binding poses were used to select the optimal grid parameters.[10]

Molecular docking calculations. The orientational matching parameter was set to 5000 and both the Rule-of-Four benchmarking set and prioritized molecules were docked at the same sampling rate. Molecules in the ultralarge docking screens (235 million lead-like molecules from the ZINC15 database[70]) were docked at a sampling rate of 1000 matches. During the generation of the benchmarking dataset, for each docked compound, 18652 orientations were calculated on average, and for each orientation, an average of 1654 conformations were sampled. The best scoring pose of each ligand was optimized using a simplex rigid-body minimizer. In total, more than 493 trillion protein-ligand complexes were calculated to generate the benchmarking datasets.

Training of machine learning classifiers. Classifiers were built and trained in combination with the CP framework. The docking scores of the datasets were used to label molecules as virtual actives (top 1%, virtual active) and virtual inactives (bottom 99%, virtual inactive), unless stated otherwise. The scikit-learn 0.24.2 package was used to stratified split the datasets in proper training sets (80% of training set), calibration sets (20% of training set), and test sets, maintaining the ratio between virtual actives and inactives.[71] This procedure was repeated using different random seeds to obtain independent sets. The CatBoost 0.26 Python package was used for building and training the corresponding classifiers. PyTorch 1.7.1 package combined with the RangerLars optimizer was used for training the DNN's.[72,73] The RoBERTa classifier was implemented from the simpletransformers 0.61.6 package.[62] Skorch 0.10.0 package was used to connect the scikit-learn and PyTorch frameworks.[74] A detailed description of the hyper parameters used in each classifier is provided in the Supporting Information.

Metrics for performance evaluation. The following metrics were used to assess the performance of the classifiers. The sensitivity was defined as:

$$sensitivity = \frac{TP}{AP}$$

where TP (true positives) were true active molecules correctly classified by the CP framework, i.e., in the predicted virtual active and both sets. AP (all positives) were all molecules with a score better than the threshold. The precision was defined as:

$$precision = \frac{TP}{TP + FP}$$

where FP (false positives) were true inactive molecules incorrectly classified by the CP framework, i.e., in the predicted virtual active and null sets. The efficiency was defined as:

$$efficiency = \frac{\{1\} + \{0\}}{AP + AN}$$

where {1} are the predicted virtual actives and {0} the predicted virtual inactives. AN (all negatives) were all molecules with a score worse than or equal to the threshold. The overall error rate was defined as:

$$overall\ error\ rate = \frac{FP + FN}{AP + AN}$$

where FN (false negatives) are true virtual active molecules incorrectly classified by the CP framework, i.e., in the predicted virtual inactives and null sets. The error rate for the virtual actives was defined as:

$$actives\ error\ rate = \frac{FN}{AP}$$

The error rate for the virtual inactives was defined as:

$$inactives\ error\ rate = \frac{FP}{AN}$$

Acknowledgments

J.C. received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement: 715052), the Swedish Cancer Society, and the Swedish Strategic Research Program eSSENCE. I.C.d.V was funded by a post-doctoral fellowship provided by the Sven och Lilly Lawski foundation. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at NSC and UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. A.L., I.C.d.V., and J.C. thank OpenEye Scientific Software for the use of OEToolkits at no cost. A.L., I.C.d.V., and J.C. thank ChemAxon for the use of the Marvin package at no cost. A.L., I.C.d.V., and J.C. thank Enamine Ltd. for sharing the REAL space database. The authors thank Jin Zhang for providing the initial Deep Neural Network code.

Author contributions

A.L., I.C.d.V., U.N., and J.C. designed the study. A.L. performed the molecular docking and machine learning calculations under the supervision of J.C. and U.N. A.L., I.C.d.V., L.S., and U.N. developed the protocol. A.L. and I.C.d.V. wrote the final version of the code. U.N. provided support with critical evaluation of the protocol. A.L., and I.C.d.V., and J.C. wrote the manuscript with contributions from the other authors.

Additional Information

Supplementary information is available for this paper. Complete datasets used in this work are available at https://doi.org/10.5281/zenodo.7903161. The code is freely available at GitHub (https://github.com/carlssonlab/conformalpredictor)

References

1. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* vol. 16 3–50 (1996).
2. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* vol. 46 3–26 (2001).
3. Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des* vol. 27 675–9 (2013).
4. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery | Journal of Chemical Information and Modeling. https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c00675.
5. Grygorenko, O. O. *et al.* Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* vol. 23 101681 (2020).
6. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* vol. 566 224–229 (2019).
7. Stein, R. M. *et al.* Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* vol. 579 609–614 (2020).
8. Fink, E. A. *et al.* Structure-based discovery of nonopioid analgesics acting through the α. *Science* vol. 377 eabn7065 (2022).
9. Luttens, A. *et al.* Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J Am Chem Soc* vol. 144 2905–2920 (2022).
10. Bender, B. J. *et al.* A practical guide to large-scale docking. *Nat Protoc* vol. 16 4799–4832 (2021).

11. Gorgulla, C. *et al.* An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).
12. Fink, E. A. *et al.* Large library docking for novel SARS-CoV-2 main protease non-covalent and covalent inhibitors. 2022.07.05.498881 Preprint at https://doi.org/10.1101/2022.07.05.498881 (2022).
13. Singh, I. *et al.* Structure-based discovery of inhibitors of the SARS-CoV-2 Nsp14 N7-methyltransferase. 2023.01.12.523677 Preprint at https://doi.org/10.1101/2023.01.12.523677 (2023).
14. Lyu, J., Irwin, J. J. & Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nat Chem Biol* (2023).
15. Neves, B. J. *et al.* QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front. Pharmacol.* **9**, (2018).
16. Spiegel, J. & Senderowitz, H. Evaluation of QSAR Equations for Virtual Screening. *Int. J. Mol. Sci.* **21**, 7828 (2020).
17. Adeshina, Y. O., Deeds, E. J. & Karanicolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl. Acad. Sci.* **117**, 18477–18488 (2020).
18. Langevin, M. *et al.* Impact of applicability domains to generative artificial intelligence. Preprint at https://doi.org/10.26434/chemrxiv-2022-mdhwz (2022).
19. Yang, Y. *et al.* Efficient Exploration of Chemical Space with Docking and Deep Learning. *J Chem Theory Comput* vol. 17 7106–7119 (2021).
20. Gentile, F. *et al.* Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **6**, 939–949 (2020).

21. Graff, D. E. *et al.* Self-Focusing Virtual Screening with Active Design Space Pruning. *J. Chem. Inf. Model.* **62**, 3854–3862 (2022).
22. Gentile, F. *et al.* Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **17**, 672–697 (2022).
23. Sivula, T. *et al.* Machine Learning-Boosted Docking Enables the Efficient Structure-Based Virtual Screening of Giga-Scale Enumerated Chemical Libraries. Preprint at https://doi.org/10.26434/chemrxiv-2023-g34tx (2023).
24. Shafer, G. & Vovk, V. A tutorial on conformal prediction. Preprint at https://doi.org/10.48550/arXiv.0706.3188 (2007).
25. Norinder, U., Carlsson, L., Boyer, S. & Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **54**, 1596–1603 (2014).
26. Carlsson, L., Eklund, M. & Norinder, U. Aggregated Conformal Prediction. in *Artificial Intelligence Applications and Innovations* (eds. Iliadis, L., Maglogiannis, I., Papadopoulos, H., Sioutas, S. & Makris, C.) 231–240 (Springer, 2014). doi:10.1007/978-3-662-44722-2_25.
27. Norinder, U. & Boyer, S. Binary classification of imbalanced datasets using conformal prediction. *J. Mol. Graph. Model.* **72**, 256–265 (2017).
28. Norinder, U., Spjuth, O. & Svensson, F. Synergy conformal prediction applied to large-scale bioactivity datasets and in federated learning. *J. Cheminformatics* **13**, 77 (2021).
29. Eklund, M., Norinder, U., Boyer, S. & Carlsson, L. The application of conformal prediction to the drug discovery process. *Ann. Math. Artif. Intell.* **74**, 117–132 (2015).
30. Zhang, J., Norinder, U. & Svensson, F. Deep Learning-Based Conformal Prediction of Toxicity. *J. Chem. Inf. Model.* **61**, 2648–2657 (2021).

31. Svensson, F., Norinder, U. & Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *J. Chem. Inf. Model.* **57**, 439–444 (2017).
32. Ahmed, L. *et al.* Efficient iterative virtual screening with Apache Spark and conformal prediction. *J. Cheminformatics* **10**, 8 (2018).
33. Hancock, J. T. & Khoshgoftaar, T. M. CatBoost for big data: an interdisciplinary review. *J. Big Data* **7**, 94 (2020).
34. Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci. Rep.* **11**, 321 (2021).
35. Kim, J., Park, S., Min, D. & Kim, W. Comprehensive Survey of Recent Drug Discovery Using Deep Learning. *Int. J. Mol. Sci.* **22**, 9983 (2021).
36. Meli, R., Morris, G. M. & Biggin, P. C. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. *Front. Bioinforma.* **2**, (2022).
37. Carlsson, J. *et al.* Structure-Based Discovery of A2A Adenosine Receptor Ligands. *J. Med. Chem.* **53**, 3748–3755 (2010).
38. Liu, W. *et al.* Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions. *Science* **337**, 232–236 (2012).
39. Wang, S. *et al.* Structure of the D2 dopamine receptor bound to the atypical antipsychotic drug risperidone. *Nature* **555**, 269–273 (2018).
40. Kampen, S. *et al.* Structure-Guided Design of G-Protein-Coupled Receptor Polypharmacology. *Angew. Chem. Int. Ed.* **60**, 18022–18030 (2021).
41. Visnes, T. *et al.* Small-molecule inhibitor of OGG1 suppresses proinflammatory gene expression and inflammation. *Science* vol. 362 834–839 (2018).

42. Beatty, J. W. *et al.* Discovery of Potent and Selective Non-Nucleotide Small Molecule Inhibitors of CD73. *J. Med. Chem.* **63**, 3935–3955 (2020).
43. Knapp, K. *et al.* Crystal Structure of the Human Ecto-5'-Nucleotidase (CD73): Insights into the Regulation of Purinergic Signaling. *Structure* **20**, 2161–2173 (2012).
44. Davies, T. G. *et al.* Monoacidic Inhibitors of the Kelch-like ECH-Associated Protein 1: Nuclear Factor Erythroid 2-Related Factor 2 (KEAP1:NRF2) Protein–Protein Interaction with High Cell Potency Identified by Fragment-Based Discovery. *J. Med. Chem.* **59**, 3991–4006 (2016).
45. Begnini, F. *et al.* Importance of Binding Site Hydration and Flexibility Revealed When Optimizing a Macrocyclic Inhibitor of the Keap1-Nrf2 Protein-Protein Interaction. *J Med Chem* (2022).
46. Stachel, S. J. *et al.* Identification of potent inhibitors of the sortilin-progranulin interaction. *Bioorg. Med. Chem. Lett.* **30**, 127403 (2020).
47. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. Preprint at https://doi.org/10.48550/arXiv.1706.09516 (2019).
48. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).
49. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at https://doi.org/10.48550/arXiv.1907.11692 (2019).
50. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

51. Winter, R., Montanari, F., Noé, F. & Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).
52. Linusson, H., Norinder, U., Boström, H., Johansson, U. & Löfström, T. On the Calibration of Aggregated Conformal Predictors. in *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications* (eds. Gammerman, A., Vovk, V., Luo, Z. & Papadopoulos, H.) vol. 60 154–173 (PMLR, 2017).
53. Krstajic, D. Critical Assessment of Conformal Prediction Methods Applied in Binary Classification Settings. *J. Chem. Inf. Model.* **61**, 4823–4826 (2021).
54. Alvarsson, J., Arvidsson McShane, S., Norinder, U. & Spjuth, O. Predicting With Confidence: Using Conformal Prediction in Drug Discovery. *J. Pharm. Sci.* **110**, 42–49 (2021).
55. Tamma, P. D. *et al.* A Primer on AmpC β-Lactamases: Necessary Knowledge for an Increasingly Multidrug-resistant World. *Clin. Infect. Dis.* **69**, 1446–1455 (2019).
56. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* **30**, 127377 (2020).
57. ChEMBL: a large-scale bioactivity database for drug discovery | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/40/D1/D1100/2903401.
58. RDKit. http://www.rdkit.org/.
59. Salle, A. terashuf. (2023).
60. Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J. & Shoichet, B. K. Ligand pose and orientational sampling in molecular docking. *PLoS One* vol. 8 e75992 (2013).

61. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. Preprint at https://doi.org/10.48550/arXiv.2010.09885 (2020).
62. Rajapakse, T. C. Simple Transformers. (2019).
63. Bank, R. P. D. RCSB PDB - 6W63: Structure of COVID-19 main protease bound to potent broad-spectrum non-covalent inhibitor X77. https://www.rcsb.org/structure/6w63.
64. Schrödinger, L. & DeLano, W. PyMOL. (2020).
65. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* vol. 285 1735–47 (1999).
66. Weiner, S. J. *et al.* A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784 (1984).
67. Gallagher, K. & Sharp, K. Electrostatic contributions to heat capacity changes of DNA-ligand binding. *Biophys J* vol. 75 769–76 (1998).
68. Mysinger, M. M. & Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J Chem Inf Model* vol. 50 1561–73 (2010).
69. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13**, 505–524 (1992).
70. Sterling, T. & Irwin, J. J. ZINC 15--Ligand Discovery for Everyone. *J Chem Inf Model* vol. 55 2324–37 (2015).
71. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

72. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems 32* 8024–8035 (Curran Associates, Inc., 2019).
73. Grankin, M. Optimizers and tests. (2023).
74. Tietz, M., Fan, T. J., Nouri, D., Bossan, B., & skorch Developers. *skorch: A scikit-learn compatible neural network library that wraps PyTorch*. (2017).

Supporting Information:

Rapid Traversal of Ultralarge Chemical Space using Machine Learning Guided Docking Screens

Andreas Luttens[1], Israel Cabeza de Vaca[1], Leonard Sparring[1], Ulf Norinder[2,3,4,*], Jens Carlsson[1,*]

[1]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, BMC, Box 596, SE-75124 Uppsala, Sweden

[2]Department of Pharmaceutical Biosciences, Uppsala University, Box 591, SE-75124, Uppsala, Sweden

[3]Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-16407, Kista, Sweden

[4]MTM Research Centre, School of Science and Technology, Örebro University, SE-70182, Örebro, Sweden

Table of Contents

Supplementary Table S1. Model hyperparameters. Key hyperparameters used during training of models.

Method	CatBoost	DNN	RoBERTa
Parameters	• nr_trees = 500 • metric = AUC • weights = balanced	• learning_rate = 1e-4 • weight_decay = 1e-2 • batch_size = 200 • max_epochs= 100 • patience = 10 • optimizer = RangerLars • class_weights = balanced	• learning_rate = 4e-7 • max_epochs = 10 • seyonec/PubChem10M

Supplementary Table 2a. Sensitivity and training set size - EFCP4. Sensitivity values obtained at optimal efficiency for different sizes of the training set.

Method	Target	Sensitivity[a]					
		25K	50K	100K	200K	500K	1M
CatBoost	$A_{2A}R$	0.754 ± 0.011	0.799 ± 0.005	0.820 ± 0.007	0.856 ± 0.004	0.873 ± 0.003	0.891 ± 0.002
	AmpC	0.857 ± 0.014	0.909 ± 0.009	0.921 ± 0.005	0.936 ± 0.001	0.945 ± 0.000	0.955 ± 0.001
	5'-NT	0.719 ± 0.025	0.773 ± 0.005	0.783 ± 0.008	0.811 ± 0.002	0.834 ± 0.001	0.849 ± 0.001
	D_2R	0.793 ± 0.002	0.813 ± 0.016	0.854 ± 0.006	0.883 ± 0.002	0.910 ± 0.001	0.917 ± 0.001
	KEAP1	0.688 ± 0.011	0.732 ± 0.008	0.777 ± 0.008	0.795 ± 0.005	0.819 ± 0.002	0.833 ± 0.003
	M^{PRO}	0.588 ± 0.010	0.650 ± 0.003	0.681 ± 0.003	0.705 ± 0.006	0.743 ± 0.003	0.765 ± 0.005
	OGG1	0.720 ± 0.014	0.770 ± 0.004	0.782 ± 0.001	0.815 ± 0.002	0.836 ± 0.006	0.853 ± 0.001
	SORT1	0.656 ± 0.011	0.703 ± 0.004	0.733 ± 0.003	0.773 ± 0.001	0.804 ± 0.004	0.821 ± 0.004
	Average	**0.722 ± 0.017**	**0.768 ± 0.015**	**0.794 ± 0.014**	**0.822 ± 0.014**	**0.845 ± 0.012**	**0.860 ± 0.012**
DNN	$A_{2A}R$	0.744 ± 0.034	0.789 ± 0.010	0.814 ± 0.013	0.833 ± 0.008	0.831 ± 0.007	0.841 ± 0.002
	AmpC	0.781 ± 0.013	0.836 ± 0.003	0.859 ± 0.004	0.897 ± 0.004	0.903 ± 0.001	0.919 ± 0.003
	5'-NT	0.731 ± 0.009	0.753 ± 0.018	0.782 ± 0.016	0.788 ± 0.010	0.807 ± 0.002	0.804 ± 0.002
	D_2R	0.747 ± 0.005	0.769 ± 0.012	0.803 ± 0.009	0.838 ± 0.005	0.861 ± 0.003	0.873 ± 0.003
	KEAP1	0.697 ± 0.035	0.766 ± 0.011	0.768 ± 0.014	0.784 ± 0.004	0.790 ± 0.004	0.796 ± 0.006
	M^{PRO}	0.677 ± 0.045	0.675 ± 0.014	0.682 ± 0.007	0.699 ± 0.003	0.713 ± 0.006	0.726 ± 0.002
	OGG1	0.728 ± 0.030	0.772 ± 0.012	0.787 ± 0.012	0.791 ± 0.006	0.806 ± 0.002	0.817 ± 0.002
	SORT1	0.704 ± 0.025	0.702 ± 0.007	0.729 ± 0.010	0.749 ± 0.010	0.760 ± 0.002	0.782 ± 0.004
	Average	**0.726 ± 0.010**	**0.758 ± 0.010**	**0.778 ± 0.011**	**0.797 ± 0.012**	**0.809 ± 0.012**	**0.820 ± 0.011**

[a] Each test set contained ten million molecules. ECFP4 descriptors were used as features of the molecules. Three independent calculations (training and prediction) were performed for each target and error bars correspond to the standard error of the mean.

Supplementary Table 2b. Precision and training set size - ECFP4. Precision values obtained at optimal efficiency for different sizes of the training set.

Method	Target	Precision[a]					
		25K	50K	100K	200K	500K	1M
CatBoost	$A_{2A}R$	0.043 ± 0.002	0.045 ± 0.001	0.052 ± 0.002	0.055 ± 0.000	0.068 ± 0.001	0.074 ± 0.001
	AmpC	0.090 ± 0.003	0.100 ± 0.005	0.117 ± 0.002	0.138 ± 0.007	0.180 ± 0.002	0.202 ± 0.001
	5'-NT	0.035 ± 0.002	0.036 ± 0.001	0.039 ± 0.001	0.041 ± 0.000	0.046 ± 0.000	0.052 ± 0.000
	D_2R	0.047 ± 0.000	0.060 ± 0.005	0.065 ± 0.003	0.079 ± 0.002	0.093 ± 0.000	0.106 ± 0.000
	KEAP1	0.034 ± 0.001	0.034 ± 0.001	0.036 ± 0.000	0.040 ± 0.000	0.044 ± 0.000	0.047 ± 0.001
	M^{PRO}	0.020 ± 0.000	0.022 ± 0.000	0.023 ± 0.000	0.025 ± 0.000	0.028 ± 0.000	0.030 ± 0.000
	OGG1	0.035 ± 0.001	0.037 ± 0.001	0.040 ± 0.001	0.042 ± 0.000	0.048 ± 0.001	0.052 ± 0.000
	SORT1	0.024 ± 0.001	0.027 ± 0.000	0.030 ± 0.001	0.033 ± 0.001	0.039 ± 0.001	0.044 ± 0.001
	Average	**0.041 ± 0.004**	**0.045 ± 0.005**	**0.050 ± 0.006**	**0.057 ± 0.007**	**0.068 ± 0.010**	**0.076 ± 0.011**
DNN	$A_{2A}R$	0.039 ± 0.003	0.041 ± 0.001	0.044 ± 0.002	0.048 ± 0.001	0.054 ± 0.001	0.057 ± 0.000
	AmpC	0.042 ± 0.002	0.052 ± 0.002	0.067 ± 0.002	0.080 ± 0.003	0.097 ± 0.002	0.104 ± 0.001
	5'-NT	0.030 ± 0.001	0.034 ± 0.001	0.035 ± 0.001	0.039 ± 0.001	0.041 ± 0.001	0.043 ± 0.000
	D_2R	0.029 ± 0.001	0.040 ± 0.003	0.045 ± 0.003	0.053 ± 0.002	0.060 ± 0.001	0.068 ± 0.001
	KEAP1	0.029 ± 0.002	0.030 ± 0.001	0.034 ± 0.001	0.035 ± 0.001	0.039 ± 0.001	0.040 ± 0.001
	M^{PRO}	0.018 ± 0.001	0.021 ± 0.000	0.023 ± 0.000	0.024 ± 0.000	0.026 ± 0.000	0.027 ± 0.000
	OGG1	0.034 ± 0.001	0.037 ± 0.001	0.038 ± 0.001	0.041 ± 0.001	0.043 ± 0.000	0.044 ± 0.000
	SORT1	0.022 ± 0.001	0.025 ± 0.001	0.028 ± 0.001	0.030 ± 0.001	0.033 ± 0.000	0.035 ± 0.000
	Average	**0.031 ± 0.002**	**0.035 ± 0.002**	**0.039 ± 0.003**	**0.044 ± 0.003**	**0.049 ± 0.004**	**0.052 ± 0.005**

[a] Each test set contained ten million molecules. ECFP4 descriptors were used as features of the molecules. Three independent calculations (training and prediction) were performed for each target and error bars correspond to the standard error of the mean.

Supplementary Table 2c. Sensitivity and training set size - CDDD. Sensitivity values obtained at optimal efficiency for different sizes of the training set.

Method	Target	Sensitivity[a]					
		25K	50K	100K	200K	500K	1M
CatBoost	$A_{2A}R$	0.784 ± 0.013	0.806 ± 0.013	0.819 ± 0.003	0.845 ± 0.002	0.852 ± 0.003	0.870 ± 0.004
	AmpC	0.847 ± 0.013	0.893 ± 0.008	0.903 ± 0.004	0.919 ± 0.003	0.931 ± 0.001	0.937 ± 0.002
	5'-NT	0.747 ± 0.012	0.790 ± 0.007	0.793 ± 0.005	0.815 ± 0.004	0.828 ± 0.003	0.832 ± 0.002
	D_2R	0.805 ± 0.014	0.839 ± 0.004	0.847 ± 0.008	0.875 ± 0.001	0.888 ± 0.001	0.896 ± 0.002
	KEAP1	0.716 ± 0.014	0.759 ± 0.008	0.784 ± 0.009	0.799 ± 0.002	0.816 ± 0.003	0.827 ± 0.001
	M^{PRO}	0.605 ± 0.004	0.658 ± 0.004	0.682 ± 0.003	0.699 ± 0.005	0.728 ± 0.003	0.737 ± 0.007
	OGG1	0.745 ± 0.007	0.776 ± 0.005	0.776 ± 0.006	0.809 ± 0.001	0.816 ± 0.002	0.833 ± 0.003
	SORT1	0.676 ± 0.006	0.691 ± 0.011	0.722 ± 0.004	0.749 ± 0.005	0.772 ± 0.004	0.792 ± 0.001
	Average	**0.741 ± 0.015**	**0.777 ± 0.015**	**0.791 ± 0.014**	**0.814 ± 0.014**	**0.829 ± 0.012**	**0.840 ± 0.012**
DNN	$A_{2A}R$	0.755 ± 0.005	0.832 ± 0.013	0.818 ± 0.008	0.851 ± 0.004	0.850 ± 0.003	0.862 ± 0.002
	AmpC	0.841 ± 0.016	0.871 ± 0.011	0.892 ± 0.007	0.908 ± 0.005	0.923 ± 0.004	0.941 ± 0.004
	5'-NT	0.777 ± 0.016	0.812 ± 0.000	0.816 ± 0.015	0.811 ± 0.007	0.822 ± 0.002	0.836 ± 0.003
	D_2R	0.786 ± 0.021	0.874 ± 0.005	0.857 ± 0.006	0.871 ± 0.002	0.884 ± 0.003	0.897 ± 0.003
	KEAP1	0.753 ± 0.022	0.792 ± 0.015	0.791 ± 0.008	0.810 ± 0.007	0.824 ± 0.005	0.820 ± 0.001
	M^{PRO}	0.657 ± 0.025	0.707 ± 0.004	0.689 ± 0.013	0.724 ± 0.002	0.723 ± 0.008	0.737 ± 0.006
	OGG1	0.772 ± 0.017	0.798 ± 0.003	0.800 ± 0.007	0.802 ± 0.014	0.821 ± 0.005	0.828 ± 0.002
	SORT1	0.684 ± 0.042	0.728 ± 0.009	0.761 ± 0.003	0.772 ± 0.012	0.775 ± 0.003	0.793 ± 0.006
	Average	**0.753 ± 0.013**	**0.802 ± 0.012**	**0.803 ± 0.012**	**0.819 ± 0.012**	**0.828 ± 0.012**	**0.839 ± 0.012**

[a] Each test set contained ten million molecules. Continuous-Data-Driven Descriptors were used as features of the molecules. Three independent calculations (training and prediction) were performed for each target and error bars correspond to the standard error of the mean.

Supplementary Table 2d. Precision and training set size - CDDD. Precision values obtained at optimal efficiency for different sizes of the training set.

Method	Target	Precision[a]					
		25K	50K	100K	200K	500K	1M
CatBoost	$A_{2A}R$	0.046 ± 0.002	0.047 ± 0.001	0.051 ± 0.001	0.052 ± 0.001	0.059 ± 0.000	0.062 ± 0.000
	AmpC	0.079 ± 0.001	0.085 ± 0.003	0.093 ± 0.000	0.100 ± 0.002	0.113 ± 0.001	0.128 ± 0.002
	5'-NT	0.042 ± 0.002	0.039 ± 0.001	0.044 ± 0.001	0.043 ± 0.001	0.046 ± 0.000	0.050 ± 0.000
	D_2R	0.052 ± 0.002	0.057 ± 0.001	0.060 ± 0.003	0.063 ± 0.000	0.071 ± 0.000	0.079 ± 0.001
	KEAP1	0.038 ± 0.001	0.037 ± 0.001	0.038 ± 0.001	0.040 ± 0.000	0.042 ± 0.000	0.045 ± 0.000
	M^{PRO}	0.021 ± 0.000	0.022 ± 0.001	0.024 ± 0.000	0.025 ± 0.000	0.026 ± 0.000	0.027 ± 0.000
	OGG1	0.035 ± 0.000	0.038 ± 0.000	0.041 ± 0.000	0.041 ± 0.000	0.044 ± 0.000	0.046 ± 0.000
	SORT1	0.026 ± 0.000	0.027 ± 0.001	0.028 ± 0.000	0.030 ± 0.000	0.034 ± 0.001	0.037 ± 0.000
	Average	**0.042 ± 0.003**	**0.044 ± 0.004**	**0.047 ± 0.004**	**0.049 ± 0.005**	**0.054 ± 0.005**	**0.059 ± 0.006**
DNN	$A_{2A}R$	0.054 ± 0.002	0.047 ± 0.001	0.056 ± 0.000	0.053 ± 0.001	0.064 ± 0.001	0.067 ± 0.001
	AmpC	0.089 ± 0.004	0.101 ± 0.002	0.105 ± 0.003	0.126 ± 0.005	0.135 ± 0.002	0.140 ± 0.003
	5'-NT	0.042 ± 0.002	0.041 ± 0.001	0.043 ± 0.002	0.047 ± 0.001	0.049 ± 0.000	0.050 ± 0.000
	D_2R	0.062 ± 0.004	0.057 ± 0.001	0.068 ± 0.001	0.071 ± 0.002	0.079 ± 0.001	0.084 ± 0.001
	KEAP1	0.037 ± 0.001	0.038 ± 0.002	0.042 ± 0.000	0.044 ± 0.001	0.045 ± 0.001	0.049 ± 0.000
	M^{PRO}	0.023 ± 0.001	0.023 ± 0.000	0.025 ± 0.001	0.026 ± 0.000	0.028 ± 0.000	0.029 ± 0.000
	OGG1	0.038 ± 0.001	0.040 ± 0.001	0.042 ± 0.001	0.045 ± 0.001	0.048 ± 0.000	0.050 ± 0.000
	SORT1	0.029 ± 0.002	0.029 ± 0.000	0.030 ± 0.001	0.034 ± 0.001	0.038 ± 0.001	0.040 ± 0.001
	Average	**0.047 ± 0.004**	**0.047 ± 0.005**	**0.051 ± 0.005**	**0.055 ± 0.006**	**0.061 ± 0.007**	**0.064 ± 0.007**

[a] Each test set contained ten million molecules. Continuous-Data-Driven Descriptors were used as features of the molecules. Three independent calculations (training and prediction) were performed for each target and error bars correspond to the standard error of the mean.

Supplementary Table 2e. Sensitivity and training set size - RoBERTa. Sensitivity values obtained at optimal efficiency for different sizes of the training set.

Method	Target	Sensitivity[a]					
		25K	50K	100K	200K	500K	1M
RoBERTa	$A_{2A}R$	0.765 ± 0.007	0.781 ± 0.006	0.806 ± 0.007	0.848 ± 0.007	0.861 ± 0.006	0.879 ± 0.002
	AmpC	0.808 ± 0.005	0.872 ± 0.005	0.890 ± 0.004	0.916 ± 0.003	0.939 ± 0.002	0.944 ± 0.002
	5'-NT	0.735 ± 0.011	0.784 ± 0.007	0.778 ± 0.005	0.808 ± 0.007	0.827 ± 0.003	0.841 ± 0.003
	D_2R	0.737 ± 0.003	0.817 ± 0.004	0.841 ± 0.007	0.863 ± 0.002	0.884 ± 0.003	0.901 ± 0.000
	KEAP1	0.727 ± 0.011	0.764 ± 0.006	0.797 ± 0.005	0.805 ± 0.006	0.822 ± 0.005	0.830 ± 0.000
	M^{PRO}	0.627 ± 0.005	0.657 ± 0.013	0.689 ± 0.005	0.703 ± 0.002	0.729 ± 0.001	0.745 ± 0.000
	OGG1	0.728 ± 0.007	0.751 ± 0.005	0.783 ± 0.003	0.805 ± 0.004	0.819 ± 0.005	0.837 ± 0.004
	SORT1	0.662 ± 0.014	0.690 ± 0.010	0.730 ± 0.001	0.757 ± 0.003	0.782 ± 0.001	0.805 ± 0.004
	Average	**0.724 ± 0.011**	**0.764 ± 0.013**	**0.789 ± 0.012**	**0.813 ± 0.013**	**0.833 ± 0.012**	**0.848 ± 0.012**

[a] Each test set contained ten million molecules. Internal RoBERTa descriptors were used as features of the molecules. Three independent calculations (training and prediction) were performed for each target and error bars correspond to the standard error of the mean.

Supplementary Table 2f. Precision and training set size - RoBERTa. Precision values obtained at optimal efficiency for different sizes of the training set.

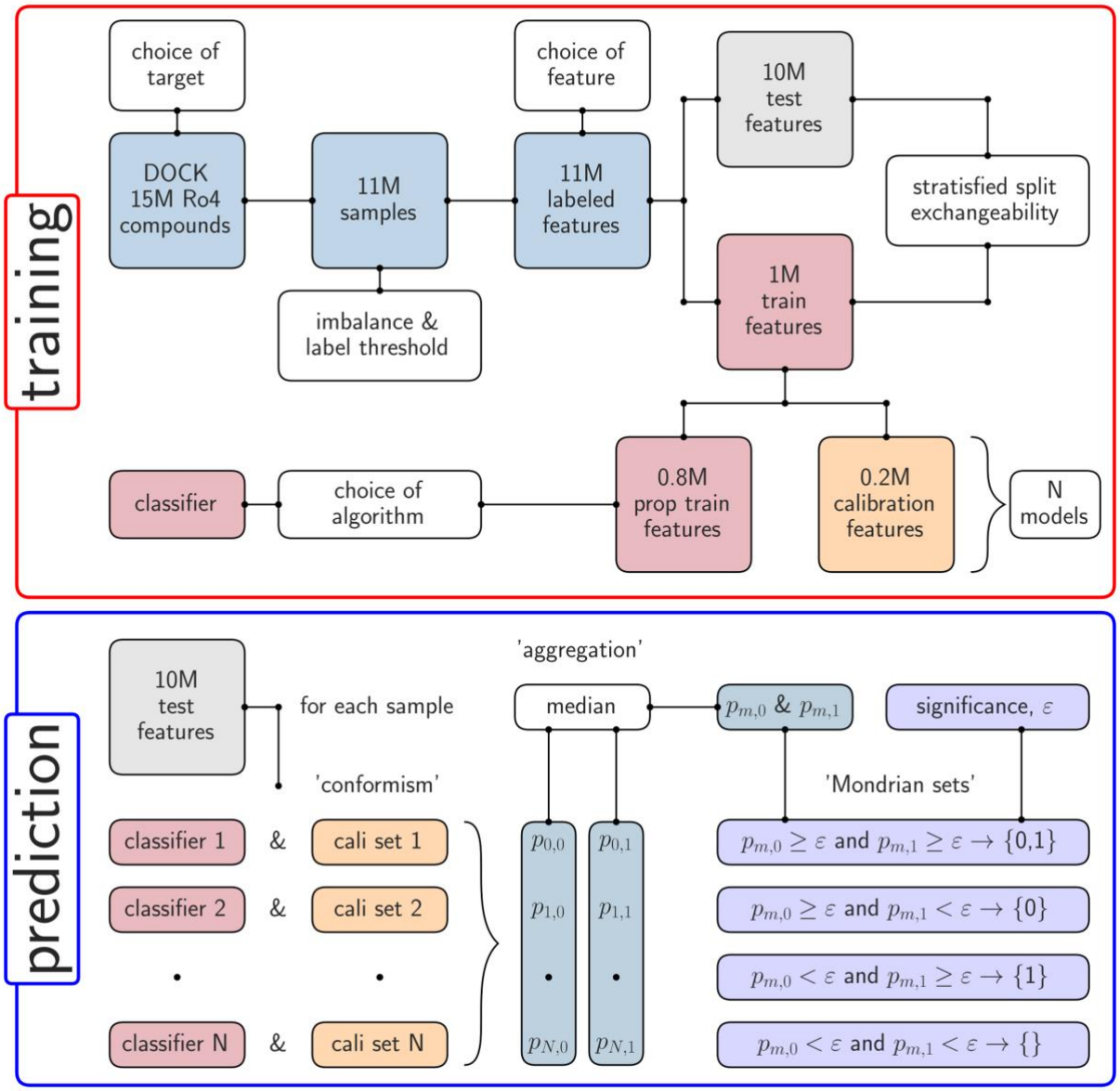
Method	Target	Precision[a]					
		25K	50K	100K	200K	500K	1M
RoBERTa	$A_{2A}R$	0.034 ± 0.001	0.042 ± 0.001	0.050 ± 0.001	0.054 ± 0.001	0.064 ± 0.002	0.070 ± 0.000
	AmpC	0.052 ± 0.000	0.065 ± 0.001	0.084 ± 0.002	0.111 ± 0.002	0.143 ± 0.004	0.181 ± 0.004
	5'-NT	0.032 ± 0.001	0.035 ± 0.001	0.042 ± 0.000	0.045 ± 0.001	0.050 ± 0.001	0.054 ± 0.001
	D_2R	0.034 ± 0.001	0.048 ± 0.001	0.058 ± 0.001	0.066 ± 0.002	0.082 ± 0.001	0.094 ± 0.001
	KEAP1	0.035 ± 0.001	0.038 ± 0.001	0.040 ± 0.000	0.043 ± 0.001	0.047 ± 0.001	0.050 ± 0.000
	M^{PRO}	0.018 ± 0.000	0.021 ± 0.000	0.023 ± 0.000	0.025 ± 0.000	0.028 ± 0.000	0.031 ± 0.000
	OGG1	0.030 ± 0.000	0.035 ± 0.000	0.039 ± 0.000	0.042 ± 0.001	0.048 ± 0.001	0.051 ± 0.000
	SORT1	0.022 ± 0.001	0.026 ± 0.000	0.029 ± 0.000	0.031 ± 0.000	0.038 ± 0.000	0.043 ± 0.000
	Average	**0.032 ± 0.002**	**0.039 ± 0.003**	**0.046 ± 0.004**	**0.052 ± 0.005**	**0.062 ± 0.007**	**0.072 ± 0.009**

[a] Each test set contained ten million molecules. Internal RoBERTa descriptors were used as features of the molecules. Three independent calculations (training and prediction) were performed for each target and error bars correspond to the standard error of the mean.

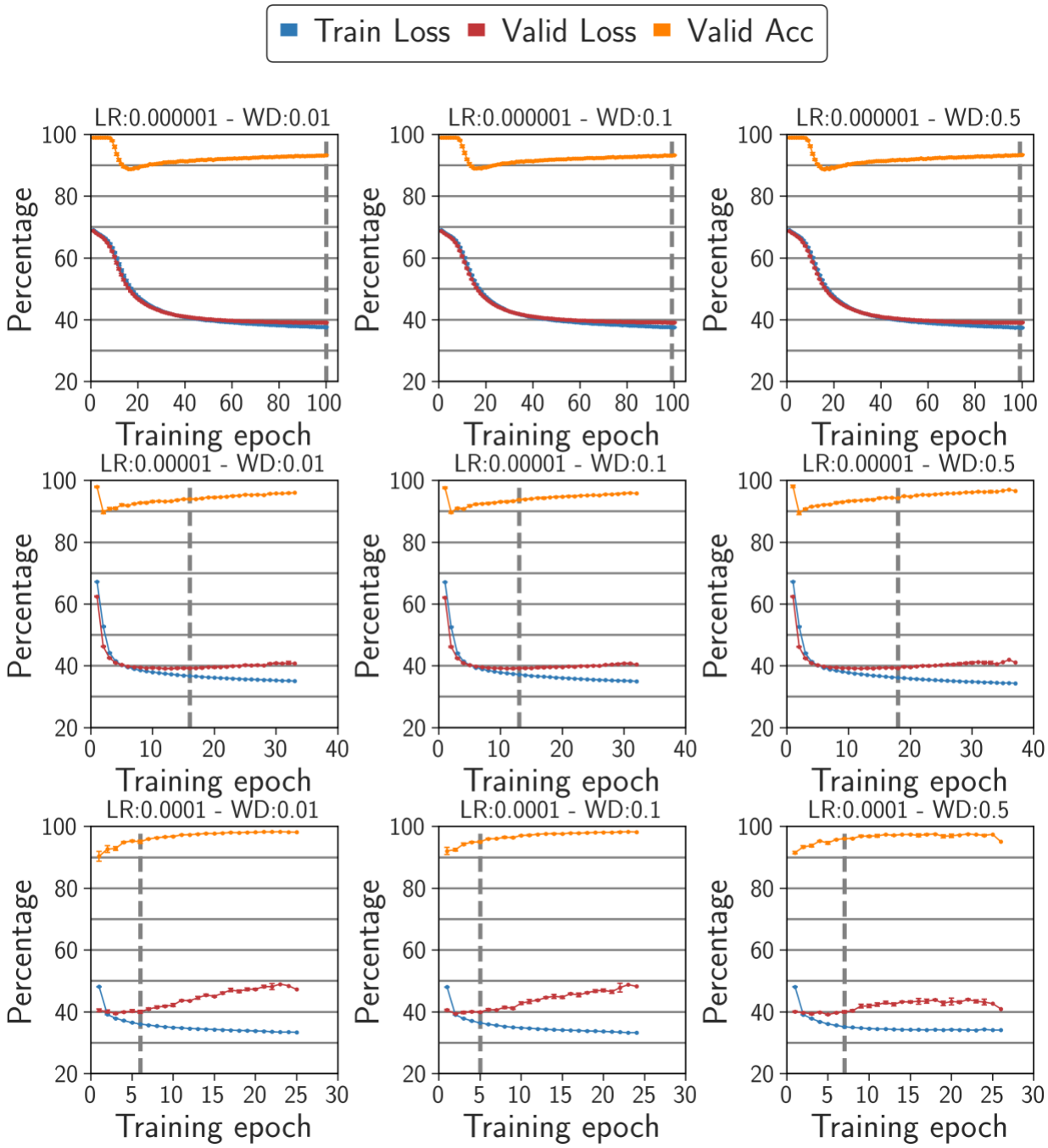
Supplementary Table 3. Protein preparation for molecular docking.

Target	Crystal structure[a]	Tarted residues[b]	Histidine protonation states	Number of matching spheres	Electrostatic radius[c]	Desolvation radius
$A_{2A}R$	4EIY	N253	δ: 155, 230 ε: 75, 250, 306 δ+ε: 264, 278	45	1.2 Å	0.3 Å
AmpC	6DPT	S64, Q120 N152, A318	ε: 13, 108, 186, 210, 314	45	1.2 Å	0.2 Å
5'-NT	6XUE	N390	δ: 33, 38, 220, 304, 440 ε: 103, 243, 375, 383, 437, 456, 518 δ+ε: 118	44	1.2 Å	0.4 Å
D_2R	6CM4	None	δ: 393, 398 ε: 106	45	1.2 Å	0.25 Å
KEAP1	5FNU	S363, Q530, S555, S602	δ: 436 ε: 424, 432, 437, 451, 516, 552, 553, 562, 575	45	1.4 Å	0.2 Å
M^{pro}	6W63	H163, G143, E166	δ: 64, 80 ε: 41, 163, 164, 172, 246	64	1.2 Å	0.3 Å
OGG1	6C3Y	G42	δ: 10, 13, 54, 97, 112, 179, 185, 195, 270, 276, 282 ε: 119, 237	45	Default (1.9 Å)	None
SORT1	6X48	Y318	δ: 68, 98, 360, 458, 490 ε: 70, 182, 220, 295, 331, 406, 428, 430, 506, 590, 664	45	1.6 Å	None

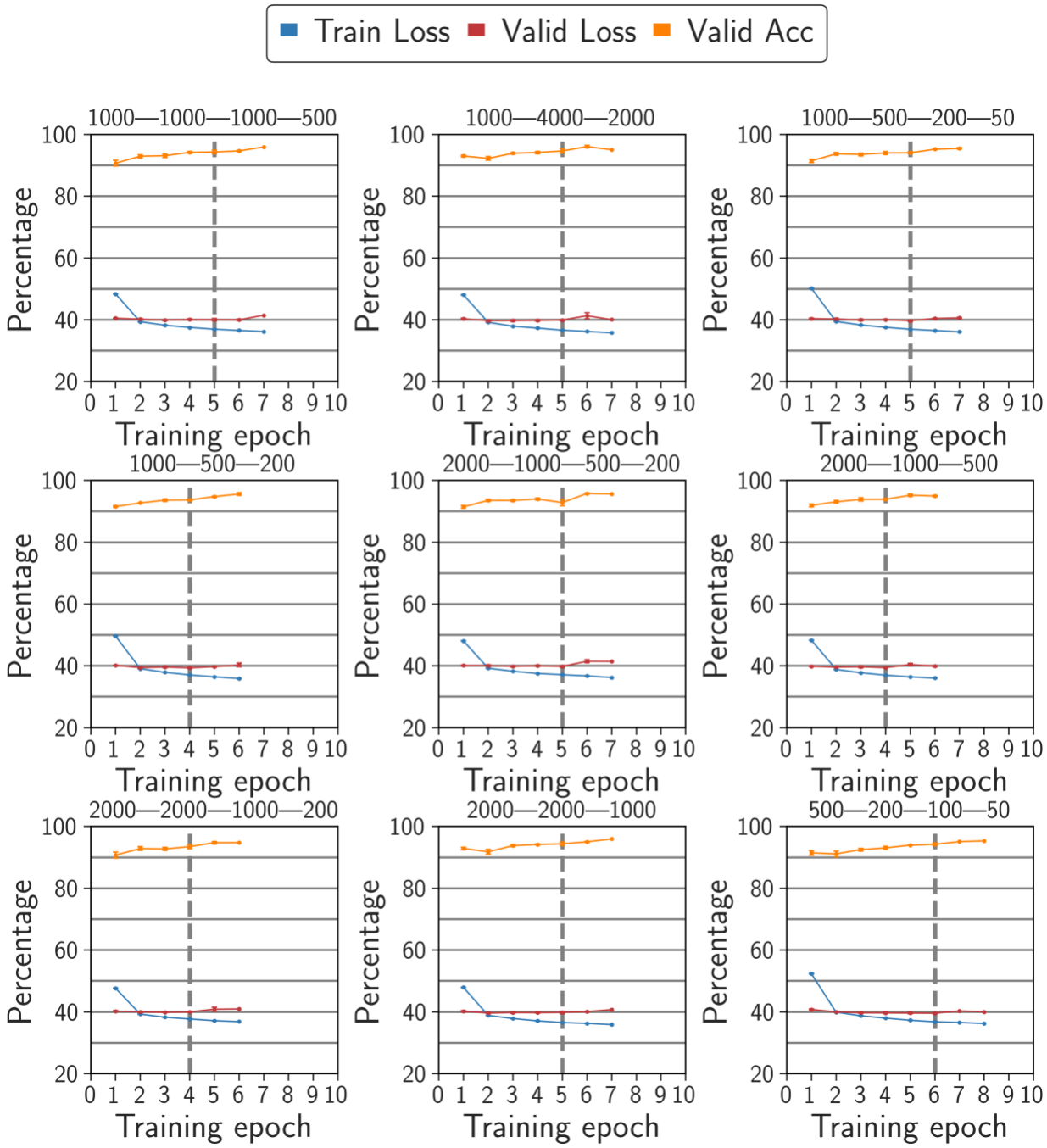
[a] PDB accession code. [b] Increase of dipole moments by adding partial charges to atoms, without altering the total charge of the system. [c] Tangent thin sphere radius. Default refers to low dielectric spheres made by blastermaster's SPHGEN program prior to thin sphere protocols.



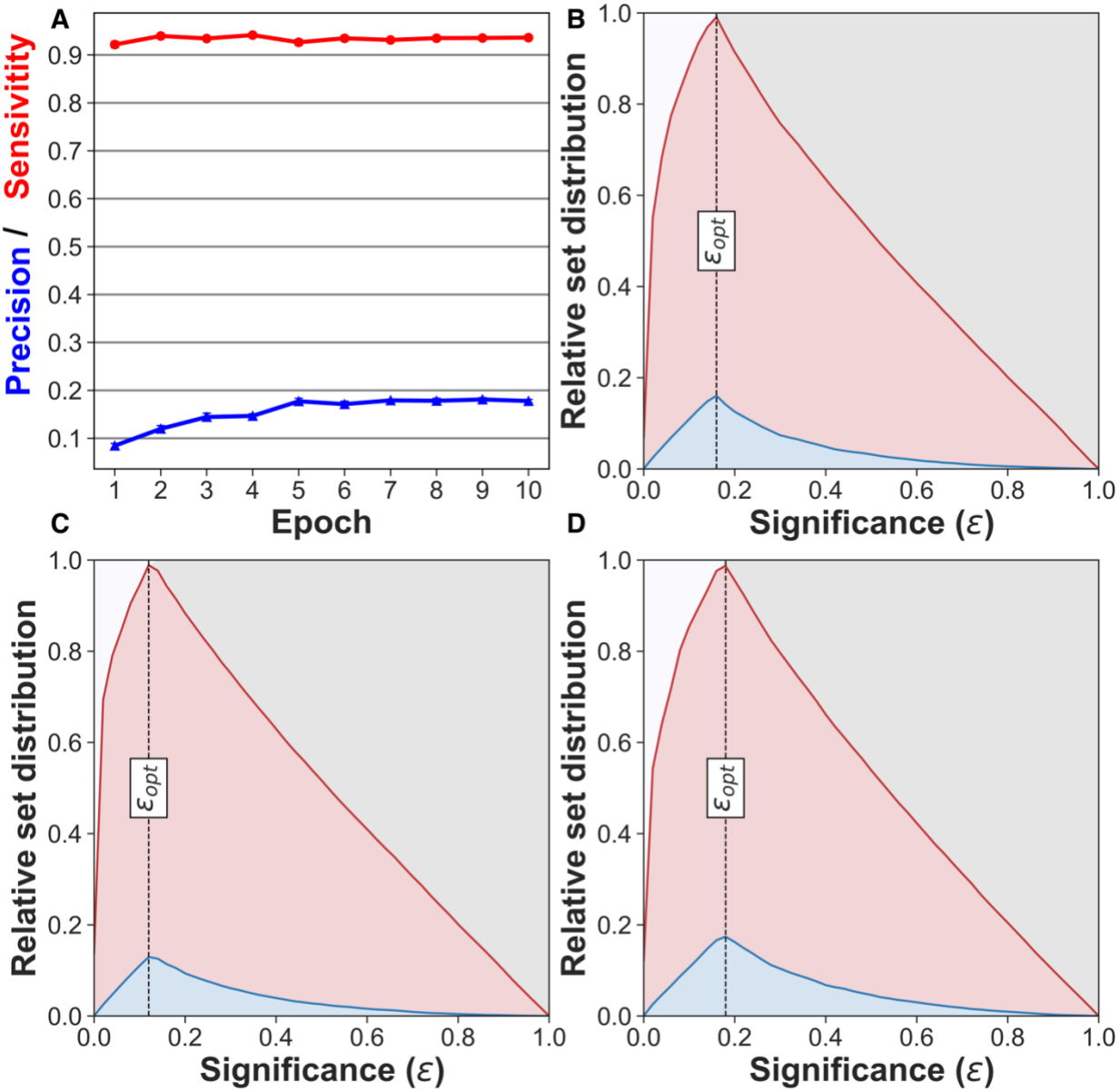
Supplementary Figure S1. Overview of the conformal prediction workflow. After docking to a target of interest, machine learning datasets are obtained through selection of a score threshold, followed by labeling and featurization of samples. Training and test sets are assumed to be exchangeable. The training set is split into a proper training and calibration set, and this process is repeated for each independent model that has to be trained. After training the classifiers, each sample in the test set is predicted. The corresponding calibration sets help normalize the outputs given by the classifiers. A pair of p-values (p_1 referring to the confidence the sample belongs to the virtual actives and p_0 referring to the confidence the sample belongs to the virtual inactives class) is obtained after aggregating model outputs by taking median values. After selecting a significance threshold, the sample can be assigned to a set prediction. For binary classifications, Mondrian conformal prediction has four sets a sample can be categorized into: virtual active {1}, virtual inactive {0}, both = virtual active or inactive {0,1}, and null = no class assignment {}.



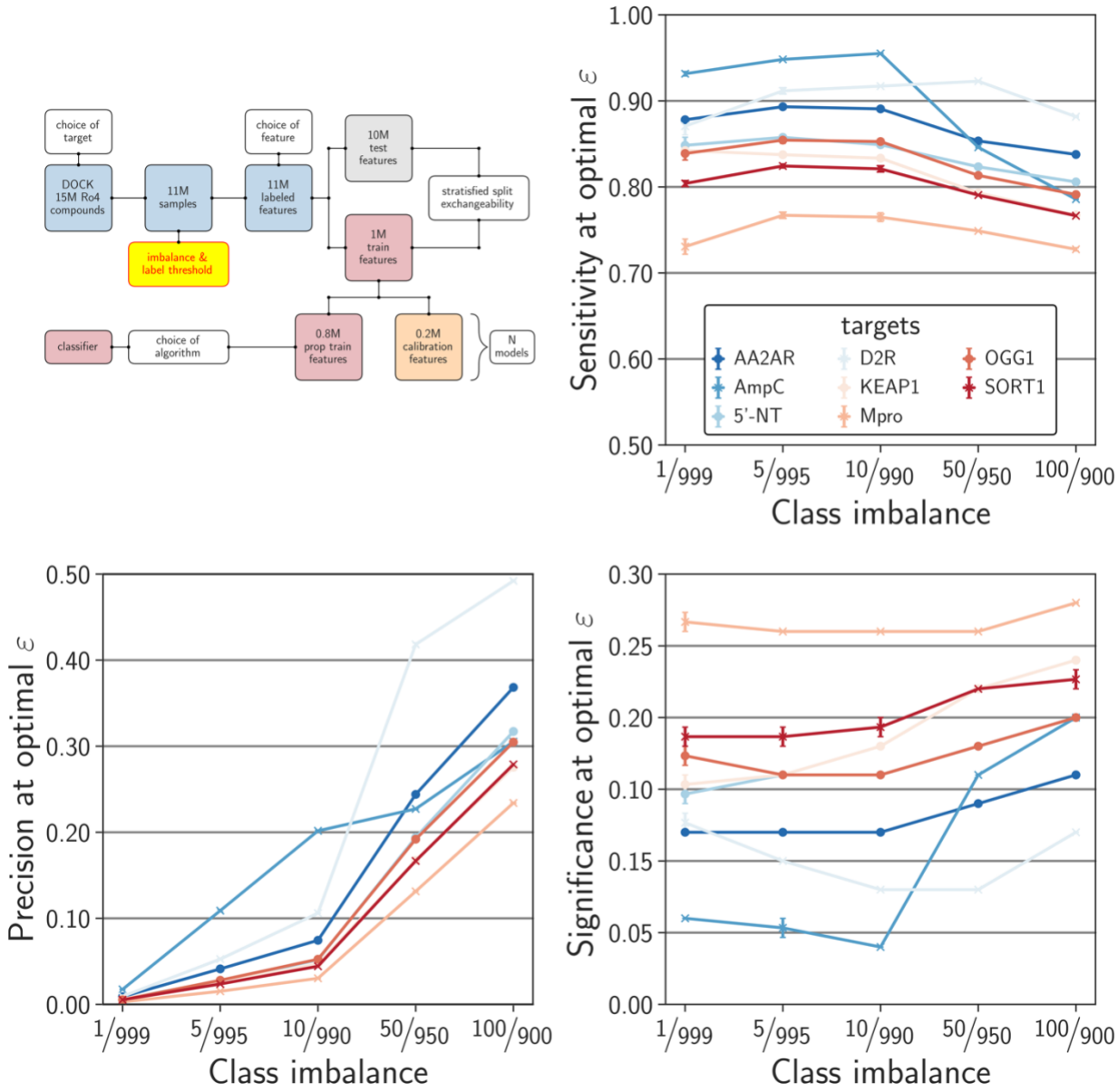
Supplementary Figure S2. Learning rate and weight decay analysis for deep neural networks. The changes in training loss, valid loss, valid accuracy, and speeds during training were monitored for deep neural networks with learning rates (LR) and weight decays (WD). Models were trained on one million molecules of the AmpC dataset represented by ECFP4 descriptors, and hence the input dimension was set to 1024. The output dimension was set to two for binary classification (virtual active and virtual inactive). The early stop patience for valid loss was set to 3, after which the best performing checkpoint (grey dashed line) was stored as final model. The default learning rate was then set to 1e-4 and the default weight decay was set to 1e-2. See Supporting Table 1.



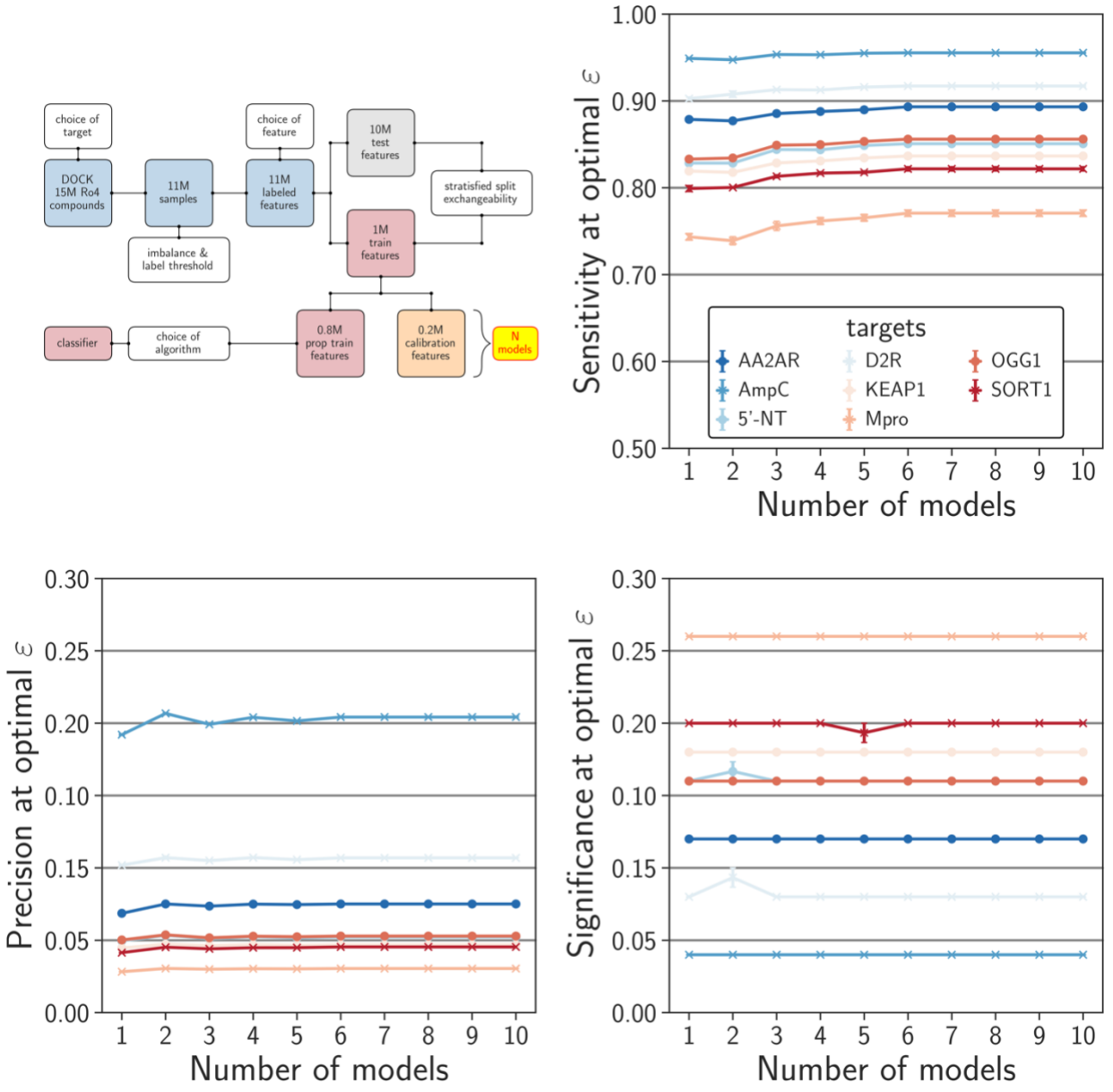
Supplementary Figure S3. Architecture analysis for deep neural networks. The changes in training loss, valid loss, and valid accuracy during training were monitored for deep neural networks with different architectures, which are shown above each subplot. Models were trained on one million molecules of the AmpC dataset represented by ECFP4 descriptors, and hence the input dimension was set to 1024. The output dimension was set to two for binary classification (virtual active and virtual inactive). The learning rate was set to 1e-4 and the weight decay was set to 1e-2. The early stop patience for valid loss was set to 3, after which the best performing checkpoint (grey dashed line) was stored as final model. The [input]-[1000]-[4000]-[2000]-[2] architecture was then selected as default.



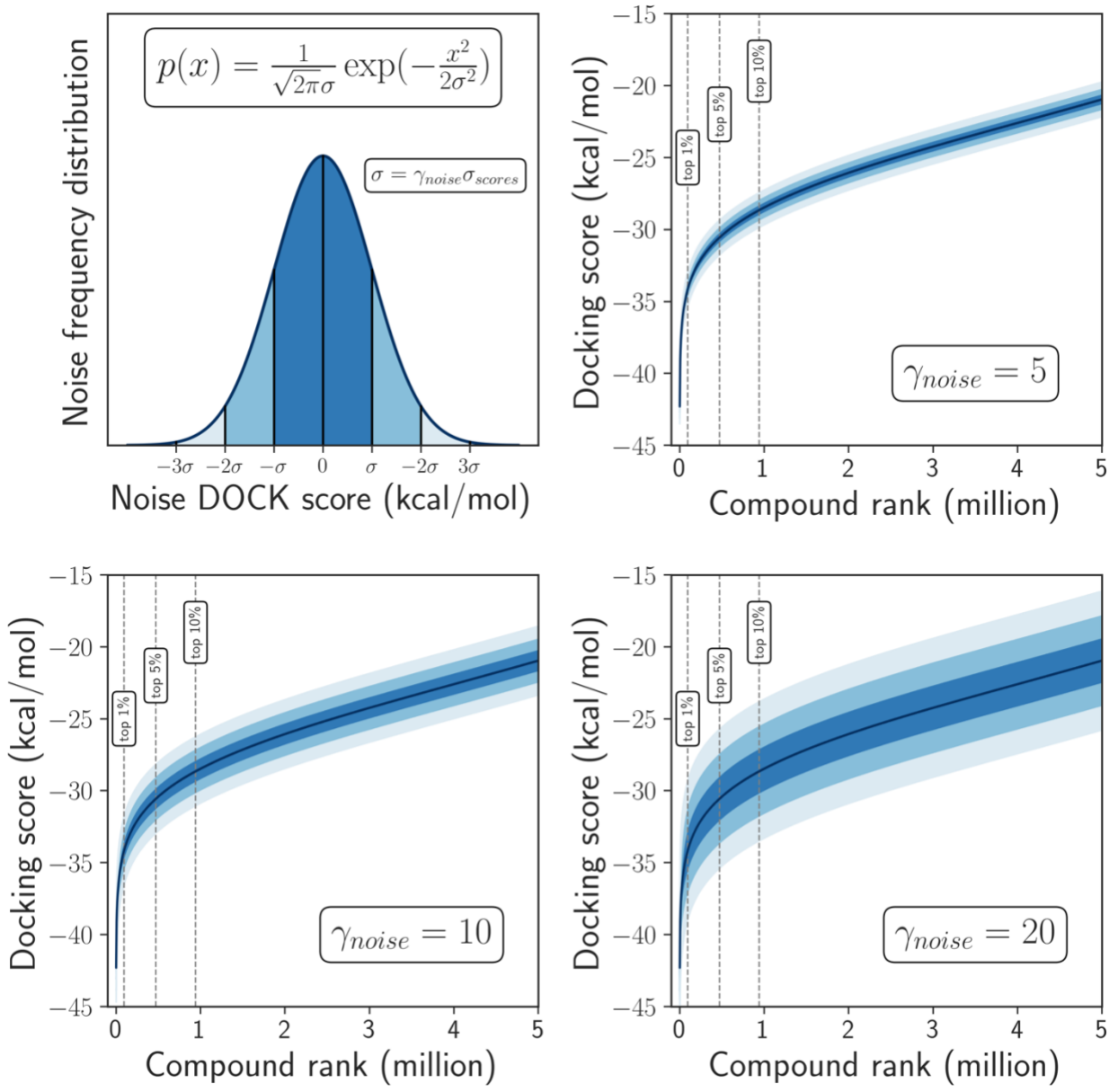
Supplementary Figure S4. Learning rate analysis for RoBERTa. (A) The changes in sensitivity and precision during training were monitored for the RoBERTa classifiers. Models were trained on one million AmpC molecules using RoBERTa's internal descriptors. A small external test set of 200000 molecules was used to obtain the sensitivity and precision metrics. Three independent calculations were carried out. The default number of epochs was set to ten in all other calculations. (B) RoBERTa models were trained on one million A_2AR molecules with three different learning rates: 1e-5 (B), 4e-6 (C), and 4e-8 (D). The relative set distributions for different significance values are shown, together with the significance at which the predict achieves highest efficiency. The default learning rate was then set to 4e-7 for training RoBERTa models. See Supporting Table 1.



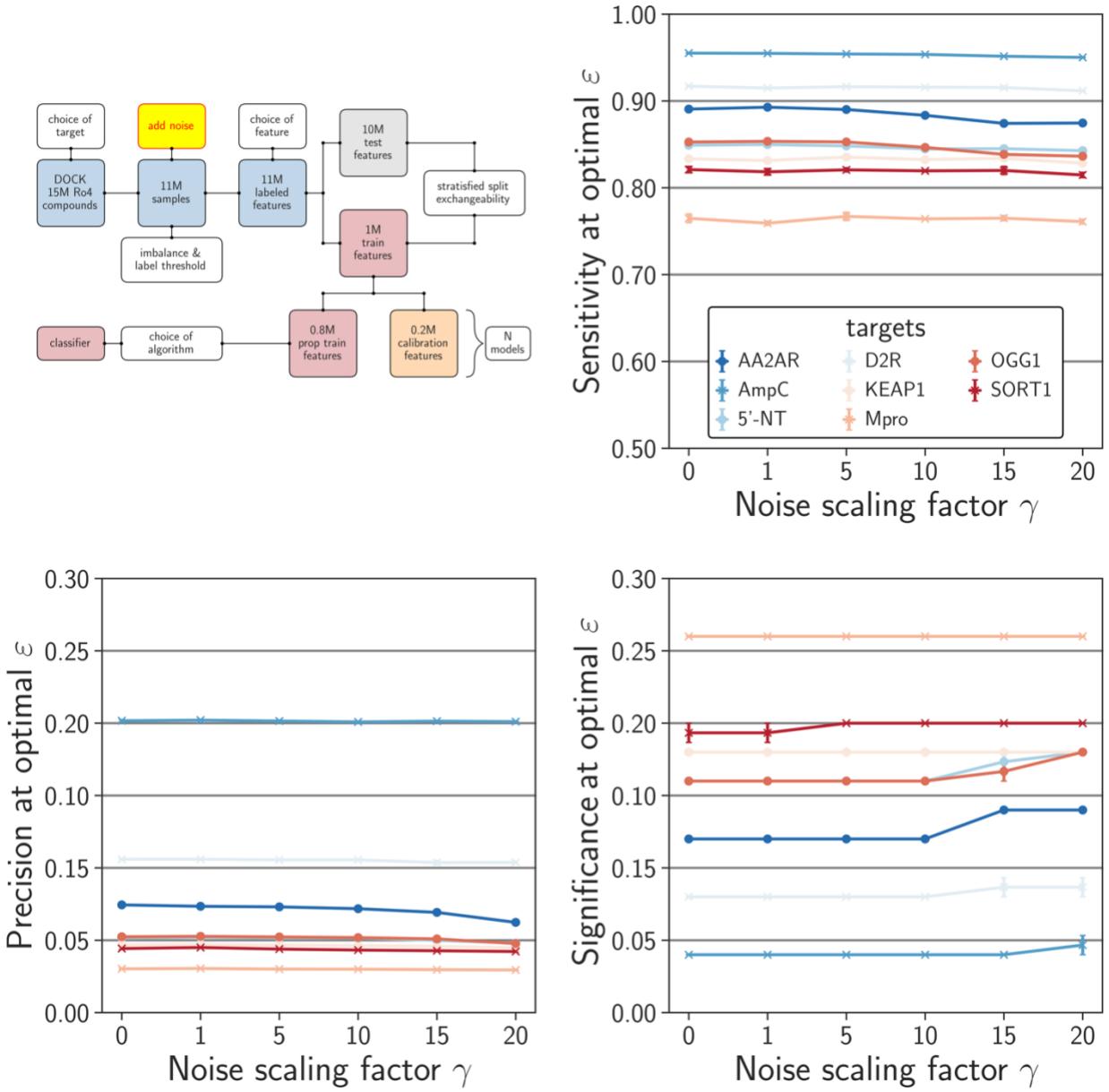
Supplementary Figure S5. Performance on imbalanced datasets. Sensitivity and precision at optimal efficiency were analyzed for different class imbalances. Five independent CatBoost models were trained on one million molecules represented by ECFP4 descriptors. Each test set contained ten million molecules. Three independent calculations (training and prediction) were performed for the eight targets and error bars correspond to the standard error of the mean.



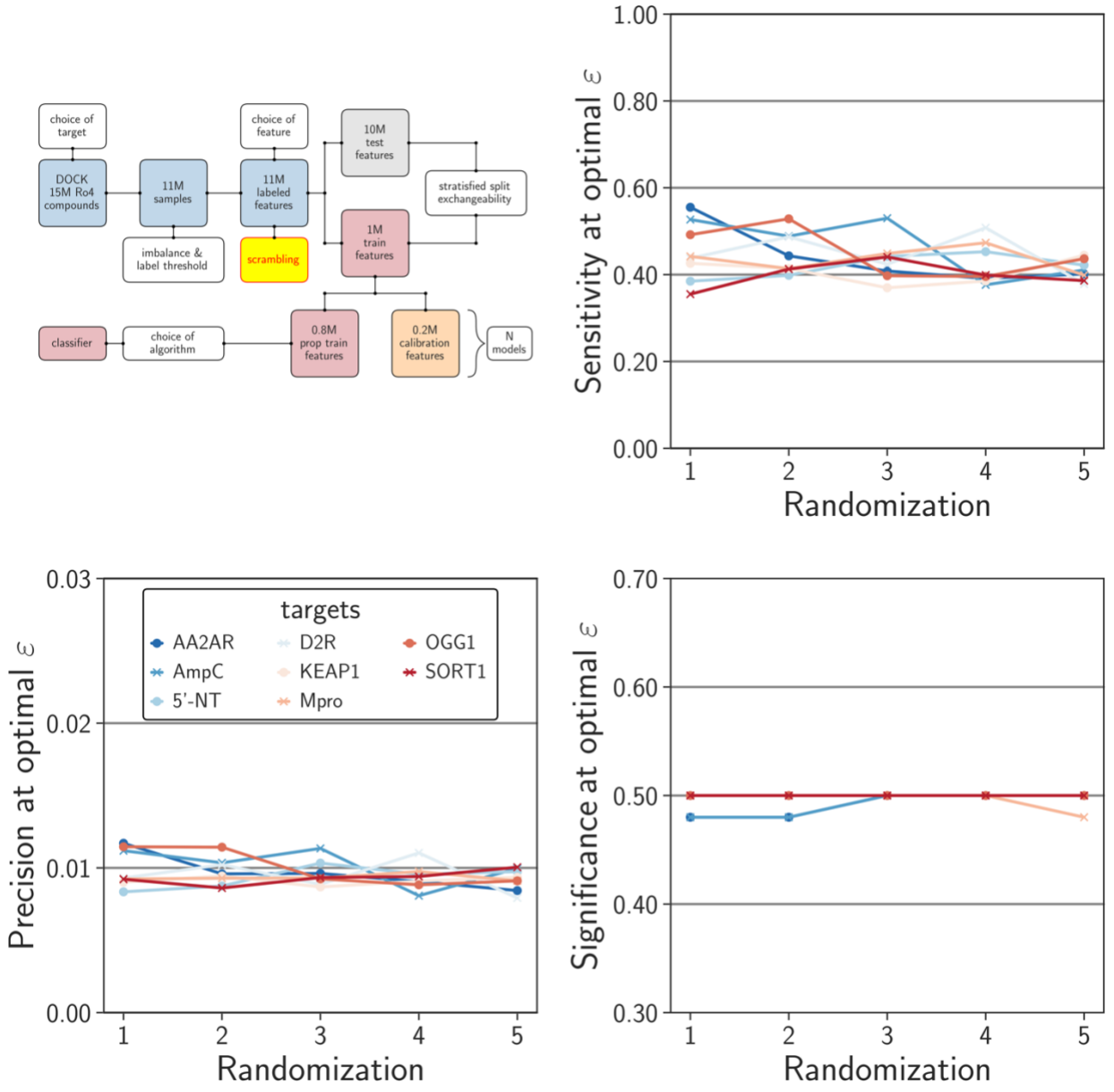
Supplementary Figure S6. Performance on number of aggregated models. Sensitivity and precision at optimal efficiency were analyzed for a different number of models during aggregation. Five independent CatBoost models were trained on one million molecules represented by ECFP4 descriptors. Each test set contained ten million molecules. Three independent calculations (training and prediction) were performed for the eight targets and error bars correspond to the standard error of the mean.



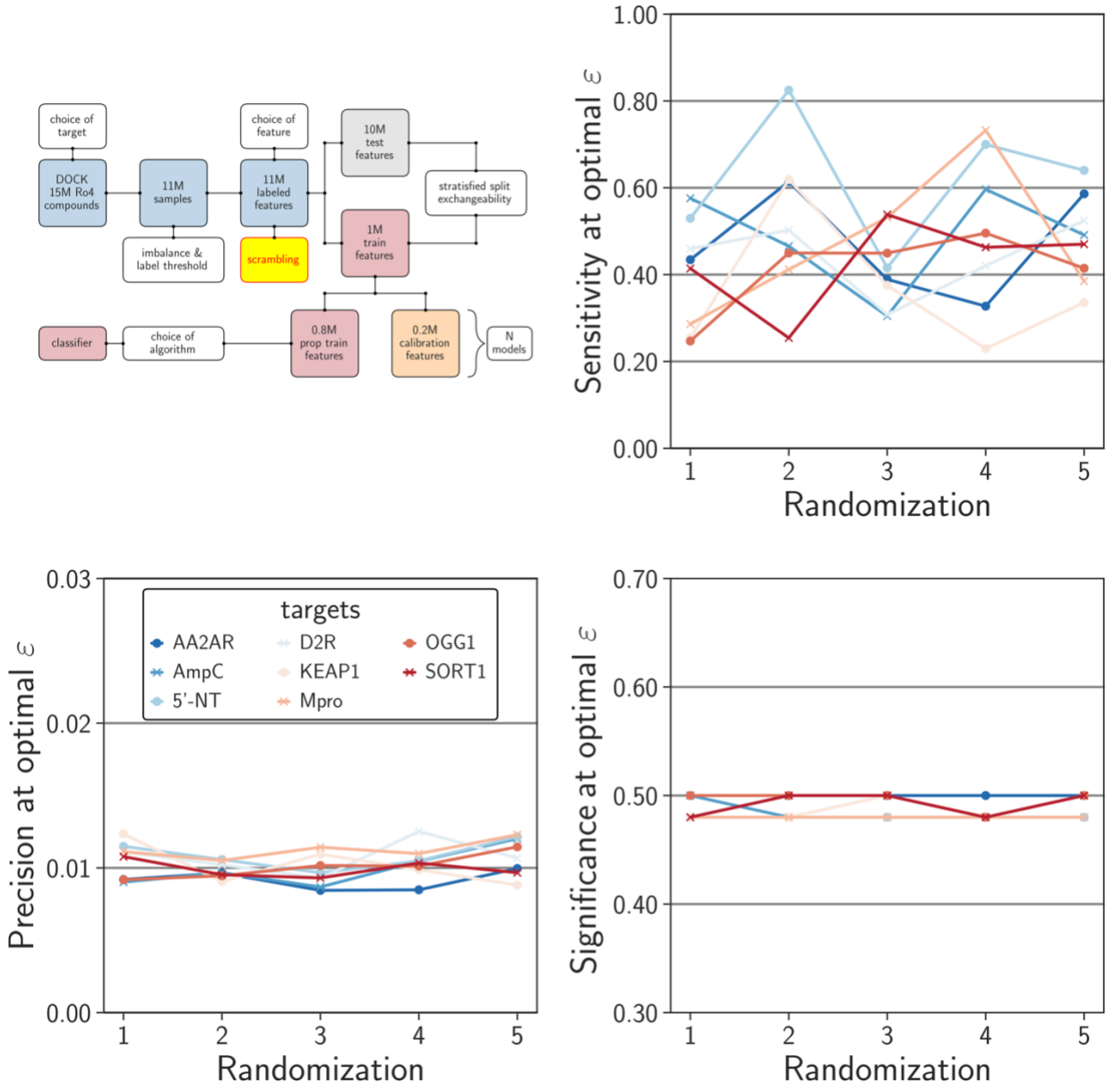
Supplementary Figure S7a. Overview of noise addition. A zero-centered normal distribution was constructed using the standard deviation (σ_{scores}) of the docking score distribution and a noise scaling factor (γ_{noise}). Noise was added to the score of each sample by taking a sample from the corresponding noise distribution. Large noise scaling factors led to wide distributions and increased perturbations of the initial docking score distributions.



Supplementary Figure S7b. Performance on noisy datasets. Sensitivity and precision at optimal efficiency were analyzed for datasets generated with different noise scaling factors (γ_{noise}). Five independent CatBoost models were trained on one million molecules represented by ECFP4 descriptors. Each test set contained ten million molecules. Three independent calculations (training and prediction) were performed for the eight targets and error bars correspond to the standard error of the mean.



Supplementary Figure S8a. Performance on non-sensical datasets - labels. Sensitivity and precision at optimal efficiency were analyzed for datasets where the labels were scrambled without affecting the class imbalance. Five independent CatBoost models were trained on one million molecules represented by ECFP4 descriptors. Each test set contained ten million molecules. Five independent calculations (training and prediction) were performed for the eight targets. When the CP operates at an optimal efficiency significance of 50%, has a sensitivity averaging around 50%, and a precision close to the class imbalance (1%), the performance will correspond to random classification.



Supplementary Figure S8b. Performance on non-sensical datasets - features. Sensitivity and precision at optimal efficiency were analyzed for datasets where the feature vectors were shuffled. Five independent CatBoost models were trained on one million molecules represented by ECFP4 descriptors. Each test set contained ten million molecules. Five independent calculations (training and prediction) were performed for the eight targets. When the CP operates at an optimal efficiency significance of 50%, has a sensitivity averaging around 50%, and a precision close to the class imbalance (1%), the performance will correspond to random classification.