J Comput Aided Mol Des (2018) 32:287–297
DOI 10.1007/s10822-017-0065-y

Optimal affinity ranking for automated virtual screening validated in prospective D3R grand challenges

Bentley M. Wingert[1] · Rick Oerlemans[2] · Carlos J. Camacho[1]

Received: 15 June 2017 / Accepted: 8 September 2017 / Published online: 16 September 2017

Abstract The goal of virtual screening is to generate a substantially reduced and enriched subset of compounds from a large virtual chemistry space. Critical in these efforts are methods to properly rank the binding affinity of compounds. Prospective evaluations of ranking strategies in the D3R grand challenges show that for targets with deep pockets the best correlations (Spearman $\rho \sim 0.5$) were obtained by our submissions that docked compounds to the holo-receptors with the most chemically similar ligand. On the other hand, for targets with open pockets using multiple receptor structures is not a good strategy. Instead, docking to a single optimal receptor led to the best correlations (Spearman $\rho \sim 0.5$), and overall performs better than any other method. Yet, choosing a suboptimal receptor for crossdocking can significantly undermine the affinity rankings. Our submissions that evaluated the free energy of congeneric compounds were also among the best in the community experiment. Error bars of around 1 kcal/mol are still too large to significantly improve the overall rankings. Collectively, our top of the line predictions show that automated virtual screening with rigid receptors perform better than flexible docking and other more complex methods.

Keywords D3R · Drug Design Data Resource · Virtual screening · Affinity ranking · Pose prediction

Introduction

Predicting protein–ligand binding affinities remains a difficult and important area of research in the field of drug design. As massive libraries of small molecules are being developed and synthesized [1, 2], it is increasingly necessary that accurate ranking of compounds' binding affinities be a central part of virtual screening and drug development. To aid in the evaluation and progression of the field of drug discovery, the NIH in partnership with the University of California San Diego (UCSD) initiated the Drug Design Data Resource (D3R) project in 2015 [3]. The challenges thus far have been broken into three sub-challenges that evaluate strategies for pose prediction, ranking affinities, and relative free energy evaluation. When trying to evaluate the ability of a protein to bind to a small molecule, a likely pose must first be generated. This is the first step of docking methods. Docking strategies generally fall into three categories: stochastic methods such as Monte Carlo, systematically searching all available degrees of freedom, and simulation using molecular dynamics methods [4]. The next step is pose scoring, which is often done as part of the docking program. Scoring functions can be classified as belonging to three groups: empirical scoring functions, force-field based scoring functions, and knowledge-based scoring functions [4]. These strategies can provide reasonable accuracy when evaluating large sets of diverse compounds. However, once compounds are identified, ranking and evaluating binding affinities of congeneric compounds remains an open problem in the field.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-017-0065-y) contains supplementary material, which is available to authorized users.

✉ Carlos J. Camacho
ccamacho@pitt.edu

[1] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15261, USA

[2] Department of Drug Design, University of Groningen, Groningen, The Netherlands

In previous efforts, the Camacho lab has developed a number of tools and strategies to aid in rational drug discovery that have successfully been validated in prospective drug discovery challenges. In the 2011 Community Structure-Activity Resource (CSAR), we developed Smina [5], an open-source fork of AutoDock Vina [6] that provides enhanced support for minimization and scoring. In the ACS 2012 Teach-Discover-Treat (TDT) experiment we utilized our virtual screening server ZincPharmer [7] and Smina to predict the best bound structure of a non-triazolopyrimidine inhibitor and most active compounds [8]. We also have developed a number of strategies aimed at identifying ideal receptor structure(s) for docking and/or affinity prediction [3]. Our results from previous CSAR and D3R competitions have shown that selection of an optimal receptor structure(s) is target dependent and an important step for both pose and affinity prediction, particularly for flexible receptors that exhibit diverse conformations. Furthermore, we showed that our rigid receptor docking and/or minimization and scoring functions like Smina can outperform flexible and other more complex methods submitted to these community-wide challenges [3, 9, 10].

The 2015 D3R grand challenge allowed us to develop a number of strategies aimed at identifying ideal receptor structure(s) for pose prediction and/or affinity prediction [9]. Strategies included methods that utilize all available receptor/ligand co-crystals (referred to as "close" methods), all available ligands and a single holo-receptor structure ("min-cross"), or a single receptor/ligand co-crystal ("cross"). The first grand challenge tasked participants with predicting (i) binding poses, (ii) affinity rankings, and (iii) relative affinity values for compounds that interact with two protein targets: heat shock protein 90 (HSP90) and Mitogen-activated protein kinase kinase kinase kinase 4 (MAP4K4). Based on these comprehensive approach to pose and ranking prediction using rigid body docking/minimization, the Camacho group obtained the most accurate poses and best overall affinity and free energy rankings [3, 9].

Here, we present the results of the most recent challenge, the D3R grand challenge 2, with similar tasks as the 2015 grand challenge but for a new protein target, Farnesoid X Receptor (FXR). FXR provided a more challenging structure than the previous proteins due to the flexibility of its hydrophobic binding pocket that displayed significant conformational changes upon ligand binding. The challenges were each split into two stages, with the first stage including all three aforementioned tasks and the second stage consisting of (ii) affinity ranking and (iii) free energy prediction after the release of 36 crystal structures for compounds in the pose prediction problem of stage one. Despite the differences in the target, our approaches again predicted the best overall ranking and absolute free energies. Based on a rigid receptor structure approach, Smina docking and/or minimization of compounds aligned to most chemically similar known bound ligands yielded the best affinity ranking when compared with other methods, including flexible docking. On the other hand, our community best overall free energy evaluation of congeneric compounds entailed a more detailed mapping of interactions that required simulations of receptor flexibility, and a scoring function that explicitly evaluates the solvation of hydrophilic and hydrophobic contacts [11]. These efforts led to slightly better ranking relative to our best performing method in these limited data sets, motivating the inclusion of flexibility to predict more accurate binding free energies.

Methods

Data preparation

A total of 102 compounds were provided from D3R in SMILES format and converted to 3D structures using Open Babel [12]. On our side, we used publicly available ligand-bound structures of human FXR were downloaded from the Protein Data Bank (PDB) [13] (Table S1). These compounds formed the training set for pose prediction evaluation and receptor selection prior to submission. All structures were aligned to the D3R-provided apo structure using the *align* command in PyMOL 1.7.4.5 [14]. This was repeated in stage two when crystal structures for compounds from the pose prediction section of stage one were released. Available IC50 data for compounds (total of 8 unique compounds) was acquired from BindingDB [15] using the DISCO cross-docking server (http://drugquery.csb.pitt.edu/disco/). Each test compound as well as the 27 training compounds were characterized and fell into five different chemical classes: benzimidazoles, isoxazoles, sulfonamides, spiros, and miscellaneous. Figure 1a shows a breakdown of the number of compounds that fell into each category for different datasets. Additionally, examples of scaffolds for each class are shown in Fig. 1b.

Upon alignment of the available crystal structures, two main binding conformations were identified (Fig. 1c). The first was a near-native like conformation which was observed primarily in receptors docked to isoxazole and miscellaneous compounds (left). The second conformation was observed in receptors bound to benzimidazole compounds and is characterized by a shift in two α-helices adjacent to bound compounds (right). While no human FXR structures were available bound to sulfonamide compounds, homologous structures were available (mainly ROR co-crystals such as 5ETH [16], 4WPF [17], and 4WLB [18]) and had binding modes similar to that seen in benzimidazole-bound FXR.

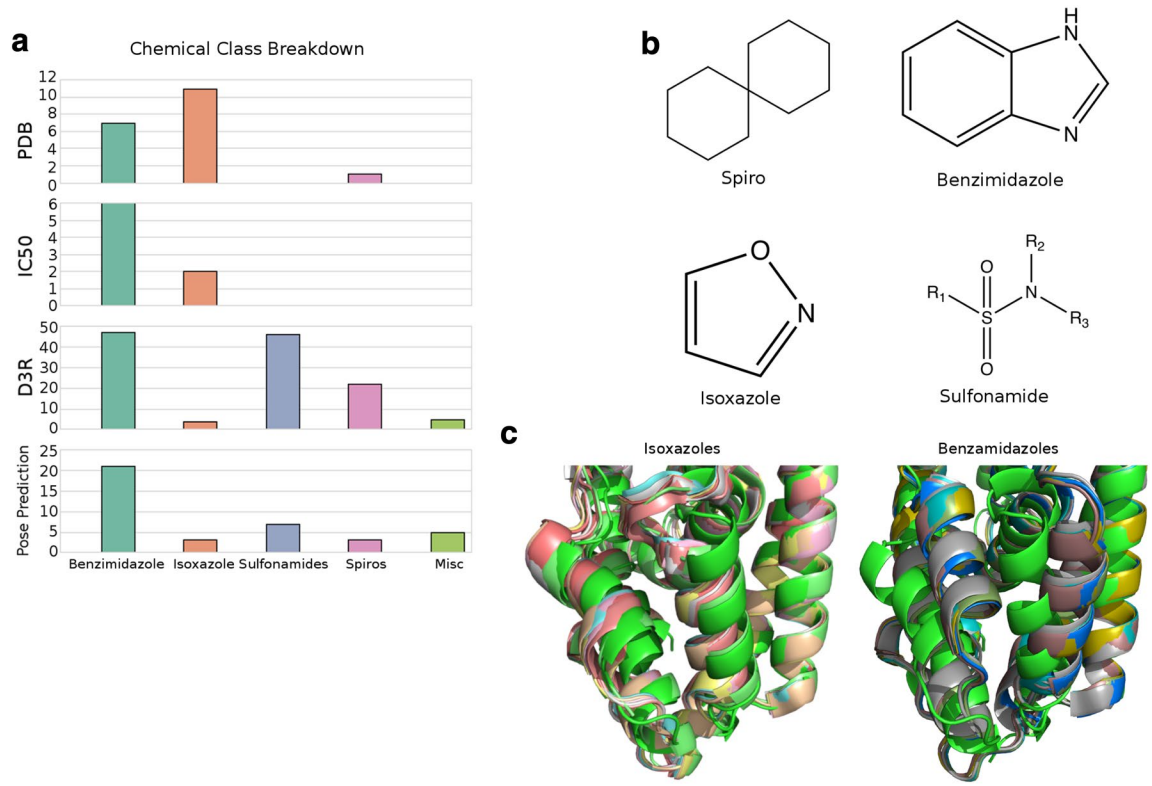

Fig. 1 Available data used for training. **a** Breakdown of number of compounds in each class that were in: (top) publicly available from PDB, second from (top) PDB structures with IC50 data, second from (bottom) test compounds from D3R, and (bottom) compounds for pose prediction challenge. **b** Training and test compounds were from four main chemical class: (1) spiro—upper left, (2) benzimidazole—upper right, (3) isoxazole—bottom left, and (4) sulfonamide—bottom right. **c** Overlaid structures of publicly available FXR structures and provided apo structure (apo structure shown in green in both). On left is apo-like binding mode seen in isoxazoles and on right is shifted binding mode seen in benzimidazoles

Affinity ranking

The main ranking submissions were generated using *align-close, dock-close, min-cross, align-cross*, and *dock-cross* methods [9]. For the "close" methods, each test compound is scored in the receptor corresponding to the most chemically similar training compound; whereas the "cross" methods place each test compound in the same receptor (Table 1). For each test set compound, the most chemically similar training compound was identified using Babel 2.3.2 [12] using Tanimoto score FP3. For align and min methods, 20 conformers for each compound were generated using Omega [19] and then aligned to the target ligand using Open3DALIGN 2.282 [20]. Affinity values were generated by either minimization (align and min methods) or docking (dock methods) using Smina [5]. For docking, up to 20 poses were generated for each compound (--num_modes flag). For both methods, search was constrained to area of receptor centered on the known ligand (--autobox_ligand flag). Compounds were then ranked by the pose with best predicted score. For all software, default parameters and settings were used unless otherwise noted.

As discussed below, choosing the optimal receptor in cross methods is a difficult and extremely important

Table 1 Descriptions of methods for automated affinity ranking

Method	Receptor choice	Pose selection
Align-close	Most similar training compound	Conformers aligned to reference receptor ligand and Smina minimization
Dock-close	Most similar training compound	Smina docking
Min-cross	Same receptor for all compounds	Conformers aligned to closest training ligand and Smina minimization
Align-cross	Same receptor for all compounds	Conformers aligned to reference receptor ligand and Smina minimization
Dock-cross	Same receptor for all compounds	Smina docking

decision. For each cross method (dock, align, or min), we selected receptors from our training set based on three criteria: Spearman coefficient when ranking training compounds with known IC50s, R^2 value with known IC50 compounds, and the percent of training compounds posed within 2.0 Å of the crystal pose. Additionally, rankings were tested both with waters present in the crystal structures and with waters removed. No conserved waters were identified in training structures and removing waters from the receptors generally gave better results on training data and so final submissions were done with no crystal waters present. For dock-cross, receptors were chosen based on best Spearman ρ and best R^2, for min-cross, only best Spearman ρ was chosen, and for align-cross, a receptor was chosen that gave best combination of all three criteria. Results of training evaluation for both stages are shown in Fig. 2.

Free energy evaluation

For stage two of the competition, the challenge consisted of evaluating the relative free energies of binding of two sets of congeneric compounds (Set 1 and Set 2). Co-crystal structures for the compounds from the pose prediction challenge were released (FXR1-36), and each free energy prediction group (Table S2) contained a compound with a solved co-crystal structure (Fig. 3). These compounds were used as templates to build bound models for the full set of congenerics. Both Set 1 and 2 were analyzed in the following manner. Force field parameters for each compound were generated using Antechamber [21] from AMBER14 [22]. Fifty nanosecond molecular dynamics simulations were then run for each compound in the corresponding crystal structure using AMBER14. Simulations were then analyzed and compounds were characterized according to solvation of observed contacts (i.e., hydrogen bonds and hydrophobic interactions) and their solvation (fully, partially, or de-solvated). Relative free energy values for compounds were then assigned based on observed contacts for each simulation using the parametrized contact potential described in [11, 23].

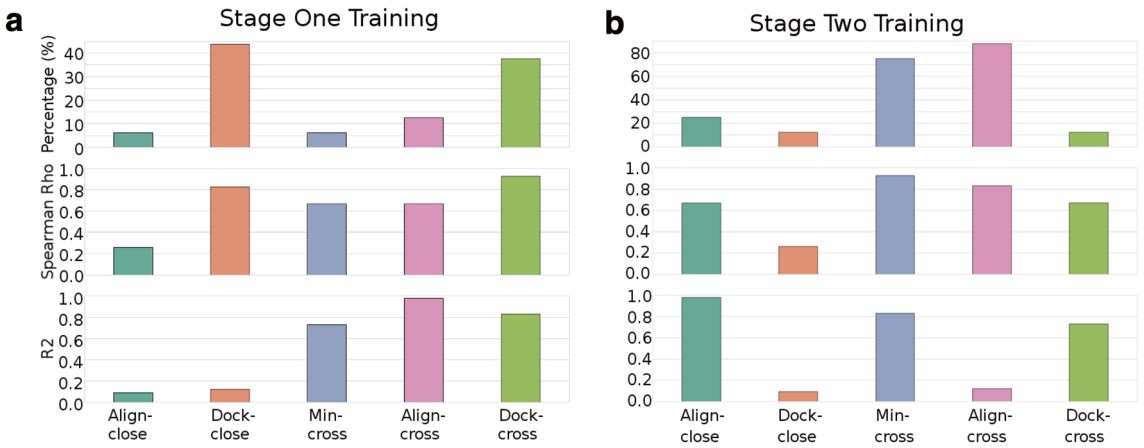

Fig. 2 Training data for **a** stage one and **b** stage two. Methods were evaluated based on Spearman correlation, R^2, and the percent of compounds within 2.0 Å of the cocrystal pose

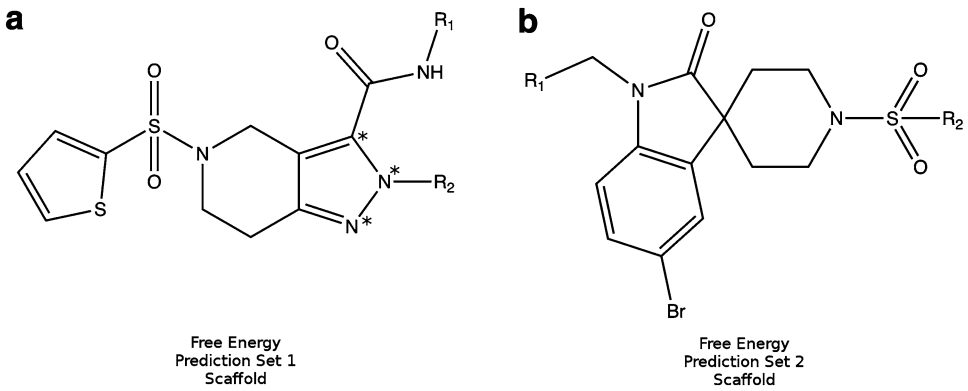

Fig. 3 Compounds used for basis of comparison for prediction of relative free energies of binding for **a** free energy prediction set one (FXR17 scaffold) and **b** free energy prediction set two (FXR10 scaffold). R-group modifications for each set are shown in Tables 2 and 3 respectively

Table 2 SLN representation for R-groups for free energy prediction Set 1

Compound	R1	R2
FXR17	O=C(OCC)C[5]=CC=C(N)C=C@6	C[0]=CC=CC=C@1
FXR45	O=C(OCC)C[5]=CC=C(N)C=C@6	FC(F)(F)OC[4]=CC=CC=C@5
FXR46	NC(C[5]=CC=CC=C@6)=O	C[0]=CC=CC=C@1
FXRF7	NC[2]=CC(C(OCC)=O)=CC=C@3	C[0]=CC=CC=C@1
FXR48	NC[2]=CC=C(CC(OCC)=O)C=C@3	C[0]=CC=CC=C@1
FXR49	NC[2]=CC=C(C(C)=O)C=C@3	C[0]=CC=CC=C@1
FXR91	NC[2]=CC=CC=C@3	C[0]=CC=CC=C@1
FXR93*	NC[2]=CC=CC=C@3	C[0]=CC=CC=C@1
FXR95	NC[2]=CC=C(NC(C)=O)C=C@3	C[0]=CC=CC=C@1
FXR96	NC[2]=CC=C(C(N(C)C)=O)C=C@3	C[0]=CC=CC=C@1
FXR98	NC[2]=CC=C(C(NC)=O)C=C@3	C[0]=CC=CC=C@1
FXR99	NC[2]=CC=C(OC)C=C@3	C[0]=CC=CC=C@1
FXR100	NC[2]=CC=C(S(=O)(N)=O)C=C@3	C[0]=CC=CC=C@1
FXR101	NC[2]=CC=C(C(O)=O)C=C@3	C[0]=CC=CC=C@1
FXR102	NC[2]=CC=C(C(N[9]CCOCC@10)=O)C=C@3	C[0]=CC=CC=C@1

FXR93 has same R-groups as FXR17 but has nitrogen and carbon atoms switched as marked by * in Fig. 3a

Table 3 SLN representation for R-groups for free energy prediction Set 2

Compound	R1	R2
FXR10	CC[2]=CC=C(C(O)=O)C=C@3	C[0]=CC=CS@1
FXR12	CC[2]=CC=C(C(O)=O)C=C@3	ClC[0]=CC=CC=C(@1)S
FXR38	CC[2]=CC=C(C(OC)=O)C=C@3	C[0]=CC=CS@1
FXR41	CC[2]=CC=C(C(OC)=O)C=C@3	ClC[0]=CC=CC=C(@1)S
FXR73	CC[2]=CC=C(O)C=C@3	C[0]=CC=CS@1
FXR74	CC[2]=CC=C(C(O)=O)C=C@3	BrC[0]=CC=CC=C(@1)S
FXR75	CC[2]=CC=NC=C@3	C[0]=CC=CS@1
FXR76	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=CC=CC=C@6
FXR77	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=CC=CC(Cl)=C(@6)Cl
FXR78	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=C(Cl)C=CC=C(@6)Cl
FXR79	CC[2]=CC=CC(C(O)=O)=C@3	C[0]=CC=CS@1
FXR81	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=CC=CC(Cl)=C(@6)C
FXR82	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=CC=CC(Cl)=C(@6)F
FXR83	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=C(Cl)C=CC(Cl)=C@6
FXR84	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=C(F)C=CC=C@6
FXR85	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=C(C)C=CC=C@6
FXR88	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=C(C(F)(F)F)C=CC=C@6
FXR89	CC[2]=CC=C(C(O)=O)C=C@3	SC[5]=CC=C(Cl)C=C@6

Results

2015 grand challenge: affinity ranking

The 2015 grand challenge involved two targets. These were HSP90 and MAP4K4, which had test compound groups of sizes 180 and 18, respectively. The 2015 challenge was also split into two stages, with new co-crystal structures released in stage two, the results of which are shown in Fig. 4. As shown in Fig. 4a, six of the seven best rankings for HSP90 were submitted by our lab. For MAP4K4, we submitted the 5th best ranking, though our methodology could have predicted a better ranking if we would have selected the optimal receptor for screening (see below). This was in part due to the small set of available data, only 8 MAP4K4 structures had IC50 data whereas HSP90 had 69 compounds with IC50 data.

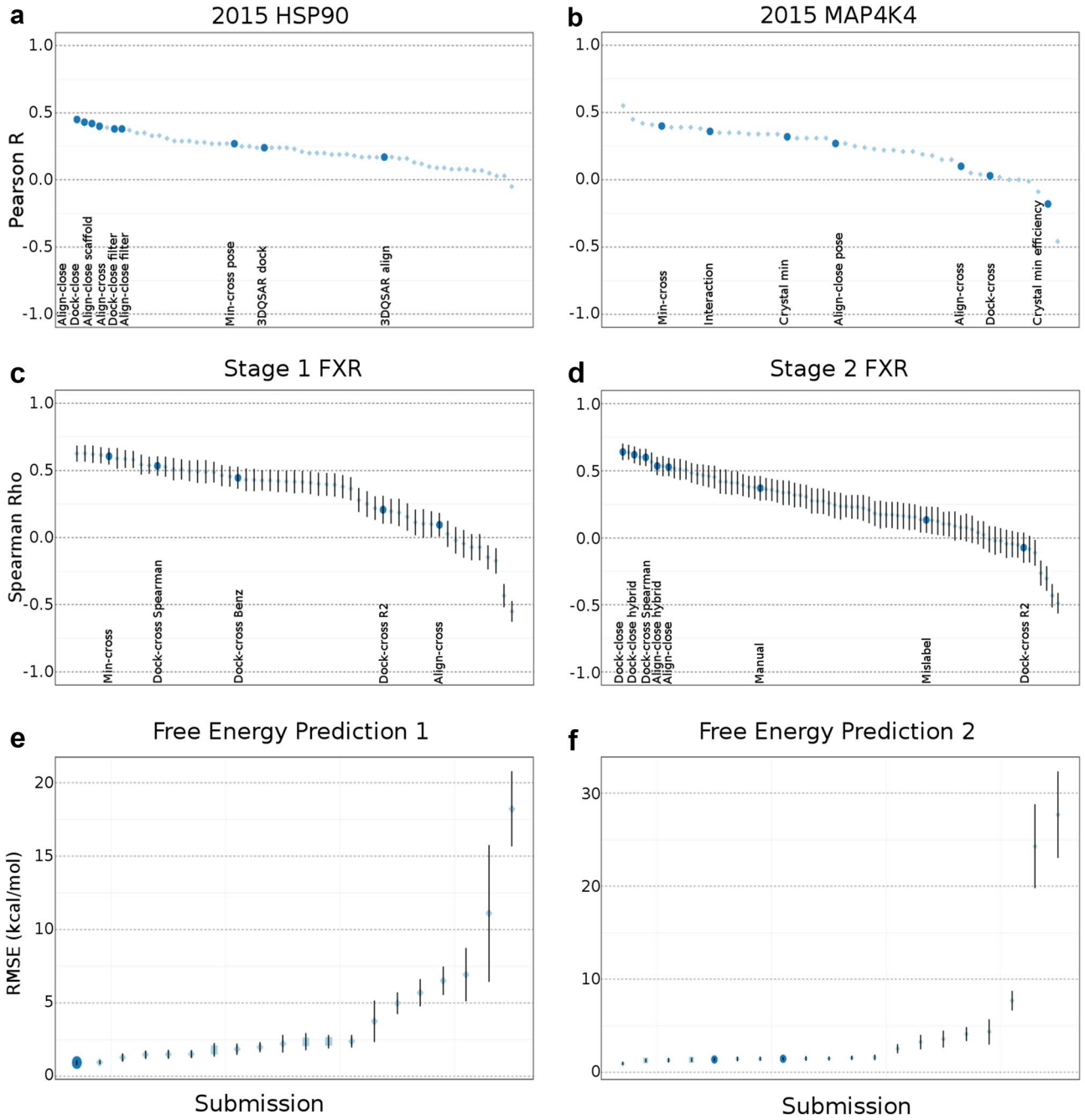

Fig. 4 Results of D3R grand challenges affinity ranking sub-challenge. Our submissions shown as large circles, others as small diamonds (or squares for incomplete submissions). **a** Results for HSP90 challenge. **b** Results for MAP4K4 rankings challenge. **c** Stage one ranking results. **d** Stage two rankings results. **e** Free energy set one prediction results. **f** Free energy set two prediction results

2016 grand challenge: affinity ranking

Given a set of 102 compounds targeting FXR, the challenge was to rank them based on predicted binding affinity. The binding pocket of FXR is large and significantly hydrophobic, including five Met residues that contact known ligands. This present a challenge for pose prediction and affinity ranking as many scoring functions place a large weight on hydrogen bonds, whereas calibration of different hydrophobic contacts such as halogens remains challenging [24, 25]. Stage two differed from stage one in that the 36 co-crystal structures from stage one pose prediction were made available to participants.

Because "cross" methods greatly outperformed "close" methods in our stage one training, we submitted five different rankings for stage one predictions (Fig. 4c). Methods

submitted were align-, dock-, and min-cross methods using receptor chosen from training data as having the best Spearman correlation. Also submitted was dock-cross with receptor chosen for best R^2 value, and dock-cross with best Spearman correlation but using only subset of training data from benzimidazole compounds. The min-cross and dock-cross using best overall Spearman receptors performed the best of our methods in this stage, with both overlapping error bars with the top overall predictions.

For stage two we submitted seven predictions, including dock- and align-close, dock-cross with Spearman and R^2 maxing receptors. Additionally, dock- and align-close lists with rankings of free energy prediction compounds reordered to match our rankings from the free energy evaluation challenge were also submitted. And finally, a ranking was submitted where predicted poses were analyzed and re-ranked manually based on predictions of important interactions observed in free energy prediction analysis. As shown in Fig. 4d, we predicted the best overall prediction and three out of seven top rankings. These three were all dock methods, with the top two being dock-close variants and the third being dock-cross with docking against the Spearman-maximizing receptor.

2016 grand challenge: free energy prediction

Two groups of test compounds were designated for prediction of relative binding affinities. Compounds FXR10, FXR12, and FXR17 had solved crystal structures released for stage two, allowing for comparison of compound behavior in receptor environments that should be close to ideal. As shown in Fig. 4e, f, our results for both groups were amongst the best predictions in the competition, with RSMDs of 0.95 kcal/mol for free energy group one being the top score of that section, and 1.39 kcal/mol for free energy group two being the third best score.

Discussion

The D3R grand challenges have served as an informative view at the current state-of-the-art strategies used in the community for common drug design problems. These challenges are broken down into three sub-challenges that are key problems in the field of rational drug design. The challenge of pose prediction is at the root of this field. A meaningful pose, say, <2 Å is necessary in order for a scoring function to have some hope to select the compound in a virtual screen. The next problem is affinity ranking. Given a library of compounds, sort them by the strength of their interaction with the target of interest. This is an increasingly important challenge as our ability to design and create drug-like compounds improves. With an ever-increasing array of possible drug compounds [1, 19], it is necessary to accurately distinguish quality compounds. Finally, the hit-to-lead problem requires meaningful predictions of relative binding free energies to improve potency and selectivity of hits. The grand challenges provide quality blinded datasets for evaluation and comparison of the wide variety of methods tested by participants. The Camacho lab has taken part in both grand challenges, consistently obtaining best affinity rankings using unbiased strategies. Here we discuss our predictions in the 2016 grand challenge and compare them with similar techniques successfully applied in the 2015 grand challenge.

Affinity ranking

The scoring problem for rational drug design efforts remains a challenge because the accuracy of scoring functions remains incremental (see, e.g., Fig. 4). In previous community-wide competitions it has been shown that top-of-the-line results can be generated with established scoring functions [26] and automated strategies that make appropriate use of known co-crystal structures [3, 4]. Using our previously described strategies we were able to predict affinity rankings with high accuracy. In particular, our dock-close and dock-cross methods had Kendall's tau values of ≥ 0.4 as reported by D3R. Surprisingly, in this year's challenge we found that our dock methods outperformed align methods. This might have been expected for the first stage of the challenge since compound similarity was low in stage one, with an average Tanimoto similarity of 0.58. However, it was also true for the second phase where average Tanimoto similarity increased to 0.94. We would have expected that align and cross methods would improve more when more similar compounds are available. What we found, however, is that dock methods improve the most between stages. This was the case for FXR because docking is a better alternative than minimization in a fully buried rigid pocket. As shown in Fig. 4d, for stage 2, docking against binding pockets with similar ligands (dock-close) led to high quality predictions for ranking compounds based on binding affinity relative to simply minimizing the compounds aligned to same ligands (align-close).

Retrospective analysis

To see if our choice of receptors for cross methods was optimal, we retrospectively calculated Spearman correlation coefficients for all receptors for cross methods against the actual affinity values released after the end of the challenge (Fig. 5). Analysis of dock-cross receptor choice is shown in Fig. 5a. For stage one receptors PDBs 3OLF [27] and 1OSV [28] were chosen due to having best Spearman correlation and R^2 on the training data, respectively. For stage two, PDB 3OLF again resulted in best ranking of training data,

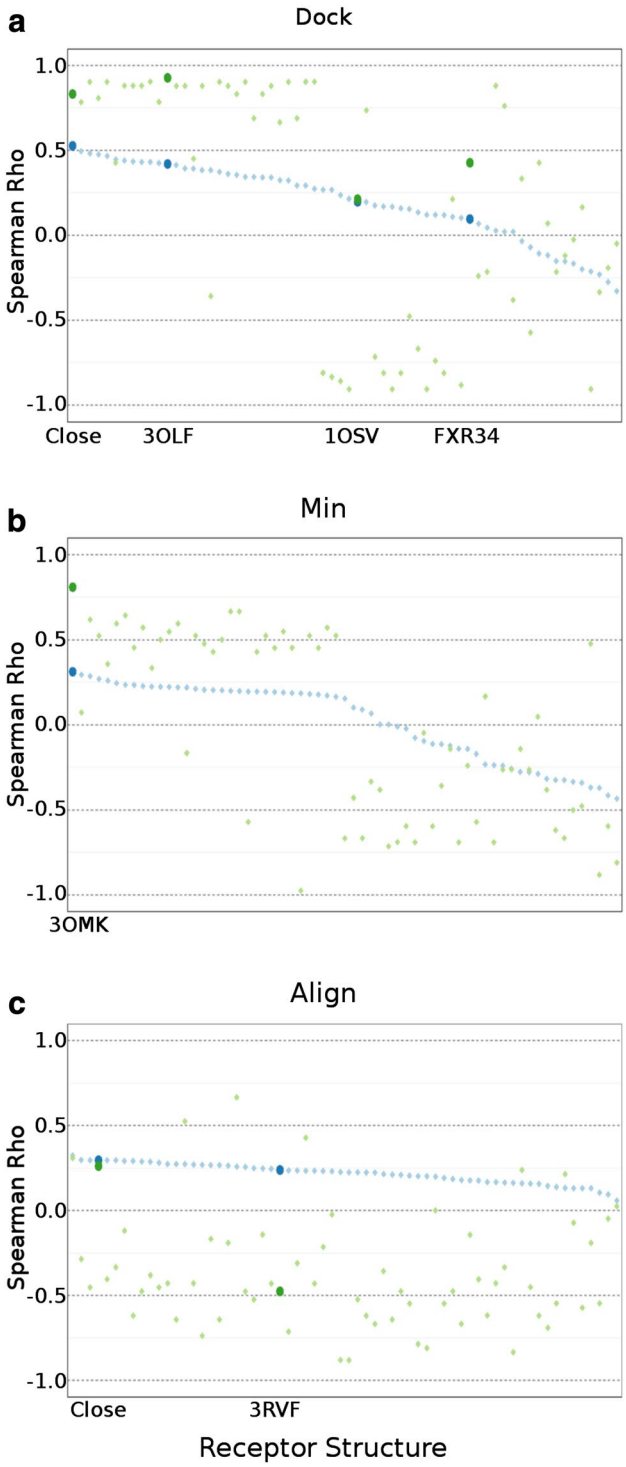

Fig. 5 Retrospective analysis of optimal receptor selection of **a** dock-cross, **b** min-cross, and **c** align-cross methods. Retrospective scoring against test data shown in blue, training data shown in green. Large circles represent receptors submitted to D3R and are labeled along x-axis, light diamonds are all other possible receptors

however FXR34 had best R^2 of training data. The receptor which would have given the best ranking of the FXR compounds for dock-cross was FXR13, which was tied for tenth highest Spearman on training set. The receptor we selected for min-cross (PDB 3OMK [27]) was the one which resulted in best Spearman when ranking our training data prospectively and retrospectively (Fig. 5b). This receptor selection was the best available, even with the 36 newly released structures for stage two. Finally, Spearman correlation coefficients for FXR compounds against every available cocrystal structure and scored using align-cross are shown in Fig. 5c. For this method we took a hybrid approach and picked the receptor with the best combination of Spearman correlation, pose prediction (% of training compounds within 2.0 Å), and R^2, which for align-cross was PDB 3RVF [29]. However, this led to poor ranking prediction. Retrospectively, the best receptor for align-cross would have been PDB 3OOF [27], which had the fourth-highest Spearman ρ prospectively.

Additionally, we calculated average root-mean-square deviation (RMSD) values for our align-close and min-cross methods to compare to our submitted ones for dock-close. We found that align-close and min-cross performed similarly at pose prediction, with average first-pose RMSD values of 4.67 and 4.69 Å respectively, significantly higher than the 3.37 Å for dock-close. This makes sense given the similarities in how poses are generated for each method. We also calculated Spearman correlation values for FXR1-36 and FXR37-102 subsets of our dock-close submission to see if having true crystal structures provided significant improvement over holo-like structures. We found that these subsets had Spearman ρ of 0.482 and 0.486 respectively. This shows that while having a true co-crystal structure provides a good framework for pose prediction, the ability of force fields used in docking methods still has significant area for improvement.

Optimal strategies for virtual screening

Table 4 summarizes prospective and retrospective analysis for the 2015 [9] and 2016 (here) grand challenges for the automated methods listed in Table 1. For *prospective* rankings, we found that dock-close was the best performing method over the course of the two grand challenges (average Spearman $\rho = 0.43$). We see that while dock-close performed the best for FXR and HSP90 affinity ranking challenges, it was about average for prospective ranking of MAP4K4 and the worst at retrospective ranking. For *retrospective* rankings, we find that dock-cross performed the best (average Spearman $\rho = 0.49$). This is interesting because dock-cross didn't perform the best for any of the targets. Yet, overall dock-cross rankings using the optimal receptor always yields near-optimal correlations.

Table 4 Prospective and retrospective analysis of ranking strategies in D3R grand challenges

Method	FXR prospective	FXR retrospective	HSP90 prospective	HSP90 retrospective	MAP4K4 prospective	MAP4K4 retrospective
Align-close	0.30	0.30	0.46	0.46	0.33	0.33
Dock-close	**0.53**	**0.53**	**0.50**	**0.50**	0.25	0.25
Min-cross	0.31	0.31	0.30	0.47	**0.41**	0.51
Align-cross	0.24	0.32	0.37	0.44	0.11	**0.57**
Dock-cross	0.42	0.50	0.40	0.47	0.06	0.51

Shown is value of Spearman ρ for each method in prospective and retrospective analysis for each receptor target in past two grand challenges. Best method for each receptor shown in bold. Note that close methods are the same for prospective and retrospective columns because they use the receptor of most similar compound, which doesn't change retrospectively

The difference between these targets is that overall dock-close performs better when targets have a well-defined binding pocket, such as those of FXR and HSP90 (see Fig. 6), whereas MAP4K4 is a large open pocket where ranking is much more dependent on scoring. Given the known limitations of scoring functions, reliance on scoring is not a good strategy and we find that for MAP4K4 optimal rankings were obtained for local minimization methods (align and min). These methods align conformers to a cocrystal ligand which ensure a reasonable starting pose. Using the optimal strategy for each target, our approaches would have been able to yield a top-of-the-line average Spearman $\rho = 0.53$. We note that these correlations are significantly superior to those reported in earlier community challenges [3, 9, 30] and they should provide a meaningful enrichment in virtual screening.

While it appears that dock-close and dock-cross are the best methods for situations such as the D3R grand challenges where you have months to work on rankings, an attractive application of these strategies is automation for virtual screening. An additional factor to consider when selecting which strategy to use is the amount of time necessary for each method. Each compound minimization takes only a few seconds using Smina, compared to 30 s–1 min for each docking. While these timescales are relatively quick for close methods, the time required rapidly increases for cross methods when you have many receptors to score against. Because of this, depending on the application or specific system of interest (receptor structures and compounds) it might be better to use a slightly less accurate method such as min-cross or align-cross (average retrospective Spearman ρ of 0.43 and 0.44 respectively).

An attractive application of these strategies is automation for virtual screening. Indeed, methods shown in Table 4 do not require human intervention, and can result is the absolute best ranking for all targets (as compared to other methods). An additional factor to consider when selecting virtual screening strategies is the amount of time necessary for each method. Align- and min-cross methods are limited by the minimization step that takes only a few seconds using Smina. On the other hand, dock-methods are limited by docking that takes about 30 per compound. Depending on the size of the compound library, a fast but perhaps less accurate strategy could also be min-cross or align-cross

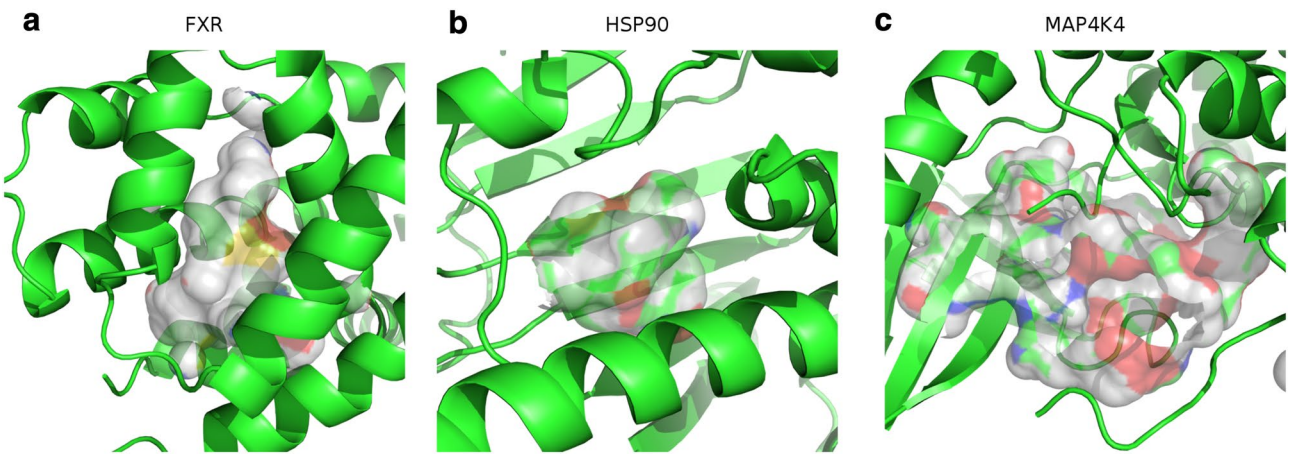

Fig. 6 Examination of binding pockets of **a** FXR (provided by D3R), **b** HSP90 (PDB 4YKY [35]), and **c** MAP4K4 (PDB 4OBO [36])

(average retrospective Spearman ρ of 0.43 and 0.44 respectively). How good is the best predicted Spearman ρ of 0.53 (dock-close)? To answer this question, we examined how well we were able to provide enrichment for the top 25 best affinity compounds. Out of the 25 best compounds in D3R, we predicted 14 of them in the top 25 of our ranking (56%), while a random ranking would be 6 (see Fig. 7). This shows that even with moderate Spearman correlation values, we are able to provide significant enrichment in predicting relative binding affinities of compounds.

Free energy prediction

Accurate prediction of relative binding a difficult problem. A variety of methods were used in the previous grand challenge [3], including docking [31], MM/GBSA [31, 32], Glide [33], and QM/MM [31]. These methods span a wide range of both computational intensity and accuracy, including free energy perturbation [34]. In this category, we used a combination of molecular dynamics simulations for modeling protein ligand interactions, which then were evaluated based on a contact potential that is modulated according with the solvation of these contacts [11]. These predictions were among the most accurate in the competition with root-mean-square error (RMSE) values of predictions for both groups of around 1 kcal/mol. This evaluation led to slightly better rankings than those predicted by, dock-close, the overall best method. Namely, free energy evaluation and dock-close method predicted Spearman ρ of (0.186 and 0.51) and (0.075 and 0.52), for Set 1 and Set 2, respectively. This modest improvement is encouraging since ultimately more accurate free energy evaluations must account for receptor flexibility.

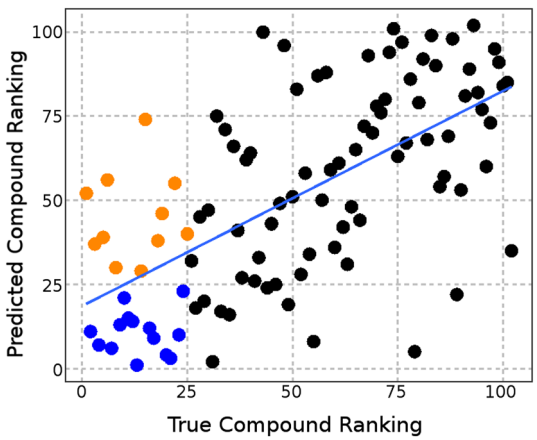

Fig. 7 Enrichment of virtual screening of FXR compounds by dock-close method. Actual ranking of compounds shown along x-axis. Compounds in top 20 correctly predicted in top 20 shown in blue, incorrect predictions of top 20 in orange, non-top 20 compounds in black. Linear regression fit line shown ($R^2 = 0.41$)

The D3R grand challenges have provided an excellent opportunity for the evaluation of tools and strategies for rational drug design. We previously presented strategies for optimal pose prediction evaluated in the 2015 grand challenge [9]. Here we discussed the application of our strategies to the problem of ranking the relative affinity of a set of compounds against a protein target. We again showed that the selection of the receptor structure (or structures) used for docking or minimization is important to obtain an optimal prediction. We found that methods which take into account all available structural information (close methods) perform best for targets with constrained binding sites; whereas for targets with open binding pockets or highly variable binding modes, methods that use only a single receptor structure (cross methods) perform better.

Acknowledgements We thank OpenEye Scientific for providing an academic license for their software. The work was funded by U.S. National Institutes of Health Grant Numbers GM097082 and T32EB009403.

References

1. Koes D et al (2012) Enabling large-scale design, synthesis and validation of small molecule protein-protein antagonists. PLoS ONE 7(3):e32839
2. Domling A, Wang W, Wang K (2012) Chemistry and biology of multicomponent reactions. Chem Rev 112(6):3083–3135
3. Gathiaka S et al (2016) D3R grand challenge 2015: evaluation of protein-ligand pose and affinity predictions. J Comput Aided Mol Des 30(9):651–668
4. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3(11):935–949
5. Koes DR, Baumgartner MP, Camacho CJ (2013) Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model 53(8):1893–1904
6. Trott O, Olson AJ (2009) Software news and update AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461
7. Koes DR, Camacho CJ (2012) ZINCPharmer: pharmacophore search of the ZINC database. Nucleic Acids Res 40(W1):W409–W414
8. Koes DR, Pabon NA, Deng X, Phillips MA, Camacho CJ, Wang S (2015) A Teach-Discover-Treat application of ZincPharmer: an online interactive pharmacophore modeling and virtual screening tool. PLoS ONE 10(8):e0134697
9. Ye Z, Baumgartner MP, Wingert BM, Camacho CJ (2016) Optimal strategies for virtual screening of induced-fit and flexible target in the 2015 D3R grand challenge. J Comput Aided Mol Des 30(9):695–706
10. Smith RD et al (2016) CSAR benchmark exercise 2013: evaluation of results from a combined computational protein design, docking, and scoring/ranking challenge. J Chem Inf Model 56(6):1022–1031
11. Temiz NA, Camacho CJ (2009) Experimentally based contact energies decode interactions responsible for protein? DNA affinity and the role of molecular waters at the binding interface. Nucleic Acids Res 37(12):4076–4088

12. O'boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminform 3:33
13. Berman HM et al (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
14. The PyMOL molecular graphics system, version 1.8 Schrödinger, LLC. [Online]. https://www.pymol.org/citing. Accessed 02 May 2017
15. Chen X, Liu M, Gilson MK (2001) BindingDB: a web-accessible molecular recognition database. Comb Chem High Throughput Screen 4(8):719–725
16. Enyedy IJ et al (2016) Discovery of biaryls as RORγ inverse agonists by using structure-based design. Bioorg Med Chem Lett 26(10):2459–2463
17. René O et al (2015) Minor structural change to tertiary sulfonamide RORc ligands led to opposite mechanisms of action. ACS Med Chem Lett 6(3):276–281
18. van Niel MB et al (2014) A reversed sulfonamide series of selective RORc inverse agonists. Bioorg Med Chem Lett 24(24):5769–5776
19. Hawkins P. C. D., Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model 50(4):572–584
20. Tosco P, Balle T, Shiri F (2011) Open3DALIGN: an open-source software aimed at unsupervised ligand alignment. J Comput Aided Mol Des 25(8):777–783
21. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25(2):247–260
22. Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, Cheatham TE, Darden TA, Duke RE, Gohlke H, Goetz AW, Gusarov S, Homeyer N, Janowski P, Kaus J, Kolossváry I, Kovalenko A, Lee TS, LeGrand S, Luchko T, Luo R, Madej B, Merz KM, Paesani F, Roe DR, Roitberg A, Sagui C, Salomon-Ferrer R, Seabra G, Simmerling CL, Smith W, Swails J, Walker, Wang J, Wolf RM, Wu X, Kollman PA (2014) Amber 14. University of California, San Francisco
23. Temiz NA, Trapp A, Prokopyev OA, Camacho CJ (2009) Optimization of minimum set of protein-DNA interactions: a quasi exact solution with minimum over-fitting. Bioinformatics 26(3):319–325
24. Kolář M, Hobza P (2012) On extension of the current biomolecular empirical force field for the description of halogen bonds. J Chem Theory Comput 8(4):1325–1333
25. Harder E et al (2016) OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J Chem Theory Comput 12(1):281–296
26. Trott O, Olson AJ (2009) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2):455–461
27. Richter HGF et al (2011) Optimization of a novel class of benzimidazole-based farnesoid X receptor (FXR) agonists to improve physicochemical and ADME properties. Bioorg Med Chem Lett 21(4):1134–1140
28. Mi LZ et al Structural basis for bile acid binding and activation of the nuclear receptor FXR. Mol Cell 11:1093–1100
29. Akwabi-Ameyaw A et al (2011) Conformationally constrained farnesoid X receptor (FXR) agonists: alternative replacements of the stilbene. Bioorg Med Chem Lett 21(20):6154–6160
30. Smith RD et al (2011) CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions. J Chem Inf Model 51(9):2115–2131
31. Misini Ignjatovic M, Caldararu O, Dong G, Munoz-Gutierrez C, Adasme-Carreno F, Ryde U (2016) Binding-affinity predictions of HSP90 in the D3R grand challenge 2015 with docking, MM/GBSA, QM/MM, and free-energy simulations. J Comput Aided Mol Des 30(9):707–730
32. Deng N et al (2016) Large scale free energy calculations for blind predictions of protein-ligand binding: the D3R grand challenge 2015. J Comput Aided Mol Des 30(9):743–751
33. Friesner RA et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47:1739–1749
34. Cole DJ, Tirado-Rives J, Jorgensen WL (2014) Enhanced Monte Carlo sampling through replica exchange with solute tempering. J Chem Theory Comput 10:565–571
35. Kang YN, Stuckey JA, Heat shock protein 90 bound to CS319. http://www.rcsb.org/pdb/explore.do?structureId=4yky
36. Crawford TD et al (2014) Discovery of selective 4-amino-pyridopyrimidine inhibitors of MAP4K4 using fragment-based lead identification and optimization. J Med Chem 57(8):3484–3493