JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

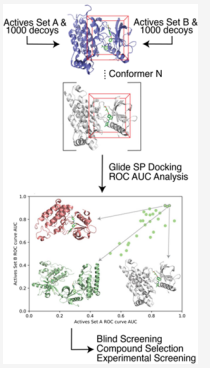
Actives-Based Receptor Selection Strongly Increases the Success Rate in Structure-Based Drug Design and Leads to Identification of 22 Potent Cancer Inhibitors

Eric R. Hantz and Steffen Lindert*

Cite This: J. Chem. Inf. Model. 2022, 62, 5675−5687 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Computer-aided drug design, an important component of the early stages of the drug discovery pipeline, routinely identifies large numbers of false positive hits that are subsequently confirmed to be experimentally inactive compounds. We have developed a methodology to improve true positive prediction rates in structure-based drug design and have successfully applied the protocol to twenty target systems and identified the top three performing conformers for each of the targets. Receptor performance was evaluated based on the area under the curve of the receiver operating characteristic curve for two independent sets of known actives. For a subset of five diverse cancer-related disease targets, we validated our approach through experimental testing of the top 50 compounds from a blind screening of a small molecule library containing hundreds of thousands of compounds. Our methods of receptor and compound selection resulted in the identification of 22 novel inhibitors in the low μM−nM range, with the most potent being an EGFR inhibitor with an IC_{50} value of 7.96 nM. Additionally, for a subset of five independent target systems, we demonstrated the utility of Gaussian accelerated molecular dynamics to thoroughly explore a target system's potential energy surface and generate highly predictive receptor conformations.



INTRODUCTION

In 2019 alone, the pharmaceutical industry spent $83 billion USD on drug research and development.[1] Despite this large investment, there is a critical need for methodological improvement in all aspects of the drug discovery pipeline. Computational methods are a central part of the early stages of the drug discovery pipeline, with a focus on computer-aided drug discovery (CADD) methods such as structure-based drug design (SBDD) and ligand-based drug design (LBDD).[2] SBDD utilizes the three-dimensional structure of a protein target obtained through structural biology methods such as X-ray crystallography, nuclear magnetic resonance (NMR), or cryo-electron microscopy (cryo-EM) to identify possible small molecule binding sites and interactions that are important to biological function. Potential small molecule inhibitors are subsequently designed utilizing the structural information to disrupt biological pathways essential for the survival of the targeted pathogen or host proteins.[3] Proteins are intrinsically flexible entities. Thus, there exists a multitude of potential structures or conformations that may be relevant for SBDD for all drug targets where the predominant mechanism underlying ligand binding is conformational selection.[4] It is possible to elucidate these conformations through structural biology techniques or molecular dynamics (MD) simulations. However, determining which of these target conformations should be used in SBDD drug screenings is nonobvious and the choice of target conformation is crucial for the success of identifying small molecule inhibitors.

The identification of high performing receptors for SBDD has been the focus of several studies. It is commonly accepted that the use of multiple receptor conformations generally leads to better performance when compared to a single receptor conformation. The relaxed complex scheme (RCS)[5,6] has been developed to screen against multiple conformations and account for the flexibility of both the receptor and docked ligands. There have been many attempts in creating guidelines for selecting the best performing subset of conformers. Rueda and coworkers found no correlation between receptor performance and binding site volume, number of atomic contacts, X-ray resolution, B-factors, or flexibility descriptors obtained from an elastic network normal mode analysis.[7] Others have attempted to generate high performing receptor conformations through the use of molecular dynamics.[8−10] Swift and colleagues created three methods for selecting structure-based ensembles.[11] The common performance metrics for virtual screening of single or multiple receptor conformations have been receiver operating characteristic (ROC) curves and enrichment factors[7−14] In a virtual screening application, ROC curves evaluate the performance of a specific conformation by calculating the true positive rate (identification of known inhibitors) and false positive rate

Received: July 7, 2022
Published: November 2, 2022

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

(identification of known/assumed decoy molecules) based on the ranked ordering by the docking score of all compounds. The diagnostic ability of this metric informs on the predictability of a single conformation in virtual screening. Conformers that perform better will have a higher area under the ROC curve (AUC) value, with the maximum value equaling 1. Additionally, the enrichment factor measures the number of active compounds found within a defined early recognition fraction of the ordered list relative to that of a random distribution.

In this work, we examined the effect of conformational selection on success in SBDD and present a simplified use of the ROC AUC metric to streamline the selection of top performing receptor conformations. We then validated the success of our approach with a blind screening and experimental follow-up on a subset of targets. This method was applied to 20 target systems identified through the Database of Useful Decoys Enhanced (DUD-E). Experimentally determined protein conformations were retrieved from the Protein Data Bank (PDB) for all target systems. Co-crystalized inhibitors (known actives) were separated into two sets (set A and set B) of similar average molecular weight and screened with a set of decoy small molecules to calculate the ROC AUC. Predictiveness of receptor conformations was initially calculated with the actives in set A and then confirmed independently with the actives in set B. The top three performing conformers were selected based on their AUC values. From the 20 targets for which this method was applied, we identified five cancer-related drug targets for further blind screening and experimental follow-up. We performed blind virtual screenings into the three top performing conformers using a diverse library of over 500,000 druglike and leadlike compounds. The compound library was prefiltered utilizing a cheminformatics approach in order to streamline the docking process. We then created a ranking of the top 50 docked compounds for experimental testing based on an averaged ligand Z-score across all receptors. We performed radiometric HotSpot kinase assays to measure the effects of our proposed inhibitors on kinase activity and obtain IC_{50} values for all identified inhibitors. The methodology described in this work led to the identification of 22 novel inhibitors in the low μM−nM range. Our method resulted in an 8.8% success rate across all five targets, with the highest success rate of any one target being 24%. Furthermore, for a subset of five targets, we explored the use of Gaussian accelerated Molecular Dynamics (GaMD) in order to create additional receptor conformations for actives/decoys screening. For three of the five systems, we created clustered conformers that, based on the ROC AUC metric, are among the top three most predictive receptor conformations to identify known binders. We also identified general trends of the predictability of clustered GaMD conformations and hypothesized methods for the selection of generating more highly predictive conformations.

METHODS

Protein Target Selection and Preparation. We identified 20 different target systems through a randomized selection of the targets available in the Database of Useful Decoys: Enhanced (DUD·E).[15] The 20 target systems used in our studies were 11-beta-hydroxysteroid dehydrogenase (11b-HSD1), acetylcholinesterase (hAChE), aldose reductase (ALDR), coagulation factor X (FA10), epidermal growth factor receptor erbB-1 (EGFR), estrogen receptor alpha (ESR1), fatty acid binding protein adipocyte (FABP4), fibroblast growth factor receptor 1 (FGFR1), heat shock protein 90 (HSP90), histone deacetylase 8 (HDAC8), human immunodeficiency virus type 1 reverse transcriptase (HIVRT), inhibitor of apoptosis protein 3 (XIAP), insulin-like growth factor I receptor (IGF1R), macrophage colony stimulating factor receptor (CSF1R), MAP kinase-activated protein kinase 2 (MAPK2), rho-associated protein kinase 1 (ROCK1), serine/threonine-protein kinase (AKT1), stem cell growth factor receptor (KIT), thyroid hormone receptor beta-1 (THB), and vascular endothelial growth factor receptor 2 (VEGFR2). A comprehensive search of experimental receptor structures was performed for each target. We collected the structures listed in DUD·E and added additional, more recent structures deposited in the RCSB PDB.[16] Known inhibitors, co-crystallized with their target protein, are henceforth referred to as actives. The number of receptors and actives per target, along with respective PDB codes, is summarized in Table S1.

All structures were imported into Schrödinger's Maestro[17] and prepared using Schrödinger's Protein Preparation Wizard.[18] For each structure, the C-terminus was capped by the addition of an *N*-methyl amide and the N-terminus with the addition of an acetyl group. The protonation states of all titratable residues were assigned using EPIK[19] with a pH constraint of 7.4 ± 1.0.[20]

Receptor Grid Generation. Receptor grids for structures with co-crystalized ligands were generated by selecting the ligand within the Maestro workspace. For structures without a ligand bound, the center of the search space was determined by submitting a PDB file of the apo receptor to the FTMap Server,[21] where fragments were globally docked into the protein structure to identify potential small molecule binding sites. The resulting output of the receptor and docked fragments was then imported into PyMOL,[22] where the align function was utilized to overlay the FTMap PDB file and that of an experimentally derived co-crystalized protein−ligand structure of the same target protein system. The coordinates for the center of the receptor grid were subsequently obtained by extracting the center of mass of one of the fragments in the FTMap generated docking sites which overlayed the coordinates of the ligand from the ligand-bound structure. For the ligand-free structures, these three-dimensional coordinates were manually entered into the receptor grid generation tool. The search area was centered on the ligand (or the manually entered coordinates for ligand-free structures) and allowed the centroids of any docked species to fully explore a 10 × 10 × 10 $Å^3$ inner search space, while the periphery of the ligand was able to extend out to 20 × 20 × 20 $Å^3$. The OPLS3e forcefield[23] was used to generate the desired search grid. All hydroxyl groups were selected to be freely rotatable in the search area.

Ligand Preparation. The LigPrep[24] tool of the Schrödinger Software Suite was used to prepare each ligand for docking. All protomers, tautomers, and stereoisomers were generated for each ligand. Protonation states were assigned using EPIK with a pH value of 7.4 ± 1.0.[20,25] The coordinates of all co-crystallized ligands were extracted from their respective PDBs. Small molecules from the Schrödinger decoy sets and the ChemBridge EXPRESS-Pick Collection used in our docking protocol originated from SDF files containing three-dimensional coordinates.

Active/Decoy Screening and Receptor Performance Analysis. For each of the 20 targets, we identified between 9 and 68 active compounds. For each target system respectively, active compounds were evenly separated into two sets of similar average molecular weights, active sets A and B. Diversity of small molecules between sets A and B was confirmed using the mutual

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

Tanimoto coefficients[26] with respect to all compounds. Set A was considered as known actives used for receptor identification, and set B was considered to be unknown actives for independent verification. To quantify how well actives rank in virtual screening, we assembled two decoy sets. The two sets of small molecules considered to be decoy ligands were obtained from Schrödinger, with average molecular weights of each set being 360 and 400 g/mol, respectively.[27] For each target system, we used the decoy set whose average molecular weight was closest to that of the average molecular weight of the active compounds. All compounds were subject to ligand preparation as detailed above. Docking of actives set A, B, and decoy compounds post-LigPrep was performed using Schrödinger's Glide SP.[27−29] Default parameters were maintained for this method as implemented in the Schrödinger 2018-3 release. The resulting docked poses were ranked by their docking score, with the top scoring pose of each protomer/stereoisomer being kept. For every receptor in every target system, the true positive rates (TPRs) and false positive rates (FPRs) were calculated to generate ROC curves for actives sets A and B, respectively. The area under the ROC curve (AUC) was calculated using Python's scikit-learn library (ver. 0.22.1).[30] The AUC of actives set A was compared against the AUC of actives set B for all receptors in each target protein system. This was done to determine the predictability of the receptor based on both sets of actives. Additionally, we averaged the AUC values for both sets of actives for each receptor conformation. We used this average AUC value to propose the top three most predictive receptor conformations for each target for potential further utilization in blind screenings. As a second metric to confirm the high predictability of the top three performing receptor conformations, the enrichment factor (EF) was calculated for each set of actives per conformation for a subset of the targets according to the following equation:

$$\mathrm{EF} = \frac{N_{\text{actives in top 50}}}{50} \times \frac{N_{\text{decoys}} + N_{\text{actives}}}{N_{\text{actives}}}$$

where $N_{\text{decoys}} = 1000$ and N_{actives} correspond to the number of actives in either set A or B for each of the target systems.

Enhanced Sampling with GaMD. In addition to using experimental structures from the DUD·E and PDB databases, we wanted to explore whether nonexperimental conformations can exhibit high predictability as well. In order to account for protein conformational flexibility that may not be sufficiently represented in the PDB, additional protein receptor conformations were obtained from 300 ns GaMD production simulations performed with Amber20.[31−33] A subset of five target protein systems were selected based on having a broad range of average AUC values across their experimental receptor conformations. The target systems selected were hAChE, FABP4, HSP90, HDAC8, and HIVRT. For each target, five receptors were selected for GaMD simulations; two receptors with the lowest average AUC values, two receptors with the highest average AUC values, and one receptor with an AUC closest to the mean average AUC value. For receptors with a small molecule bound, the small molecule was parameterized using the second generation of the generalized amber forcefield (GAFF2)[34] and AMBER's Antechamber[31] software. In order to reduce computational expense, any homo-multimeric protein structure was reduced to a single monomer while maintaining the integrity of the ligand binding site. The proteins or protein−ligand complexes were solvated with TIP3P[35] water molecules in a 10 Å octahedron and neutralized with sodium ions. All GaMD simulations were performed using ff14SB.[32]

All systems were minimized with strong restraints on the protein and small molecule (if applicable) using 2500 steps of steepest decent minimization followed by 2500 steps of conjugate gradient descent. A second, unrestrained minimization was performed using 2500 steps of steepest decent minimization followed by 2500 steps of conjugate gradient descent. The system was subsequently heated to 310 K over a span of 1 ns using the Langevin thermostat.[36] For each system, a short equilibration conventional MD simulation of 10 ns was performed at a constant temperature (310 K) and pressure (1 bar) prior to the GaMD preparation simulations. During the GaMD preparation runs, statistics to calculate appropriate boosts to apply to the dihedral and total potential energies were collected. Statistics were obtained from a second 10 ns conventional MD run, initial boosts applied, and subsequently updated during a 50 ns GaMD biasing run. The final GaMD restart parameters (VmaxP, VminP, VavgP, sigmaVP, VmaxD, VminD, VavgD, and sigmaVD) were then read in for 300 ns GaMD production runs. The upper limit for the dihedral and total boost potentials was set to 6 kcal/mol. All simulations were run with a 12 Å cutoff for electrostatic and van der Waals interactions and used a 2 fs timestep with the SHAKE algorithm.[37] Periodic boundary conditions were under an *NPT* ensemble with a pressure set at 1 bar using a Berendsen barostat[38] and Langevin thermostat. Coordinates were saved every 4 ps, resulting in 75,000 frames.

The final structures used for docking studies were obtained by clustering each 300 ns GaMD simulation individually. For the clustering analysis of each trajectory, all waters, ions, and ligands were stripped, and every frame was analyzed, resulting in 75,000 frames available for clustering. The density-based clustering algorithm (DBScan)[39] implemented in AMBER's CPPTRAJ was used to cluster the processed trajectories to obtain approximately 10 new conformations. Trajectories were clustered using the backbone atoms of residues in the ligand binding site. Residues in the ligand binging site were identified using the ligand interaction preset in PyMOL and cross-referenced using the ligand interaction tool in the Maestro workspace. Residues used for clustering in each target system can be found in Table S2. Clustered conformers (284) were created for further docking studies. For each clustered conformation, actives set A, B, and the appropriate decoy set for that target were docked using Glide SP. ROC curves were generated for each clustered conformation, and the AUC of actives set A was compared against the AUC of actives set B for all conformations in each target protein system. The AUC value for both sets of actives for each conformer was averaged in order to assess the predictability of the clustered conformer.

Cancer Target Subset Selection for Blind Screening. To experimentally verify the increased success rate of the identified most predictive receptor conformations in true blind screening scenarios, we selected five targets that were related to various types of cancers: AKT1, CSF1R, EGFR, FGFR1, and VEGFR2. These five targets were selected based on low mutual sequence identities, existence of receptors with AUC values showing high predictability, and commercial availability from Reaction Biology Corporation (RBC). The mutual sequence identities were calculated using the LALIGN/PLALIGN server provided by the University of Virginia.[40] Mutual sequence identities for the five targets are shown in Figure S1.

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

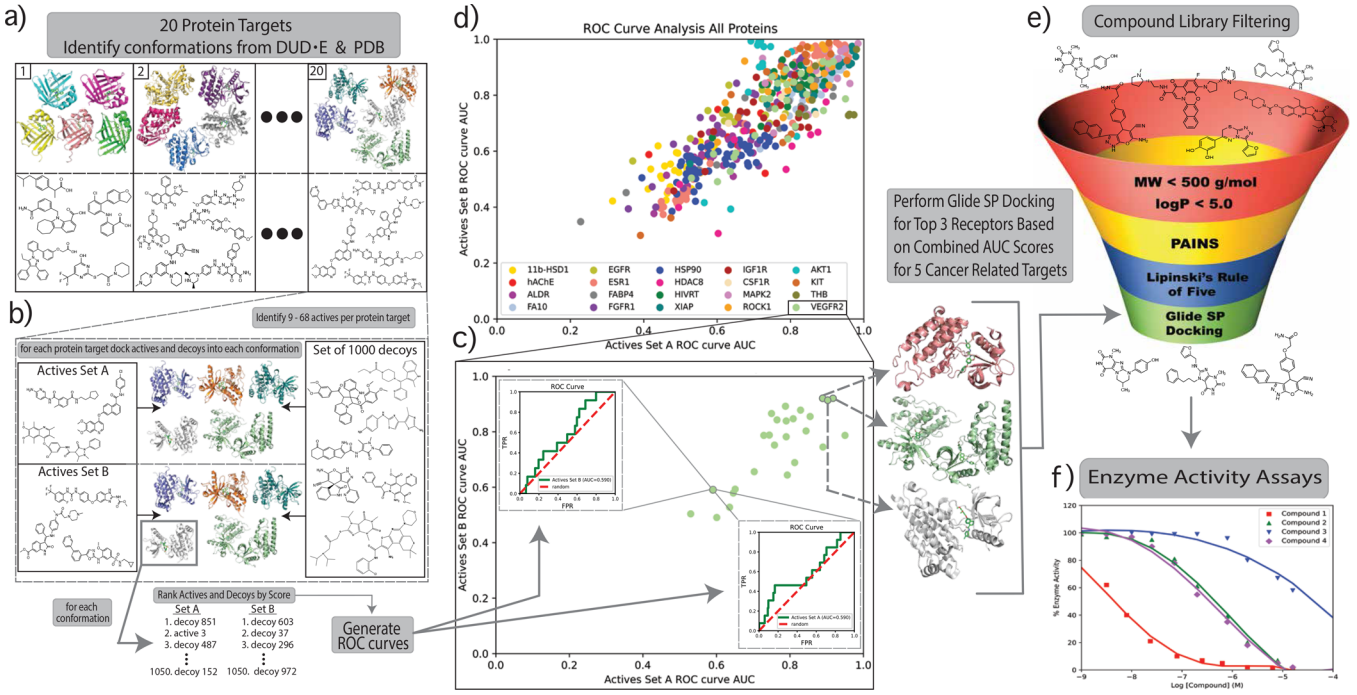


Figure 1. Project workflow. (a) Identification of conformations and actives for 20 target protein systems. (b) Illustration of docking active set A, B, and decoy set into each conformation for all targets and ranking the small molecules based on the Glide SP docking score to generate ROC curves. (c) AUC of set B plotted against the AUC of set A for all conformations of target VEGFR2. The three most predictive conformations used for Glide AP docking of filtered ChemBridge EXPRESS-Pick Library are circled and shown on the right. (d) AUC of set B plotted against AUC of set A for all conformations and all target systems. (e) Filtering criteria for the docking of the ChemBridge EXPRESS-Pick Library. (f) Experimental Hot Spot Kinase Assays on suggested top scoring ChemBridge compounds.

Small Molecule Library Selection and Blind Screening. To identify novel inhibitors for the five cancer targets, we selected the ChemBridge EXPRESS-Pick Collection for screening in this study. It contained 501,916 small druglike molecules. We prefiltered the compounds of this collection based on molecular weight (MW) and predicted solubility (logP), while additionally excluding compounds with functional groups implicated as pan-assay inference compounds,[41] and those violating Lipinski's rule of five.[42] The prescreening of this collection of compounds was done to increase the efficiency of our docking process and was performed using the 2020.09.1 release of the RDKit[43] package implemented through Python 3.7. Compounds with a MW over 500 g/mol were removed, in order to maintain an average compound MW closer to that of the known actives for our set of targets. LogP parameters were calculated in RDKit using the Wildman and Crippen's model,[44] and compounds with a predicted logP value over 5.0 were discarded. PAINS filters A, B, and C were used to remove potentially promiscuous compounds from our database.[41] Compounds found to have more than five hydrogen bond donors and more than ten hydrogen bond acceptors were removed from the library in accordance with the remaining conditions of Lipinski's rule of five which were not previously imposed as hard cut-offs. Upon implementation of these filters, the EXPRESS-Pick library was reduced to 409,672 compounds. The initial filtration based on MW and logP values removed 60,945 compounds. The PAINS filter removed an additional 30,556 compounds, and filtering the remaining compounds based on Lipinski's rule of five removed another 743 compounds. The remaining 409,672 compounds were prepared with Schrödinger's LigPrep module, resulting in 633,076 stereoisomers/enantiomers used for screening.

All resulting compounds (633,076) were docked into the three most predictive receptor conformations as determined by the average AUC value for all five target systems (AKT1, CSF1R, EGFR, FGFR1, and VEGFR2). For each docked compound, a Z-score of the ligand was calculated based on the respective docking scores as described by Kim et al.[45] The Z-score was calculated for each docked compound in every receptor conformation. The compound's Z-score was then averaged across all three receptors to create an unbiased ranking of docked compounds for the target system. This was done to avoid any bias in compound selection as a result of different ranges of docking scores across the receptors. The top 50 ranked compounds by averaged Z-score were ordered directly from ChemBridge and tested in vitro by Reaction Biology Corporation.

Radiometric Hot Spot Kinase Assays. Experimental testing of the identified 250 potential cancer target inhibitors was performed by RBC using their HotSpot assay, a miniaturized assay which significantly reduces the consumption of radioisotope materials, kinase targets, substrates, and compounds, making this method highly appropriate for high-throughput screening.[46] Substrates were prepared in a base reaction buffer consisting of 10 mM Hepes (pH 7.5), 10 mM $MgCl_2$, 1 mM EGTA, 0.01% Brij35, 0.02 mg/mL BSA, 0.1 mM Na_3VO_4, 2 mM DTT, and 1% DMSO. For CSF1R, EGFR, FGFR1, and VEGFR2 cofactors, $MnCl_2$ was then added to the substrate solution at a concentration of 0.2 mg/mL. The kinase was then added to the substrate solution and gently mixed. Enzyme and substrate specific conditions for all five cancer-related targets are listed in Table S3. Compounds were received as power stock from ChemBridge and dissolved to 10 mM in DMSO. All compounds are >95% pure by liquid chromatog-

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

raphy−mass spectroscopy (LC−MS) analysis. Compounds were initially tested in single dose duplicate mode at a concentration of 10 μM and delivered into the kinase reaction mixture by Acoustic technology (Echo550; nanoliter range) and incubated for 20 min at room temperature.[47] Compounds that resulted in an average percent enzyme activity relative to DMSO controls of less than 55% were determined to be inhibitors. For the 22 identified inhibitors, 10-dose IC_{50} values were obtained. All compounds were tested in a 10-dose IC_{50} mode with a three-fold serial dilution starting at 50 μM, except for compound 7572363 which was tested at a three-fold dilution starting at 10 μM because of the compound's high potency. Control compound, Staurosporine, was tested in 10-dose IC_{50} mode with four-fold serial dilution starting at 20 μM. Experimental conditions for the IC_{50} experiments can be found in Table S4.

RESULTS AND DISCUSSION

Here, we are presenting novel methodology to reliably and straightforwardly identify receptor conformations with a significantly improved success rate in virtual screening. Based on the knowledge of a few known binders, we employed a custom use of ROC curve AUC values to identify and confirm an ensemble of three highly predictive receptor conformations for every investigated target system. To independently verify the strength of our approach, we conducted blind virtual screening of a large compound library on five cancer targets, selected promising potential small molecule inhibitors, and performed radiometric HotSpot kinase assays to test their inhibition of enzyme activity. Through the utilization of this method, we identified several high-affinity, novel inhibitors for five cancer-related target systems (AKT1, CSF1R, EGFR, FGFR1, and VEGFR2). A summary of the workflow for this work can be found in Figure 1.

Knowledge of Only a Few Active Compounds Can Confidently Identify Highly Predictive Receptor Conformation. We first investigated whether knowledge of known binders enables reliable identification of predictive receptor conformations for SBDD. Utilizing the Glide SP docking methodology, we evaluated 20 individual target systems and a total of 533 receptor conformations (between 9 and 68 per target, see Table S1) obtained from the PDB (see Figure 1a). For each target system, 9−68 known small molecule binders ("actives") were obtained from the co-crystallized structures and separated into two unique sets of similar molecular weight (actives set A and B). This allowed us to test whether knowledge of as little as five actives is sufficient for predictive receptor identification. Diversity between the two sets of actives was confirmed using Tanimoto coefficients, with none of the target systems having a coefficient above 0.476. The average mutual Tanimoto coefficient of all actives within each target is summarized in Table S5. Actives sets A and B were docked alongside 1000 assumed small molecule decoys (see Figure 1b). It has been shown that lack of similarity between sets of actives and decoys leads to artificially high enrichment in the docking process.[48] Therefore, we evaluated the similarity between the structure-based known actives and the Schrödinger decoy sets based on multiple chemical properties: molecular weight, formal charge, and solubility (see Figure S2). We found that the actives and decoy sets agreed well in all evaluated properties, thereby avoiding any inflated enrichment and supporting our choice of decoy sets. Active/decoy screening was performed for all receptor conformations across the 20 target systems. We calculated the respective true positive rates (TPRs) and false positive rates (FPRs) (see Figure 1c) and plotted the AUC of the generated ROC curves for set A and B of all 533 receptor conformations, as shown in Figure 2. We observed a strong correlation such that target conformations with favorable screening results (high AUCs) for actives of set A also exhibited high success rates (high AUCs) for completely independent actives of set B. Additionally, it was also true that target conformations with poor screening results (low AUCs) for actives of set A exhibited similarly low success rates for independent actives set B. This strongly suggested that the knowledge of as few as five known actives allowed for reliable identification of strongly predictive receptor conformations for virtual screening of unknown compounds. Highly predictive receptor ensembles were identified for each target consisting of the three best performing receptor conformations based on active/decoy screening (Table 1). With a few exceptions, we were able to identify conformations with average AUC > 0.8 for almost all target systems. Ligand-bound conformers accounted

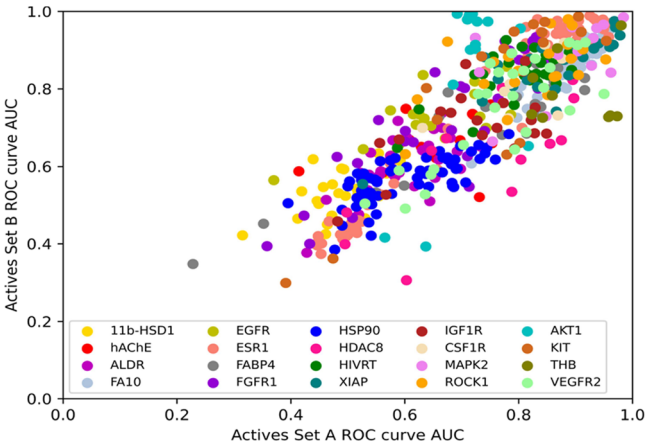


Figure 2. AUC of actives set B plotted against AUC of actives set A for all 533 receptor conformations.

Table 1. Top Three Performing Receptor Conformations Resulting from Active/Decoy Screens

target system	top performing receptor conformations	average AUCs
11b-HSD1	3D4N, 1XU7, 1XU9	0.604, 0.603, 0.594
hAChE	6U34, 3LII, 4EY6	0.756, 0.728, 0.727
ALDR	1X96, 2NVC, 1XGD	0.702, 0.693, 0.683
FA10	1IQI, 1IQN, 1IQL	0.961, 0.953, 0.946
EGFR	1XKK, 4RJ7, 2J6M	0.827, 0.818, 0.771
ESR1	6VNN, 6VMU, 3DT3	0.967, 0.964, 0.962
FABP4	5D4A, 3P6F, 5D47	0.872, 0.862, 0.857
FGFR1	6C19, 6C18, 3TT0	0.883, 0.876, 0.874
HSP90	3EKR, 1YC3, 2BYH	0.734, 0.697, 0.693
HDAC8	6ODA, 6ODB, 3F07	0.859, 0.837, 0.823
HIVRT	1TKT, 1TL3, 1TKZ	0.892, 0.876, 0.871
XIAP	3HL5, 3CM2, 2JK7	0.972, 0.969, 0.965
IGF1R	3LVP, 1K3A, 4D2R	0.842, 0.827, 0.791
CSF1R	3BEA, 3DPK, 2I0V	0.900, 0.880, 0.877
MAPK2	3R30, 3KA0, 3M42	0.985, 0.958, 0.947
ROCK1	5WNG, 5WNE, 5UZJ	0.934, 0.929, 0.915
AKT1	3MVH, 3OW4, 3L9M	0.880, 0.860, 0.857
KIT	6XV9, 6GQK, 4U0I	0.946, 0.939, 0.929
THB	1NQ0, 1Q4X, 1NAX	0.972, 0.887, 0.852
VEGFR2	3B8Q, 2RL5, 6GQQ	0.919, 0.908, 0.905

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

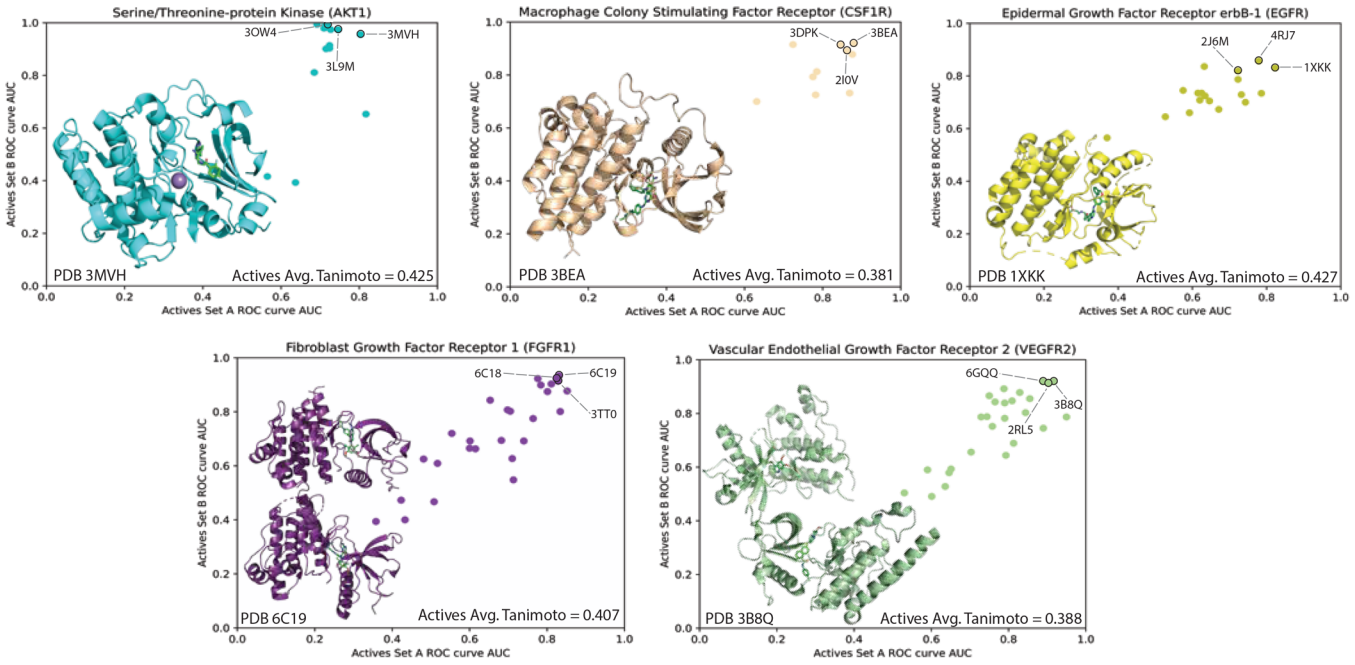


Figure 3. AUC of actives set B plotted against AUC of actives set A for all receptor conformations for five cancer target systems. The best performing receptor conformation is shown in cartoon representation with its respective PDB code. The three most predictive conformations are labeled and the average Tanimoto coefficient is displayed in the bottom right corner.

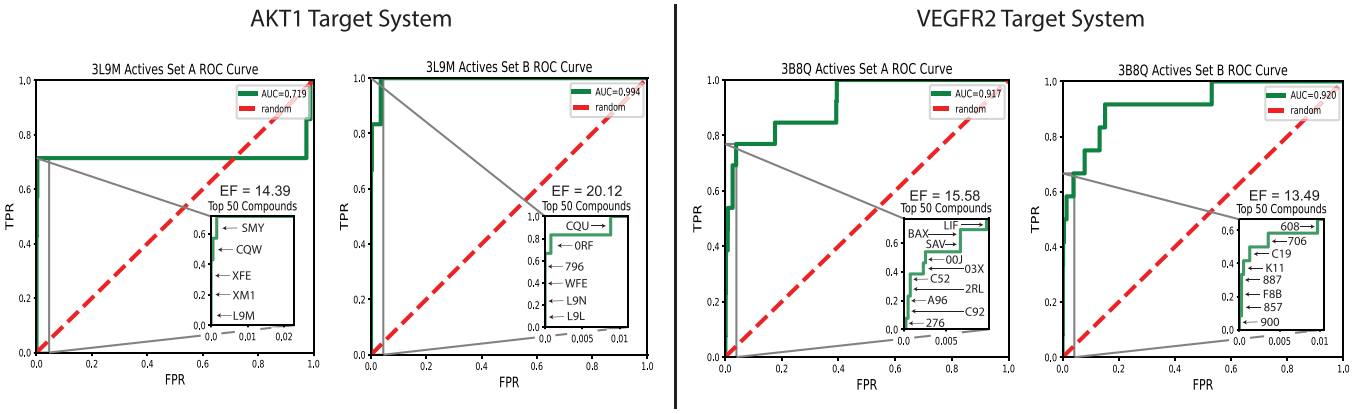


Figure 4. ROC curves depicting receptor conformer performance in active/decoy screening using Glide SP. The TPR is plotted against the FPR showing where the docking algorithm placed the known active compounds with respect to the decoys in each conformer. The inset region shows the TPR and FPR of the top 50 docked compounds, the calculated enrichment factor (EF), and highlights the true positives, along with their identities designated as the respective RCSB ligand ID code.

for 97% (58/60) of the top performing conformations among all target systems. Of the 20 target systems, apo (ligand unbound) conformers were identified for 14 targets. However, for only two targets (hAChE and ALDR), a single apo conformer ranked in the top three performing conformations. Conformer 3LII ranked second for hAChE with an averaged AUC of 0.728 and conformer 1XGD ranked third for ALDR with an averaged AUC of 0.683. These results support previous work that concluded ligand-bound conformations are significantly more suited for virtual screening studies than apo conformers.[7] Individual AUC plots of all target systems with the top three performing conformations labeled are shown in Figure S3.

The ChEMBL database is a manually curated database of bioactive molecules with druglike properties.[49,50] It contains molecules that have been shown to bind a specific target with data routinely extracted from seven core journals.[51] As such, there are a significantly larger number of actives reported in ChEMBL vs PDB. For two targets, ALDR and FABP4, we obtained all associated small molecules from the ChEMBL database and filtered the compounds to remove violators of two or more rules of Lipinski's rule of five. In addition, we filtered the compounds for any identified PAINS compounds (Table S6). In order to explore how a larger set of actives impacts receptor selection in our protocol, we examined the impact of active/decoy docking for the ALDR and FABP4 target systems with associated active small molecules from the ChEMBL database. We performed active/decoy docking with a filtered set of ChEMBL actives and the Schrödinger decoy set (400 MW) for all receptor conformations. We observed a similar trend as with the structure-based actives. Receptor conformations with favorable screening results (high AUCs) for actives of set A also exhibited high success rates (high AUCs) for completely independent actives of set B. Furthermore, we observed an overlap in the ten most predictive receptor conformations

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

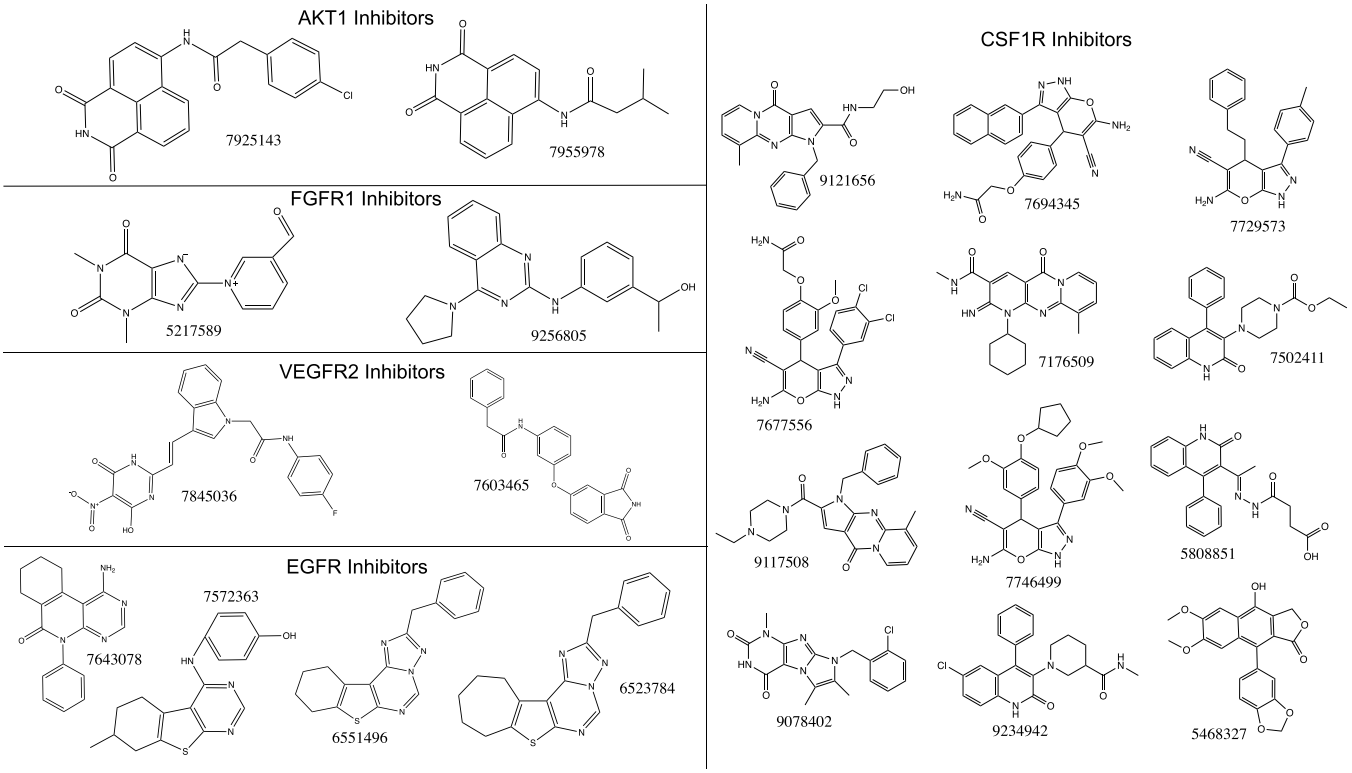


Figure 5. 2D structures of inhibitors along with their unique ChemBridge IDs.

identified based on average ROC AUC for both the structure-based actives and ChEMBL actives. Target system ALDR had an overlap of 50% of the top ten receptor conformations, while target system FABP4 had an overlap of 60% of the top ten receptor conformers for both the structure-based actives and ChEMBL actives (Figure S4). Given the similar trends in the active/decoy results and the overlap of the top performing receptor conformations, we concluded that our smaller structure-based active sets are a sufficient representation of active chemical space to identify predictive receptor conformations.

Receptor Selection Strategy Successfully Identified 22 Novel Cancer Inhibitors. After identification of the top three predictive receptor conformations for all 20 target protein systems, we sought to experimentally validate our receptor selection method through a blind screening of five cancer-related targets with the goal of identifying novel kinase inhibitors. Five targets (AKT1, CSF1R, EGFR, FGFR1, and VEGFR2) were selected for the blind screening based on low mutual sequence identities and high AUC values. The respective sequence identities of all five targets compared to one another are summarized in Figure S1. With the exception of one target (EGFR, average AUC range 0.827−0.771), all receptors utilized for the blind screening were found to have high average AUC values: AKT1 0.880−0.857, CSF1R 0.900−0.877, FGFR1 0.883−0.874, and VEGFR2 0.919−0.905. Individual plots of the active/decoy screening performance of all receptor conformations of the five cancer targets are shown in Figure 3. The top performing conformers identified by AUC values also showed high enrichment factors (EFs). Averaged EFs for a single receptor ranged from 6.02 to 17.26 across all five cancer targets, with as many as 10 true actives being identified within the top 50 predicted compounds (PBD 3B8Q). Example ROC curves, including EF analysis, are shown in Figure 4, whereas all ROC curves of the top three performing conformations for the five cancer-related targets are provided in Figure S5. Additionally, the averaged ROC AUC and averaged EF values for all receptor conformations of the five cancer related targets are provided in Table S7. While receptor selection was based on the ROC AUC values, the EF served as a secondary validation metric to support our choice of conformers. For instance, in four of the target systems (AKT1, CSF1R, EGFR, and VEGFR2) two of the top three conformers based on ROC AUC overlapped with the top three conformers based on EF. For the target FGFR1, the top three conformers were identical for both metrics.

We utilized the Glide SP docking algorithm to dock 633,076 Lig-Prepped small molecules from the ChemBridge EXPRESS-Pick Collection into the top three receptor conformations for each of the five cancer targets. Previous work by our group has found that Glide SP provided the greatest accuracy based on ligand self-docking compared to Glide XP and AutoDock Vina.[52−54] Glide SP is also known to be less computationally expensive compared to Glide XP. Therefore, given the size of our ligand set, the Glide SP protocol was a suitable choice. The small-molecule library was prefiltered based on MW and logP, PAINS functional groups, and Lipinski's rule of five (see Figure 1e). After the docking simulations, we created a ranked ordering of compounds based on the averaged Z-score of each ligand across all receptor conformations for each target. The Z-score metric is based on the following equation:

$$\text{Z score} = \frac{x_{\text{ligand}} - \overline{x}_{\text{receptor}}}{\sigma_{\text{receptor}}}$$

where x_{ligand} corresponds to the docking score of an individual ligand, $\overline{x}_{\text{receptor}}$ is the average docking score of all ligands in the respective receptor conformation, and σ_{receptor} corresponds to

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

the standard deviation of the docking scores of the receptor conformation. Using the Z-score ranking prevented compound selection bias based on different docking score ranges of individual receptor conformations. Another popular method of compound selection is consensus scoring,[55] where the results from multiple docking score functions are combined by averaging the score or the rank of each ligand obtained from the individual programs. However, this method can fail if just one of the docking programs results in poor performance due to training-set dependencies and score function parameterization.[56] Additionally, depending on the number of score functions utilized, consensus scoring quickly becomes computationally expensive for large databases of ligands. Furthermore, consensus scoring does not address the potential for bias in compound selection if one receptor conformer scores compounds remarkably well compared to other screened receptor conformations. Therefore, to remove bias and balance computational costs, we utilized the Z-score metric for compound selection.

The top 50 compounds for each of the five cancer targets were ordered from ChemBridge and subsequently tested in vitro using Radiometric HotSpot Kinase Assays (see Figure 1f). Radiometric based filtration binding assays are well suited for detecting kinase reactions.[46] We utilized RBC's HotSpot kinase assay, a miniaturized assay platform optimized for high-throughput screening. In total, 250 compounds (50 compounds per target system) were initially tested in a single dose duplicate at a concentration of 10 μM. Compounds which showed an average enzyme activity relative to DMSO controls of less than 55% were identified as promising inhibitor hits. In total, 22 compounds were identified as hits (see Figure 5). We subsequently obtained IC_{50} values for all 22 compounds using a 10-dose measurement. The average percent enzyme activity and IC_{50} values for all inhibitors are reported in Table 2.

We successfully identified a total of 22 compounds that exhibited strong (low μM−nM) inhibition. For target systems AKT1, FGFR1, and VEGFR2, we identified two novel potent inhibitors each. We identified hit compounds with low and sub-μM inhibition (7955978 IC_{50} = 6.47 μM and 7925143 IC_{50} = 0.17 μM) for AKT1. Interestingly, compound 7925143 was one of the most potent known inhibitors for this specific kinase. For FGFR1, we identified inhibitors with low μM affinity, compound 5217589 (IC_{50} = 33.1 μM), and compound 9256805 (IC_{50} = 10.8 μM). Additionally, for kinase VEGFR2, we identified novel inhibitors with low μM IC_{50} values. Compounds 7845036 and 7603465 exhibited IC_{50} values of 6.24 and 5.38 μM, respectively.

We identified a total of four highly potent inhibitors for target system EGFR, with IC_{50} values ranging from 7.96 nM to 8.2 μM. Compound 7572363 was our most potent inhibitor (IC_{50} = 7.96 nM) and compound 7643078 also displayed sub-μM inhibition (IC_{50} = 0.252 μM). Additionally, compounds 6551496 and 6523784 exhibited low micromolar inhibition with IC_{50} values of 5.7 and 8.2 μM, respectively. We were extraordinarily successful at identifying small molecule inhibitors for target system CSF1R, where 24% of the tested compounds showed inhibition. Of the 12 identified CSF1R inhibitors, nine compounds possessed IC_{50} values under 10 μM, ranging from 1.41 to 9.34 μM. The most potent inhibitors for this kinase (5468327, 7729573, and 7502411) had IC_{50} values of 1.41, 1.42, and 1.76 μM, respectively. Furthermore, compounds 9078402, 7694345, 9234942, 9117508, 7176509, and 9121656 possessed IC_{50} values of 2.30, 2.51, 4.16, 5.06, 5.58, and 9.34 μM, respectively. Additionally, three CSF1R inhibitors exhibited low micromolar IC_{50} values in the range of 10−15 μM (compound 5808852 [IC_{50} = 10.8 μM], compound 7746499 [IC_{50} = 12.8 μM], and compound 7677556 [IC_{50} = 14.7 μM]). All compounds were docked into the ATP binding site of their respective target protein. Therefore, we would expect the identified inhibitors to be a mixture of type I and type II kinase inhibitors.[57] However, further structure information is required to confirm the specific protein−ligand interactions.

Interestingly, the majority of identified inhibitors for target system CSF1R belong to three clusters. Based on the docked poses of the ligands, we hypothesized general structure activity relationships. Cluster I was composed of compounds 7694345, 7729573, 7677556, and 7749499. Hydrogen-bond interactions were observed for all compounds in this cluster with residues K616, T663, E664, Y665, and C666 depending on the individual compound's docked pose. For compound 7729573 (the most potent of this cluster), we observed cation−Pi interactions between the positively charged residue R801 and the polarizable electron cloud of the benzene ring. Additionally, the nitrile functional group was also able to form H-bond interactions with residues A800 and K616 depending on the docked orientation. Compound 7694345 was slightly less potent than 7729573; however, 7694345 consistently formed a Pi−Pi interaction with F797 and one of the benzene rings. Compound 7694345 was also more solvent-exposed when compared to 7729573, presumably leading to a decrease in potency. Compounds 7677556 and 7746499 also formed the same Pi−Pi interaction as 7694345. However, these two compounds on average formed less hydrogen bond interactions, contributing to the decrease in their respective potencies. The ligand interaction diagrams for the docking poses of this cluster of small molecules are displayed in Figure S6. Cluster II consisted of compounds 5808851, 9234942, and 7502411. All three compounds formed hydrogen bond interactions primarily with residues K616, E664, and C666 depending on the docked pose. Additionally, all three

Table 2. Inhibitor Compound HotSpot Data

system	ChemBridge compound ID	average % enzyme activity	IC_{50} (M)
AKT1	7925143	34.11	1.74×10^{-7}
AKT1	7955978	47.88	6.47×10^{-6}
EGFR	7643078	17.14	2.52×10^{-7}
EGFR	6551496	35.38	5.70×10^{-6}
EGFR	6523784	43.25	8.20×10^{-6}
EGFR	7572363	1.12	7.96×10^{-9}
FGFR1	5217589	50.53	3.31×10^{-5}
FGFR1	9256805	48.06	1.08×10^{-5}
CSF1R	7694345	16.42	2.51×10^{-6}
CSF1R	7677556	41.29	1.47×10^{-5}
CSF1R	7502411	34.39	1.76×10^{-6}
CSF1R	7746499	46.30	1.28×10^{-5}
CSF1R	9121656	45.33	9.34×10^{-6}
CSF1R	7729573	16.69	1.42×10^{-6}
CSF1R	5808851	42.84	1.08×10^{-5}
CSF1R	7176509	40.27	5.58×10^{-6}
CSF1R	9234942	28.28	4.16×10^{-6}
CSF1R	9078402	30.52	2.30×10^{-6}
CSF1R	5468327	26.06	1.41×10^{-6}
CSF1R	9117508	42.53	5.06×10^{-6}
VEGFR2	7845036	29.57	6.24×10^{-6}
VEGFR2	7603465	53.96	5.38×10^{-6}

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

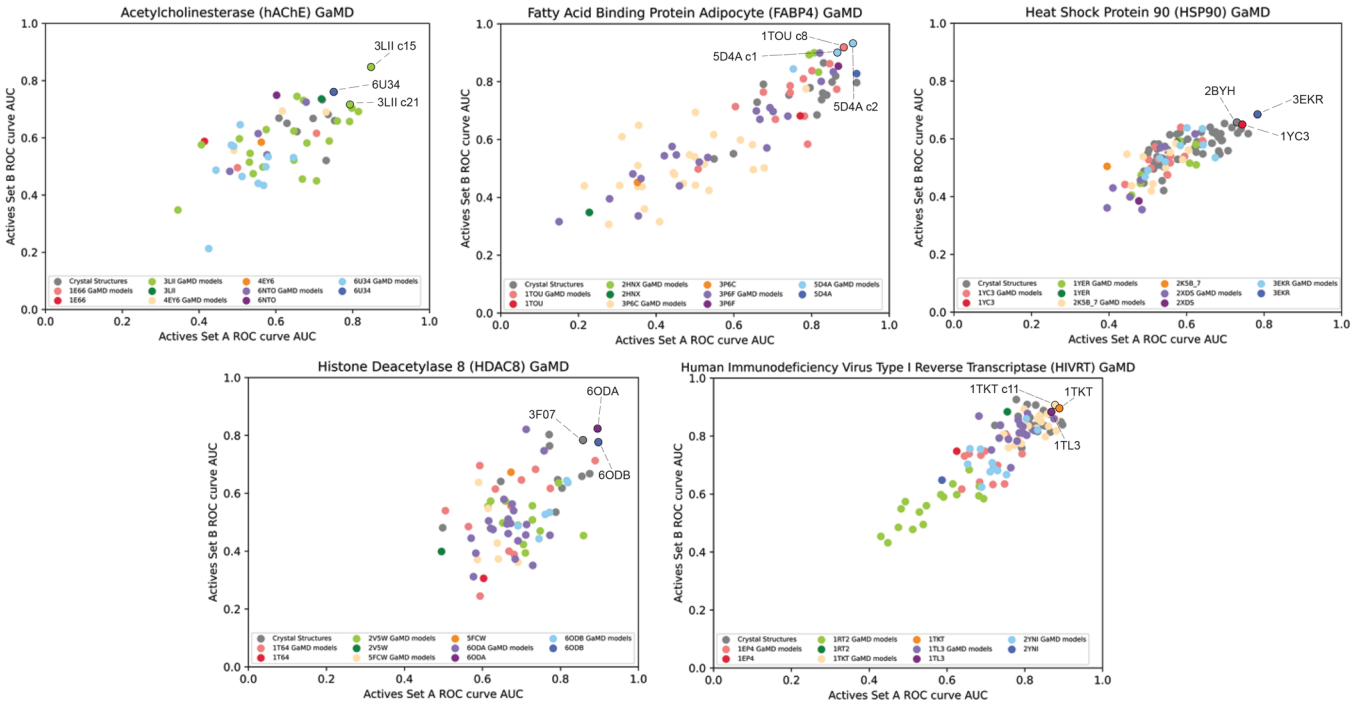


Figure 6. AUC of actives set B plotted against AUC of actives set A for all receptor conformations (crystal structures (gray) and GaMD clustered conformations (colors)) for five target systems. Three most predictive conformations based on the averaged AUC are labeled. Crystal conformers that served as the initial structure for 300 ns GaMD are labeled in a darker shade and the clustered conformers of the corresponding GaMD simulations are displayed in a lighter shade of the same color.

compounds were predicted to have Pi−Pi interactions with residue F797. Compound 5808851 was more solvent-exposed compared to the other compounds in this cluster, which may influence the decreased potency. Compounds 7502411 and 9234942 have very similar protein interactions; however, we suggest that the increased potency of compound 7502411 was a result of the carboxylic oxygen atom being in a more favorable position for hydrogen bonding with residue D670. The carboxyl group in compound 9234942 was out of plane of the piperidine ring, making the hydrogen bond interaction less likely. The ligand interaction diagrams for the docking poses of cluster II small molecules are displayed in Figure S7. Cluster III consisted of compounds 9117508 and 9121656. Both compounds formed hydrogen bond interactions with residue C666; however, discerning the difference in potency based on docking pose alone was difficult. Therefore, more structural information is needed. The ligand interaction diagrams for the docking poses of cluster III small molecules are displayed in Figure S8.

In addition to assessing the potency of the identified inhibitors, we also characterized their structural uniqueness. This property was determined by calculating the Tanimoto coefficients between each inhibitor and all known actives utilized in the active/decoy screening process with respect to the individual target systems. Additionally, we calculated the Tanimoto coefficient for each inhibitor to all associated small molecules in the ChEMBL database for each respective target system. The Tanimoto coefficients for each inhibitor and the active compound with the highest similarity can be found in Table S8. Tanimoto coefficients of $\geq$0.6 are considered to be structurally similar, while coefficients below this threshold are considered to be dissimilar. Only one hit compound was found to have a coefficient above 0.6 when compared to the most similar structure-based active (CSF1R 7176509). Whereas eight hit compounds were found to have a coefficient below 0.6 when compared to the most similar ChEMBL actives. While the identified inhibitors were rather dissimilar from the structure-based actives, the majority of the identified inhibitors had their respective chemotypes represented in the ChEMBL actives. However, all identified compounds are newly identified inhibitors for their respective targets and have the potential for further lead optimization.

For all five tested kinases, the methods described in this work have led to the identification of potent inhibitors in the low nM−μM range. For the AKT1 target system, the inhibitors identified in this work were more potent than compounds in previous studies.[58,59] The hit compounds identified for target system FGFR1 were rather dissimilar to the structure-based actives as well as the actives in the ChEMBL database. The highest Tanimoto coefficient for FGFR1 hit compound 5217589 was 0.439 and 0.546 for the structure-based actives and ChEMBL actives, respectively. For FGFR1 hit compound 9256805, the highest Tanimoto coefficients were 0.429 and 0.572 with respect to the structure-based actives and ChEMBL actives, respectively. These inhibitors potentially offer new chemotypes for lead compounds. Additionally, for target system EGFR, we identified four small molecule inhibitors that are considerably more potent that compounds identified in previous screenings.[60] For the target system CSF1R, we identified the highest number of low μM inhibitors. However, the chemotypes of our identified hit compounds were already represented in the ChEMBL database. Nonetheless, the compounds in this work are still viable candidates for further lead optimization. For VEGFR2, we successfully identified inhibitors in the low μM range. The potency of our compounds are within a similar range of a virtual screening conducted by Lee and coworkers,[61] who screened five different small molecule databases and identified 10 small molecule inhibitors with IC_{50} values in the low μM range. Based on the success rate of our receptor and compound

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

selection methodology for five independent kinase targets, we believe that our protocol can be successfully applied to variety of target systems.

GaMD Generated Highly Predictive Receptor Conformations. In addition to using experimental structures obtained from the protein data bank, we explored the use of enhanced computational sampling techniques to create potentially highly predictive receptor conformations that may not be represented by current experimental structures. Thus, for a subset of five target systems (hAChE, FABP4, HSP90, HDAC8, and HIVRT), we performed 300 ns GaMD simulations on five crystal structures, respectively. The crystal structures used in the GaMD simulations were based on their respective actives/decoys screening performance. Of the five conformers per target, we selected the two highest average AUC values, the two lowest average AUC and the conformer with the average AUC closest to the mean of all receptors for that target system. In total, we performed 25 separate 300 ns GaMD simulations and clustered the resulting trajectories. Through clustering, we identified an additional 284 conformations for active/decoy screening that were distinct from those of the crystal structures. Figure 6 shows AUC values of actives sets A and B for all receptor conformations (crystal structures and GaMD clustered conformations). Table S9 contains the mean ROC AUC for each initial structure for the GaMD simulations (crystal or NMR model) and the mean ROC AUC for the top three clustered models for each simulation. We successfully generated GaMD conformers with a higher average AUC for 60% of the original crystal or NMR structures. For kinases hAChE, FABP4, and HIVRT, we successfully generated conformers that ranked among the top three most predictive receptor conformations. For the hAChE system, the clustered conformers 3LII c15 and 3LII c21 scored as the most and third most predictive conformation based on averaged ROC AUC, respectively. Meanwhile, three clustered conformers of target FABP4 scored as most predictive receptor conformations based on the averaged ROC AUC. Conformer 5D4A c2 ranked the highest, followed by conformers 1TOU c8 and 5D4A c1. For target system HIVRT, one clustered conformer ranked second in the top three receptor conformations (1TKT c11). This impressively demonstrates that GaMD simulations have the potential to generate conformations that are highly predictive for drug discovery and even outperform the best experimental structures. For the majority of receptor conformations, we were able to use GaMD simulations to generate improved conformers based on the averaged ROC AUC regardless of the initial performance of the crystal or NMR conformation. Therefore, we believe GaMD to be a valuable tool in generating highly predictive receptor conformations.

The results of our receptor characterization via active/decoy docking showed that 98% of the highly predictive conformers (AUC > 0.8) were ligand-bound experimental structures or clustered GaMD conformers where a ligand-bound experimental structure served as the initial frame of the simulation. However, ligand-bound structures are not always available for a target of interest. We investigate the utility of GaMD enhanced sampling to generate highly predictive conformers by simulating a ligand-bound conformation of an originally apo crystal structure. We examined the active/decoy docking results for four receptor conformations across three targets: ALDR−1XGD, hAChE−3LII, HSP90−1YER, and 1YES. For each apo crystal structure, we obtained the docked pose of the highest ranked structure-based active, which served as the initial frame for our GaMD simulations. The simulations were performed for 300 ns; we then generated clustered conformers based on clustering the trajectories utilizing all frames. After clustering, we performed active/decoy docking for all generated GaMD conformers and obtained ROC AUC for actives sets A and B, respectively (Figure S9). For the ALDR and hAChE targets, we found GaMD conformers that ranked as the best receptor conformation based on the averaged ROC AUC when compared to the original apo crystal structure. The most predictive conformer for ALDR was 1XGD c3 with an average AUC of 0.831, and for hAChE, the most predictive conformer was 3LII c10 with an average AUC of 0.769. For target HSP90, we generated models that outperformed the crystal structure of 1YER; however, none of the conformers generated for 1YES outperformed the apo crystal structure. These results indicate that even in cases where there are no ligand-bound crystal structures, GaMD methods can generate highly predictive conformers based on a docked ligand pose in an original apo crystal structure.

CONCLUSIONS

Here, we explored the role of conformational selection in virtual screening. We have successfully demonstrated that knowledge of known actives significantly improves virtual screening. Previously, we had employed a similar strategy for a single target system, cardiac troponin with noteworthy success.[52,62] In this work, we verified the strength and generalizability of this approach over a diverse selection of target systems. We successfully developed an easy-to-follow protocol of assessing receptor conformation predictability based on knowledge of a few known actives for a particular target protein. We verified our protocol for 538 conformers obtained via X-ray crystallography, NMR, and cryo-EM across 20 diverse target systems. For all 20 targets, the top three most predictive conformers were identified based on the averaged ROC curve AUC from two independent sets of known actives. A blind screening using the ChemBridge EXPRESS-Pick library was performed for five cancer-related targets, with experimental testing of the top 50 ranked compounds via radiometric based filtration binding assays. Twenty-two novel kinase inhibitors were identified in the low μM−nM range, with several compounds being strong candidates for further lead optimization. The inhibitors identified in this study were not only shown to be highly potent but also structurally unique compared to the known actives utilized in the active/decoy screenings. Additionally, we demonstrated the effectiveness of enhanced sampling methods such as GaMD for creating highly predictive clustered conformers. For a subset of five distinct target systems, an additional 284 clustered conformers were created from 25 independent 300 ns GaMD simulations. For three targets (hAChE, FABP4, and HIVRT), clustered conformers ranked in the top three performing conformers.

While this work was focused on improving selection and sampling of a target's conformational space, we acknowledge that conformational selection may not be the driving force in ligand binding for all target systems. Knowledge of a specific target's biological function is crucial to the success of any virtual screening study and other mechanisms of enzyme−substrate interaction, such as induced fit, which may play an important part in governing ligand binding. In future work, we plan to evaluate the impact of alternate docking strategies such as induced fit docking on the success of virtual screens. However, our protocol has been shown to work very well for almost all

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

targets in the benchmark set, suggesting that conformational selection is a crucial mechanism of ligand−protein interactions for many receptors. The simplicity and adaptability of this work permits the protocol to be applied to any system of interest, with confidence of identifying novel inhibitors. We have provided our protocols, analysis scripts, and clustered models in pdb format as the Supporting Information to allow users to follow a similar protocol to identify the most predictive conformations for their targets of interest. Additionally, we have provided quality control data of the tested ChemBridge compounds in the Supporting Information. This includes LC−MS spectra for a representative number of hit compounds identified from our in vitro experiments and ^{1}H NMR spectra for all ordered compounds.

Data and Software Availability. The protein structural information for the 20 target systems was gathered from the Protein Data Bank at https://www.rcsb.org/. We have included the exact PDB IDs for the respective structures in the Supporting Information (SI Table 1). Additionally, all generated protein receptor conformations from our GaMD simulations are available for download in the SI_information.zip file. All docking simulations were performed using Schrödinger's Glide SP software. We have also included example scripts for data processing and plotting the ROC AUC in the SI_information.zip file. Quality control data of the ordered compounds from ChemBridge (LC−MS spectra and NMR spectra) can also be found in the SI_information.zip file. Additionally, we have included PDB formatted models for all hit compounds docked into their respective receptor conformations and their respective ligand interaction diagrams.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00848.

Additional experimental details and methods (PDF)

Protocols and analysis scripts; all generated protein receptor conformations from GaMD simulations; quality control data of the ordered compounds from ChemBridge; and docked poses of identified hit compounds (ZIP)

VSP inhibitor SMILES (CSV)

■ AUTHOR INFORMATION

Corresponding Author

Steffen Lindert − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0002-3976-3473; Phone: 614-292-8284; Email: lindert.1@osu.edu; Fax: 614-292-1685

Author

Eric R. Hantz − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0002-2588-7619

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00848

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank members of the Lindert Lab for helpful discussions relating to this work. Additionally, we would like to thank the Ohio Supercomputer Center[63] for valuable resources. We would also like to thank the staff of Reaction Biology Corporation for their help and support.

■ REFERENCES

(1) U.S. Senate. Committee on Finace. *Research and Development in the Pharmaceutical Industry.* Available from: Congressional Budget Office (Accessed August 06, 2021).

(2) Leelananda, S. P.; Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **2016**, *12*, 2694−2718.

(3) Yu, W.; MacKerell, A. D. Computer-Aided Drug Design Methods. *Methods Mol. Biol.* **2017**, *1520*, 85−106.

(4) Hammes, G. G.; Chang, Y. C.; Oas, T. G. Conformational selection or induced fit: a flux description of reaction mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 13737−13741.

(5) Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J. Am. Chem. Soc.* **2002**, *124*, 5632−5633.

(6) Amaro, R. E.; Baron, R.; McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 693−705.

(7) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* **2010**, *50*, 186−193.

(8) Nichols, S. E.; Baron, R.; Ivetac, A.; McCammon, J. A. Predictive power of molecular dynamics receptor structures in virtual screening. *J. Chem. Inf. Model.* **2011**, *51*, 1439−1446.

(9) Ellingson, S. R.; Miao, Y.; Baudry, J.; Smith, J. C. Multi-conformer ensemble docking to difficult protein targets. *J. Phys. Chem. B* **2015**, *119*, 1026−1034.

(10) Xu, M.; Lill, M. A. Utilizing experimental data for reducing ensemble size in flexible-protein docking. *J. Chem. Inf. Model.* **2012**, *52*, 187−198.

(11) Swift, R. V.; Jusoh, S. A.; Offutt, T. L.; Li, E. S.; Amaro, R. E. Knowledge-Based Methods To Train and Optimize Virtual Screening Ensembles. *J. Chem. Inf. Model.* **2016**, *56*, 830−842.

(12) Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J. F.; Montes, M. Multiple structures for virtual ligand screening: defining binding site properties-based criteria to optimize the selection of the query. *J. Chem. Inf. Model.* **2013**, *53*, 293−311.

(13) Yoon, S.; Welsh, W. J. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88−96.

(14) Choi, J.; Choi, K. E.; Park, S. J.; Kim, S. Y.; Jee, J. G. Ensemble-Based Virtual Screening Led to the Discovery of New Classes of Potent Tyrosinase Inhibitors. *J. Chem. Inf. Model.* **2016**, *56*, 354−367.

(15) 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. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(17) *Schrödinger Release 2021−1: Maestro*; Schrödinger, LLC: New York, NY, 2021.

(18) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221−234.

(19) *Schrödinger Release 2021−1: Epik*; Schrödinger, LLC: New York, NY, 2021.

(20) 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.-Aided Mol. Des.* **2007**, *21*, 681−691.

(21) (a) Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25*, 621−627. (b) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **2015**, *10*, 733−755.
(22) *The PyMOL Molecular Graphics System*; Schrödinger, LLC.
(23) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281−296.
(24) *Schrödinger Release 2021−1: LigPrep*; Schrödinger, LLC: New York, NY, 2021.
(25) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591−604.
(26) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Aust. J. Chem.* **2015**, *7*, 20.
(27) 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. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(28) 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.; et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(29) 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. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(30) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, 2825−2830.
(31) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E., III; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, M. K. H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Jin, C.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xue, Y.; York, D. M.; Zhao, S.; Kollman, P. A.*Amber20*; University of California: San Francisco, 2020.
(32) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(33) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.* **2015**, *11*, 3584−3595.
(34) Vassetti, D.; Pagliai, M.; Procacci, P. Assessment of GAFF2 and OPLS-AA General Force Fields in Combination with the Water Models TIP3P, SPCE, and OPC3 for the Solvation Free Energy of Druglike Organic Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1983−1995.
(35) Jorgensen, W. L.; Madura, J. D. Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water. *J. Am. Chem. Soc.* **1983**, *105*, 1407−1413.
(36) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **1992**, *32*, 523−535.
(37) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(38) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(39) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.*A density-based algorithm for discovering clusters in large spatial databases with noise*. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96); 1996, pp 226−231.
(40) *UVa FASTA Server: LALIGN/PLALIGN*; University of Virgina, 2014.
(41) 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. *J. Med. Chem.* **2010**, *53*, 2719−2740.
(42) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technol.* **2004**, *1*, 337−341.
(43) Landrum, G.*RDKit: Open-Source Cheminformatics Software*; 2016.
(44) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(45) Kim, S. S.; Aprahamian, M. L.; Lindert, S. Improving inverse docking target identification with Z-score selection. *Chem. Biol. Drug Des.* **2019**, *93*, 1105−1116.
(46) Ma, H.; Deacon, S.; Horiuchi, K. The challenge of selecting protein kinase assays for lead discovery optimization. *Expert Opin. Drug Discovery* **2008**, *3*, 607−621.
(47) Anastassiadis, T.; Deacon, S. W.; Devarajan, K.; Ma, H.; Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29*, 1039−1045.
(48) Stein, R. M.; Yang, Y.; Balius, T. E.; O'Meara, M. J.; Lyu, J.; Young, J.; Tang, K.; Shoichet, B. K.; Irwin, J. J. Property-Unmatched Decoys in Docking Benchmarks. *J. Chem. Inf. Model.* **2021**, *61*, 699−714.
(49) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612−W620.
(50) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.
(51) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(52) Coldren, W. H.; Tikunova, S. B.; Davis, J. P.; Lindert, S. Discovery of Novel Small-Molecule Calcium Sensitizers for Cardiac Troponin C: A Combined Virtual and Experimental Screening Approach. *J. Chem. Inf. Model.* **2020**, *60*, 3648−3661.
(53) Lu, Y.; Papa, J. L.; Nolan, S.; English, A.; Seffernick, J. T.; Shkolnikov, N.; Powell, J.; Lindert, S.; Wozniak, D. J.; Yalowich, J.; et al. Dioxane-Linked Amide Derivatives as Novel Bacterial Topoisomerase Inhibitors against Gram-Positive. *ACS Med. Chem. Lett.* **2020**, *11*, 2446−2454.
(54) Lu, Y.; Vibhute, S.; Li, L.; Okumu, A.; Ratigan, S. C.; Nolan, S.; Papa, J. L.; Mann, C. A.; English, A.; Chen, A.; et al. Optimization of TopoIV Potency, ADMET Properties, and hERG Inhibition of 5-Amino-1,3-dioxane-Linked Novel Bacterial Topoisomerase Inhibitors: Identification of a Lead with. *J. Med. Chem.* **2021**, *64*, 15214−15249.
(55) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687

Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(56) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(57) Roskoski, R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacol. Res.* **2016**, *103*, 26−48.
(58) Fratev, F.; Gutierrez, D. A.; Aguilera, R. J.; Tyagi, A.; Damodaran, C.; Sirimulla, S. Discovery of new AKT1 inhibitors by combination of. *J. Biomol. Struct. Dyn.* **2021**, *39*, 368−377.
(59) Chuang, C. H.; Cheng, T. C.; Leu, Y. L.; Chuang, K. H.; Tzou, S. C.; Chen, C. S. Discovery of Akt kinase inhibitors through structure-based virtual screening and their evaluation as potential anticancer agents. *Int. J. Mol. Sci.* **2015**, *16*, 3202−3212.
(60) Lee, J. H.; Lin, W. C.; Wen, T. K.; Wang, C.; Lin, Y. T. Inhibiting two cellular mutant epidermal growth factor receptor tyrosine kinases by addressing computationally assessed crystal ligand pockets. *Future Med. Chem.* **2019**, *11*, 833−846.
(61) Lee, K.; Jeong, K. W.; Lee, Y.; Song, J. Y.; Kim, M. S.; Lee, G. S.; Kim, Y. Pharmacophore modeling and virtual screening studies for new VEGFR-2 kinase inhibitors. *Eur. J. Med. Chem.* **2010**, *45*, 5420−5427.
(62) Aprahamian, M. L.; Tikunova, S. B.; Price, M. V.; Cuesta, A. F.; Davis, J. P.; Lindert, S. Successful Identification of Cardiac Troponin Calcium Sensitizers Using a Combination of Virtual Screening and ROC Analysis of Known Troponin C Binders. *J. Chem. Inf. Model.* **2017**, *57*, 3056−3069.
(63) Ohio Supercomputer Center; *Ohio Technology Consortium of the Ohio Board of Regents*, 1987.

https://doi.org/10.1021/acs.jcim.2c00848
J. Chem. Inf. Model. 2022, 62, 5675−5687