Practical model selection for prospective virtual screening

Shengchao Liu,[†,‡,#] Moayad Alnammi,[†,‡,#] Spencer S. Ericksen,[¶] Andrew F. Voter,[§] James L. Keck,[§] F. Michael Hoffmann,[¶,‖] Scott A. Wildman,[¶] and Anthony Gitter[*,⊥,†,‡]

†*Department of Computer Sciences, University of Wisconsin-Madison*

‡*Morgridge Institute for Research*

¶*Small Molecule Screening Facility, University of Wisconsin Carbone Cancer Center*

§*Department of Biomolecular Chemistry, University of Wisconsin School of Medicine and Public Health*

‖*McArdle Laboratory for Cancer Research, University of Wisconsin-Madison*

⊥*Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison*

#*Co-first author*

E-mail: gitter@biostat.wisc.edu

Abstract

Virtual (computational) high-throughput screening provides a strategy for prioritizing compounds for experimental screens, but the choice of virtual screening algorithm depends on the dataset and evaluation strategy. We consider a wide range of ligand-based machine learning and docking-based approaches for virtual screening on two protein-protein interactions, PriA-SSB and RMI-FANCM, and present a strategy for choosing which algorithm is best for prospective compound prioritization. Our workflow identifies a random forest as the best algorithm for these targets over more sophisticated neural network-based models. The top 250 predictions

from our selected random forest recover 40 of the 62 active compounds from a library of 25,279 new molecules assayed on PriA-SSB. We show that virtual screening methods that perform well in public datasets and synthetic benchmarks, like multi-task neural networks, may not always translate to prospective screening performance on a specific assay of interest.

1 Introduction

Drug discovery is time consuming and expensive. After a specific protein or mechanistic pathway is identified to play an essential role in a disease process, the search begins for a chemical or biological ligand that can perturb the action or abundance of the disease target in order to mitigate the disease phenotype. A standard approach to discover a chemical ligand is to screen thousands to millions of candidate compounds against the target in biochemical or cell-based assays via a process called high-throughput screening (HTS), which produces vast sets of valuable pharmacological data. Even though HTS assays are highly automated, screens of thousands of compounds sample only a small fraction of the millions of commercially-available, drug-like compounds. Cost and time preclude academic laboratories and even pharmaceutical companies from blindly testing the full set of millions of drug-like compounds in HTS assays. Thus, there is a crucial need for an effective virtual screening (VS) process as a preliminary step in prioritizing compounds for HTS assays.

Virtual screening comprises two categories: structure-based[1,2] and ligand-based methods.[3,4] Structure-based methods require that the target protein's molecular structure be known so that the 3D interactions between the target and each chemical compound (binding poses) may be predicted *in silico*. These interactions are given numerical scores, which are then used to rank compounds for potential binding to the target. These methods do not require or typically make use of historical screening data in compound scoring. In contrast, ligand-based methods require no structural information about the target but use data generated from testing molecules in biochemical or functional assays of the target to fit empirical models that relate compound attributes to assay outcomes.

For targets with abundant assay data or where a druggable binding site is not well-defined,

such as the targets considered here, ligand-based methods are generally superior to structure-based methods.[5–7] Confronted with the variety of ligand-based model building methods (e.g., regression models, random forests, support vector machines, etc.),[8] compound input representations, and performance metrics, how should one proceed? The Merck Molecular Activity Challenge[9] incited the development of many ligand-based deep learning VS methods,[10–14] as recently reviewed.[15] These methods are often assessed with cross-validation on existing HTS data, but there is presently little experimental evidence on the best option for prioritizing new compounds given a fixed screening budget.

We critically evaluated a collection of VS algorithms that include both structure-based and ligand-based methods. We present a VS workflow that first uses available HTS training data to systematically prune the specific versions of the algorithms and calculate their cross-validation performance on a variety of evaluation metrics. Based on the cross-validation results and analysis of the various evaluation metrics, we selected a single virtual screening algorithm. The selected method, a random forest model, was the best option for prioritizing a small number of compounds from a new library, as verified by experimental screening. These model selection and evaluation strategies can guide VS practitioners to select the best model for their target even as the landscape of available VS algorithms continues to evolve.

2 Methods

2.1 Datasets

Our case studies were on new and recently generated datasets[16,17] for the targets PriA-SSB and RMI-FANCM. The PriA-SSB interaction is important in bacterial DNA replication and is a potential target for antibiotics.[18] The RMI-FANCM interaction is involved in DNA repair that is induced in human cancer cells to confer chemoresistance to cytotoxic DNA-cross-linking agents, making it an attractive drug target.[19] We previously screened these targets with a library of compounds obtained from Life Chemicals Inc. (LC) in different assay formats. In addition, we screened new

LC compounds on the PriA-SSB target to evaluate our VS models. The four datasets derived from these screens are described below and summarized in Table 1.

PriA-SSB AlphaScreen: PriA-SSB was initially screened using an AlphaScreen (AS) assay in 1536-well format[17] on 72,423 LC compounds at a single concentration (33.3 μM), with data reported as % inhibition compared to controls. We refer to these continuous values as "PriA-SSB % inhibition". Those compounds that tested above a certain activity threshold ($\geq$ 35% inhibition) and passed PAINS chemical structural filters[20,21] were retested in the same AS assay. Compounds that were confirmed in the AS retest screen ($\geq$ 35% inhibition) were marked as actives, creating the binary dataset **PriA-SSB AS**.

PriA-SSB fluorescence polarization: Compounds that had PriA-SSB % inhibition $\geq$ 35% and passed the PAINS filters were also tested in a fluorescence polarization (FP) assay as a secondary screen. Those compounds with FP inhibition $\geq$30% were labeled as actives, creating the binary dataset **PriA-SSB FP**, with all other compounds in the screening set labeled inactive.

RMI-FANCM fluorescence polarization: The RMI-FANCM interaction was initially screened with a subset of 49,796 compounds from the same LC library as PriA-SSB.[16] This FP assay was run at a single compound concentration (32 μM). We refer to these continuous values as "RMI-FANCM % inhibition". Those compounds that demonstrated activity $\geq$ 2 standard deviations (SD) above the assay mean and passed PAINS filters were marked as actives in the binary dataset **RMI-FANCM FP**.

PriA-SSB prospective: In subsequent screens used for prospective testing, we tested an additional 25,279 compounds, purchased from LC after the initial screening was complete. There was no overlap between these compounds and the 72,423 LC compounds described above that were used to train VS models. As with the initial library, the PriA-SSB AS assay was used in the same 1536-well format at a single concentration (33.3 μM) to test the additional 25,279 LC compounds. Actives were defined with the same criteria used for the binary dataset **PriA-SSB AS**. Compounds with at least 35% inhibition that passed the PAINS filters were retested with the AS assay. Those with at least 35% inhibition in the AS retest were labeled as actives, creating the binary dataset

PriA-SSB prospective.

Because secondary screens and structural filters were used to define the active compounds, there was no single primary screen % inhibition threshold that separated the actives from the inactives. Some compounds exhibiting high % inhibition values were labeled as inactive because they did not satisfy the structural requirements or were not active in the secondary screen.

Table 1: Summary statistics for the four binary datasets.

Stage	Dataset	% inhibition threshold	# actives	# inactives
Cross-validation	**PriA-SSB AS**	$\geq 35\%$	79	72,344
	PriA-SSB FP	$\geq 30\%$	24	72,399
	RMI-FANCM FP	≥ 2 SD	230	49,566
Prospective	**PriA-SSB prospective**	$\geq 35\%$	62	25,217

To help learn a better chemical representation with multi-task neural networks, we considered other screening contexts from which to transfer useful knowledge. We used a specific subset of 128 assays (AIDs) from the **PubChem BioAssay (PCBA)**[22] repository. This dataset was used in previous work on multi-task neural networks.[14] This subset contained assays for which the assays were developed to probe a specific protein target and dose-response measurements were obtained for each compound (among other assay query filters, see Appendix A). Potency and curve quality are factored into a PubChem Activity Score. Regardless of assay, compounds receiving a PubChem Score of 40 or greater (0-100) were assigned a PCBA Bioactivity outcome (label) of "Active". Compounds receiving PC Activity Scores 1-39 were labeled "Inconclusive," and those receiving 0 are labeled "Inactive" (Appendices B and C).

2.2 Compound Features

Ligand-based virtual screening methods require each chemical compound to be represented in a particular format as input to the model. We adopted two common representations. All of the ligand-based algorithms except the Long Short-Term Memory (LSTM) neural network use 1024-bit Extended Connectivity Fingerprints (ECFP4).[23] For LSTM networks, we used the Simplified Molecular Input Line Entry System (SMILES) representation,[24] where the characters were treated

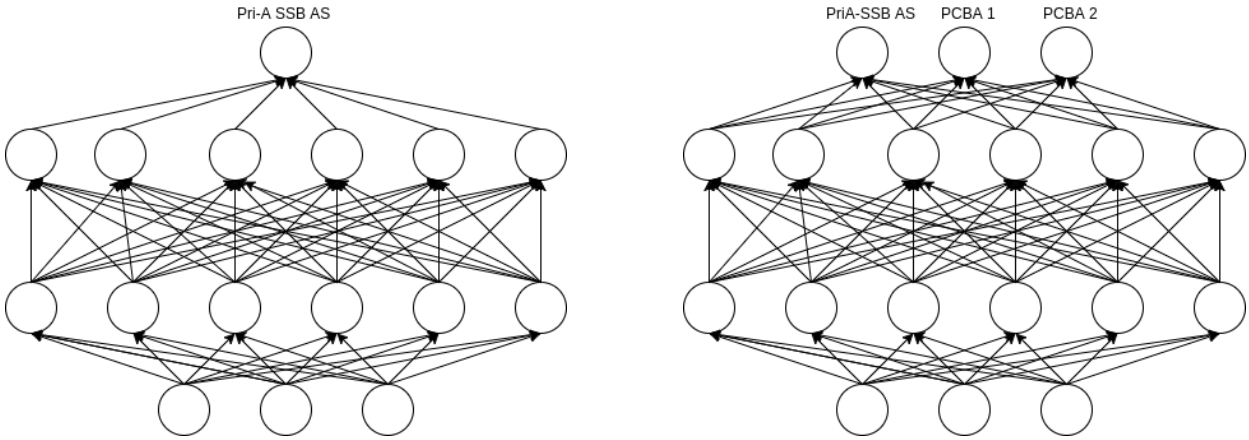
as sequential features. Both ECFP and canonical SMILES were generated in RDKit.[25]

2.3 Virtual Screening Models

We selected a variety of existing virtual screening approaches for our benchmarks and prospective predictions. These included ligand-based supervised learning approaches, structure-based docking, and a chemical similarity baseline. Table 2 summarizes the types of training data used by each algorithm. We discuss hyperparameter tuning for these models in Section 2.5.

2.3.1 Ligand-Based Neural Networks

Deep learning is a machine learning approach that encompasses neural network models with multiple hidden layer architectures and the techniques for training these models. It represents the state of the art for many predictive tasks, which has generated extensive interest in deep learning for biomedical research, including virtual screening.[15] We evaluated multiple types of established neural network architectures for virtual screening.



(a) Single-task Neural Network (b) Multi-task Neural Network

Figure 1: Neural network structures. The neural networks map the input features (e.g. fingerprints) in the input (bottom) layer to intermediate chemical representations in the hidden (middle) layers and finally to the output (top) layer, which makes either continuous or binary predictions. Figure 1(a) has only one unit in the output layer. Figure 1(b) has multiple units in the output layer representing different targets, one for our new target of interest and the others for PCBA targets.

Single-task Neural Network (STNN): A single-task neural network (Figure 1(a)) makes a

single prediction for a single target (also referred to as a task). We trained a separate model for each of the **PriA-SSB AS**, **PriA-SSB FP**, and **RMI-FANCM FP** datasets, taking a compound's ECFP4 elements (1024 bits) as the input features. We trained the neural networks using Keras[26] with the Theano backend.[27] The single-task neural networks were trained on each task to predict either the binary activity label in the classification setting (STNN-C) or the continuous % inhibition in the regression setting (STNN-R).

Multi-task Neural Network (MTNN): Multi-task neural networks make different predictions for multiple targets or tasks but share knowledge by training the first few hidden layers together. Each of our multi-task neural networks included one target task (**PriA-SSB AS**, **PriA-SSB FP**, or **RMI-FANCM FP**) and 128 tasks from PCBA. We only trained multi-task neural networks in the classification setting (MTNN-C).

Single-task Atom-level LSTM (LSTM): The LSTM is one of most prevalent recurrent neural network models,[28] which has been applied previously in virtual screening.[29] An LSTM assumes there exists a sequential pattern in the input string. We used a one hot encoding of the SMILES strings as input for the LSTM model. In a one hot encoding, each character in a SMILES string is replaced by a binary vector. The binary vector has one bit for each possible unique character in all SMILES strings. At each position in a SMILES string, the bit corresponding to the character at that position is set to 1, and all other bits are set to 0. We trained the LSTM model to predict the binary activity labels.

Influence Relevance Voter (IRV): IRV[30,31] is a hybrid between k-nearest neighbors and neural networks. Each compound's output value is a non-linear combination of the similarity scores from its most closely related compounds in the training dataset. We used ECFP4 fingerprints as the input and trained separate IRV models for each dataset.

2.3.2 Other Ligand-Based Models

Random Forest (RF): Random forests are ensembles of decision trees that are often used as a baseline in virtual screening benchmarks.[32,33] We used scikit-learn[34] to train a random forest

classifier for each binary label with ECFP4 fingerprints as features.

2.3.3 Protein-Ligand Docking

Target Preparation: Our structure-based VS approach involved the docking-based ranking of the LC library to the holo-form of PriA, using the crystal structure (PDB: 4NL8)[35] in which it is bound to a C-terminal segment of an SSB protein. A missing loop in this structure was added from the the apo-form (PBD: 4NL4), though this is not near the SSB binding site. The docking search space was limited to 8Å from the coordinates of the co-crystallized SSB C-terminal tripeptide.

Compound Preparation: LC library compounds were assigned 3D coordinates and Merck Molecular Force Field partial charges using OpenEye OMEGA and Molcharge.[36] Compounds in the LC library with ambiguous stereochemistry were enumerated in all possibilities, and the best resulting docking score was retained for each.

Docking (Dock) and Consensus Docking (CD): We ran eight different docking programs and generated nine docking scores. The docking programs and names we use for their scores are AutoDock v4.2.6[37] (Dock_ad4), Dock v6.7[38] (Dock_dock6), FRED v3.0.1[39] (Dock_fred), HYBRID v3.0.1[39] (Dock_hybrid), PLANTS v1.2[40] (Dock_plants), rDock v2013.1[41] (Dock_rdocktot and Dock_rdockint), Smina v1.1.2[42] (Dock_smina), and Surflex-Dock v3.040[43] (Dock_surflex). In addition, we calculated consensus docking scores using three traditional approaches (CD_mean, CD_median, and CD_max) and two versions of the Boosting Consensus Score (CD_efr1_opt and CD_rocauc_opt).[44] The Boosting Consensus Score method was trained using 21 of the DUD-E benchmark targets[45] without any PriA-SSB or RMI-FANCM assay data. Compounds with missing scores due to preparation or docking failures were not considered during evaluation.

2.3.4 Chemical Similarity Baseline

In the prospective screening stage, we introduced a simple baseline compound ranking method against which to compare more sophisticated VS methods. Each of the 25,279 compounds in the prospective screening set was compared to the actives in **PriA-SSB AS** using Tanimoto similarity on

the ECFP4 fingerprints. Compounds in the prospective library were ranked by their similarity, and those with Tanimoto similarity ≥ 0.45 to any active compound in the initial screen were predicted to be active.

In addition, all compounds were clustered by two separate approaches to describe chemical series. Chemical similarity based hierarchical clusters on ECFP4 using Ward's clustering are described as SIM. Maximum common substructure clusters, used to group molecules with similar scaffolds, are described as MCS. JKlustor was used for both types of clustering (JChem v17.26.0, ChemAxon).

Table 2: A summary of the virtual screening methods and which labels each model used during training. The docking and consensus docking models do not train on the PriA-SSB or RMI-FANCM datasets.

Model	Continuous % inhibition	Binary label	PCBA binary labels
Dock			
CD			
STNN-C		✓	
STNN-R	✓		
MTNN-C		✓	✓
LSTM		✓	
RF		✓	
IRV		✓	
Similarity baseline		✓	

2.4 Evaluation Metrics

Given our goal of developing VS methods that enable very small, cost-effective, productive screens, we considered how metrics weight early active retrieval. All of the VS algorithms produce a ranked list of compounds, where compounds are ordered by the probability of being active, the continuous predicted % inhibition, the docking score, or a comparable output value. For a ranked list of compounds, we can threshold the ranked list and consider all compounds above the threshold as positive (active) predictions and those below the threshold as negative (inactive). By comparing those predictions to the experimentally-observed activity, we can compute true positive (TP), false

positive (FP), true negative (TN), and false negative (FN) predictions for that ranked list at that threshold. We explored several options for summarizing how well each algorithm ranks the known active compounds.

The area under the receiver operating characteristic curve (AUC[ROC]) has been recommended for virtual screening because it is robust, interpretable, and does not depend on user-defined parameters.[46] The ROC curve plots the relationship between true positive rate (TPR, also known as sensitivity or recall) and false positive rate (FPR, equivalent to 1 - specificity), which are defined in Equation 1. As the FPR goes to 100%, all ROC curves will converge, whereas early active retrieval (a more meaningful characteristic of VS performance) can be assessed in the low FPR region of the ROC curve, which exhibits greater variability across VS methods. Thus, we also considered the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC).[47] It emphasizes the early part of the ROC curve through a scaling function α, which we set to 10 for our purposes of early enrichment up to 20%. We used the BEDROC implementation from the CROC Python package.[48]

$$TPR = \frac{TP}{TP+FN}, \qquad FPR = \frac{FP}{FP+TN} \tag{1}$$

$$Recall = \frac{TP}{TP+FN}, \quad Precision = \frac{TP}{TP+FP} \tag{2}$$

Area under the precision-recall curve (AUC[PR]) is another common metric (Equation 2). AUC[PR] has an advantage over AUC[ROC] for summarizing classifier performance when the class labels are highly-skewed, as in virtual screening where there are few active compounds in a typical library. AUC[PR] evaluates a classifier on its ability to retrieve actives (recall) and which of the predicted actives correctly classified (precision) as we vary the prediction threshold. We used the PRROC R package[49] to compute AUC[PR] because it correctly interpolates between points in the PR curve when there are ties in the compound rankings.[50]

Another VS metric is enrichment factor (EF), which is the ratio between the number of actives

found in some prioritized subset of compounds versus the expected number of actives in a random subset of same size. In other words, it assesses how much better the VS method performs over random compound selection. Let $R \in [0\%, 100\%]$ be a pre-defined fraction of the compounds from the total library of compounds screened.

$$EF_R = \frac{\text{\# actives in top } R \text{ ranked compounds}}{\text{\# actives in entire library} \times R} \tag{3}$$

$$EF_{max,R} = \frac{min\{\text{\# actives, total \# compounds} \times R\}}{\text{\# actives in entire library} \times R} \tag{4}$$

$EF_{max,R}$ represents the maximum enrichment factor possible at R. Difficulty arises when interpreting EF scores because they vary with the dataset and threshold R. We defined the normalized enrichment factor (NEF) as:

$$NEF_R = \frac{EF_R}{EF_{max,R}} \tag{5}$$

Because $NEF_R \in [0, 1]$, it is easier to compare performance across datasets and thresholds. 1.0 is the perfect normalized enrichment factor. Furthermore, we can create an NEF curve as NEF_R versus $R \in [0\%, 100\%]$ and compute the area under that curve to obtain AUC[NEF] $\in [0, 1]$. However, most models tend to exhibit similar late enrichment behavior. We are typically interested in early enrichment behavior so we computed AUC[NEF] using $R \in [0\%, 20\%]$.

Finally, we introduce the metric n_{hits} which is simply the number of actives found in a selected number of tested compounds (e.g. how many hits or actives were found in 250 tested compounds). This metric represents the typical desired utility of a screening process: retrieve as many actives as possible in the selected number of tests (denoted as n_{tests}). We compare n_{hits} at various n_{tests} to the different evaluation metrics to identify which metrics best mimic the n_{hits} utility.

2.5 Pipeline

Our virtual screening workflow contains three stages:

1. Tune hyperparameters in order to prune the model search space.

2. Train, evaluate, and compare models with k-fold cross-validation results.

3. Assess the best models' ability to prospectively identify active compounds in a new set.

In contrast to most other virtual screening studies, the experimental screen was not conducted until after all models were trained and evaluated in the cross-validation stage. For the first two stages, we first split the **PriA-SSB AS**, **PriA-SSB FP**, and **RMI-FANCM FP** datasets into five stratified folds as described in Appendix B.

2.5.1 Hyperparameter Sweeping Stage

Hyperparameters are model configurations or settings that can be set by an expert as opposed to the weights or parameters that are learned or fit during model training. For example, in neural networks the hyperparameters include the number of hidden layers, the number of hidden units in each layer, types of activation functions, drop out ratios, learning rates, etc. In random forests, hyperparameters include the number and depth of trees, the size of the subsamples, etc.

For most of the ligand-based machine learning models, the hyperparameter space was too large for exhaustive searches using the full dataset. Therefore, we applied a grid search on a pre-defined set of hyperparameters in a smaller dataset and pruned those that performed poorly. We performed a single iteration of training on the first four folds of **PriA-SSB AS** to avoid over-fitting. The hyperparameters considered are listed in Appendix D.

2.5.2 Cross-Validation Stage

To identify which VS algorithms are likely to have the best performance in a prospective screen, we applied a traditional cross-validation training strategy on datasets **PriA-SSB AS**, **PriA-SSB FP**, and

RMI-FANCM FP after reducing the hyperparameter combinations to consider. Selecting the best model is non-trivial. Ideally, the best model would have dominant performance on all evaluation metrics, but this is rarely observed with existing models. Each evaluation metric prioritizes different performance characteristics. Our cross-validation results illustrate which models consistently perform well over different metrics, the correspondence of metrics relative to a desired utility (n_{hits}), and how to choose models and evaluation metrics in order to successfully identify active compounds in a prospective screen.

Cross-validation is commonly used to avoid over-fitting when there are few training samples. We split the training data into 5 folds, 4 folds for training and 1 for testing. Models like RF and IRV that do not require a hold-out dataset for early stopping used 4 folds for training. Other models like the neural networks perform early-stopping based on a hold-out set, so we iteratively selected 1 of the 4 training data folds for this purpose. This led to a two-layer cross-validation with $5 \times 4 = 20$ trained neural networks.

2.5.3 Prospective Screening Stage

Our prospective screen used a library of 25,279 new LC compounds that were not present in the training set. We used each VS model to prioritize 250 of these compounds that are most likely to be active. This emulates virtual screening on much larger compound libraries, in which only a small fraction of all computationally scored compounds can be tested experimentally. After finalizing the models' predictions, we screened all 25,279 compounds in the wet lab and assigned actives based on a 35 % inhibition threshold and structural filters (**PriA-SSB prospective**). Finally, we evaluated how many of the wet lab actives each VS method identified in its top 250 predictions, the number of distinct chemical clusters recovered, and the number of active compounds that were not in the top 250 predictions from any of the VS algorithms. The prospective screen allowed us to assess how well the cross-validation results generalized to new compounds and further verified our conclusions from the retrospective cross-validation tests.

2.6 Data and Software Availability

Code implementing our ligand-based virtual screening algorithms is available at `https://github.com/gitter-lab/pria_lifechem`. This GitHub repository also contains additional Jupyter notebooks to reproduce the visualizations and analyses. Version 0.1.0 of our software is archived on Zenodo (doi:10.5281/zenodo.1257674). Our PriA-SSB and RMI-FANCM HTS data will be made available on PubChem. A formatted version of this dataset for training virtual screening algorithms is available on Zenodo (doi:10.5281/zenodo.1257463).

3 Results

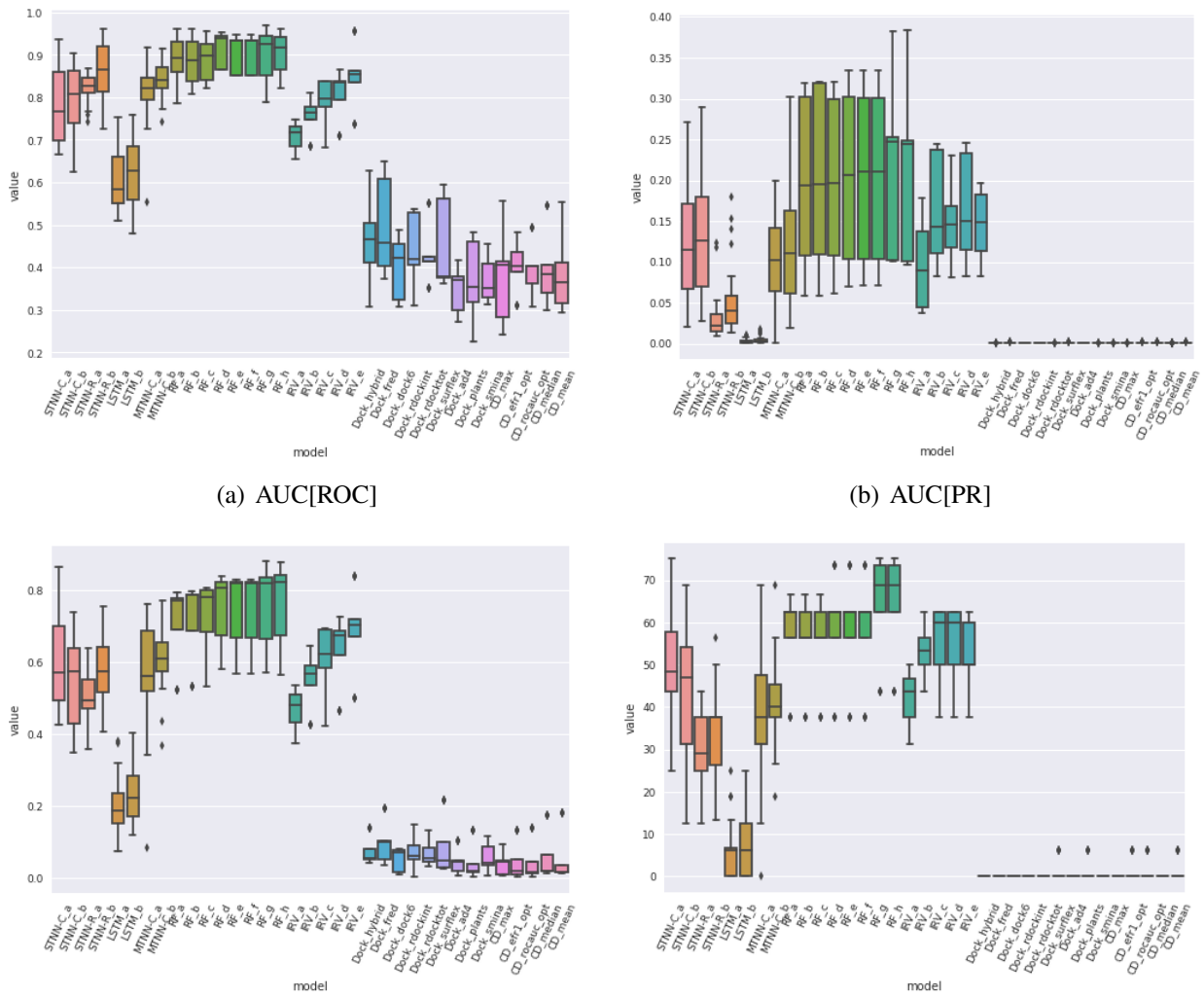
3.1 Cross-Validation Results

In the cross-validation stage we assessed 35 models: 8 neural networks (STNN-C, STNN-R, MTNN-C, and LSTM), 8 random forests (RF), 5 influence relevance voter (IRV), and 14 from docking (Dock) or consensus docking (CD). When there are multiple versions of a model that use different hyperparameters, we distinguish them with alphabetic suffixes such as "_a" and "_b". Appendix E describes the hyperparameters associated with these suffixes. We highlight the **PriA-SSB AS** dataset as a representative example, but the VS workflow is applicable for all tasks.

3.1.1 Comparing Virtual Screening Algorithms

We tested all 35 models on three datasets, and the results for four evaluation metrics on the **PriA-SSB AS** dataset are shown in Figure 2. Appendix F contains the results for **PriA-SSB FP** and **RMI-FANCM FP**. The **PriA-SSB AS** performance using AUC[ROC] was comparable for most models, where nearly all models except LSTM and the docking programs were above 0.8 AUC[ROC]. Random forest was the best model, reaching AUC[ROC] of approximately 0.9.

Random forest was again the best method for the **RMI-FANCM FP** dataset for most of the evaluation metrics (Appendix F). On the **PriA-SSB FP** dataset, STNN-R achieved the highest



(a) AUC[ROC]

(b) AUC[PR]

(c) AUC[BEDROC]

(d) EF@1%

Figure 2: The evaluation metric distributions on **PriA-SSB AS** on all models over the cross-validation folds.

scores over the majority of the metrics (Appendix F). The other types of VS models were effectively tied for most metrics (Appendix G).

3.1.2 Evaluation Metrics

Given a fixed evaluation metric, we could compare two models with a t-test to assess if one statistically outperforms the other. However, we needed to make such comparisons repeatedly between each pair of models and required a statistical test that accounts for multiple hypothesis

testing. Due to unequal variances and sample sizes (Figure 2), we used Dunnett's modified Tukey-Kramer test (DTK)[51,52] for pairwise comparison to assess whether the mean metric scores of two models were significantly different.

Using DTK results for *each metric*, we scored each model based on how many times it attained a statistically significantly better result than other models (Appendix G). For most metric-target pairs, many models have the same rank because DTK does not report a significant difference. Based on DTK alone, RF models consistently place in the top five ranks for each metric-target pair.

In a prospective screen, our goal is to maximize the number of active compounds identified by a VS algorithm given a fixed budget (number of predictions). We wanted to determine which of the evaluation metrics commonly used for VS best aligns with n_{hits}; i.e. maximizing the metric leads to maximizing n_{hits}. We assume that n_{hits} is ultimately the desired utility to maximize. Thus, we compared the model ranking induced by each metric with the model ranking induced by n_{hits} at varying number of tests.

To score the evaluation metrics, we used Spearman's rank correlation coefficient based on the model rankings induced by the metric of concern versus n_{hits} at a specific n_{tests}. We then ranked the metrics based on their correlation with n_{hits}. The correlation coefficients and metric rankings can be found in Appendix H. The metric ranking varies depending on n_{tests} and the target. Some metrics overtake one another as we increase or decrease n_{tests}. For **PriA-SSB AS**, NEF_R consistently placed in the top ranking correlations when R coincided with n_{tests}. This is evident when we focus on a single metric and see the top ranking metrics for $n_{tests} \in [100, 250, 500, 1000, 2500]$. Only for a large enough n_{tests} do metrics like AUC[ROC] that evaluate the complete ranked list become comparable. This suggests that if we know *a priori* how many new compounds we can afford to screen, then NEF_R at a suitable R is a viable metric for choosing a VS algorithm during cross-validation in the hopes of maximizing n_{hits}.

3.1.3 Selecting the Best Model

Based on these results, we selected the VS screening models that are most likely to generalize to new compounds and identify actives in our experimental screen of 25,279 new compounds. We focus on PriA-SSB for the prospective screen using models trained on **PriA-SSB AS** because the assay was more readily available for us to generate data for the new compounds.

Table 3 compares model selection based on evaluation metric means alone versus the DTK+Mean approach for multiple evaluation metrics on the three tasks. The complete model rankings for means only and DTK+Means can be found in Appendix I. DTK+Mean ranks models by statistical significance and uses the mean value only for tie-breaking. Both strategies selected the same models for a fixed evaluation metric, except for AUC[PR] on all three tasks and $NEF_{1\%}$ for **RMI-FANCM FP**. This is mainly due to DTK not detecting statistically significant differences among the models' evaluation scores, so tie-breaking by means selected the same models as ranking by means. Recall that **PriA-SSB FP** has fewer actives than **PriA-SSB AS** and **RMI-FANCM FP** (Table 1). Similar models from random forest and STNN-C were selected for **PriA-SSB AS** and **RMI-FANCM FP**. However, **PriA-SSB FP** identified STNN-R models exclusively.

Table 3: Top-ranked models by means versus DTK+Mean on the three tasks. Evaluation metric means were computed over all cross-validation folds. The prospective screening was only performed on PriA-SSB. Model names are mapped to their hyperparameter values in Appendix E.

	Best by Mean Model			**Best by DTK+Mean Model**		
Metric	**PriA-SSB AS**	**PriA-SSB FP**	**RMI-FANCM FP**	**PriA-SSB AS**	**PriA-SSB FP**	**RMI-FANCM FP**
AUC[ROC]	RF_d	STNN-R_a	RF_h	RF_d	STNN-R_a	RF_h
AUC[BEDROC]	RF_h	STNN-R_b	RF_h	RF_h	STNN-R_b	RF_h
AUC[PR]	RF_g	STNN-R_a	RF_h	STNN-C_b	STNN-R_b	STNN-C_b
AUC[NEF]	RF_h	STNN-R_b	RF_h	RF_h	STNN-R_b	RF_h
$NEF_{1\%}$	RF_h	STNN-R_b	RF_g	RF_h	STNN-R_b	RF_h

In our prospective screen, each model prioritizes 250 top-ranked compounds, approximately 1% of the new LC library. In this setting where each model has a fixed budget for the predicted compounds, NEF_R is a suitable metric. Therefore, we used $NEF_{1\%}$ with DTK+Means to choose the best models from each class. The best-in-class models were RandomForest_h, SingleClassification_a, SingleRegression_b, MultiClassification_b, LSTM_b, IRV_d, and ConsensusDocking_efr1_opt, with RandomForest_h being the strongest model overall (Appendix I).

3.2 Prospective Screening Results

After selecting the best model from each class based on cross-validation and the $NEF_{1\%}$ metric, we retrained the models on all 72,423 LC compounds to predict PriA-SSB inhibition using the same types of data shown in Table 2. This provided a single version of each model instead of one for each cross-validation fold. All models then ranked 25,279 new LC compounds that were provided blind, without activity labels. We selected the top 250 ranked new compounds from each model. Then, we experimentally screened all 25,279 new compounds to assess PriA-SSB % inhibition and defined actives based on a 35% inhibition and PAINS filters. The new binary dataset **PriA-SSB prospective** contained 62 actives.

Table 4 presents how many of the 62 actives were identified by each of the best-in-class virtual screening methods, along with the chemical structure similarity baseline. For comparison, randomly selecting 250 compounds from the **PriA-SSB prospective** dataset is expected to identify less than one active. Appendices J and K show the VS models' **PriA-SSB prospective** performance for the other evaluation metrics.

Table 4: The number of active compounds in the top 250 predictions from the seven selected models and the chemical similarity baseline compared to the number of experimentally-identified actives. These selected models are the best in each algorithm category from cross-validation. The last two columns correspond to the number of distinct chemical clusters from similarity or maximum common substructure clustering that are represented among the 62 actives. Complete hits on all models are in Appendix L.

Model	Actives	Actives not in baseline	SIM clusters	MCS clusters
Experimental	62	–	32	37
Similarity baseline	31	–	14	8
CD_efr1_opt	0	0	0	0
STNN-C_a	23	4	12	7
STNN-R_b	29	13	16	11
LSTM_b	1	1	1	1
MTNN-C_b	30	6	15	9
RF_h	40	10	16	9
IRV_d	29	5	13	7

Table 4 also lists the number of distinct chemical clusters identified by each method, with the goal of identifying as many diverse compounds as possible. The 62 experimental hits represent

32 SIM and 37 MCS clusters or chemical series. Commonly, virtual screening is followed by a medicinal chemistry effort that would be expected to identify other members of these clusters.

In general, the number of distinct chemical clusters captured in the top 250 predictions is correlated with the number of actives (Table 4), meaning that the methods selected structurally diverse hits. The best ligand-based approaches predicted at least one compound from half of the SIM clusters but a smaller fraction of the MCS clusters. RF found the most actives and the most distinct chemical clusters.

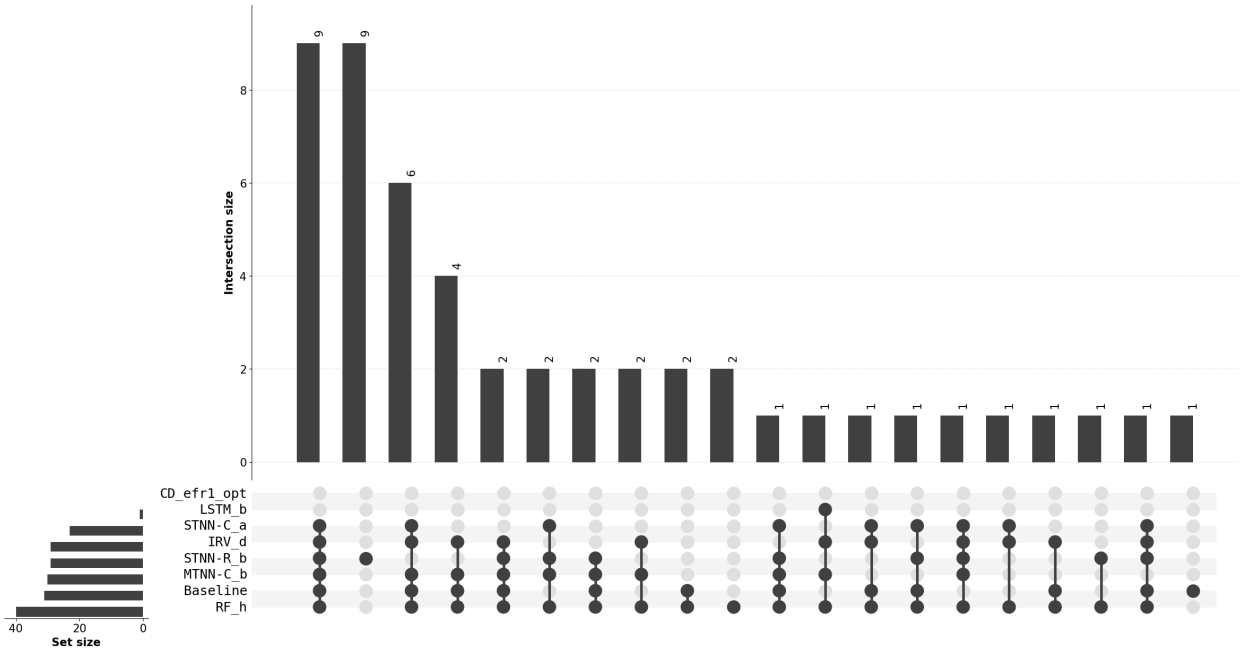


Figure 3: An UpSet plot showing the overlap between the top 250 predictions from the selected VS models and the chemical similarity baseline on **PriA-SSB prospective**. The plot generalizes a Venn diagram by indicating the overlapping sets with dots on the bottom and the size of the overlaps with the bar graph.[53] Altogether, the combined predictions from the best-in-class VS methods found 49 of the 62 actives.

On the **PriA-SSB prospective** dataset, only the RF model recovered more active compounds in its top 250 predictions than the chemical similarity baseline. Cross-validation with $NEF_{1\%}$ as the metric successfully identified the best PriA-SSB model before the prospective screen. The similarity baseline included one active compound that was excluded from the top 250 compounds from RF (Figure 3), but RF recovered a different member from this active compound's SIM and MCS clusters (Appendix M). The STNN-R and MTNN-C models identified fewer total active compounds than

the similarity baseline but more unique chemical clusters.

The ligand-based methods recovered many of the same actives as the chemical similarity baseline, but they also found actives that were missed by the baseline (Figure 3). There was a group of nine active compounds that were identified by most ligand-based methods, including the baseline model. The best model, RF, identified two unique actives that were not detected by any other methods. STNN-R was not the top performer overall but found nine unique actives and three unique chemical clusters (Appendix M).

3.2.1 Trained Models are Target-Specific

As a control, we assessed the performance of models trained on **RMI-FANCM FP** instead of **PriA-SSB AS** for making **PriA-SSB prospective** predictions on the new 25,279 compounds. As expected, the **RMI-FANCM FP** models perform poorly on **PriA-SSB prospective** (Appendix J), indicating that the best **PriA-SSB AS** models have learned compound properties that are specific to PriA-SSB.

4 Discussion

We followed a VS pipeline with the goal of maximizing the number of active compounds identified in a prospective screen with a limited number of predictions. From an initial pool of structure-based and ligand-based models, we pruned models in a hyperparameter search stage and conducted cross-validation with multiple evaluation metrics. We used DTK+Means with the $NEF_{1\%}$ metric to select the best models based on the cross-validation results and evaluated their top 250 prospective predictions from a new library of 25,279 compounds. The single best model from our pool, which was RandomForest_h for **PriA-SSB AS**, was also the top performing model on **PriA-SSB prospective**. Therefore, our overall pipeline successfully identified the best prospective model.

Metrics like AUC[ROC] can compare models in general, regardless of cost or other additional constraints.[46] However, for virtual screening in practice, one typically only experimentally tests

a small fraction of all compounds in the test library. In this setting, metrics like EF that capture early enrichment are preferable. In our prospective screen, STNN-R had higher AUC[ROC] than the random forest (Appendix J), but the random forest found 11 more active compounds in its top 250 predictions (Table 4). Our study suggests that EF_R, or its normalized version NEF_R, are the preferred metrics for identifying the best target-specific virtual screening method that maximizes n_{hits} when there is a budget for experimental testing. Other metrics like AUC[ROC] or AUC[PR], which is more appropriate for problems where the inactive compounds far outnumber the actives,[50] may still be reasonable for benchmarking virtual screening methods on large existing datasets where the entire ranked list of compounds is evaluated.[33]

Some recent studies[3,32,54] reported that deep learning models substantially outperform traditional supervised learning approaches, including random forests. Our finding that a random forest model was the most accurate in both cross-validation and our prospective screen does not refute those results. Rather, it reinforces that the ideal virtual screening method can depend on the training data available, target attributes, and other factors. Therefore, careful target-specific cross-validation is important to optimize prospective performance. One cannot assume that deep learning models will be dominant for all targets and all virtual screening scenarios. We also recommend hyperparameter exploration for all models, including traditional supervised learning methods. For example, our best random forest model contained 8000 estimators, whereas a previous benchmark considered at most 50 estimators.[3]

Ramsundar et al.[14] showed that performance improved in multi-task neural networks as they added more training compounds and tasks. Furthermore, the degree of improvement varied across the datasets and was moderately correlated with number of shared active compounds among the targets within a single dataset. Task-relatedness also affects the success of multi-task learning but is difficult to quantify.[55,56] We observed that **PriA-SSB AS**, **PriA-SSB FP**, and **RMI-FANCM FP** have no shared actives with any of the PCBA tasks, and multi-task neural networks were not substantially better than single-task neural networks in **PriA-SSB AS** cross-validation (Figure 2). The MTNN-C model outperformed the STNN-C model in the prospective evaluation (Table 4),

possibly because multi-task learning can help prevent overfitting,[57] but was still considerably worse than the random forest.

We focused on docking and well-established machine learning models instead of more recent deep learning models, such as graph-based neural networks.[33,58–61] This is because our main goal was to investigate the virtual screening principles for choosing the best model for a specific task (**PriA-SSB AS**) in a practical setting instead of broadly benchmarking virtual screening algorithms. In addition, a recent benchmark showed that conventional methods outperformed graph-based methods on most biophysics datasets.[33]

Consensus docking[44] failed to recover any actives in the **PriA-SSB prospective** dataset, even though some of the individual docking programs did. It should be noted that the individual docking and consensus methods do not make use of the initial HTS screening data like the machine learning methods do. A more direct comparison might be to re-train a custom consensus scoring function to include the initial HTS data, though this effort is out of scope for this study.

Specifically for the PriA-SSB protein-protein interaction, docking is also limited by the large, flat nature of the binding site. Many compounds that are inactive in the experimental screen have good scores and reasonable binding poses (per visual inspection) but fail to interrupt necessary specific interactions in the protein-protein interface. This will the limit overall performance by pushing true actives down the ranked list. Therefore, our docking results are not necessarily representative for other targets with other structural features, once again demonstrating the need for target-specific model selection during virtual screening.

The random forest model performed the best overall, but there were ten active compounds identified by the other methods that the random forest missed (Figure 3). The single-task regression neural network recovered nine of those ten and three unique active compound clusters (Appendix M). In addition, this regression model performed the best on **PriA-SSB FP** during cross-validation (Table 3), possibly because there are fewer binary actives in this dataset. In future work, we will explore whether ensembling classification and regression models, potentially in combination with structure-based VS algorithms, can further improve accuracy.

We emphasize our prospective performance on the new LC library, which minimizes the biases that make evaluation with retrospective benchmarks challenging.[62] There are many sources of experimental error in HTS, and the active compounds in the prospective evaluation must still be interpreted conservatively. However, a VS algorithm that can prioritize compounds with high % inhibition in primary and retest screens is valuable for further compound optimization even if not all of the actives confirm experimentally. Our study provides guidelines for selecting a target-specific VS model and complements other practical recommendations for VS pertaining to hit identification, validation, and filtering[63] as well as avoiding common pitfalls.[64] Having established that our best virtual screening model successfully prioritized new active compounds in the LC library, another future direction will be to test prospective performance on much larger, more diverse chemical libraries.

Acknowledgements

The authors acknowledge GPU hardware from NVIDIA; compute resources of the University of Wisconsin-Madison Center for High Throughput Computing; and funding from the Center for Predictive Computational Phenotyping NIH/NIAID U54 AI117924, the University of Wisconsin Carbone Cancer Center Support Grant NIH P30 CA014520, and the Morgridge Institute for Research. Additional support for this research was provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. We are grateful for the assistance and feedback from Gene E. Ananiev, Pengcheng Wang, Haozhen Wu, and many members of the Center for High Throughput Computing and the Gitter lab.

References

(1) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening

accuracy. *Journal of chemical information and modeling* **2009**, *49*, 1455–1474.

(2) Lionta, E.; Spyrou, G.; K Vassilatis, D.; Cournia, Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry* **2014**, *14*, 1923–1938.

(3) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Molecular Pharmaceutics* **2017**, *14*, 4462–4475.

(4) Tseng, Y. J.; Hopfinger, A. J.; Esposito, E. X. The great descriptor melting pot: mixing descriptors for the common good of QSAR models. *Journal of Computer-Aided Molecular Design* **2012**, *26*, 39–43.

(5) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *Journal of medicinal chemistry* **2007**, *50*, 74–82.

(6) Krüger, D. M.; Evers, A. Comparison of Structure-and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2010**, *5*, 148–158.

(7) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *Journal of chemical information and modeling* **2010**, *50*, 2079–2093.

(8) Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews. Computational Molecular Science* **2014**, *4*, 468–481.

(9) Merck, Merck Molecular Activity Challenge. `https://www.kaggle.com/c/MerckActivity` **2012**,

(10) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. *arXiv preprint arXiv:1406.1231* **2014**,

(11) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling* **2015**, *55*, 263–274.

(12) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science* **2016**, *3*, 80.

(13) Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep learning as an opportunity in virtual screening. *Advances in neural information processing systems* **2014**, *27*.

(14) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072* **2015**,

(15) Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* **2018**, *15*, 20170387.

(16) Voter, A. F.; Manthei, K. A.; Keck, J. L. A high-throughput screening strategy to identify protein-protein interaction inhibitors that block the Fanconi Anemia DNA repair pathway. *Journal of biomolecular screening* **2016**, *21*, 626–633.

(17) Voter, A. F.; Killoran, M. P.; Ananiev, G. E.; Wildman, S. A.; Hoffmann, F. M.; Keck, J. L. A High-Throughput Screening Strategy to Identify Inhibitors of SSB Protein–Protein Interactions in an Academic Screening Facility. *SLAS DISCOVERY: Advancing Life Sciences R&D* **2017**, 94–101.

(18) Nordmann, P.; Cuzon, G.; Naas, T. The real threat of Klebsiella pneumoniae carbapenemase-producing bacteria. *The Lancet infectious diseases* **2009**, *9*, 228–236.

(19) Manthei, K. A.; Keck, J. L. The BLM dissolvasome in DNA replication and repair. *Cellular and molecular life sciences* **2013**, *70*, 4067–4084.

(20) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry* **2010**, *53*, 2719–2740.

(21) Lagorce, D.; Sperandio, O.; Baell, J. B.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs3: a web server for compound property calculation and chemical library design. *Nucleic acids research* **2015**, *43*, W200–W207.

(22) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Research* **2017**, *45*, D955–D963.

(23) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

(24) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* **1989**, *29*, 97–101.

(25) RDKit: Open-source cheminformatics. `http://www.rdkit.org`

(26) Chollet, F. Keras. `https://github.com/fchollet/keras` **2015**,

(27) The Theano Development Team, et al. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* **2016**,

(28) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.

(29) Jastrzębski, S.; Leśniak, D.; Czarnecki, W. M. Learning to SMILE (s). *arXiv preprint arXiv:1602.06289* **2016**,

(30) Swamidass, S. J.; Azencott, C.-A.; Lin, T.-W.; Gramajo, H.; Tsai, S.-C.; Baldi, P. Influence relevance voting: an accurate and interpretable virtual high throughput screening method. *Journal of chemical information and modeling* **2009**, *49*, 756–766.

(31) Lusci, A.; Fooshee, D.; Browning, M.; Swamidass, J.; Baldi, P. Accurate and efficient target prediction using a potency-sensitive influence-relevance voter. *Journal of cheminformatics* **2015**, *7*, 63.

(32) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is multitask deep learning practical for pharma? *Journal of chemical information and modeling* **2017**, *57*, 2068–2076.

(33) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**,

(34) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(35) Bhattacharyya, B.; George, N. P.; Thurmes, T. M.; Zhou, R.; Jani, N.; Wessel, S. R.; Sandler, S. J.; Ha, T.; Keck, J. L. Structural mechanisms of PriA-mediated DNA replication restart. *Proceedings of the National Academy of Sciences* **2014**, *111*, 1373–1378.

(36) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2010**, *50*, 572–584.

(37) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **2009**, *30*, 2785–2791.

(38) Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry* **2015**, *36*, 1132–1156.

(39) McGann, M. FRED and HYBRID docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design* **2012**, *26*, 897–906.

(40) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *Journal of Chemical Information and Modeling* **2009**, *49*, 84–96.

(41) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLOS Computational Biology* **2014**, *10*, e1003571.

(42) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling* **2013**, *53*, 1893–1904.

(43) Cleves, A. E.; Jain, A. N. Knowledge-guided docking: accurate prospective prediction of bound configurations of novel ligands using Surflex-Dock. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 485–509.

(44) Ericksen, S. S.; Wu, H.; Zhang, H.; Michael, L. A.; Newton, M. A.; Hoffmann, F. M.; Wildman, S. A. Machine learning consensus scoring improves performance across targets in structure-based virtual screening. *Journal of Chemical Information and Modeling* **2017**, *57*, 1579–1590.

(45) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry* **2012**, *55*, 6582–6594.

(46) Nicholls, A. What do we know and when do we know it? *Journal of Computer-Aided Molecular Design* **2008**, *22*, 239–255.

(47) Truchon, J.-F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling* **2007**, *47*, 488–508.

(48) Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* **2010**, *26*, 1348–1356.

(49) Grau J, K. J., Grosse I PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **2015**, *31*, 2595–2597.

(50) Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. 2006; pp 233–240.

(51) Lau, M. DTK: Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test adjusted for unequal variances and unequal sample sizes. *R package* **2013**, *3*.

(52) Dunnett, C. W. Pairwise Multiple Comparisons in the Unequal Variance Case. *Journal of the American Statistical Association* **1980**, *75*, 796–800.

(53) Lex, A.; Gehlenborg, N.; Strobelt, H.; Vuillemot, R.; Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* **2014**, *20*, 1983–1992.

(54) Lenselink, E. B.; Dijke, N. t.; Bongers, B.; Papadatos, G.; Vlijmen, H. W. T. v.; Kowalczyk, W.; IJzerman, A. P.; Westen, G. J. P. v. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics* **2017**, *9*, 45.

(55) Kearnes, S.; Goldman, B.; Pande, V. Modeling industrial ADMET data with multitask networks. *arXiv preprint arXiv:1606.08793* **2016**,

(56) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Central Science* **2017**, *3*, 283–293.

(57) Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv preprint arXiv:1706.05098* **2017**, arXiv: 1706.05098.

(58) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. Advances in neural information processing systems. 2015; pp 2224–2232.

(59) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 595–608.

(60) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling* **2017**, *57*, 1757–1772.

(61) Matlock, M. K.; Dang, N. L.; Swamidass, S. J. Learning a Local-Variable Model of Aromatic and Conjugated Systems. *ACS Central Science* **2018**, *4*, 52–62.

(62) Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *Journal of Chemical Information and Modeling* **2018**,

(63) Zhu, T.; Cao, S.; Su, P.-C.; Patel, R.; Shah, D.; Chokshi, H. B.; Szukala, R.; Johnson, M. E.; Hevener, K. E. Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based on a Critical Literature Analysis. *Journal of Medicinal Chemistry* **2013**, *56*, 6560–6572.

(64) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *Journal of Chemical Information and Modeling* **2012**, *52*, 867–881.