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

Eric R. Hantz[1], Steffen Lindert[1,*]
[1]Department of Chemistry and Biochemistry, Ohio State University, Columbus, OH, 43210

* Correspondence to:
Department of Chemistry and Biochemistry, Ohio State University
2114 Newman & Wolfrom Laboratory, 100 W. 18th Avenue, Columbus, OH 43210
614-292-8284 (office), 614-292-1685 (fax)
lindert.1@osu.edu

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 $\mu M - nM$ 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 non-obvious 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 [8,9,10]. 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 [7-12, 13, 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 (identification of known/assumed decoy molecules) based on the ranked ordering by 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 drug-like and lead-like 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^{TM} 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 $\mu M - nM$ range. Our method resulted in a 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 two 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 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), 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 protein data bank [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 a 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 3-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 $10 \times 10 \times 10$ Å^3 inner search space, while the periphery of the ligand was able to extend out to $20 \times 20 \times 20$ 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 3-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, actives set A and B. Diversity of small molecules between sets A and B was confirmed using the mutual 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 g/mol 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 subjected 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, 28]. 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 receiver operating characteristic (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) [29]. 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:

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

where $N_{decoys} = 1{,}000$ and $N_{actives}$ corresponds to the number of actives in either set A or B for each of the target systems.

Enhanced Sampling with Gaussian Accelerated Molecular Dynamics (GaMD)

In addition to using experimental structures from the DUD·E and PDB databases, we wanted to explore whether non-experimental conformations can exhibit high predictability as well. In order to account for protein conformational flexibility that may not be sufficiently represented in the protein data bank, additional protein receptor conformations were obtained from 300 ns Gaussian accelerated Molecular Dynamics (GaMD) production simulations performed with Amber20 [30-32]. 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)[33] and AMBER's Antechamber [30] 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[34] water molecules in a 10 Å octahedron and neutralized with sodium ions. All GaMD simulations were performed using ff14SB[31].

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[35]. 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 300ns 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[36]. Periodic boundary conditions were under an NPT ensemble with a pressure set at 1 bar using a Berendsen barostat[37] 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 fourth frame was analyzed resulting in 18,750 frames available for clustering. The density-based clustering algorithm (DBScan)[38] 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**. 238 clustered conformers 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 [39]. Mutual sequence identities for the five targets are shown in **Figure S1**.

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 drug-like 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[40], and those violating Lipinski's rule of five[41]. 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 RDKit [42] 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 [43], 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[40]. 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, 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.* [44]. 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 HotSpot^{TM} Kinase Assays

Experimental testing of the identified 250 potential cancer target inhibitors was performed by RBC using their HotSpot^{TM} 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[45]. Substrate was prepared in a base reaction buffer consisting of 10 mM Hepes (pH 7.5), 10 mM $MgCl_2$, 1mM 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 cofactor $MnCl_2$ 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. 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 minutes at room temperature [46]. 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 3-fold serial dilution starting at 50 μM, except for compound 7572363 which was tested at a 3-fold dilution starting at 10 μM due to the compound's high potency. Control compound, Staurosporine, was tested in 10-dose IC_{50} mode with 4-fold serial dilution starting at 20 μM. Experimental conditions for the IC_{50} experiments can be found in **Table S4**.

Results and Discussion

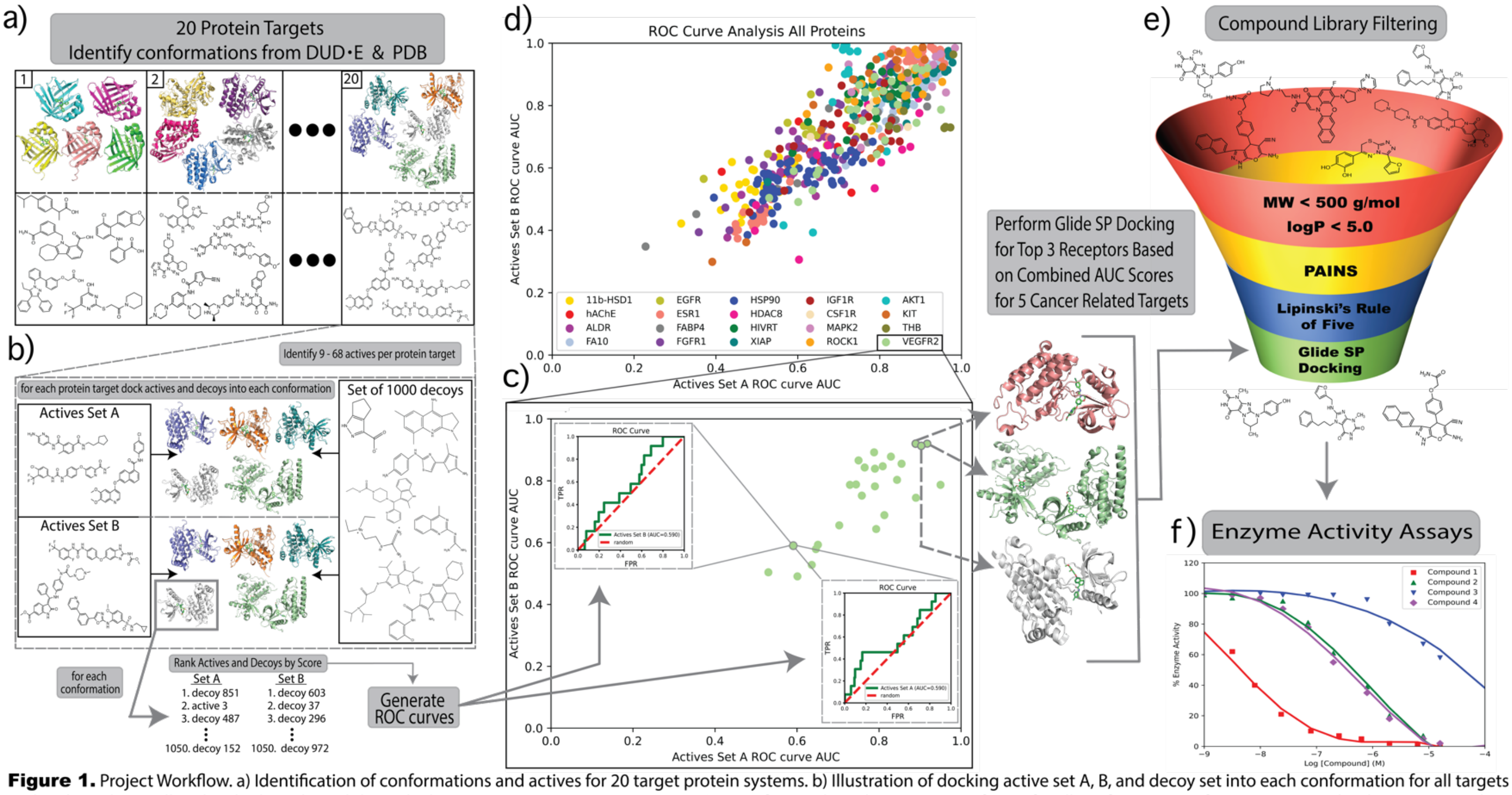

Figure 1. Project Workflow. a) Identification of conformations and actives for 20 target protein systems. b) Illustration of docking active set A, B, and decoy set into each conformation for all targets and ranking the small molecules based on Glide SP docking score to generate ROC curves. c) AUC of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

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 anticancer agents for five target systems (AKT1, CSF1R, EGFR, FGFR1, 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 protein data bank (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**). 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

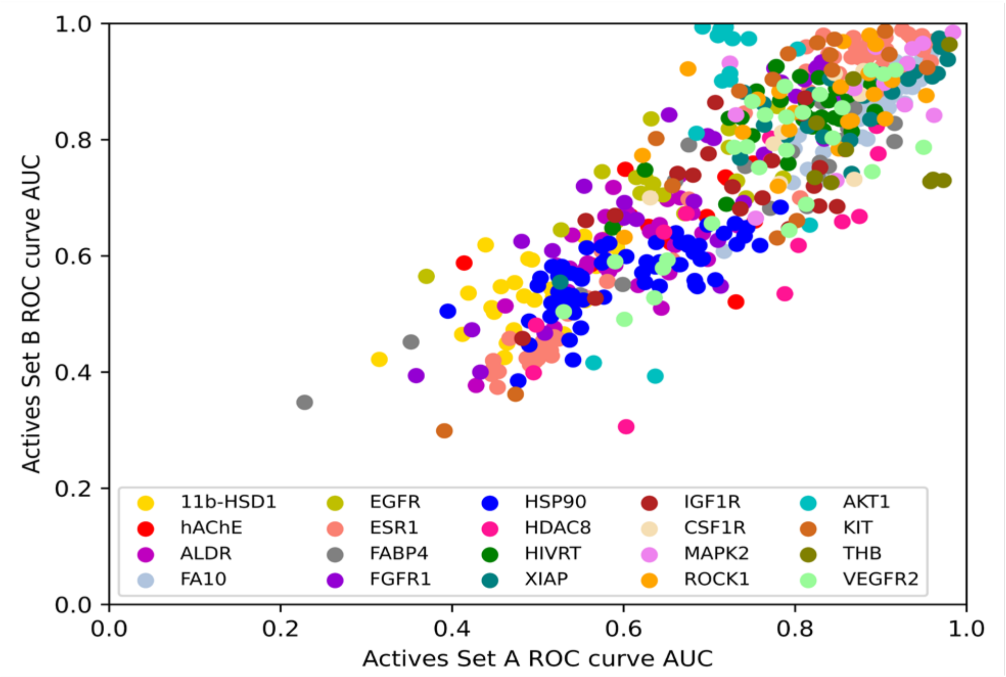

Figure 2. AUC of actives set B plotted against AUC of actives set A for all 533 receptor conformations.

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 for 97% (58/60) of the top performing conformations amongst 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 S2**.

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

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 anticancer agents. Five

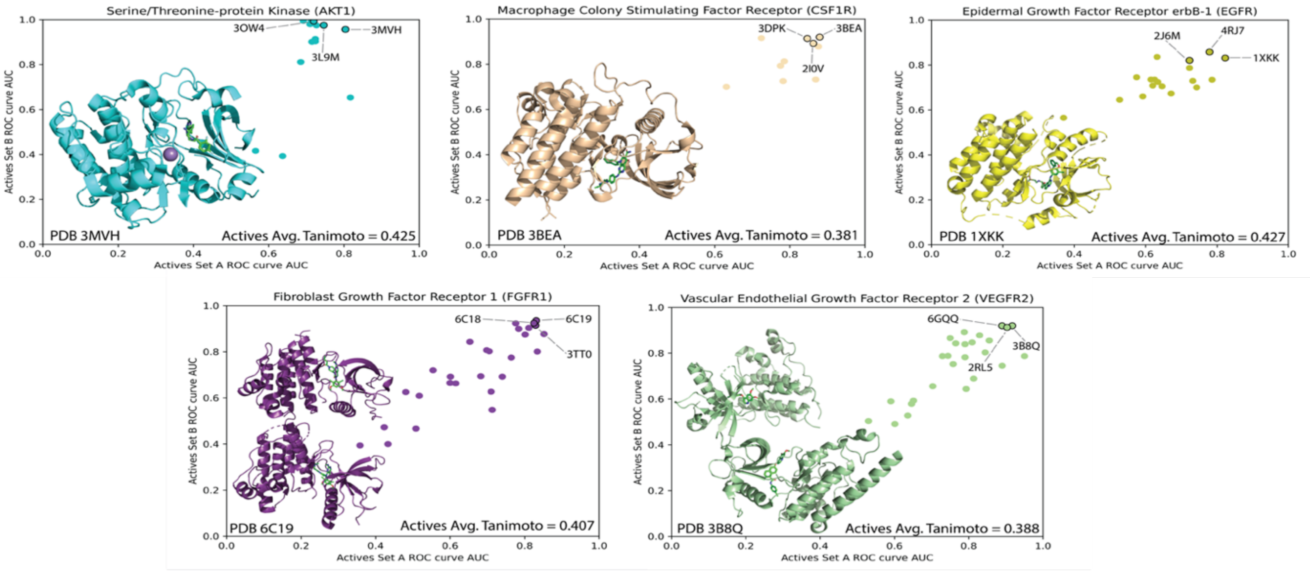

Figure 3. AUC of actives set B plotted against AUC of actives set A for all receptor conformations for five cancer target systems. The best performing receptor conformation is shown in cartoon representation with its respective PDB code. The three most predictive conformations are labeled and the averaged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

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 – 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 S3**.

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. 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

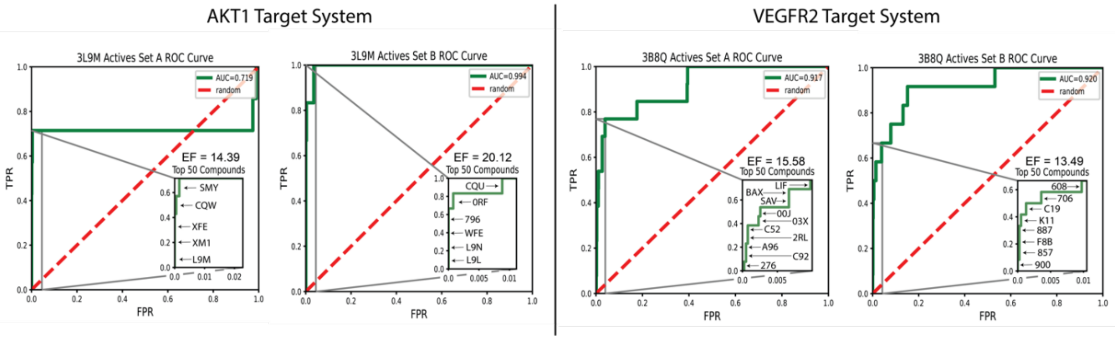

Figure 4. ROC curves depicting receptor conformer performance in active/decoy screening using Glide SP. The TPR is plotted against the FPR showing where the docking algorithm placed the known active compounds with respect to the decoys in each conformer. The inset region shows the TPR and FPR of the 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.

each ligand across all receptor conformations for each target. Using the Z-score ranking prevented compound selection bias based on different docking score ranges of individual receptor conformations. 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[45]. We utilized RBC's HotSpot^{TM} 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 $\mu M - nM$) 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 μM and 5.38 μM, respectively.

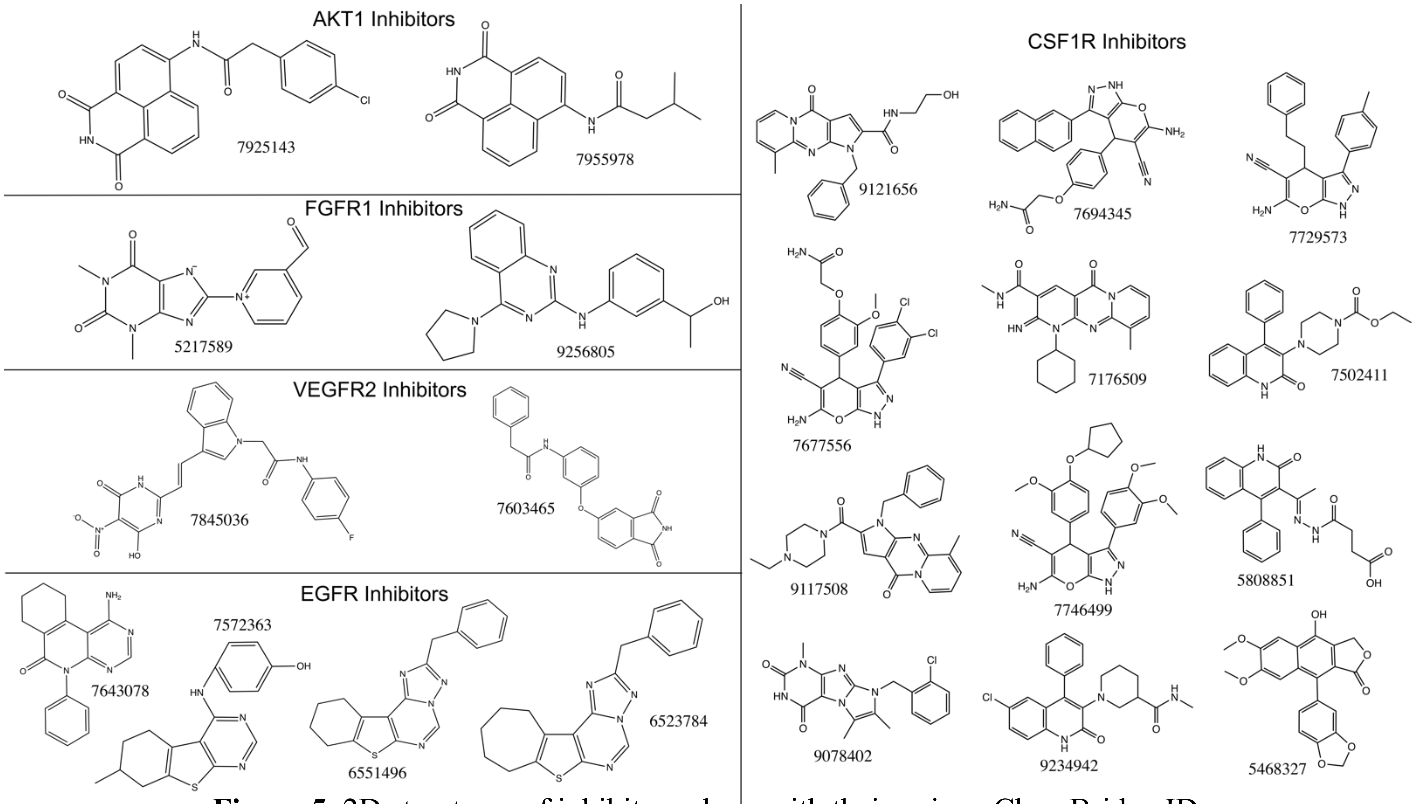

Figure 5. 2D structures of inhibitors along with their unique ChemBridge IDs.

We identified a total of four highly potent inhibitors for target system EGFR., with IC_{50} values ranging from 7.96 nM – 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 μM 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 – 9.34 μM. The most potent inhibitors for this kinase (5468327, 7729573, and 7502411) had IC_{50} values of 1.41 μM, 1.42 μM, and 1.76 μM, respectively. Furthermore, compounds 9078402, 7694345, 9234942, 9117508, 7176509, and 9121656 possessed IC_{50} values of 2.30 μM, 2.51 μM, 4.16 μM, 5.06 μM, 5.58 μM, 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]).

Table 2. Inhibitor Compounds HotSpot^{TM} 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}

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. The average Tanimoto coefficients for each inhibitor can be found in **Table S6**. Tanimoto coefficients of ≥ 0.6 are considered to be structurally similar, while coefficients below this threshold are considered to be dissimilar. The greatest coefficient was found to be 0.428 for compounds 7572363 and 7176509, and the lowest coefficient being 0.299 for compound 7955978. This data suggests that all of our 22 novel inhibitors are representing significantly novel actives.

For three of the tested kinases, the methods described in this work have led to a meaningful improvement for virtual screening performance as compared to published virtual screens and using the same stringent criteria for defining compounds as inhibitors (hits) as detailed above. For instance, Fretev and coworkers identified 0/9 tested compounds for AKT1 showing inhibition rates greater than 45% at compound concentrations of 10 μM [47]. Chuang, *et al.* also performed a virtual screening to identify AKT1 inhibitors [48]. Applying our hit selection criteria to this study results in the identification of only two hit compounds (a46 and a48) with IC_{50} values of 11.1 μM and 9.5 μM, respectively. Notably, the hit compounds identified here are slightly more potent. Ravindranathan and coworkers performed a virtual screening for FGFR1 which initially identified one hit compound when using our inhibitor criteria [49]. The EGFR kinase was utilized in a virtual screening study by Lee and coworkers[50]. This study proposed a significantly more intensive methodology using consensus scoring across 11 different scoring functions, resulting in the identification of four compounds having low μM IC_{50} values ranging from $1.53 \pm 0.15\ \mu M$ – $12.52 \pm 0.37\ \mu M$. In comparison, while we also identified four inhibitors, our methods were considerably less costly and time intensive as we utilized only one scoring metric. Additionally, the inhibitors identified in the present work are considerably more potent. Despite an extensive literature search, we could not locate any comparable virtual screens with subsequent experimental verification of potential inhibitors for target system CSF1R. For VEGFR2, a virtual screening was performed by Lee and coworkers [51]. Five different small molecule databases were used for screening (Key Organics, Maybridge, OTAVA, Life Chemicals, and Asinex), and 10 compounds were found to show low μM inhibition with IC_{50} values ranging from 1.6 – 10 μM. The ten inhibitors spanned three databases with four compounds coming from the OTAVA library, three belonging to the Asinex library, and three compounds originating from the Life Chemicals library. Therefore, we found our methods to have comparable success rates identifying compounds of similar potency (IC_{50} values below 10 μM) when evaluating a single library. Further analysis was performed to determine the novelty of our hits compared to the compounds identified by Lee. Tanimoto coefficients were calculated between the two sets of hit compounds (summarized in **Table S7**), with the averaged Tanimoto coefficients for ChemBridge compounds 7845036 and 7603465 being 0.395 and 0.335, respectively. Thus, supporting the assertion that the inhibitors identified in this work are novel actives. Additionally, compared to other recent virtual screening studies across numerous target systems, our method showed a higher virtual screening success rate given its more rigorous definition of inhibitor compounds [52] [53] [54] [55] [56] [57] [58] [59] [60] [61] [62].

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

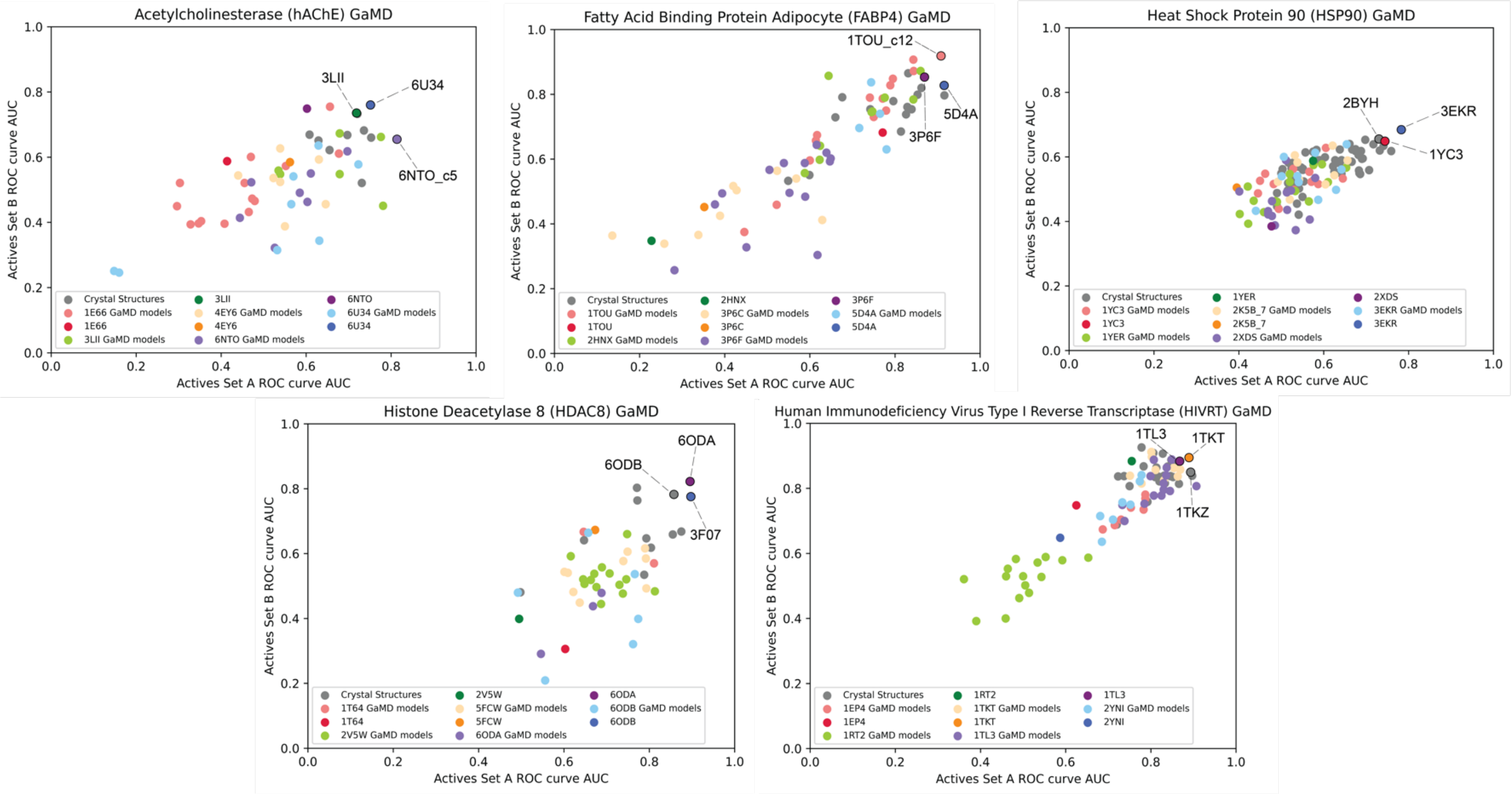

Figure 6. AUC of actives set B plotted against AUC of actives set A for all receptor conformations (crystal structures (grey) and GaMD clustered conformations (colors)) for five target systems. The three most predictive conformations based on averaged AUC are labeled. The crystal conformers which served as the initial structure for 300ns 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.

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 238 conformations for actives/decoys 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). For kinases hAChe and FABP4, we successfully generated conformers that ranked among the top three most predictive receptor conformations. For the hAChe system, the clustered conformation 6NTO_c5 was the second most predictive conformer; while for the FABP4 system, the clustered conformation 1TOU_c12 was the most predictive receptor conformer based on average AUC. 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. We also observed general trends with respect to the predictability of the GaMD conformations. For instance, if the crystal structure used as the initial starting structure of the GaMD simulation performed well (AUC ≥ 0.8), then we always observed a decrease in performance for all clustered conformations. However, clustered conformations resulting from crystal structures that had low or near-mean average AUC values tended to have a wide range of predictability performance with some conformers performing considerably better in the actives/decoys screening than the original crystal structure. Conformers 6NTO_c5 and 1TOU_c12 both derived from the original crystal structure that was closest to the mean average AUC value for their respective target systems. Based on these observations, we hypothesize that selecting crystal structures with near mean actives/decoys screening performance for advanced sampling methods of the target's potential energy surface such as GaMD, would have the highest chance to generate new highly predictive conformers. Therefore, we recommend utilizing conformers with near mean AUC values for initial starting structures for GaMD.

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[63,64]. 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 few know 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. 22 novel kinase inhibitors were identified in the low $\mu M - nM$ 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 238 clustered conformers were created from 25 independent 300ns GaMD simulations. For two targets (hAChE and FABP4) clustered conformers ranked in the top three performing conformers. We also suggest the use of near mean average AUC conformers to serve as initial starting structures of GaMD simulations for the greatest probability of clustering highly predictive receptor conformations.

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, may play an important part in governing ligand binding. However, our protocol has been shown to work very well for almost all targets in the benchmark set, suggesting that conformational selection is a crucial mechanism of ligand-protein interaction for many receptors. Our methods have proven to significantly improve the success rate of virtual screening compared to previous studies of numerous drug targets. 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 supporting information to allow users to follow a similar protocol to identify the most predictive conformations for their targets of interest.

Acknowledgements

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 [65] 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: 06/08/21

(2) Leelananda, S. P.; Lindert, S. Computational methods in drug discovery. *Beilstein J Org Chem* **2016**, *12*, 2694-2718. DOI: 10.3762/bjoc.12.267.

(3) Yu, W.; MacKerell, A. D. Computer-Aided Drug Design Methods. *Methods Mol Biol* **2017**, *1520*, 85-106. DOI: 10.1007/978-1-4939-6634-9_5.

(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* (33), 13737-13741. DOI: 10.1073/pnas.0907195106.

(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* (20), 5632-5633. DOI: 10.1021/ja0260162.

(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* (9), 693-705. DOI: 10.1007/s10822-007-9159-2.

(7) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the selection of experimental protein conformations for virtual screening. *J Chem Inf Model* **2010**, *50* (1), 186-193. DOI: 10.1021/ci9003943.

(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* (6), 1439-1446. DOI: 10.1021/ci200117n.

(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* (3), 1026-1034. DOI: 10.1021/jp506511p.

(10) Xu, M.; Lill, M. A. Utilizing experimental data for reducing ensemble size in flexible-protein docking. *J Chem Inf Model* **2012**, *52* (1), 187-198. DOI: 10.1021/ci200428t.

(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* (5), 830-842. DOI: 10.1021/acs.jcim.5b00684.

(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* (2), 293-311. DOI: 10.1021/ci3004557.

(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* (1), 88-96. DOI: 10.1021/ci0341619.

(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* (2), 354-367. DOI: 10.1021/acs.jcim.5b00484.

(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. *Journal of Medicinal Chemistry* **2012**, *55* (14), 6582-6594. DOI: 10.1021/jm300687e.

(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 Research* **2000**, *28* (1), 235-242. DOI: 10.1093/nar/28.1.235 (acccessed 5/6/2021).

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

(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* (3), 221-234. DOI: 10.1007/s10822-013-9644-8.
(19) ***Schrödinger Release 2021-1*** *: Epik*; Schrödinger, LLC: New York, NY, 2021 (accessed.
(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* (12), 681-691. DOI: 10.1007/s10822-007-9133-z.
(21) 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* (5), 621-627. DOI: 10.1093/bioinformatics/btp036. 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* (5), 733-755. DOI: 10.1038/nprot.2015.043.
(22) *The PyMOL Molecular Graphics System*; Schrödinger, LLC: (accessed.
(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* (1), 281-296. DOI: 10.1021/acs.jctc.5b00864.
(24) ***Schrödinger Release 2021-1*** *: LigPrep*; Schrödinger, LLC: New York, NY, 2021. (accessed.
(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. *Journal of Computer-Aided Molecular Design* **2010**, *24* (6), 591-604. DOI: 10.1007/s10822-010-9349-1.
(26) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015**, *7* (1), 20. DOI: 10.1186/s13321-015-0069-3.
(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* (7), 1750-1759. DOI: 10.1021/jm030644s.
(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* (7), 1739-1749. DOI: 10.1021/jm0306430. 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* (21), 6177-6196. DOI: 10.1021/jm051256o.
(29) 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. *The Journal of Machine Learning Research* **2011**, (12), 2825-2830.
(30) *Amber20*; University of California, San Francisco. , 2021. (accessed.
(31) 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. *Journal of Chemical Theory and Computation* **2015**, *11* (8), 3696-3713. DOI: 10.1021/acs.jctc.5b00255.

(32) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of Chemical Theory and Computation* **2015**, *11* (8), 3584-3595. DOI: 10.1021/acs.jctc.5b00436.
(33) 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. *Journal of Chemical Theory and Computation* **2019**, *15* (3), 1983-1995. DOI: 10.1021/acs.jctc.8b01039.
(34) Jorgensen, W. L.; Madura, J. D. Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water. *Journal of the American Chemical Society* **1983**, *105* (6), 1407-1413. DOI: 10.1021/ja00344a001.
(35) 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* (5), 523-535. DOI: 10.1002/bip.360320508.
(36) 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. *Journal of Computational Physics* **1977**, *23* (3), 327-341. DOI: https://doi.org/10.1016/0021-9991(77)90098-5.
(37) 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. *The Journal of Chemical Physics* **1984**, *81* (8), 3684-3690. DOI: 10.1063/1.448118 (acccessed 2021/03/24).
(38) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. Second Int. Conf. Knowledge Disc. Data Mining (KDD-96)* **1996**, 226-231.
(39) *UVa FASTA Server: LALIGN/PLALIGN*; University of Virgina, 2014. (accessed.
(40) 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* (7), 2719-2740. DOI: 10.1021/jm901137j.
(41) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* **2004**, *1* (4), 337-341. DOI: 10.1016/j.ddtec.2004.11.007.
(42) Landrum, G. RDKit: Open-Source Cheminformatics Software. **2016**.
(43) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **1999**, *39* (5), 868-873. DOI: 10.1021/ci990307l.
(44) Kim, S. S.; Aprahamian, M. L.; Lindert, S. Improving inverse docking target identification with Z-score selection. *Chem Biol Drug Des* **2019**, *93* (6), 1105-1116. DOI: 10.1111/cbdd.13453.
(45) Ma, H.; Deacon, S.; Horiuchi, K. The challenge of selecting protein kinase assays for lead discovery optimization. *Expert Opin Drug Discov* **2008**, *3* (6), 607-621. DOI: 10.1517/17460441.3.6.607.
(46) 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* (11), 1039-1045. DOI: 10.1038/nbt.2017.
(47) 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* (1), 368-377. DOI: 10.1080/07391102.2020.1715835.

(48) 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* (2), 3202-3212. DOI: 10.3390/ijms16023202.
(49) Ravindranathan, K. P.; Mandiyan, V.; Ekkati, A. R.; Bae, J. H.; Schlessinger, J.; Jorgensen, W. L. Discovery of novel fibroblast growth factor receptor 1 kinase inhibitors by structure-based virtual screening. *J Med Chem* **2010**, *53* (4), 1662-1672. DOI: 10.1021/jm901386e.
(50) 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* (8), 833-846. DOI: 10.4155/fmc-2018-0525.
(51) 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* (11), 5420-5427. DOI: 10.1016/j.ejmech.2010.09.002.
(52) de Sousa, A. C. C.; Combrinck, J. M.; Maepa, K.; Egan, T. J. Virtual screening as a tool to discover new β-haematin inhibitors with activity against malaria parasites. *Sci Rep* **2020**, *10* (1), 3374. DOI: 10.1038/s41598-020-60221-0.
(53) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure* **2002**, *10* (7), 1013-1023. DOI: 10.1016/s0969-2126(02)00799-2.
(54) Shoda, M.; Harada, T.; Kogami, Y.; Tsujita, R.; Akashi, H.; Kouji, H.; Stahura, F. L.; Xue, L.; Bajorath, J. Identification of structurally diverse growth hormone secretagogue agonists by virtual screening and structure-activity relationship analysis of 2-formylaminoacetamide derivatives. *J Med Chem* **2004**, *47* (17), 4286-4290. DOI: 10.1021/jm040103i.
(55) Ballester, P. J.; Mangold, M.; Howard, N. I.; Robinson, R. L.; Abell, C.; Blumberger, J.; Mitchell, J. B. Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. *J R Soc Interface* **2012**, *9* (77), 3196-3207. DOI: 10.1098/rsif.2012.0569.
(56) Cai, H.; Liu, Q.; Gao, D.; Wang, T.; Chen, T.; Yan, G.; Chen, K.; Xu, Y.; Wang, H.; Li, Y.; et al. Novel fatty acid binding protein 4 (FABP4) inhibitors: virtual screening, synthesis and crystal structure determination. *Eur J Med Chem* **2015**, *90*, 241-250. DOI: 10.1016/j.ejmech.2014.11.020.
(57) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J Med Chem* **2000**, *43* (3), 401-408. DOI: 10.1021/jm990408a.
(58) Nagarajan, S.; Doddareddy, M.; Choo, H.; Cho, Y. S.; Oh, K. S.; Lee, B. H.; Pae, A. N. IKKbeta inhibitors identification part I: homology model assisted structure based virtual screening. *Bioorg Med Chem* **2009**, *17* (7), 2759-2766. DOI: 10.1016/j.bmc.2009.02.041.
(59) David, B.; Schneider, P.; Schäfer, P.; Pietruszka, J.; Gohlke, H. Discovery of new acetylcholinesterase inhibitors for Alzheimer's disease: virtual screening and. *J Enzyme Inhib Med Chem* **2021**, *36* (1), 491-496. DOI: 10.1080/14756366.2021.1876685.
(60) Li, X.; Zhang, X. X.; Lin, Y. X.; Xu, X. M.; Li, L.; Yang, J. B. Virtual Screening Based on Ensemble Docking Targeting Wild-Type p53 for Anticancer Drug Discovery. *Chem Biodivers* **2019**, *16* (7), e1900170. DOI: 10.1002/cbdv.201900170.
(61) Huang, Y. X.; Zhao, J.; Song, Q. H.; Zheng, L. H.; Fan, C.; Liu, T. T.; Bao, Y. L.; Sun, L. G.; Zhang, L. B.; Li, Y. X. Virtual screening and experimental validation of novel histone deacetylase inhibitors. *BMC Pharmacol Toxicol* **2016**, *17* (1), 32. DOI: 10.1186/s40360-016-0075-8.

(62) Park, H.; Chi, O.; Kim, J.; Hong, S. Identification of novel inhibitors of tropomyosin-related kinase A through the structure-based virtual screening with homology-modeled protein structure. *J Chem Inf Model* **2011**, *51* (11), 2986-2993. DOI: 10.1021/ci200378s.
(63) 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* (12), 3056-3069. DOI: 10.1021/acs.jcim.7b00536.
(64) 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* (7), 3648-3661. DOI: 10.1021/acs.jcim.0c00452.
(65) Ohio Supercomputer Center. *Ohio Technology Consortium of the Ohio Board of Regents,* 1987.