Efficient Exploration of Chemical Space with Docking and Deep-Learning

*Ying Yang[1], Kun Yao[2], Matthew Repasky[3], Karl Leswing[2], Robert Abel[2], Brian Shoichet[1], Steven V. Jerome[4]**

1. Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California 94158, United States

2. Schrödinger Inc., 120 West 45th Street, 17th floor, New York, New York 10036, United States

3. Schrödinger Inc., 101 SW Main Street, #1300, Portland, Oregon, 97239, United States

4. Schrödinger Inc., 10201 Wateridge Cir Suite 220, San Diego, California, 92121, United States

*Corresponding author: Steven Jerome: steven.jerome@schrodinger.com;

ABSTRACT

With the advent of make-on-demand commercial libraries, the number of purchasable compounds available for virtual screening and assay has grown explosively in recent years, with several libraries eclipsing one billion compounds. Today's screening libraries are larger and more diverse, enabling discovery of more potent hit compounds and unlocking new areas of chemical space, represented by new core scaffolds. Applying physics-based in-silico screening methods in an exhaustive manner, where every molecule in the library must be enumerated and evaluated independently, is increasingly cost-prohibitive. Here, we introduce a protocol for machine learning-enhanced molecular docking based on active learning to dramatically increase throughput over traditional docking. We leverage a novel selection protocol that strikes a balance between two objectives: (1) Identifying the best scoring compounds and (2) exploring a large region of chemical space, demonstrating superior performance compared to a purely greedy approach. Together with automated redocking of the top compounds, this method captures nearly all the high scoring scaffolds in the library found by exhaustive docking. This protocol is applied to our recent virtual screening campaigns against the D4 and AMPC targets that produced dozens of highly potent, novel inhibitors, and a blinded test against the MT1 target. Our protocol recovers more than 80% of the experimentally confirmed hits with a 14-fold reduction in compute cost, and more than 90% of the hit scaffolds in the top 5% of model predictions, preserving the diversity of the experimentally confirmed hit compounds.

INTRODUCTION

The number of organic compounds containing 30 heavy atoms has been estimated to exceed 10^{60} molecules[1]. According to the most recent release of GDB-17, a partial enumeration of possible compounds containing up to 17 heavy atoms and the elements C, N, O, and S yields 166 billion molecules[2]. Until recently, the number of these compounds readily available to biology and chemistry has been far less than these staggering numbers, perhaps around 10 million molecules. Very recently, however, the number of commercially available molecules has grown substantially, owing to the advent of make-on-demand libraries. For instance, the ZINC database (http://zinc15.docking.org)[3], now exceeds one billion readily synthesizable molecules from commercial vendors, with 448 million in the lead-like range (Date of Access: 10/25/2020). In addition, several commercial vendors offer DNA-encoded libraries comprising billions of compounds for laboratory screening[4, 5].

In recent studies[6-9], structure-based, *in-silico* virtual screening has achieved high hit-rates by leveraging the new ultra-large chemical libraries that have emerged from the make-on-demand chemistry. Here, we use *ultra-large* screening libraries to mean those containing more than 100M distinct compounds, distinguishing them from traditional, in-stock libraries, which may typically contain about 3 to 10 million compounds. The early proof-of-concept studies suggest two benefits of docking these ultra-large libraries. First, their size and high diversity suggests that, if only by chance, ultra-large libraries will fortuitously sample more ligands that can tightly bind to a particular target. Second, the early studies support the idea that, for all their liabilities, empirical docking scoring functions, such as DOCK3.7[10-12] or Glide SP[13-15] and others[16-19] can prioritize these better ligands from the sea of decoy molecules inevitably also sampled in the ultra-large libraries, preventing the true actives from being drowned out by false positives.

Docking campaigns utilizing ultra-large libraries remain relatively uncommon in the literature, reflecting the substantial computational cost required. Recently, two teams reported structure-based virtual screens of over one billion compounds. The first such screen was performed by OpenEye Scientific on their Orion cloud computing platform, leveraging vast quantities of CPUs to explicitly predict the pose and score of each compound[20] (detailed experimental outcomes were not provided). A second billion compound screen was performed with VirtualFlow, a workflow that relies on cheaper, arguably less accurate methods to triage compounds[8]. Docking an ultra-large library of one billion compounds on 1000 CPUs would take approximately eleven days with DOCK3.7 (at 1 sec/lig), 300 days with Glide SP (at 30 sec/lig), and 173 days with Virtual Flow (at 15 sec/lig). This does not account for the cost to prepare the library for screening, which typically includes the generation of low energy ionization and tautomeric states as well as stereochemical states for unspecified centers. It is expected that on-demand, synthesized libraries will continue to rapidly grow as further reagents and reactions are included. For example, the ZINC library comprised roughly 100 Million compounds in April 2016, 500 million in July 2019 and 1 billion in January 2020. If we extrapolate this growth rate, it's not unreasonable to anticipate libraries of tens of billions of compounds in the next five years.

As chemical libraries continue to grow, explicit docking solutions are expected to become impractical due to cost and compute resource requirements. Hierarchical workflows are similarly unattractive as they compromise scoring function accuracy for the method that screens all compounds in order to achieve acceptable throughput. Accordingly, there is a need for docking workflows that can find, with minimal loss of accuracy, the best scoring ligands in multi-billion molecule libraries that exceed our capacities to screen or perhaps even build them explicitly, through combinatorial enumeration.

Advances in machine learning, especially deep learning, have provided great opportunities in drug discovery[21-24]. Studies range from the prediction of compound properties, including on-target activity, de novo design of chemical structures to target-specific property spaces, reaction predictions and retro-synthetic analysis, and the prediction of protein-ligand interactions via convolutional neural networks. Active-Learning is a category of supervised machine learning methods that train an increasingly accurate model to act as a stand-in for a difficult or costly to compute scoring function. Many applications of active learning in the area of drug discovery have been reported recently, including applications to molecular docking[25, 26] and free energy calculations[27, 28]. These studies have demonstrated the ability for ligand-based QSAR models to "learn" a docking score over a specific domain, greatly reducing the computational cost to screen a large library. What has been missing to date, however, are comparisons of hit rates and chemical diversity between traditional docking programs and active learning protocols using data from real hit-discovery campaigns, which is the subject of this work. To our knowledge, this study presents, for the first time, a comparison of different strategies for selecting compounds for active learning (selection rule) in a virtual screening context. We introduce a novel selection rule defined by choosing the most uncertain of the top scoring compounds, striking a balance in the classic explore-exploit tradeoff, and demonstrate that it outperforms a selection rule defined by either the best scoring molecules or the most uncertain compounds.

Here we present a retrospective study examining the feasibility of an active learning-based approach to focus docking screens on the most productive areas of chemical space. A small percentage of docking results of D4 dopamine receptor (D4) and AmpC β-lactamase (AmpC) from a previous publication[6] were used to train AutoQSAR/DeepChem (AQ/DC) models, followed by a prediction of DOCK3.7 scores on the entire library. A subsequent blind test was carried out on

MT1 melatonin receptor (MT1)[7], where the docking results were not accessible to those training the ML method. In addition to prioritizing new chemical space for DOCK3.7 to focus on, we demonstrate the versatility of the active learning approach by extending it to the Glide SP docking program using the fully-automated Active Learning Glide Program available in the Schrödinger Suite[29].

MATERIALS AND METHODS

Datasets

Three protein systems were used in this retrospective study. Docking results of 99,459,562 molecules to AmpC β-lactamase (AmpC), and 138,312,677 molecules to the D4 dopamine receptor (D4) were obtained from ref 6. A blind test was carried out on 150,927,915 molecules docked against the MT_1 melatonin receptor (MT1)[7], where only randomly selected training sets were provided for the training of the model. These systems were chosen because hundreds of docking predictions have been experimentally tested, thus experimentally confirmed actives can be used to evaluate the method. It is worth mentioning that AmpC preferentially binds anionic molecules, while D4 and MT1 favor cationic and neutral molecules, respectively.

AutoQSAR/DeepChem (AQ/DC)

Here, we evaluate our implementation of active learning for molecular docking using both the DOCK3.7 and Glide SP programs. All machine-learning models in this work were developed using the AutoQSAR/DeepChem package available in Schrödinger Suite beginning with the 2019-1 release. AQ/DC is an extension of our previously reported AutoQSAR[30] engine for building best-practices QSAR models without the need for specialized knowledge in machine learning or cheminformatics.

AutoQSAR, as originally implemented, is limited to modest size data sets (fewer than 5,000 data-points), making it unsuitable for the active-learning approach presented in this work. AQ/DC implements an automated approach to neural-networks, by implementing the Graph-Convolutional Neural Networks (GCNN) available in the DeepChem package of Pande et. al[31]. The GCNN model has been described as a sort of neural fingerprint that operates on a molecular graph. In addition to GCNN neural networks, AQ/DC also constructs random forest models based on Morgan Fingerprints from the RDKit[32] cheminformatics package.

In AQ/DC, a user-adjustable parameter, the model search time, is used to define the length of model search and selection process. AQ/DC performs a random search of hyper parameters during this time. Examples of the hyperparameters sampled during training include the model algorithm, number of layers in the GCNN or DNN, the number of training epochs, and the normalization scheme applied to the raw response variable. AQ/DC performs 5-fold cross validation scheme to select hyper parameters. The output of this procedure is 5 models, each trained to a different fold of the data, for a given hyperparameter set. When the search time has been reached, this process terminates, and the best performing model sets are combined into a three-membered ensemble, where each member is trained on one of five folds, comprising of a total of 15 models. Compound evaluation using the trained ensemble is performed by averaging the scores of all 15 models. Relevant for this work, a standard deviation across the 15-member ensemble is provided as a measure of uncertainty.

Docking Calculations

Molecular docking using Glide and DOCK3.7 require prepared 3D structures for all input molecules. 3D inputs for Glide calculations were generated using the LigPrep program[33] of Schrodinger Suite starting from a SMILES representation of each molecule, using the following

settings: (1) Tautomers were generated using Schrodinger's Epik[34] with a target pH of 7.0+- 1.0 and (2) a maximum of 16 stereoisomers were generated for each molecule. Glide was run with SP precision using all default docking parameters. 3D inputs for DOCK3.7 where generated with the ligand building pipeline in ZINC15[35]. In brief, ChemAxon's CXCALC program was used to generate protonation and tautomer states at physiologically relevant pH from a SMILES representation of each compound[36]. Protomers were then converted to 3D via CORINA[37] and conformational ensembles were generated using OpenEye's Omega[38]. The details of docking settings of DOCK3.7 for AmpC and D4 have been described previously[6].

General Workflow

The general workflow of the hybrid approach is depicted in **Figure 1**. First, a random subset of the ligand library is selected and docked. The docking scores of the subset are used to train an AQ/DC model, which is subsequently used to predict docking scores across the entire ligand database. This is followed by one or multiple rounds of active learning, where an attempt is made to improve the model by making a selection of the top scoring compounds, according to a selection rule, and training a new model using the union of the previously docked compounds and this additional batch of docked ligands. For these retrospective experiments, the performance of AQ/DC is evaluated by the percentage of top ranked, explicit docking-prioritized molecules recalled by the AQ/DC model. Here, we define the top 10K compounds by docking score in the explicitly docked library as *virtual hit* compounds. Different models can be compared by the Receiver-Operator Characteristics (ROC) metric where docking hits are the true positives.

Here, all AQ/DC model training and prediction were done in triplicate, where training sets were selected with different random seeds, i.e. Run 1, Run 2, Run 3. As shown in **Figure S1**, the overlap among molecules in the three randomly selected training sets is minimal.

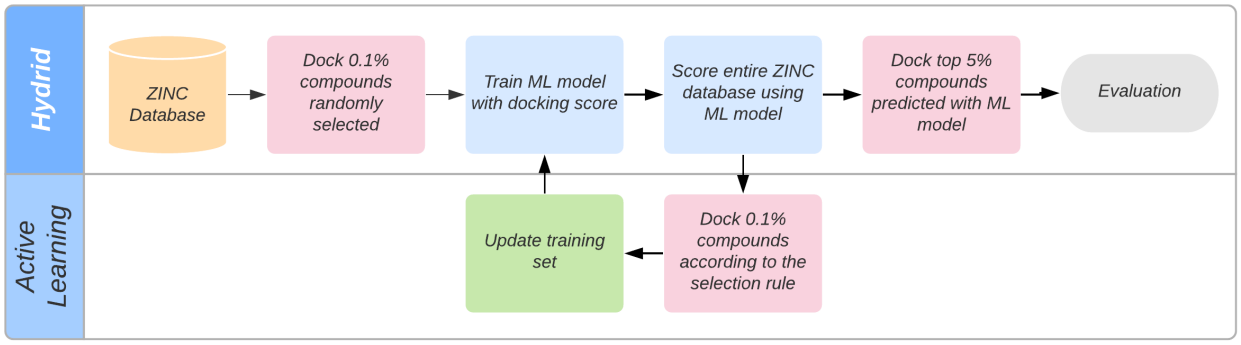


Figure 1. General workflow of the presented approach in the study with the option of active learning.

Chemical Libraries

Libraries for virtual screening can vary greatly in both size and chemical diversity. We have evaluated the active-learning protocol against different libraries to provide coverage along both of these dimensions. We evaluated ~100M subsets of the ZINC15 library of commercially available compounds and the Sigma-Aldrich Market-Select collection (~8M compounds) [39]. While smaller in size, every product in the Sigma-Aldrich type of library should be in-stock. This compares to ZINC15, where only about 0.5% of the database is expected to be in-stock. The bulk of ZINC15, like other virtual libraries of greater than 100 million compounds, is composed of molecules generated by a computational enumeration of building blocks and a reaction database. Of the one billion compounds in ZINC15, roughly 11 million are available in commercial in-stock collections, while the rest are expected to be readily synthesizable. The combinatorial design of these libraries makes them well suited to the protocol described here, where the higher expected density around virtual hit compounds in chemical space could allow for more robust training of the ML models in the active-learning workflow as compared to in-stock libraries, where compounds are expected to be more sparsely distributed in chemical space.

Null Model based on 2D Chemical Similarity

A simple null model based on 2D topological similarity was constructed to assess the effectiveness of the active learning workflow for recovering virtual hit compounds. As with the workflow described in **Fig. 1**, an initial random selection of the ligand database is docked. The top compounds by docking score are selected as probe molecules. Pairwise similarities are computed for each compound in the database against each of the probe molecules. The final score for each compound according to this model is defined as the maximum similarity over the set of probe molecules. Considering the size of the library (~100 million compounds), this analysis requires roughly 100 billion pairwise fingerprint similarity calculations against 1000 probe ligands.

These vast calculations were made possible by the GPU-accelerated FPsimGPU application available in Schrödinger Suite. FPsimGPU consists of a cloud server populated with molecular fingerprints for each compound in the library. To realize the GPU compute performance for similarity operations, the fingerprints of each library compound are stored in GPU memory on the server. Molecular fingerprints are "folded-down" to maximize information density and ensure that the entire database can fit in memory. This process can be described as follows: For each cycle of the folding process the fingerprint size is reduced by a factor of two, lowering the sparsity of the fingerprint, but increasing the rate of expected collisions, where a single bit is mapped to multiple chemical substructures[40,41]. After comparing all folded fingerprints on the GPU, the highest scoring matches are rescored on the CPU using the original unfolded fingerprints to get fully accurate results

The server can perform more than 4 billion similarity operations per second. We use Morgan Fingerprint with Radius=2 as implemented in the RDKit cheminformatics package.

RESULTS AND DISCUSSION

The AQ/DC model was initially trained against a set of 0.1% molecules randomly selected from the full dataset for AmpC and D4 systems, respectively. The model was then applied to the entire dataset, where an ML score was predicted for each molecule for comparison with the true docking score.

The overall correlation between the docking score and the ML score is shown in Figure S2, where the Pearson correlation coefficients are 0.78 and 0.81 for DOCK3.7 and Glide SP, respectively, for the D4 receptor and 0.81 and 0.72 for DOCK3.7 and Glide SP, respectively, for AmpC. When focusing only on the region with the best true docking scores, the correlation coefficient drops substantially; the ML models are clearly an imperfect stand-in for the true docking scoring function. From the correlation statistic, it is not immediately clear how useful the ML models are in a high-throughput screening context. We can reframe the analysis by asking how many of the top scoring compounds from the ML model would need to be docked to recover a specified number of the virtual hit compounds. The new Active-Learning Glide program implements an automated rescoring step where a fraction of the top-ranked compounds by the ML predictions are rescored with explicit docking for this purpose. What is really needed then, for the models to be useful in reducing the computational cost of the screen, is simply to be sufficiently effective in separating compounds that will yield good docking scores from those with poor docking scores, so that the desired recovery can be achieved while docking a far smaller number of compounds in total. Here, we focus on the recovery of virtual hit compounds as a function of the library screened by the ML model. This quantity represents the number of compounds that must be rescored in order to recover a given number of virtual hit compounds.

Recovery of virtual hits and experiment confirmed actives from AQ/DC model predicted rank list

The plots in **Figure 2** present the recovery of the (1) top 10K virtual hits from DOCK3.7 for D4 and AmpC (**Figure 2A & 2B**); (2) the experimentally verified hit compounds (actives) (**Figure 2C & 2D**); and (3) the top 10K virtual hits from Glide SP (**Figure 2E & 2F**).

In sharp contrast with the prediction of docking scores for top-ranking compounds, the classification of top-ranking compounds by the models was strong. Most top-ranking compounds by both DOCK3.7 and Glide SP can be recovered very early by the AQ/DC model. Docking the top 5% of the library from the AQ/DC model, 80% and 98% of the virtual hits from DOCK3.7 and Glide SP can be recovered for the D4 screen. For AmpC, 97% and 70% of the virtual hits from DOCK3.7 and Glide can be recovered. Recovery performance using DOCK3.7 is noticeably worse for D4 than AmpC, particularly with respect to early enrichment. The opposite behavior is seen for models based on Glide SP, where recovery performance is improved, relative to AmpC. For AmpC, the docking parameters for DOCK3.7 were adjusted based on benchmark calculations [6]. These target-specific changes largely impacted the treatment of electrostatic interactions in the binding site, promoting compounds with plausible warheads. Glide SP calculations were run without any adjustment from the default parameters, which may account for the observed difference.

Recovery of experimentally confirmed actives by the AQ/DC model trained to the original DOCK3.7 scores is likewise very good. These plots suggest that the ML models are able to separate compounds with good docking scores from those with poor docking scores.

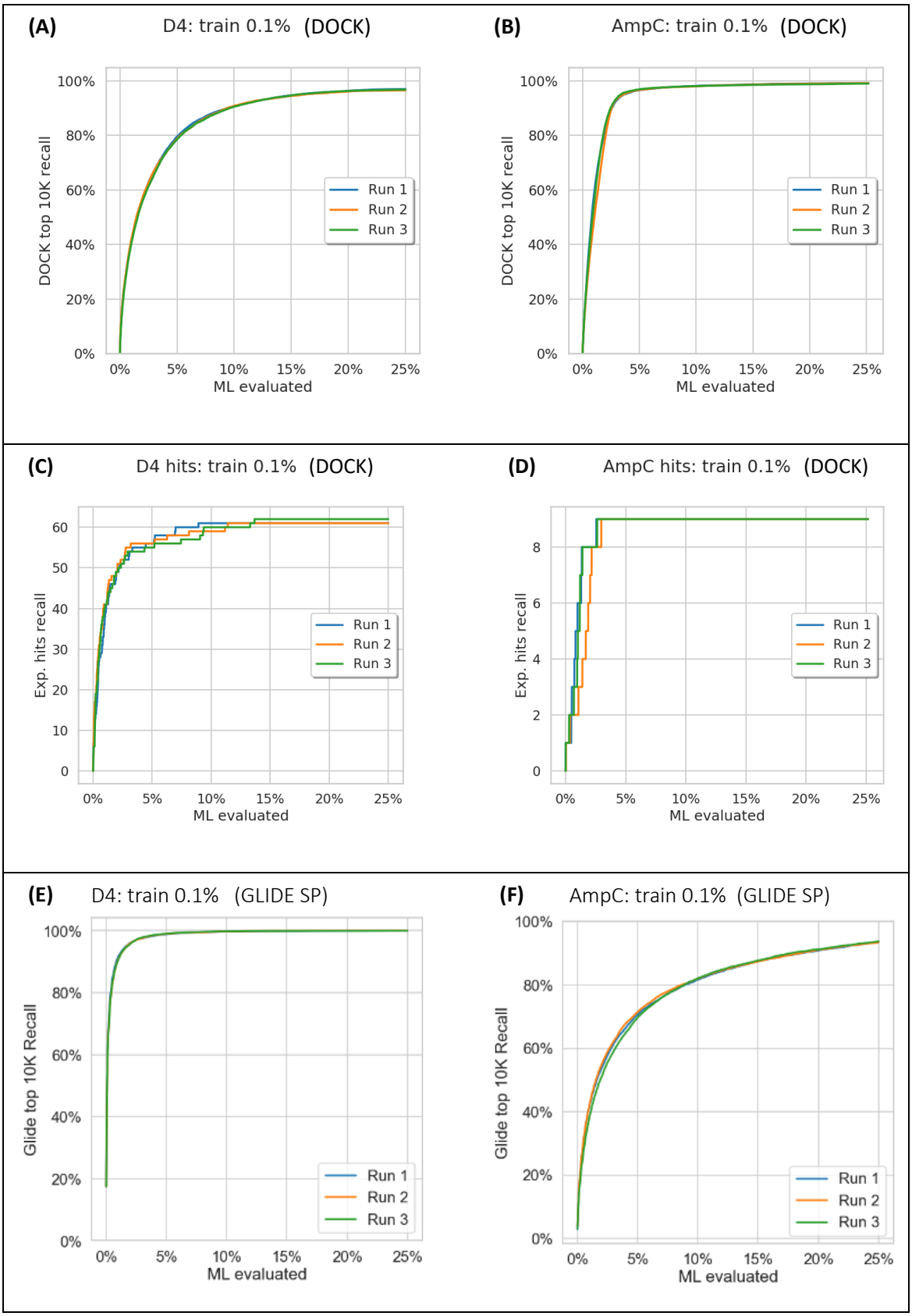

(A)
D4: train 0.1% (DOCK)
DOCK top 10K recall
ML evaluated
Run 1
Run 2
Run 3
(B)
AmpC: train 0.1% (DOCK)
DOCK top 10K recall
ML evaluated
Run 1
Run 2
Run 3
(C)
D4 hits: train 0.1% (DOCK)
Exp. hits recall
ML evaluated
Run 1
Run 2
Run 3
(D)
AmpC hits: train 0.1% (DOCK)
Exp. hits recall
ML evaluated
Run 1
Run 2
Run 3
(E)
D4: train 0.1% (GLIDE SP)
Glide top 10K Recall
ML evaluated
Run 1
Run 2
Run 3
(F)
AmpC: train 0.1% (GLIDE SP)
Glide top 10K Recall
ML evaluated
Run 1
Run 2
Run 3

Figure 2. ROC curves showing the percent of top 10K virtual hits and experiment hits recovered by the AQ/DC model. The true positives are the top 10K virtual hits from DOCK for D4 (A), AmpC (B), experiment confirmed actives for D4 (C), AmpC (D), and top 10K virtual hits from Glide SP for D4 (E), AmpC (F) respectively. Three replicates of random selection of training set are indicated by Run 1, Run 2, and Run 3, where no significant difference is observed.

We considered two routes to improve the performance of our model while maintaining a reasonable training time. First, we increased the size of the training data set, from 0.1% to 0.2% and 0.5% of ZINC screening subset. For the workflow presented in this work, a larger data set should always be beneficial for model accuracy, eventually reaching a point of diminishing returns. The other route is to increase the information content of the training set to gain higher accuracy. In the following study, we show the impact of 1) the size of the training set of docked compounds, 2) the selection rule used in the active-learning scheme to select the compounds for model training.

Role of Training Set Size

We varied the training set size for the AQ/DC model with set sizes of 0.01% to 0.02%, 0.05%, 0.1%, 0.2%, and 0.5%. As the size of the training set is increased, recovery of virtual hit compounds is improved (**Table S1**). For D4, evaluating the top 2% scoring compounds according to the AQ/DC model, the percent recall of DOCK3.7 virtual hits significantly improved from 23% to 67%; similarity, when evaluating the top 5% compounds by ML-score, the percent recall increased from 47% to 86%. Although great early recovery of virtual hits was seen for AmpC, improvement in percent recall of virtual hits was still observed.

Selection rules used in active learning

Next, we consider the impact of the selection rule used to select the training set. In the AL workflow of the present work, an ML model is first built using a small training set (0.1% of the ligand database). After the prediction, another 0.1% of molecules are selected and docked. The docking scores of these compounds are combined with the initial selection to form a new training set used to retrain the AQ/DC model. Four types of selection rules have been tested here: 1) top 0.1% by ML score; 2) most uncertain 0.1% according to the ML model, as defined by the standard deviation of the ensemble predictions for a given compound; 3) 0.1% randomly selected compounds from top 10% by ML score; and 4) most uncertain 0.1% from the top 5% compounds by ML score. The detailed recall for each selection rule is reported in **Table S1 and Figure S3**. The addition of the top 0.1% of compounds by ML score leads to almost no improvement in the percent recall of virtual hit compounds as compared to the initial ML model trained to 0.1% of the library, while the addition of the most uncertain 0.1% compounds according to the AQ/DC model actually decreases recall. This result suggests that the compounds selected for retraining the model should be high-scoring according to original AQ/DC model. Impressively, when the selection rule is defined by random selection of 0.1% molecules from the top 10% scoring molecules according to the ML model, performance is similar to training an AQ/ML model in to 0.2% of the dataset, in a single pass. Finally, using a selection rule defined by the 0.1% most uncertain molecules from the top 5% according to ML yields the best recovery of virtual hits. This leads to an important conclusion from our experiments. Selecting the most uncertain compounds for the training set is only effective when they are also among the highest-scoring compounds according to the initial ML model. This powerful combination of score and uncertainty provides a recipe for the most effective way to choose the compounds for active learning.

Figure 3 contains bar plots comparing recovery performance across three different models: (1) AQ/DC model trained with a random 0.1% subset of the database (blue; train 0.1%), (2) with a larger 0.2% training set (orange; train 0.2%), (3) AL protocol with 0.1% most uncertain of top 5% from ML prediction (green; active learn 0.1% + 0.1%), (4) even larger 0.5% training set (red; train 0.5%). ROC curves of the detail recovery are shown in **Figure S4** and **S5** for D4 and AmpC, respectively.

The effect of increasing the size of training set is most pronounced early in the list. In **Figure 3A and 3B**, for example, the AQ/DC model built using a smaller training set (0.1% of the database) retrieves 50% of the top 10K compounds, according to DOCK, in the top 2% of compounds scored by the ML model. Recovery using a model build to a larger training set (0.5% of the database) jumps to more than 60%. The use of a larger training set improves recovery of active compounds as well (**Figure 3C** and **3D**). **Figure 3E** and **3F** show the recovery of Glide SP virtual hits, for AmpC, training a ML model to a larger fraction of the ligand database made a large improvement in the recovery power of the model. For D4, the improvement is subtle since the recovery is already 96% from the model trained using a smaller training set (0.1% of the ligand database).

The performance of the AL protocol is significantly improved from the AQ/DC model trained with a smaller training set (0.1% of the ligand database) and is within error of the performance achieved with a larger training set (0.5% of the ligand database). The total number of docking calculations required to for the AL protocol is 0.2% of the ligand database (e.g. 200,000 ligands with 100 million compound database). This compares to the model built using the larger training set, which required 0.5% of the ligand database (e.g. 500,000 ligands). This is a powerful illustration of the impact of active learning. Thus, with an active learning refinement approach, we

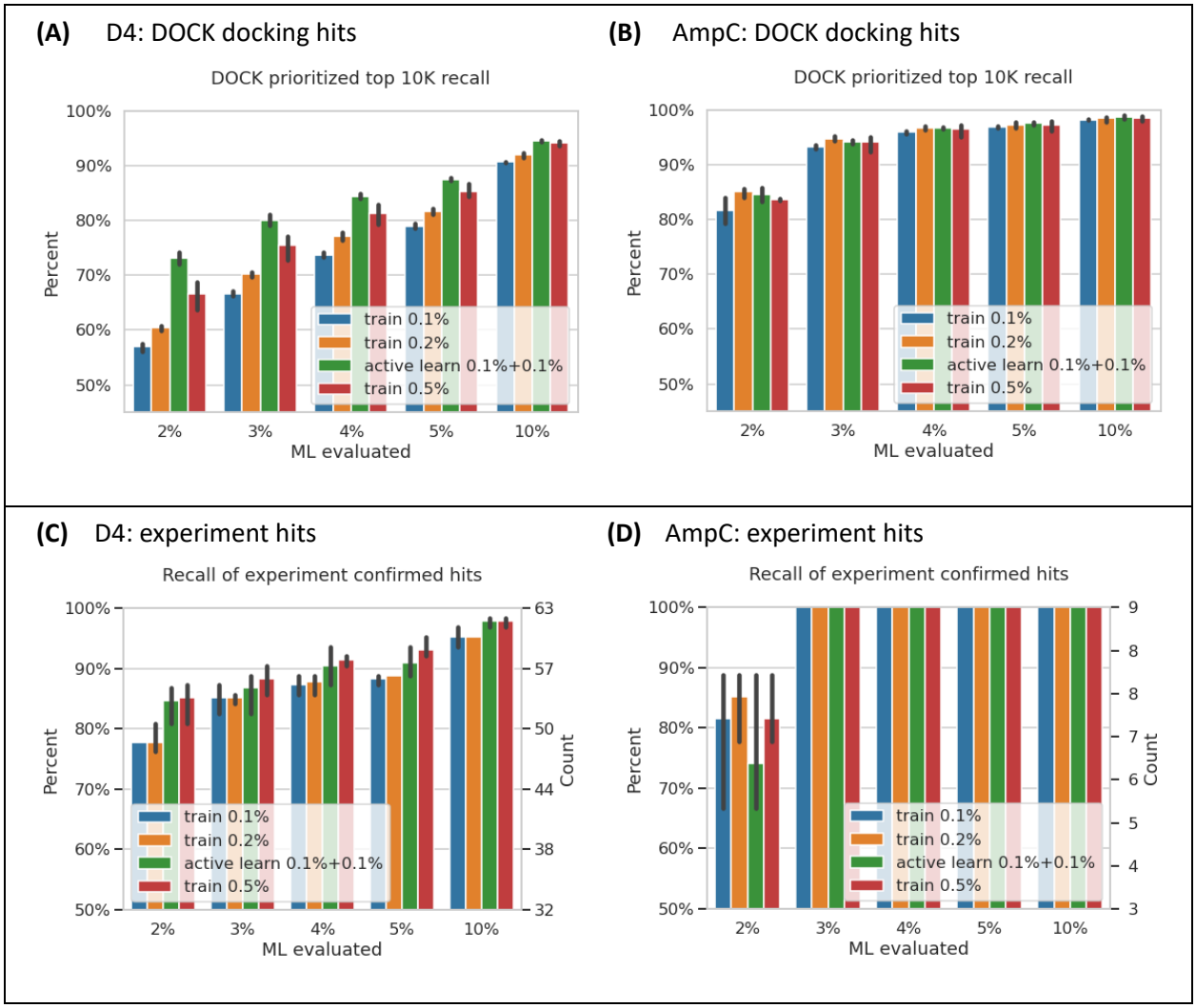
can save nearly 300,000 docking calculations. These conclusions are independent of whether DOCK or Glide SP is used.

In all cases, as expected, both having a larger data set (train 0.5%) and using active-learning to refine the model (active learn 0.1% + 0.1%) outperforms a model trained to a smaller training set. The gain in performance is especially significant in the early part of the screen.

The comparison between a model trained to 0.5% ZINC database and the active learning protocol, however, is a bit more complex. When benchmarking against DOCK, active learning shows a consistent, although small, increase in performance as compared to a model trained with a larger data set. However, this improvement is eliminated when benchmarking against experimental data as shown in **Fig 3C**. This interesting observation can be understood as follows. Active learning is built upon the existing ML model trained using DOCK, and its performance is therefore closely associated with the original model trained upon 0.1% of the ZINC database. In our previous study, the compounds selected for experimental testing were not simply the best by DOCK score. While the DOCK score was used as a guide to select the compounds, human voting was ultimately used to form the buy list from a pool of high-scoring compounds. Nonetheless, the protocol described here achieves performance on-par with standard approach that required 2.5 times more training data. This allows more rapid screening of molecules while maintaining good accuracy.

In practice, the top 10k compounds in a docking campaign are sufficient in most cases to form a voting pool for human interrogation. We have, however, in previous work, pulled virtual hits from further down the ranked list, including the top 100K and top 300K by DOCK score. This was done to include additional compounds for post filtering procedures and hit picking[6,7]. The recovery

of top 100K and 300K virtual hits by AQ/DC model ML prediction have been shown in **Figure S7**.



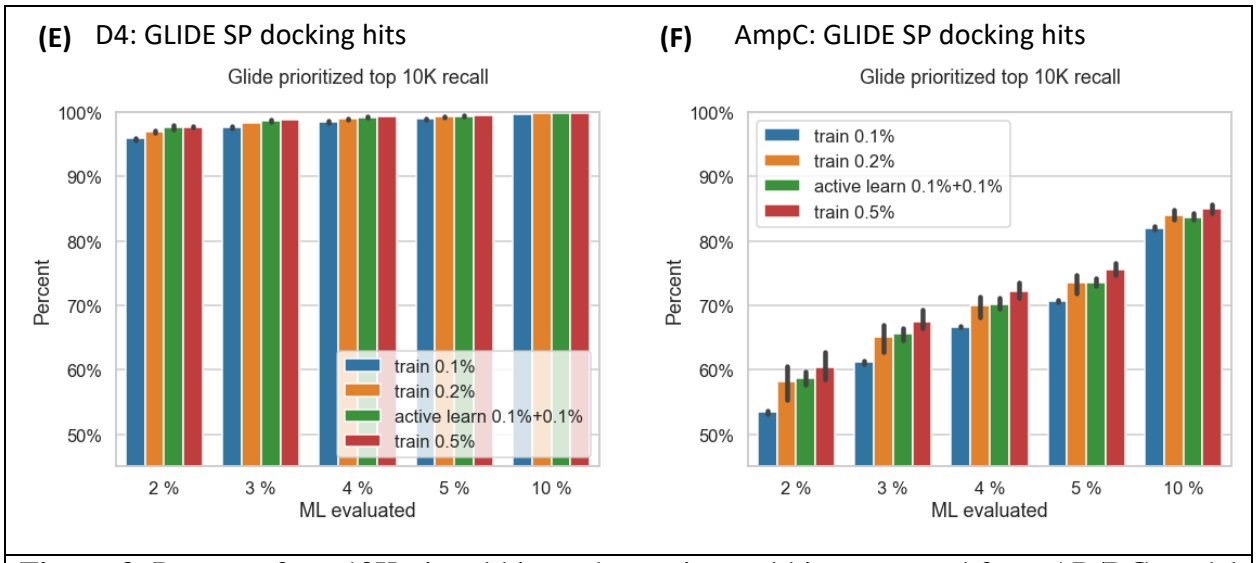


Figure 3. Percent of top 10K virtual hits and experimental hits recovered from AD/DC model with three protocols. AQ/DC model trained with a random 0.1% subset of the database (blue). With a larger 0.5% percent training set (red). Models trained to 0.2% of the library; One based on the AL Protocol (0.1% + 0.1%) (green) and the other to a random set (orange) Recovery of virtual hits from DOCK for D4 (A) and AmpC (B). Recovery of experiment confirmed hits for D4 (C) and Ampc (D). Recovery of virtual hits from Glide SP for D4 (E) and AmpC(F). The error bars were obtained from three replicates of random selection of training set with 95% confidence interval.

Chemical diversity analysis of recovered hits from AQ/DC model

GCNN models learn features from graph neighborhoods and the random forest models sampled in AQ/DC are based on Morgan fingerprints; both models are trained using a set of features derived from a simple 2D representation of molecules (SMILES). These models form the AQ/DC ML ensembles used to screen large libraries in this work. To ensure that these models are not simply learning a simple measure of ligand similarity, we assessed the ability of the AQ/DC models to

capture the diversity of the virtual hit compounds according to DOCK3.7 and Glide SP. We clustered the virtual hit compounds (Top 10,000) according to both Glide and DOCK3.7 using ECFP4 fingerprints. The clustering was done using the Butina clustering algorithm in RDKit, varying the Tanimoto coefficient (Tc) cutoff over from 0.3 to 0.6. The cluster recovery is defined as whether any ligand in the cluster was in the ML prioritized ligands. For D4 screening with DOCK3.7, more than 60% and 80% of the clusters (ECFP Tc= 0.5) were recovered if we dock top 2% and 5% of predictions of AQ/DC model with a larger training set (train 0.5%) or with active learning protocol (active learn 0.1% + 0.1%) (Figure 4A). For D4 docking with Glide SP, more than 90% of the cluster heads (ECFP4 Tc=0.5) are recovered if we dock the top 2% of AQ/DC model predictions no matter which training protocol was used (Figure 4B). A similar trend is seen clustering with Tc 0.3, 0.4, and 0.6 (Figure S9 and S10).

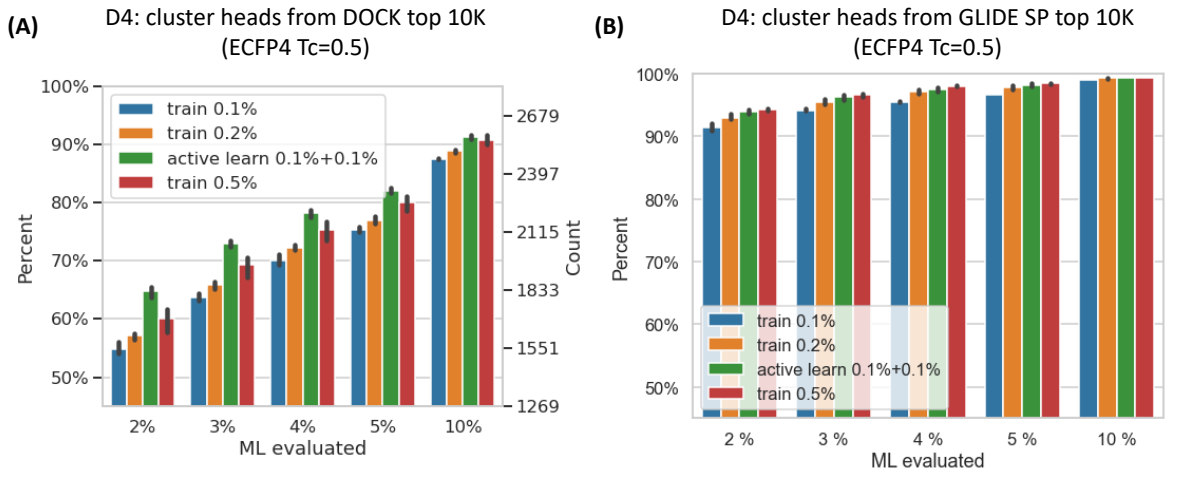
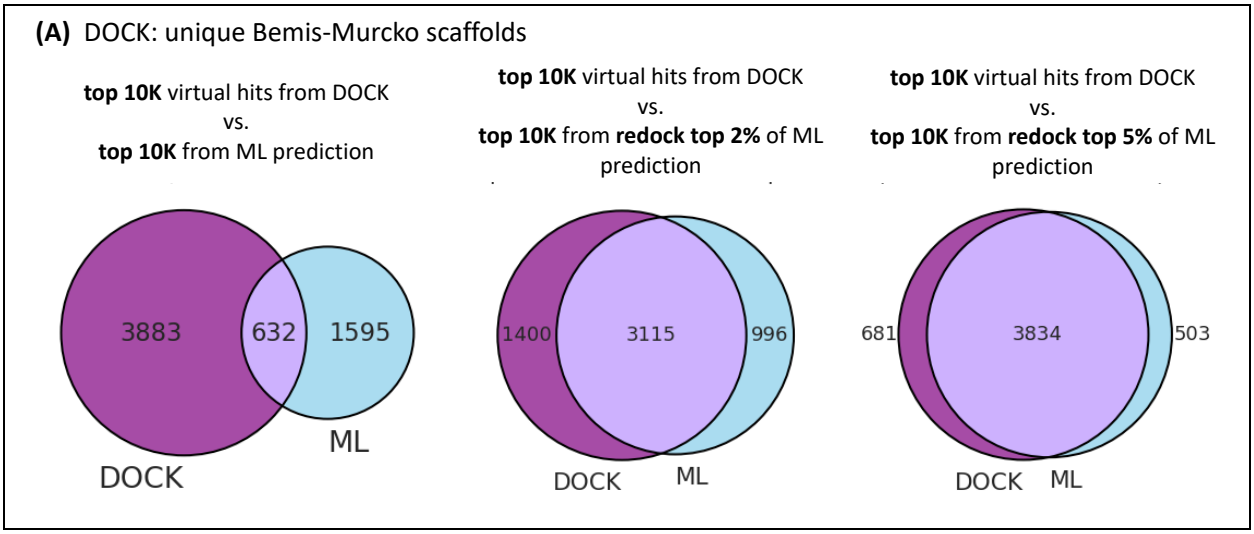


Figure 4. Percent of cluster heads recovered by AQ/DC model with three protocols. The top 10K virtual hits from DOCK (A) and Glide SP (B) were clustered based on ECFP4 fingerprint with a tanimoto coefficient of 0.5. AQ/DC model trained with a random 0.1% subset of the database (blue). With a larger 0.2% percent training set (orange). With an even larger 0.5% percent training set (red). AL protocol with 0.1% most uncertain ML prediction (green).

Another way to assess intrapopulation chemical diversity is to count unique Bemis-Murcko scaffolds. We plot the number of scaffolds from top 10K virtual hits of DOCK3.7 against D4 (Figure 5A). A total of 4,515 unique scaffolds are identified from the top 10k compounds by DOCK score, compared with 1595 for the AQ/DC model, a reduction of roughly 50%. Such observation agrees with a detailed score comparison for compounds with the same Bemis-Murcko scaffold (Figure S8), which tend to have more similar ML scores and more diverse DOCK3.7 scores. If we redock the top 2% and 5% of compounds scored by the AQ/DC model, a much better overlap of common scaffolds is achieved with the new top 10K after redock. Thus, we have to redock the top 2% to 5% (roughly 2M to 5M) of the scored library by our active-learning protocol to recover a similar diversity of the chemical library. The same observation is also seen with Glide SP (Figure 5B). These results suggest strong benefits from using an active learning approach followed by rescoring the top-ranked by ML predictions with explicit docking. This result also highlights the important role physics-based methods can continue to serve in hit identification campaigns. Where pure-ML approaches rely on interpolation, physics-based methods may provide greater opportunity to expand the known chemical space for a project.



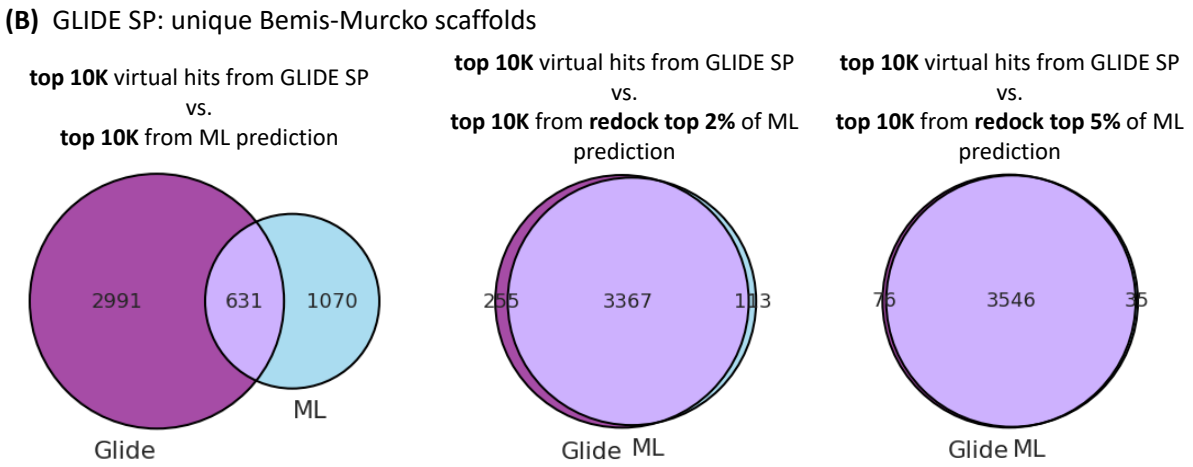


Figure 5. Venn diagram showing the overlap of common Bemis-Murcko scaffolds from top 10K prioritized molecules from DOCK (A) and Glide SP (B) for the D4 target. Scaffolds of top10K virtual hits top 10K (left panel), and new top 10K from redocking the top 2% (middle panel) and redocking the top 5% (right panel) from AQ/DC ML prediction.

Null Model

We compared the performance of our active-learning workflow versus a simple null model based on chemical similarity. This experiment is meant to determine the information gain of our ML approach beyond simply learning the training set molecules themselves. To construct the null model, a random selection of 0.2% of the library was docked. The top 100, 500 and 1000 compounds by docking score were selected as the probe molecules, forming the basis of three models. In addition to the null model, the active learning workflow was run with a 0.1% batch size and two iterations of active learning. Both the active-learning model and the null model receive the same information – the 2D chemical structures and docking scores for 0.2% of the library. Judged by recovery of the top 10k virtual hit compounds, the active learning workflow outperforms the null models (**Figure 6**). The top 1% of predictions by active learning workflow recovered 92%

and 47% top 10K compounds by Glide score for D4 and AmpC, respectively. The top 1% of predictions based on the null model only recovered around 40% and 20% of the top 10K compounds, respectively. This suggests that the ML models are learning features of the docking function itself, and not simply memorizing the training set. Fig S11 shows that the conclusion is similar when molecule chirality is considered during fingerprint generation.

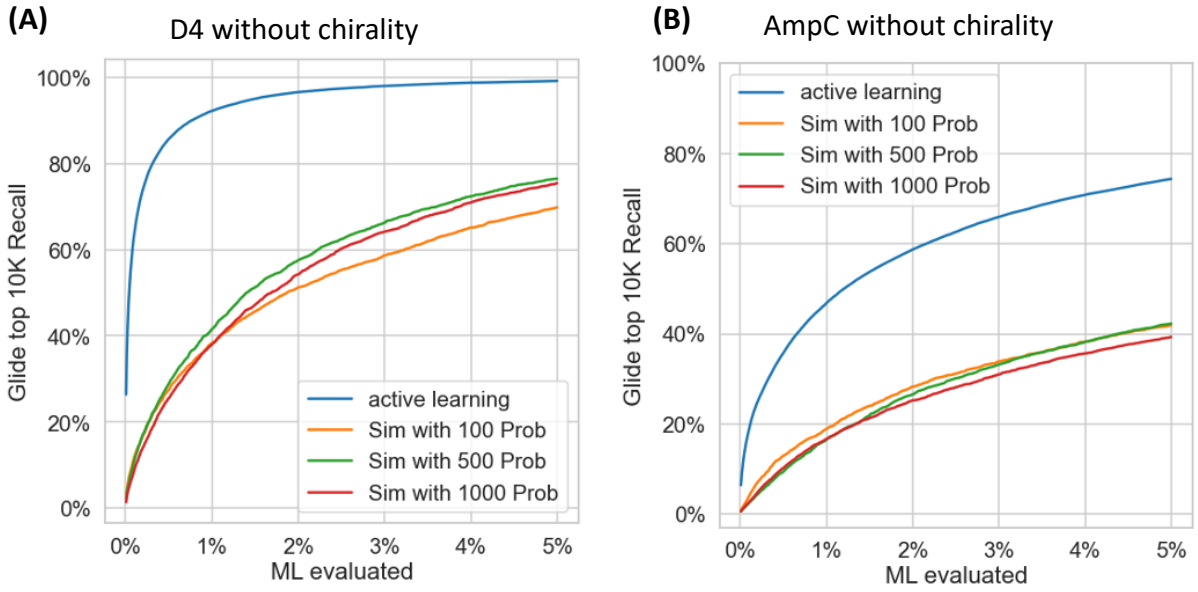


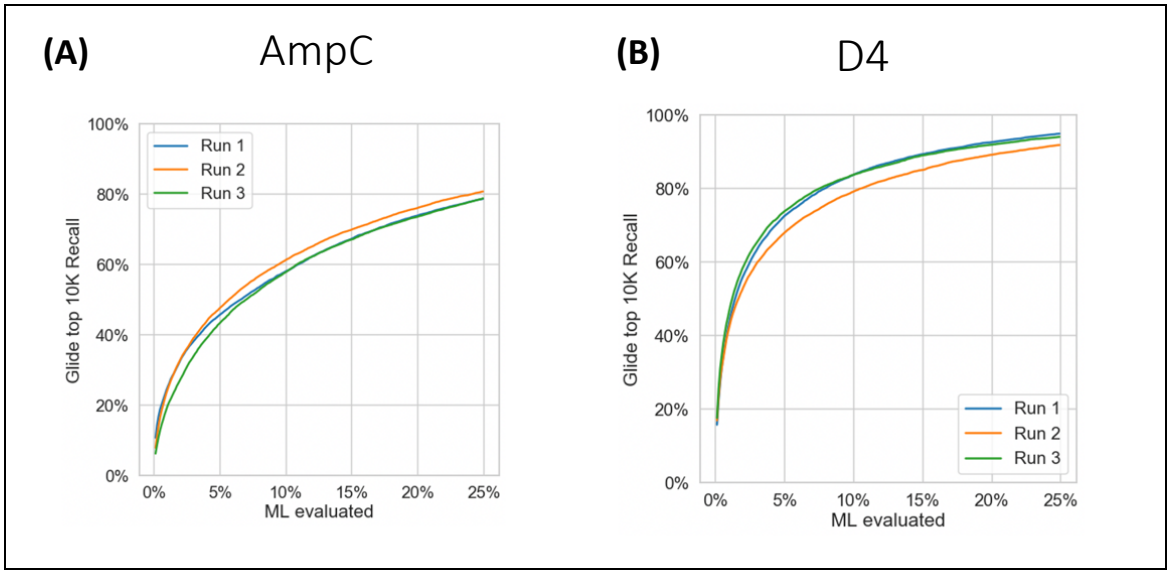
Figure 6. ROC curves showing the percent of top 10K Glide virtual hits recovered by the active learning workflow and 2D fingerprint similarity based null models with different number of probe molecules. The true positives are top 10K virtual hits from Glide for D4 (A) and AmpC (B). The active learning workflow results are the averaged over three runs with different random seed. The 2D fingerprint does not consider the chirality of a molecule.

Comparison of enrichment in on-demand synthesized and in-stock libraries

Besides the large, make-on-demand libraries such as the ZINC database, we evaluated our active-learning protocol on the Sigma Aldrich Market Select database[39], a much smaller, traditional screening library (8M unique compounds) of in-stock compounds. These comparisons were

performed using the Glide SP program with the same simulation conditions as other Glide SP dockings in this work. Results in **Fig 7A-7B** show that AQ/DC models are able recover virtual hit compounds, performing well above random. However, there is a significant degradation in performance as compared with the results on the ZINC database shown in **Fig 2E-2F**. When screening the ZINC library against AmpC, redocking the Top 2% of compounds according the AQ/DC model recovers roughly 40% of the top 10k virtual hits. When the smaller library is screened, only 20% of virtual hit compounds are recovered. A similar trend is observed for the D4 target. The ZINC screen recovers over 80% of the top 10k virtual hit compounds in the top 2% of predictions, as compared with roughly 40% when the Sigma-Aldrich library is used. We conclude that the choice of library can strongly impact the effectiveness of the workflow described in this work. Our analysis of chemical diversity in the previous section indicated that the ML models recovered some scaffolds more completely than others. This loss of diversity could explain the degraded performance on the in-stock library, compared to the virtual, enumerated library. It is intuitive that an enumerated library should feature greater density in local regions of chemical space. In order to test this theory, we selected the top 20 scoring compounds in both libraries from the D4 exhaustive docking with Glide SP. We performed a "hit expansion" of each of these virtual hit compounds by extracting the set of nearest neighbors within a specified chemical similarity to the initial compound. Pairwise similarity was defined as the Tanimoto distance between the molecular fingerprints of each compound. Neighbors with similarity greater than 0.5 were binned into three groups: (1) Between 0.5 and 0.6, (2) Between 0.6 and 0.7, and (3) Greater than 0.7. The ZINC15 subset for the D4 screening in this work is roughly 17x larger than the Sigma Aldrich library screened. As shown in the left panel of **Fig 7C**, a virtual hit compound from the ZINC subset can be expected to have 4.5x more neighbors within a chemical similarity of 0.7, 12x more

neighbors with a chemical similarity to the reference between 0.6 and 0.7, and 36x compounds that are within a chemical similarity of 0.5. The larger ZINC library does offer more close neighbors per query on average. When normalized to account for the difference in library size as showed in the right panel of **Fig 7C**, we find that tighter neighborhoods around each compound approach the size of the smaller library. The information gained from these extremely similar neighbors, > 0.6, may not amount to much for training the ML model, as compounds above this threshold are generally very similar. However, as we widen the radius to a similarity of 0.5, we find a proportional advantage to screening the larger library. The greater density of points in this region could provide a significant boost in the task of learning the parent scoring function.



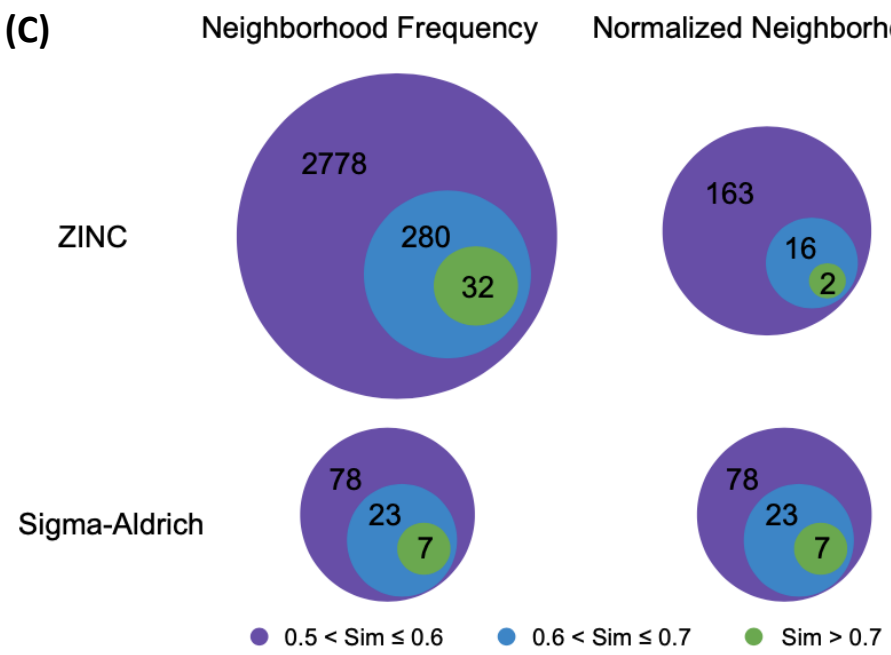
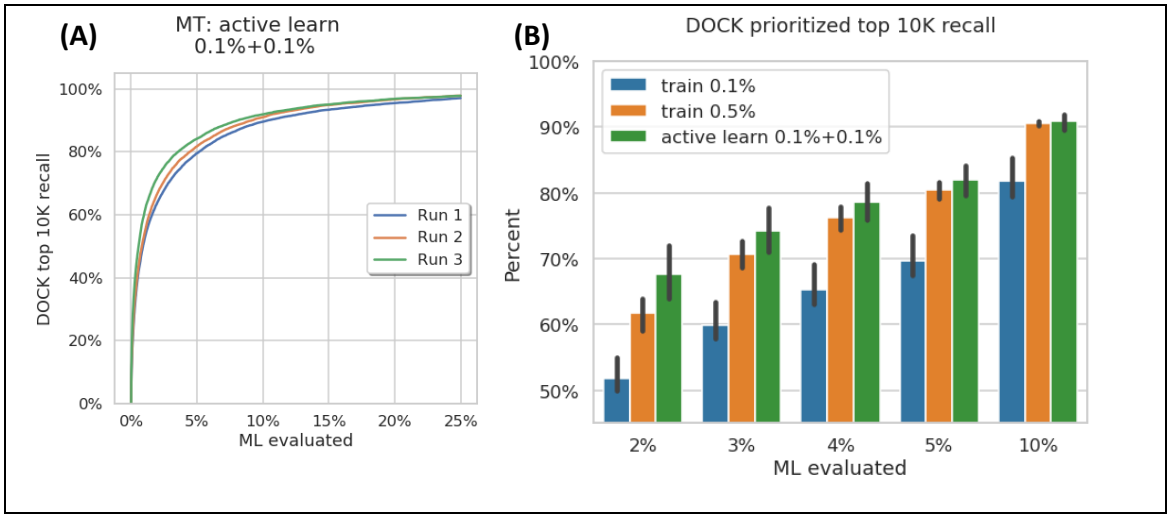


Figure 7. Results of active-learning workflow where the smaller, Sigma-Aldrich Market-Select library was screened against AmpC (A) and D4 (B). (C) Hit Expansion for the top 20 scoring compounds according to Glide SP screened against the D4 target. Average neighborhood size at varying radii in chemical space are shown for the ZINC15 subset and Sigma-Aldrich libraries screened in this work.

Blind test on MT1 system

Besides the retrospective test of the AQ/DC model on the D4 and AmpC datasets, a blind test was carried out employing DOCK3.7 on the MT1 system. At the time of model training, only the training sets were made available to the ML method. Similar to the retrospective study on D4 and AmpC, the AQ/DC model was trained with three protocols: a smaller training set (train 0.1%), a larger training set (train 0.5%), and active learning protocol (active learn 0.1% + 0.1%) where the most uncertain 0.1% from the top 5% of first round ML prediction was added to the original

randomly selected 0.1% training set. The detailed ROC curves for all three protocols are shown in **Figure S6**. Percent recovery of the top 10K virtual hits from DOCK as well as the 104 experimentally confirmed hits are shown in **Figure 8**. Roughly 80% of the virtual hits are recovered in the top 5% of predictions by the AD/DC model trained with a larger set (train 0.5%; orange) or the active learning protocol (active learn 0.1%+0.1%; green) (**Figure 8B**). The advantage of the active learning protocol is most obvious in the early recovery for the MT1 system. From the top 2% of predictions according to the AQ/DC model with active learning protocol, 68% of virtual hits are recovered, while the AQ/DC model with training sets of 0.1% and 0.5% can recover 52% and 62% respectively. A similar trend is seen for the 104 experiment confirmed hits of MT1 system, where the AQ/DC model with active learning protocol can recover 65%, 71%, 78%, and 81% of the experiment hits at the top 2%, 3%, 4% and 5% of the ML prediction (**Figure 8D**).



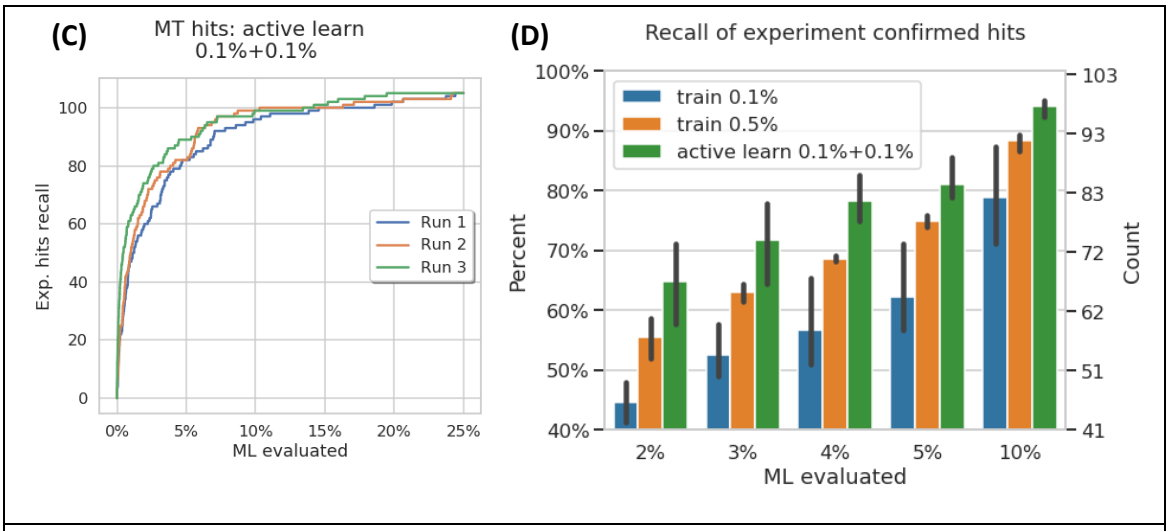


Figure 8. ROC curves and percent recovery by AQ/DC model for MT1 system as a blind test. Percent recall of top 10K DOCK virtual hits (B) and 104 experiment confirmed hits (D) showing AQ/DC model trained with active learning protocol yield best performance in recovery hits. ROC curves of the active learning protocol trained AQ/DC model prediction showing the recovery of top 10K DOCK virtual hits (A) and experimental hits (C).

Comparison of computational costs

The goal of this study is to apply AQ/DC models to complement docking programs (DOCK3.7 and Glide) in screening large data sets through the use of an active-learning scheme to identify compounds likely to score well by docking methods. By sacrificing a few good candidates, a substantial speedup can be achieved. The speed of docking with Glide and DOCK 3.7 as well as the AQ/DC model predictions was evaluated on a single core of an Intel i5-8400 CPU with a clock speed of 2.80GHz. On average, docking a single molecule with Glide and DOCK took 7.9 and 1.0 seconds respectively. The AQ/DC prediction for a single molecule took as little as 5 milliseconds. With a NVIDIA GTX 1080Ti graphics card, the AQ/DC model training took 16 hours on average.

Based on these benchmarks, it is reasonable to train an AQ/DC model with a random selection of 0.5% of the 138 million molecules ZINC subset used to screen the D4 target, followed by a re-docking the top scoring 5% of compounds by AQ/DC model prediction. The workflow offers a savings of 20-fold over explicit docking of all 138 million compounds. The reduction in computational time can be seen in **Figure 9A**, showing the full computational time for the exhaustive docking (Glide, DOCK) approach, the hybrid approach combining docking and AQ/DC modeling (ML+Glide, ML+DOCK), and the active learning protocol (AL+Glide, AL+DOCK). For both DOCK and Glide based workflows, the dominant contributor to the overall compute cost is docking, followed by model evaluation, and finally, by model training. For the active learning protocol, the total CPU wall-clock time is roughly inversely proportional to the number of CPU cores used. Both docking and ML model evaluation tasks are performed on the CPU. Only ML model training is performed on the GPU. Currently, the GPU wall-clock time cannot be reduced by utilizing additional GPUs, as the AQ/DC backend does not support parallel training. **Figure 9B** presents a typical real-world scenario, with a fixed amount of compute resources. Here, the wall-clock time is compared across protocols for a single GPU and 100 CPUs. The wall-clock time achieved using the hybrid approach is 16-fold less than exhaustive docking with Glide, and 10-fold less than with DOCK3.7. Although a small increase in wall-clock time is seen with active learning protocol-mostly from training the AQ/DC model twice, a 14-fold and 7-fold reduction in wall-clock time can be realized against the exhaustive docking approach with Glide and DOCK. Considering the slight increase in wall-clock time and the additional accuracy of the AQ/DC model, the active learning protocol has the best balance of accuracy and time savings and is recommended for ultra-large scale docking campaigns.

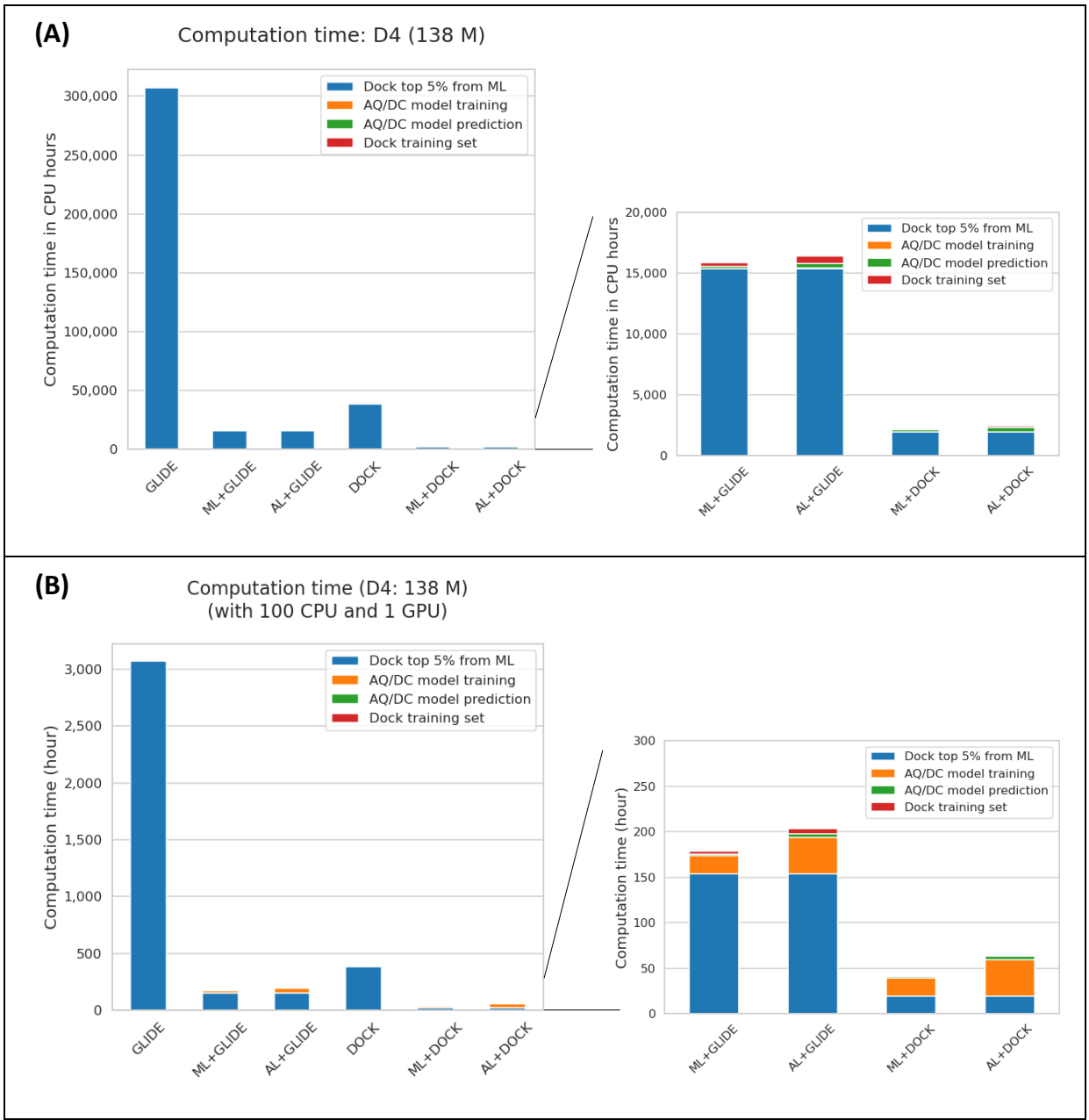


Figure 9. Wall-clock times for a 138 million molecule virtual screen of the D4 system in CPU hours (A) with one CPU and one GPU; (B) with 100 CPUs and one GPU.

Conclusion

Ultra-large libraries vastly expand the number and diversity of compounds accessible to structure-based docking screens, and indeed other virtual screens. It is clear, however, that such large libraries will soon become intractable to prepare for atomistic docking, let alone for explicitly docking to a given receptor structure. This already limits the feasibility of conventional docking in many interesting applications. Here, we investigate a more cost-effective method to dock ultra-large libraries. The active-learning workflow can act a fast stand-in for a computationally expensive docking, minimizing the number of docking calculations required. The AQ/DC models can recover compounds outside of the training-set, as evidenced by the substantial improvement over a similarity-based null model. The active-learning workflow does suffer from a loss of diversity compared to the hit list ordered by full docking score, as determined by the count of unique chemical scaffolds. This loss of diversity can be dramatically reduced by adding a simple redocking step to the end of the workflow. We propose a new selection rule for optimizing the information gain for the ML model. By combining high-scoring compounds with the ensemble uncertainty reported by our ML models, we reach a good balance of the explore vs exploit tradeoff. The workflow presented here, and others in reported in the literature[8, 25, 26], will expand the scope of typical virtual screening campaigns to billions of compounds, democratizing access to ultra-large chemical libraries.

Supporting Information Available

Supporting Information is available free of charge at http://pubs.acs. org/.

Overlap among three random selected training set (**Figure S1**), correlation between dock scores and ML scores from AQ/DC model prediction (**Figure S2**), percent recovery of docking virtual

hits from different protocols of AQ/DC model (**Table 1**), recovery performance of top 10K virtual hits for D4 and AmpC (**Figure S3**), D4 ROC curves of top 10K virtual hits and experiment hits recovered from three protocols of the AQ/DC models (**Figure S4**), AmpC ROC curves from three protocols of the AQ/DC models (**Figure S5**), MT1 ROC curves from three protocols of the AQ/DC models (**Figure S6**), recovery performance of top 100K and top 300K virtual hits for D4 and AmpC from DOCK and GLIDE SP (**Figure S7**), example score distribution from DOCK and AQ/DC model of molecules with the same Bemis-Murcko scaffold (**Figure S8**), recovery performance of D4 cluster heads from DOCK top 10K virtual hits (**Figure S9**), recovery performance of D4 cluster heads from GLIDE SP top 10K virtual hits (**Figure S10**), ROC curve of AQ/DC model and null model based on 2D fingerprint similarity (with chirality) (**Figure S11**), comparison of computational hours and cost for three protocols (**Table S2, Figure S12**).

NOTES

The authors declare the following competing interest(s): B.S is a consultant to Schrödinger, Inc., and is on the Scientific Advisory Board of Schrödinger, Inc

ACKNOWLEDGMENT The authors thank Pat Lorton for useful discussions.

ABBREVIATIONS

AMPC, Beta-lactamase; AQ/DC, AutoQSAR/DeepChem; D4, D4 Dopamine receptor; ML, machine learning; MT1, Melatonin MT1 receptor; ROC, Receiver-Operator Characteristics

REFERENCES

1. Bohacek, R. S.; McMartin, C.; Guida, W. C., The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal Research Reviews* **1996,** *16* (1), 3-50.

2. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L., Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J Chem Inf Model* **2012,** *52* (11), 2864-75.

3. Sterling, T.; Irwin, J. J., ZINC 15--Ligand Discovery for Everyone. *J Chem Inf Model* **2015,** *55* (11), 2324-37.

4. SpiroChem DEL. https://www.spirochem.com/libraries.

5. WuXi DEL. https://wuxi-rsd.com/dna-encoded-library-technology-services.

6. Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J., Ultra-large library docking for discovering new chemotypes. *Nature* **2019,** *566* (7743), 224-229.

7. Stein, R. M.; Kang, H. J.; McCorvy, J. D.; Glatfelter, G. C.; Jones, A. J.; Che, T.; Slocum, S.; Huang, X. P.; Savych, O.; Moroz, Y. S.; Stauch, B.; Johansson, L. C.; Cherezov, V.; Kenakin, T.; Irwin, J. J.; Shoichet, B. K.; Roth, B. L.; Dubocovich, M. L., Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **2020,** *579* (7800), 609-614.

8. Gorgulla, C.; Boeszoermenyi, A.; Wang, Z. F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.;

Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H., An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **2020,** *580* (7805), 663-668.

9. McGann, M., GigaDocking - Structure Based Virtual Screening of Over 1 Billion Molecules. 2019.

10. Mysinger, M. M.; Shoichet, B. K., Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *Journal of Chemical Information and Modeling* **2010,** *50* (9), 1561-1573.

11. Coleman, R. G.; Carchia, M.; Sterling, T.; Irwin, J. J.; Shoichet, B. K., Ligand Pose and Orientational Sampling in Molecular Docking. *Plos One* **2013,** *8* (10).

12. Irwin, J. J.; Shoichet, B. K., Docking Screens for Novel Ligands Conferring New Biology. *Journal of Medicinal Chemistry* **2016,** *59* (9), 4103-4120.

13. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **2004,** *47* (7), 1739-1749.

14. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry* **2006,** *49* (21), 6177-6196.

15. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of Medicinal Chemistry* **2004,** *47* (7), 1750-1759.

16. McGann, M., FRED Pose Prediction and Virtual Screening Accuracy. *Journal of Chemical Information and Modeling* **2011,** *51* (3), 578-596.

17. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins* **2003,** *52* (4), 609-623.

18. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible ligand docking. *Abstr Pap Am Chem S* **1997,** *214*, 154-Comp.

19. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996,** *261* (3), 470-489.

20. Grebner, C.; Malmerberg, E.; Shewmaker, A.; Batista, J.; Nicholls, A.; Sadowski, J., Virtual Screening in the Cloud: How Big Is Big Enough? *J Chem Inf Model* **2020,** *60* (9), 4274-4282.

21. Mitchell, J. B., Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci* **2014,** *4* (5), 468-481.

22. Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B., Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* **2018,** *23* (8), 1538-1546.

23. Jiménez-Luna, J.; Grisoni, F.; Schneider, G., Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2020,** *2* (10), 573-584.

24. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T., The rise of deep learning in drug discovery. *Drug Discovery Today* **2018,** *23* (6), 1241-1250.

25. Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A. T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A., Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent Sci* **2020,** *6* (6), 939-949.

26. Reker, D., Practical considerations for active machine learning in drug discovery. *Drug Discovery Today: Technologies* **2020**.

27. Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S., Reaction-Based Enumeration, Active Learning, and Free Energy Calculations To Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin-Dependent Kinase 2 Inhibitors. *Journal of Chemical Information and Modeling* **2019,** *59* (9), 3782-3793.

28. Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E., Less is more: Sampling chemical space with active learning. *J Chem Phys* **2018,** *148* (24).

29. *Schrödinger Release 2020-4*, Schrödinger, LLC, New York, NY, 2021.

30. Dixon, S. L.; Duan, J. X.; Smith, E.; Von Bargen, C. D.; Sherman, W.; Repasky, M. P., AutoQSAR: an automated machine learning tool for best-practice quantitative structure-activity relationship modeling. *Future Med Chem* **2016,** *8* (15), 1825-1839.

31. Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V., *Deep Learning for the Life Sciences*. O'Reilly Media: 2019.

32. *RDKit: Open-source cheminformatics*, http://www.rdkit.org.

33. *Schrödinger Release 2020-1: LigPrep*, Schrödinger, LLC, New York, NY, 2020.

34. Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M., Epik: a software program for pK (a) prediction and protonation state generation for drug-like molecules. *J Comput Aid Mol Des* **2007,** *21* (12), 681-691.

35. Sterling, T.; Irwin, J. J., ZINC 15-Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015,** *55* (11), 2324-2337.

36. Csizmadia, F., JChem: Java applets and modules supporting chemical database handling from web browsers. *J Chem Inf Comp Sci* **2000,** *40* (2), 323-324.

37. Gasteiger, J.; Hiller, C.; Rudolph, C.; Sadowski, J., Automatic-Generation of 3d-Atomic Coordinates for Organic-Molecules. *Abstr Pap Am Chem S* **1991,** *202*, 36-Cinf.

38. 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* (4), 572-584.

39. Aldrich Market Select. https://www.aldrichmarketselect.com/.

40. Daylight. https://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed 11/24/2020).

41. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010,** *50* (5), 742-754.

Table of Contents Graphic

Efficient Exploration of Chemical Space with Docking and Deep-Learning Ying Yang[1], Kun Yao[2], Matthew Repasky[3], Karl Leswing[2], Robert Abel[2], Brian Shoichet[1], Steven V. Jerome[4]*