J Comput Aided Mol Des (2016) 30:805–815
DOI 10.1007/s10822-016-9979-z

Docking-undocking combination applied to the D3R Grand Challenge 2015

Sergio Ruiz-Carmona[1] · Xavier Barril[1,2]

Received: 9 June 2016 / Accepted: 24 September 2016 / Published online: 5 October 2016

Abstract Novel methods for drug discovery are constantly under development and independent exercises to test and validate them for different goals are extremely useful. The drug discovery data resource (D3R) Grand Challenge 2015 offers an excellent opportunity as an external assessment and validation experiment for Computer-Aided Drug Discovery methods. The challenge comprises two protein targets and prediction tests: binding mode and ligand ranking. We have faced both of them with the same strategy: pharmacophore-guided docking followed by dynamic undocking (a new method tested experimentally here) and, where possible, critical assessment of the results based on pre-existing information. In spite of using methods that are qualitative in nature, our results for binding mode and ligand ranking were amongst the best on Hsp90. Results for MAP4K4 were less positive and we track the different performance across systems to the level of previous knowledge about accessible conformational states. We conclude that docking is quite effective if supplemented by dynamic undocking and empirical information (e.g. binding hot spots, productive protein conformations). This setup is well suited for virtual screening, a frequent application that was not explicitly tested in this edition of the D3R Grand Challenge 2015. Protein flexibility remains as the main cause for hard failures.

Keywords D3R · Drug discovery data resource · Grand Challenge · Docking · Dynamic · Undocking · GC2015 · Protein flexibility

Introduction

Computer-Aided Drug Discovery (CADD) methods are constantly under development anda wide spectrum of options is available to the scientific community to address each specific situation at every stage of the drug discovery process [1–3].

Independent validation experiments are extremely useful to test the different methods, try them out under different circumstances and validate them for a specific goal. For instance, there have been experiments to help the development of protein structure modeling software [4],the prediction of protein–protein interactions [5] or certain physico-chemical properties of small molecules [6].

In this direction, the D3R Grand Challenge 2015 provides an independent exercise to assess and validate CADD tools related with protein–ligand interactions. Two proteins (Hsp90 and MAP4K4) with datasets comprising different ligands with measured affinities and crystal structures are provided as blind sets. Different measures for each of the datasets were used to evaluate the performance of different methods in two situations that are common in drug discovery projects: ligand ranking and binding mode prediction.

Docking, scoring and free energy methods have been widely applied in structure-based drug discovery [7–12]

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9979-z) contains supplementary material, which is available to authorized users.

✉ Xavier Barril
xbarril@ub.edu

[1] Departament de Fisicoquímica, Facultat de Farmàcia, Institut de Biomedicina de la Universitat de Barcelona (IBUB), Universitat de Barcelona, Av. Joan XXIII s/n, 08028 Barcelona, Spain

[2] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

as they provide an excellent assistance particularly in early stages of the development of new drugs. Docking is a very common method that can be used both for predicting the binding mode of a protein–ligand complex and for virtually assaying thousands to millions of drug-like molecules in a relatively short amount of time, speeding up the finding of promising candidates and dramatically decreasing the cost in comparison with the experimental alternative [13]. However, the scoring functions employed in docking have been trained to reproduce specific data sets and are qualitative in nature. As such they are not expected to correlate with binding free energies [14]. Further limitations include receptor flexibility or the presence of water molecules that can be wither trapped or displaced by the ligand.

In our particular approach, docking is a central component tackling the D3R Grand Challenge 2015, but we aim to overcome some of its limitations with complementary tools and, whenever possible, guiding the calculations with previous knowledge about the systems. Specifically, we have used rDock software [15] as the docking engine, using pharmacophoric restraints to ensure that the predicted ligand poses fulfil certain key interaction points [16–18]. In the case of Hsp90, they correspond to a hydrogen bond with the carboxylate of Asp93 and, in case of MAP4K4, a hydrogen bond with the nitrogen atom of Cys108 in the hinge region (a short linear sequence that acts as a hinge between the N-terminal and C-terminal domains in kinases). These interaction points can be identified merely by superimposing all the available crystal structures of protein–ligand complexes for each system in the PDB and obtaining a pharmacophore definition as detailed in the Methods, which can be supplemented to rDock in order to increase its efficiency as shown in previous studies [15]. Hsp90 presents at least one water molecule that can be displaced by certain ligand classes. By excluding this water molecule, we make the receptor definition valid for all chemotypes [19]. Then, to address the protein flexibility, we took a knowledge-based approach. We investigated the effect of protein flexibility on docking performance using Hsp90 as a test set, so we are familiar with the different conformations the protein can adopt upon ligand binding. We selected the We selected the most common conformation amongst all known Hsp90 protein–ligand complexes (namely, closed lid) for running docking and revised the quality of the predictions knowing that certain chemotypes can induce a conformational change of the lid to the open or helical states [20]. In contrast, MAP4K4 is a much less well characterized protein and we took a best guess based on our previous knowledge about other kinases. As we will discuss below, the different degree of previous knowledge for each system has had a major effect on the outcome and highlights the importance of the human factor, which remains essential even as the computational tools improve.

Finally, we have introduced the use of Dynamic Undocking (DUck), a new tool used to assess the structural stability of protein–ligand complexes [21]. Here we have experimentally adopted a consensus approach, where the docking poses are re-evaluated and re-ranked based on their resistance to break the key hydrogen bonding interaction. This approach allowed us to detect not only false positives but also false negatives from docking results. DUck has been shown to be orthogonal to docking, as it evaluates structural stability as opposed to binding affinity [21]. For some ligands, re-scoring by DUck has allowed us to identify good binding poses which are *apriori* discarded due to bad docking scores. In other cases, docking and DUck selected the same pose, increasing our confidence on predicted binding modes that would be deemed doubtful if they had been backed up only by docking.

In the next sections, we will discuss in detail the methodology and the results obtained in the D3R Grand Challenge, drawing some conclusions to explain the failures and successes, as well as some recommendations for future editions of this challenge.

Methods

Selection of cavity

The D3R Grand Challenge 2015 has two differentiated objectives: predict the crystallographic poses and the affinities or rankings for a series of ligands. Both of these objectives rely on a good definition of the system and a reliable characterization of the ligand-receptor interaction is crucial. For Hsp90, 4 receptor structures from the PDB were proposed by the organizers (2JJC, 2XDX, 4YKR and 4YKY). All of them were in the so-called closed conformation of the lid with the exception of 2XDX, which had the lid in open conformation.

As most of the known ligand-Hsp90 complexes have the lid in closed conformation, 2XDX was discarded. 2JJC was also discarded because, unlike the ligands in the test set, it is a very small and may be unable to modulate the cavity for better docking performance [22, 23]. Structures 4YKY and 4YKR are very similar in all respects (both bind a ligand of the resorcinol family) and were considered equivalent. The former was selected as reference structure. In a previous study [15] we demonstrated the improvement in virtual screening applications when guiding the docking process by adding previous knowledge, with a specific example for Hsp90. Additionally, it is known that three interfacial water molecules have an important role mediating the protein–ligand contacts. For this reason, they have been included in all docking runs as structural waters in the binding site. Some ligand types (e.g. adenine) interact with a fourth interfacial water molecule, but it is displaced by others ligands (e.g. resorcinol) and cannot be kept as part of the receptor [24]. Hence, the protocol used in all the docking calculations for Hsp90 includes a pharmacophore definition of two hydrogen bonds with Asp93:OD2 and one of the water molecules (included in all the runs as non displaceable), as previously defined in [15]. For undocking, the water molecule is added explicitly to the initial structure. In case of MAP4K4, 2 receptor structures from the PDB were supplied by the organizers (4OBO and 4U44). The main difference between the conformations of the two crystals is a loop folding towards the hinge region in 4OBO, thus decreasing the size and the solvent exposure of the binding site. Due to those restrictions we decided to use 4U44 as reference for all MAP4K4 applications, which had a bigger and more accessible binding site. In order to guide docking, we performed a pharmacophore search (more details in the next section) using all crystal structures of MAP4K4-ligand complexes in the PDB. We then supplied all docking calculations with a pharmacophore defined by a hydrogen bond with Cys108:N, located in the hinge region.

Pharmacophore search

To get a reliable pharmacophore definition for the MAP4K4 system, a set of known protein–ligand 3D structures was necessary. We selected all MAP4K4 protein–ligand complexes from the PDB (4OBO, 4OBP, 4OBQ, 4RVT, 4U40, 4U41, 4U42, 4U43, 4U44, 4U45, 4ZK5 and 5DI1) and aligned them to the reference 4U44. The “Pharmacophore Search” tool of MOE was run and a hydrogen bond with Cys108:N in the hinge region was selected as pharmacophore. It was fulfilled by all 12 ligands in the PDB subset. Moreover, it was consistent with other protein–ligand interactions in the kinases family [25, 26].

Molecular docking

For all molecular docking simulations we used rDock [15, 27], a fast and reliable docking program that we released as open source several years ago. To run rDock, only a correctly prepared 3D structure of the receptor and a definition of the binding site are needed. In this work, we defined the cavity using the crystallized ligand found in both PDB structures for Hsp90 and MAP4K4, 4YKY and 4U44 respectively. Some rDock rbcavity parameters were decreased with respect to the default values in order to optimize the binding site definition: *radius* (changed from 10.0 to 6.0), which defines the region around the reference ligand that will be used to define the docking binding site and *max_cavities* (from 99 to 1), as we only want to run docking in one cavity. The pharmacophoric restraints were defined as mandatory and all the ligands unable to fulfill the definition were discarded. For the docking protocol, no modifications were made to the standard as previously published [15]: 50 individual docking runs per ligand, which is considered exhaustive sampling, in order to ensure that the lowest-energy binding mode is found.

Receptor preparation

The 3D structure of the receptor has to be provided to rDock with standard Tripos MOL2 format and atom types [28]. However, as rDock relies on the user-supplied structure, we need to provide it with correct protonation states and charges, as well as correct orientations of flexible side chains (rDock only considers as flexible atoms of the receptor the hydrogen atoms of terminal OH and NH3 + groups within 3 Å of the binding site cavity). The “Structure Preparation” tool from MOE [29] was used to protonate at pH 7.0 and correct all the issues found for Hsp90 and MAP4K4 receptors, such as chain breaks, missing loops or disulfide bonds, incorrect residue labeling or alternate conformations. The prepared structures were then saved in mol2 format and used as input for rDock.

Ligand structure

As all ligands provided by the organization were in 2D format, Ligprep from Schrödinger [30] was used to calculate the 3D structure with correct topology, bond orders and geometry of bonds, angles, dihedrals and rings. The ionisable groups were protonated at pH = 7 with a tolerance of ±1. All ligands were saved in MOL SDF format and used as input for docking.

Dynamic undocking

We used Dynamic Undocking, or DUck, as a complementary tool to molecular docking in order to improve the overall performance of docking-based virtual screening [21]. DUck is a methodology developed in our group based on Steered Molecular Dynamics (SMD). The interaction of the ligand and the receptor with the key interaction point (specified when defining the cavity and protocols for docking) is monitored with SMD. In particular, DUck simulations consist on unbiased molecular dynamics (MD) simulations of the complex and repeated SMD simulations launched at 1 ns intervals of the MD to simulate the rupture of the ligand-receptor interaction and measure the force needed to achieve a state where the interaction has just been broken or, as we named it, a Quasi-Bound state. The work profiles obtained from the SMD simulations are processed to obtain the work to achieve the Quasi-Bound state (W_{QB}), which will be used to score and rank the ligands. Moreover, in order to increase throughput and reduce the influence of peripheral interactions and focus on the desired interaction, we use a model receptor that includes only a small part of the protein of interest. This portion is created around the defined key interaction point and preserves its local environment, simplifying also the dissociation pathway and avoiding artifactual results (more details about DUck can be found in Ref. [21] and http://www.ub.edu/bl/undocking/). For Hsp90 and MAP4K4, the following protocol was set: protein models were created containing the residues with any atom within 6 Å around the key interaction points (as detailed in *Selection of the Cavity* section) and manually refined to include other important residues for the binding site environment (Figure S1; Table S2). The best-scored docking poses for each ligand were subjected to an in-house script that automatically parameterized each ligand and prepared the necessary files for running the MD and SMD simulations of DUck. Each protein–ligand complex system was placed in a cuboid box with a minimum distance between each atom and the edge of the box of 12 Å in every dimension and solvated with TIP3P water molecules and Na+ or Cl− ions were added to the solvation box depending on the charge of each of the protein–ligand complexes in order to ensure the electroneutrality of the simulated systems. Due to the artificiality of the protein models, MD simulations were run with harmonic restraints (1 kcal/mol $Å^2$) in all heavy atoms of the receptor to prevent big conformational changes. In order to preserve key hydrogen bond interaction during the equilibration part of the simulations, distances beyond 3 Å are penalized (parabolic restraint with $k = 1$ kcal/mol $Å^2$ between 3 and 4 Å; linear restraint with $k = 10$ kcal/mol Å beyond 4 Å). All unbiased MD steps were run using a Langevin thermostat with the cutoff for non-bonded interactions set to 9 Å and the collision frequency to 4 ps^{-1}. The equilibration consisted in 1000 cycles of minimization, gradual warming from 100 K to 300 K for 400 ps in the NVT ensemble and equilibration of the system for 1 ns in the NPT ensemble. At intervals of 1 ns (starting right after the equilibration), two SMD runs are executed from the same restart file (at 300 and 325 K, as described in Ref. [21]) for 500 ps. During this time, the distance of the key hydrogen bond is steered from 2.5 to 5.0 Å with a spring constant of 50 kcal/mol $Å^2$. More unbiased MD steps (1 ns each) were run to create more starting points for SMD runs to repeat the process as much as desired. All simulations were run with AMBER 14 [31] using in-house NVIDIA GeForce TITAN X GPUs or at the Barcelona Supercomputing Center using NVIDIA Tesla M2090 GPUs. AMBER forcefield 99SB was used for the protein and parm@Frosst [32] for the ligands.

Binding mode prediction

For all of the ligands where a binding mode was to be predicted, the protocol was the following: 1- Run docking as described in the “Molecular Docking” section above. 2- From the docking results, select a set of poses with a RMSD between them higher than 1 Å using the *sdrmsd* script from rDock package. 3- Run DUck to calculate the W_{QB} for all the sets of selected poses per ligand. 4- Select the pose with the highest W_{QB} as the correct binding mode and 5- visually inspect the results to check the selected poses fulfilled the defined interaction and the receptor conformation (more details in the following sections).

Ligand ranking

A few differences from the protocol for binding mode prediction were introduced in case of ligand ranking: 1- Run docking as described in the “Molecular Docking” section above. 2- From the docking results, select the top scored pose for each ligand. 3- Run DUck to calculate the W_{QB} for the selected poses. 4- For each of the ligands in the sets, the similarity to all known PDB ligands with measured affinity for the corresponding receptor (Hsp90 or MAP4K4) was calculated and taken into account to check the rankings and possible docking errors. 5- Docking score and W_{QB} from DUck were normalized for each of the sets. All ligands were ranked according to the sum of the two corresponding normalized scores. In the cases where docking was not able to find a good binding mode (i.e. the key interaction was not fulfilled), the similarity of each ligand with respect to other ligands in the challenge set and other ligands in PDB was used to assign a corrected ranking. Finally, a final step of visual analysis was carried on to check all ligands and re-rank some of them taking into account our previous experience.

Results and discussion

Following our primary hypothesis, we designed a docking protocol that would reinforce the importance of the most important binding hot spot. This was done through the introduction of pharmacophoric restraints that forced the presence of hydrogen bonding groups at specific locations (Fig. 1). The protein conformation was chosen to be as general as possible, thus for MAP4K4 we selected 4U44 as it has a bigger cavity than other structures available. For Hsp90, the biggest cavities present a ligand-induced hydrophobic sub-pocket (the PU3 cavity), but the associated protein conformation (helical) is energetically penalized and tends to downgrade the docking results [19]. For this reason, we chose a non-helical conformation (4YKY)

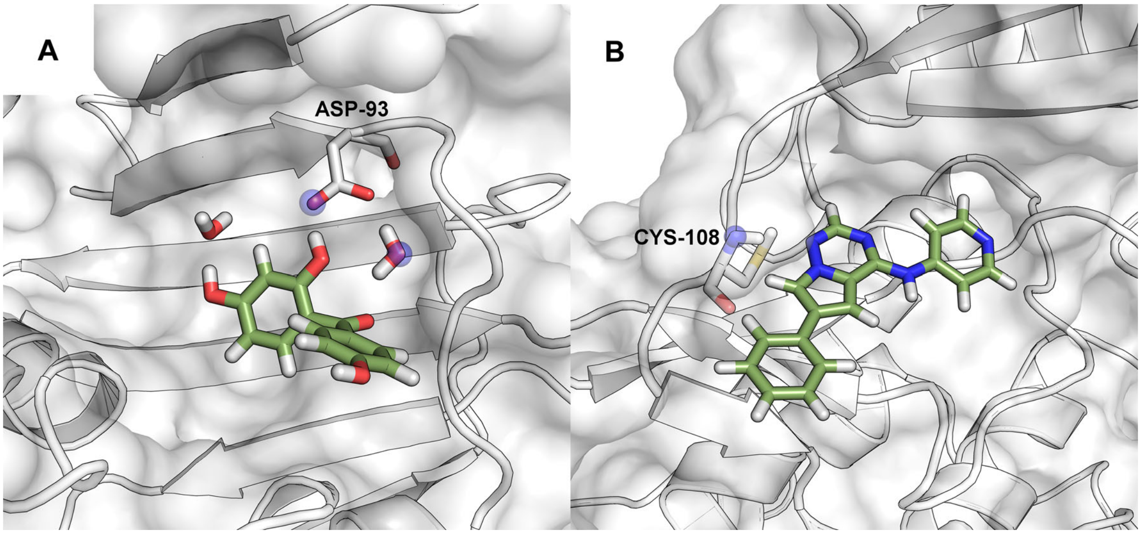

Fig. 1 **a** Hsp90 receptor definition. Asp93 and two surrounding water molecules (shown in *sticks*) define the key interaction element. The pharmacophoric points (transparent *blue spheres*) force the presence of a H-bond donor next to Asp93:OD2 and a H-bond acceptor next to the interstitial water molecule. **b** MAP4K4 receptor definition. The hinge region is a characteristic binding hot spot of protein kinases. A pharmacophoric restraint forced the presence of a hydrogen bond acceptor next to Cys108:N (transparent *blue sphere*)

taking care that the binding site was not blocked by any side-chain.

Binding mode prediction

We ran rDock to generate 50 poses per ligand. Poses with restraint penalties higher than 1 kJ/mol (indicating that the pharmacophore is not fulfilled) were discarded. After that, we selected a diverse set of the remaining poses, sorted by docking score to be re-evaluated by Dynamic Undocking (DUck). On average, 10 poses per ligand were selected for next step. DUck measures the work needed to break a given hydrogen bond (W_{QB}). We have found that true ligands in their correct binding mode, form hydrogen bonds that are much harder to break than decoys [21]. Here we employ this method to compare various binding modes of the same ligand. In the majority of cases, the binding pose with the best docking score also presented the highest W_{QB} value and was proposed as the correct solution. But often DUck provides a much more clear distinction between poses, removing uncertainty from the decision. This is illustrated with the Hsp90 ligand 40, which presented two alternative binding modes (Fig. 2). In the first binding mode, the ligand interacts with Asp93 through the resorcinol, whereas the cyclic urea plays this role in the second binding mode. Though their docking scores are relatively similar (−23.4 and −18.9 kJ/mol, respectively), the hydrogen bond formed by the second binding mode is extremely labile ($W_{QB} = 0.5$ kcal/mol), which makes this binding mode very unlikely. By contrast, the first binding mode presented a very strong hydrogen bond ($W_{QB} = 17.7$ kcal/mol) and was selected with full confidence. For Hsp90, in several cases a lower ranking pose was selected based on the DUck calculation (Table S1). This is shown in Fig. 3, where the Hsp90 ligand 73 presents a relatively similar binding mode with two different orientations. The first one (green) is the preferred one by docking (score = −20.3 kJ/mol), whereas the second one (pink) is heavily penalized due to a steric clash of the 1-chloro-3-nitrobenzene moiety (score = 1.3 kJ/mol). Dynamic undocking indicated that the latter binding mode was actually preferred ($W_{QB} = 11.6$ kcal/mol vs. 10.9 kcal/mol), which prompted us to seek a protein conformation where the second binding mode would fit without

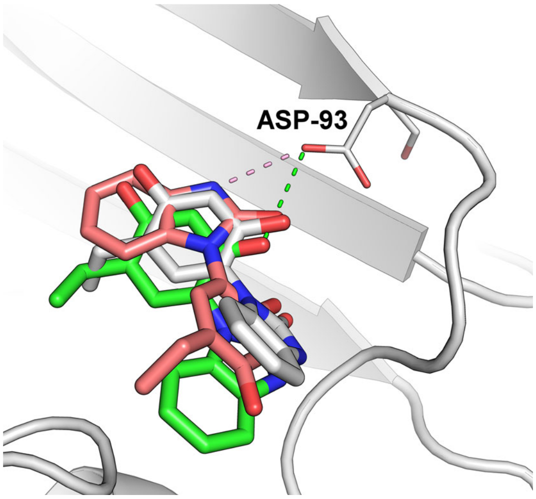

Fig. 2 Two binding modes proposed by docking for ligand 40 in the Hsp90 set (green and pink sticks; RMSD = 5.5 Å). *Dashed lines* represent the hydrogen bond between each ligand and Asp93. The crystal structure of ligand 40 is represented in white sticks for comparison

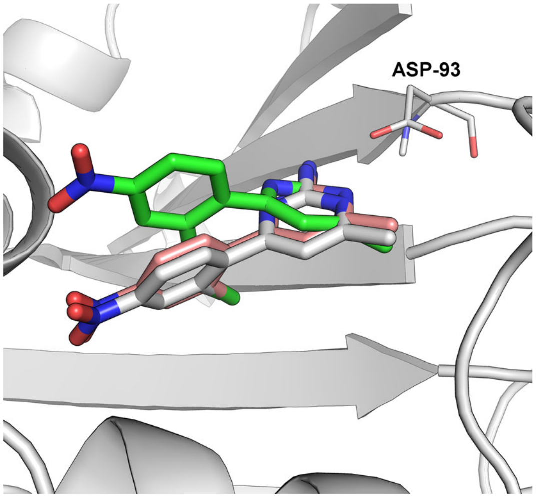

Fig. 3 Two different binding modes (RMSD: 2.5A) for ligand 73 in the Hsp90 set proposed by docking represented in green and pink sticks. The *green* one is the preferred conformation according to docking, whereas the *pink* one has a really bad score due to a clash penalization. With DUck, we could detect that the correct binding mode was the *pink* one. The crystal structure of ligand 73 is represented in *white sticks* for comparison, the RMSD with respect to the *pink binding* mode is 0.61

clashing. In this particular case, the ligand binds to helical conformation (e.g. 2WI6) where a hydrophobic pocket (the PU3 pocket) emerges [20].

The results submitted to the stage 1 of the D3R Grand Challenge are summarized in Tables 1 and 2 for Hsp90 and MAP4K4, respectively. The accuracy of binding mode prediction is generally measured in terms of root mean squared deviation (RMSD) from the crystallographic pose. It is also common to convert this value to a binary decision (correct/incorrect) based on a fixed threshold (usu. 2.0 Å). This is a debated topic, and several alternative solutions have been put forward [33, 34].

In practice, the best measure may depend on the particular problem that one is facing. For instance, a prediction that captures the main interactions is valid when dealing with a new chemotype, but inadequate at the lead optimization stage. Since our lab focuses on the hit identification stages of drug discovery, we are particularly interested in predicting the position of the central scaffold, i.e. the part of the ligand that forms the main interactions and defines the vectors of growth in the hit to lead stage. Thus, we have complemented the objective RMSD measure with a subjective binary classification telling if the prediction is sufficiently accurate to be used in the hit progression. In terms of RMSD, our average results were 1.6 ± 0.9 Å for Hsp90 (8th position among the participants of the D3R Grand Challenge 2015) and 3.7 ± 2.8 Å for MAP4K4 (3rd position). On the former set, we predict all but one ligand within 2.0 Å of RMSD. The only exception is ligand 44 (RMSD = 3.0 Å), but even then the position of the scaffold is correct and the deviation is due to the different orientation of a part of the ligand that does not engage in interactions with the protein (Figure S2).

Table 1 Summary of the results for the 6 ligands in the Hsp90 system Stage 1

Ligand ID	RMSD (Å)	Scaffold OK
40	1.71	Yes
44	3.00	Yes
73	0.61	Yes
164	1.40	Yes
175	1.94	Yes
179	0.70	Yes
Summary	RMSD (Å)	Scaffold OK
Average	1.56	6/6 (100 %)

The MAP4K4 results are much worse, but we still fared better than most participants, which highlights the difficulty of this set. Using the 2.0 Å RMSD cutoff, we only predicted 11 ligands correctly (37 %). In our subjective assessment, we predicted the position of the scaffold correctly for 18 ligands (60 %). The reason behind the poor performance is almost exclusively due to the flexibility of the protein. As this is a key issue in molecular docking, it will be discussed in detail. On the positive side, our protocol was still capable of predicting the main interaction correctly for a majority of ligands. Worthy of note, the structure of ligand 32 was originally inverted (Fig. 4). Docking, but particularly Dynamic Undocking, argued strongly against this binding mode. After consultation with the crystallographers our predicted binding mode was accepted as the proper binding mode. This is a reminder of the necessary dialogue between crystallographers and modelers, particularly where various binding modes are consistent with the observed electron density (e.g. due to tautomerism) [35].

Protein flexibility: the greatest docking challenge?

Reviewing the cause of the cases in MAP4K4 where we failed in making a good prediction, we found that using a single receptor conformation was by far the most important factor. There is a large body of literature indicating the importance of protein flexibility [36–38] but back in 2005 we demonstrated that using multiple protein conformations could actually downgrade the results, particularly in virtual screening applications [19]. Since then, other authors have suggested that judicious selection of two or three structures can produce a small but systematic improvement over the best single structure [39–41]. However, as we did not have any previous knowledge on this system, we adopted the

Table 2 Summary of the results for the 30 ligands in the MAP4K4 system Stage 1, and simulation of Stage 1 results for MAP4K4 taking into account additional conformations of Tyr36 (ligands with bad prediction only)

Lig ID	RMSD (Å)	RMSD Flexible (Å)	Scaffold OK	Lig ID	RMSD (Å)	RMSD Flexible (Å)	Scaffold OK
1	2.26	–	Yes	17	8.44	6.94[b]	No/No
2	3.59	10.44[a]	No/No	18	1.99	–	Yes
3	1.02	–	Yes	19	1.85	–	Yes
4	7.16	6.14[b]	No/No	20	1.29	–	Yes
5	8.90	1.13[a]	No/Yes	21	0.96	–	Yes
6	2.50	–	Yes	22	1.47	–	Yes
7	1.03	–	Yes	23	2.66	1.46[a]	Yes/Yes
8	1.43	–	Yes	25	3.40	3.13[a]	Yes/Yes
9	2.49	3.55[a]	No/No	26	6.77	6.93[a]	No/No
11	0.68	–	Yes	27	1.29	–	Yes
12	11.06	1.42[a]	No/Yes	28	1.68	–	Yes
13	5.71	1.06[a]	No/Yes	29	6.34	6.58[a]	No/No
14	4.95	1.50[a]	No/Yes	30	3.83	0.52[a]	Yes/Yes
15	2.14	–	Yes	31	6.64	7.25[a]	No/No
6	5.69	4.83[a]	No/Yes	32	3.09[c]	1.32[ac]	Yes/Yes

Summary	Original	Flexible
Avgerage RMSD (Å)	3.74	2.57
Scaffold OK	18/30 (60 %)	23/30 (77 %)

[a] P-loop and Tyr36 faced inwards to the cavity

[b] No hydrogen bond made with Cys108 in the hinge region

[c] Measured with respect to the corrected crystallographic pose

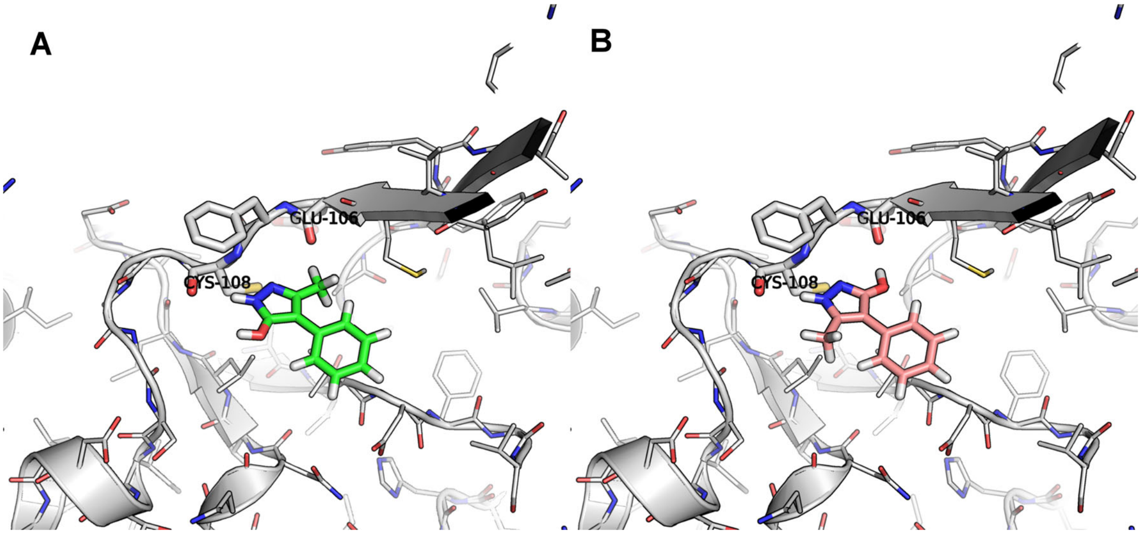

Fig. 4 Structure of MAP4K4 ligand MAP32 in the disclosed crystallographic structure (**a**) and the alternative mode we proposed (**b**). Note the different tautomers with inverted methyl and hydroxyl groups, where in the crystallographic pose there is a clash between the methyl group and Glu106, we found a well structured hydrogen bond between the hydroxyl group and Glu106

simple approach of using the biggest cavity (4U44), hoping that it would be valid for a larger proportion of ligands [42].

Once the experimental structures were disclosed, we observed that a large proportion of the ligands actually bind to a conformation where the cavity is partly occluded by the side-chain of Tyr36 in the P-loop (Fig. 5). In order to measure the impact of these effects, we ran the exact same experiments using as receptor structure 4OBO (Tyr36-IN), which has this alternative conformation. As shown in Table 2, most of the recalculated poses have an RMSD lower than the one we submitted to the D3R Grand Challenge 2015. Taking the best RMSD of the different binding modes, we obtain an average RMSD improvement of 1.1 Å (2.6 vs. 3.7 Å) with 18 ligands (60 %) below the 2.0 Å threshold and 23 ligands (77 %) with a correctly placed scaffold. While the results are still imperfect, one must consider that three structures are still insufficient to represent the whole array of conformational possibilities. In fact, we deem that there are only 2 ligands (7 %) for which the failure cannot be attributed to the conformation of the protein: Ligands 4 and 17 do not form a hydrogen bond with the backbone of Cys108 (the hinge region) and are thus incompatible with our docking and dynamic undocking protocol. On the other hand, if the relative energies of the conformational states are not properly considered, using multiple structures may cause more problems than it solves [43]. In our opinion, except for direct experimental observation of the conformational states [44], empirical knowledge gained from detailed analysis of multiple crystallographic structures is—at present—the only practical solution to this problem.

This is indeed the case for Hsp90, a system that we have studied thoroughly. Here, we were able to predict not only the structure of the ligands, but also which conformation would the protein adopt upon ligand binding. This aspect was not evaluated in the D3R Grand Challenge 2015. Considering the importance of this issue, we suggest that it should be included as a measurement of success in future editions. As shown in Table 3, the RMSD of the residues lining the binding site was below 0.4 Å in all cases, and the change in backbone conformation induced by ligand 73 could be predicted based on the DUck calculations (vide supra).

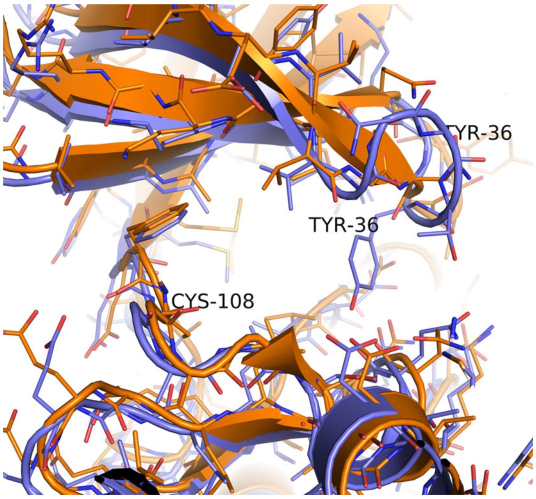

Fig. 5 Comparison between the two MAP4K4 supplied starting structures 4OBO (*blue*) and 4U44 (*orange*). In the former structure the side-chain of Tyr36 in the P-Loop is facing inwards, reducing the cavity space available for ligand binding

Virtual screening

For Stage 2 of the D3R Grand Challenge 2015, we were asked to predict the affinities or affinity rankings for 180 ligands in Hsp90 and 18 ligands in MAP4K4 systems. The tools developed and used in our group are geared towards virtual screening, where we aim to identify true ligands from huge libraries of chemical compounds. As such, our predictions are fast and qualitative and not well suited to predict binding affinities, instead our goal was to produce a ranked list enriched with potent ligands in the top positions. For this reason, we only discuss the results in terms of virtual screening performance: area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve and Enrichment Factors (EF). This type of analysis could not be performed on the MAP4K4 set because 15 out of the 18 ligands were considered as active ($IC_{50} < 1$ μM) and the other three were in a close range (1.74, 2.25 and 10 μM). The Hsp90 set presented more dispersion: 40.6 % of ligands (73 out of 180) had an IC_{50} lower than 1 μM and are considered active, the remaining are considered inactive even though 21.7 % (39 out of 180) have an IC_{50} between 1 and 10 μM. The fact that the inactive set contains molecules that are, a) true binders and, b) structurally very similar to the active ones makes this a very unusual and challenging test set. We encourage the organizers to include more standard virtual screening test sets in future editions of the challenge.

Our ranking protocol was based on an initial docking stage followed by DUck simulations of the top scoring

Table 3 RMSD (Å) between binding site residues of submitted Hsp90 receptor structures and crystal structure

PDB Code submitted	Crystal structure					
	40	44	73	164	175	179
2CCU	*0.29*	*0.30*	1.19	*0.27*	*0.25*	*0.37*
2WI6	1.05	1.05	*0.34*	1.05	1.07	1.03

List of residues defining the binding site: LEU48, ASN51, SER52, ALA55, ASP93, ILE96, GLY97, MET98, ASN106, LEU107, GLY108, PHE138, TYR139, VAL150, THR152, THR184 and VAL186

For each of the crystal structures, the submitted PDB receptor structure is highlighted in italics

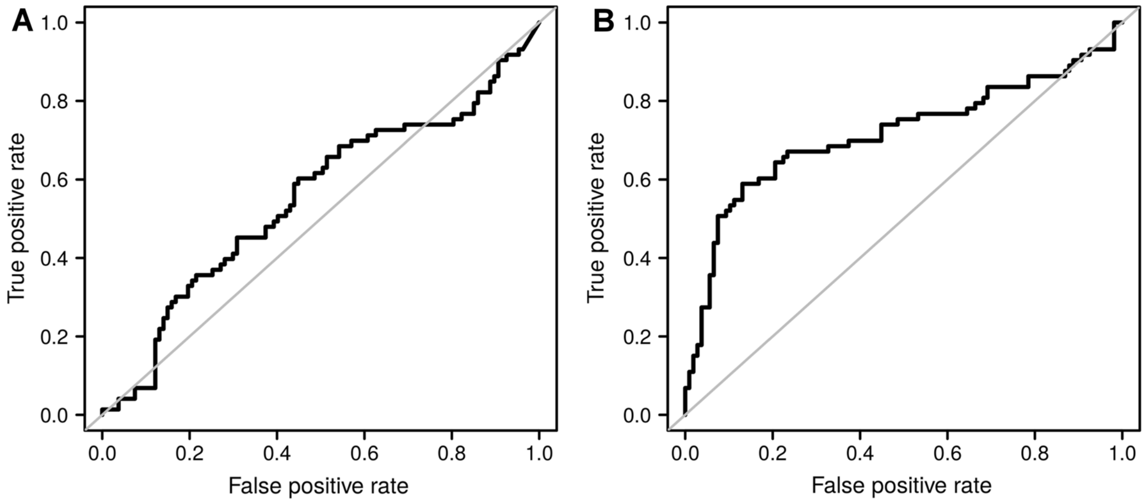

Fig. 6 ROC Curves of the 180 ligands in the Hsp90 Stage 2 Set. Ligands with an IC50 higher than 1 μM were considered as active. **a** Ranking according to rDock docking scores (AUC = 0.55). **b** Consensus ranking as submitted to the challenge (AUC = 0.71)

pose. We combined the scores obtained from docking and DUck and, following visual inspection to check all the ligands and the corresponding rankings, the final position of 49 ligands (27 %) in the ranked list was manually modified. Visual inspection introduces a subjective step that is difficult to control, but is essential in real applications to correct some of the limitations of docking. In our case, we used it mostly to rescue compounds that were predicted as inactive because they had an incorrect binding mode (e.g. ligands binding to the helical conformation that could not fit in the docking cavity). Considering the qualitative nature of our approach, the ROC curve (Fig. 6) demonstrates a very good performance, as do the corresponding enrichment factors (Table 4). To assess the effect of consensus ranking and visual inspection, we also plotted the ROC curve that would be obtained after the first stage (docking) and without the visual inspection (Figure S3). The AUC was much better for the combined ranking (0.71 vs. 0.55) and the enrichments were also higher for the combined ranking. This was the best performance across participants in this metric. Unexpectedly, we also ranked well in terms of Spearman correlation (0.39). This was surprising because both Docking and Dynamic Undocking are designed to discriminate between active and inactive compounds, rather than to obtain a quantitative assessment of their (relative) binding free energies. In part, this reflects our knowledge about this particular system, where we can anticipate from previous experience the conformational changes that take place in the protein and the ligand features that contribute to binding affinity. However, this correlation should not be considered a success, as it is likely insufficient to drive drug design. Instead it indicates that ranking ligands using structure-based methods is particularly challenging. In fact, many ligands in the test set have analogues with published binding affinity and we anticipate that a purely ligand-based strategy might have provided very good results. We suggest that the performance of one such knowledge-based approach would be useful as a benchmark of the performance of all participants in the contest.

Table 4 Summary of statistics for Hsp90 system stage 2 results

Ranking	AUC[a]	Enrichment[b]		
		1 %	10 %	20 %
Docking score	0.55	1.23	0.68	1.37
Combination	0.71	2.47	1.91	1.99

[a] Max.value for AUC = 1.00

[b] Max.enrichment possible = 2.47

Conclusions

Through the participation in the D3R Grand Challenge 2015, we have been able to validate the methods developed and used in our lab. We must emphasize that our main focus is virtual screening, an application that has not been considered explicitly in the challenge. Binding mode prediction is a first essential step for any subsequent prediction, so we had a particular interest on this part of the challenge. Binding affinity prediction (or ligand ranking) is much more demanding than virtual screening, and we participated in this part of the challenge somewhat reluctantly, expecting a clear underperformance compared to free energy methods.

We used a combination of qualitative techniques that, together, have worked much better than any of them separately. Namely, we used rDock for molecular docking with pharmacophoric restraints and DUck, a new technique based on molecular dynamics. For Stage 1, we were able to correctly predict how the ligands bind, particularly the position of the central scaffold forming the main interactions with the protein: for Hsp90 5 out of 6 ligands had an RMSD lower than 2 Å and 100 % of the scaffolds were correctly predicted; for MAP4K4 11 out of 30 ligands had an RMSD lower than 2 Å and 60 % of the scaffolds were correctly predicted. This figures would have increased to 18 out of 30 ligands and 77 % of the scaffolds if one single additional conformation (Tyr46-IN) would have taken into account. Retrospectively, we performed additional experiments to understand the failures, finding that protein flexibility was the major factor limiting the quality of the results. Predicting protein conformations is feasible, but increasing the number of conformation generally leads to decreased docking performance [19] and even when few conformations are considered, their relative energies must be considered to avoid artifacts [44]. This is a tall order that we have by-passed by employing previous knowledge about the system, which enabled us to predict the most likely receptor conformation for each Hsp90 ligand purely based on chemical structure. The fact that we did not have this information for MAP4K4 explains the difference in performance between both systems. It should be possible to extract this type of knowledge automatically from existing crystal structures deposited in the PDB, but we are not aware of any tool capable of doing this task. Forcing certain interactions during the docking process is equally important because it corrects some of the limitations of the scoring functions. Fortunately, in this case, the main pharmacophoric points can be extracted easily and automatically with existing tools. In the absence of known ligands, binding hot spots can be identified from molecular simulations [45].

In Stage 2, for Hsp90 we performed much better than expected considering the qualitative nature of our methods. The results were biased by our previous knowledge on this system, which had an important effect on the final performance, but this reflects the typical situation in drug discovery, where expert users combine tools and previous knowledge whenever possible. Our relative success highlights the challenges that free energy methods are still facing, but also indicates that there is a lot of potential in combining relatively simple structure-based tools with knowledge-based approaches. No doubt, machine learning will play an increasingly important role in the future, driven both by the growing body of public data [46, 47] and major advances in the field [48, 49].

Finally, we have several suggestions to improve future editions of the challenge. Namely, the prediction of protein conformation as a measure of success in binding mode prediction, the inclusion of a virtual screening prediction set and the introduction of an automated ligand-based approach as a baseline for measuring success of ligand ranking applications. We consider that all these aspects may improve what is already an extremely useful and necessary exercise.

References

1. Barril X, Javier Luque F (2012) Molecular simulation methods in drug discovery: a prospective outlook. J Comput Aided Mol Des 26:81–86. doi:10.1007/s10822-011-9506-1
2. Bajorath J (2015) Computer-aided drug discovery. F1000Research. doi:10.12688/f1000research.6653.1
3. Sliwoski G, Kothiwale S, Meiler J, Lowe EWJ (2014) Computational methods in drug discovery. Pharmacol Rev 61:67–75. doi:10.1016/j.vascn.2010.02.005
4. Moult J, Fidelis K, Kryshtafovych A et al (2014) Critical assessment of methods of protein structure prediction (CASP)—round x. Proteins 82(Suppl 2):1–6. doi:10.1002/prot.24452
5. Janin J (2005) Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci 14:278–283. doi:10.1110/ps.041081905
6. Muddana HS, Fenley AT, Mobley DL, Gilson MK (2014) The SAMPL4 host-guest blind prediction challenge: an overview. J Comput Aided Mol Des 28:305–317. doi:10.1007/s10822-014-9735-1
7. Ferreira L, dos Santos R, Oliva G, Andricopulo A (2015) Molecular docking and structure-based drug design strategies. Molecules. doi:10.3390/molecules200713384
8. Michel J, Essex JW (2010) Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. J Comput Aided Mol Des 24:639–658. doi:10.1007/s10822-010-9363-3
9. Steinbrecher TB, Dahlgren M, Cappel D et al (2015) Accurate binding free energy predictions in fragment optimization. J Chem Inf Model 55:2411–2420. doi:10.1021/acs.jcim.5b00538
10. Wang L, Wu Y, Deng Y et al (2015) Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. J Am Chem Soc. doi:10.1021/ja512751q
11. Chodera JD, Mobley DL, Shirts MR et al (2011) Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struct Biol 21:150–160. doi:10.1016/j.sbi.2011.01.011
12. Christ CD, Fox T (2014) Accuracy assessment and automation of free energy calculations for drug design. J Chem Inf Model 54:108–120. doi:10.1021/ci4004199
13. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432:862–865. doi:10.1038/nature03197
14. Mobley DL, Graves AP, Chodera JD et al (2007) Predicting absolute ligand binding free energies to a simple model site. J Mol Biol 371:1118–1134. doi:10.1016/j.jmb.2007.06.002
15. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N et al (2014) rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol 10:e1003571. doi:10.1371/journal.pcbi.1003571

16. Joseph-McCarthy D, Thomas BE, Belmarsh M et al (2003) Pharmacophore-based molecular docking to account for ligand flexibility. Proteins 51:172–188. doi:10.1002/prot.10266
17. Hindle SA, Rarey M, Buning C, Lengaue T (2002) Flexible docking under pharmacophore type constraints. J Comput Aided Mol Des 16:129–149
18. Good AC, Cheney DL, Sitkoff DF et al (2003) Analysis and optimization of structure-based virtual screening protocols. 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success. J Mol Gr Model 22:31–40. doi:10.1016/S1093-3263(03)00124-4
19. Barril X, Morley SD (2005) Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. J Med Chem 48:4432–4443. doi:10.1021/jm048972v
20. Wright L, Barril X, Dymock B et al (2004) Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. Chem Biol 11:775–785. doi:10.1016/j.chembiol.2004.03.033
21. Ruiz-Carmona S et al (2016) Dynamic undocking and the Quasi-Bound state as tools for drug design. Nat Chem, In press
22. McGovern SL, Shoichet BK (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 46:2895–2907. doi:10.1021/jm0300330
23. Verdonk ML, Mortenson PN, Hall RJ et al (2008) Protein-ligand docking against non-native protein conformers. J Chem Inf Model 48:2214–2225. doi:10.1021/ci8002254
24. Barril X, Hubbard RE, Morley SD (2004) Virtual screening in structure-based drug discovery. Mini Rev Med Chem 4:779–791
25. Bavi R, Kumar R, Choi L, Woo Lee K (2016) Exploration of novel inhibitors for Bruton's tyrosine kinase by 3D QSAR modeling and molecular dynamics simulation. PLoS One 11:e0147190. doi:10.1371/journal.pone.0147190
26. Quesada-Romero L, Mena-Ulecia K, Tiznado W, Caballero J (2014) Insights into the interactions between maleimide derivates and GSK3β combining molecular docking and QSAR. PLoS One 9:e102212. doi:10.1371/journal.pone.0102212
27. Morley SD, Afshar M (2004) Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. J Comput Aided Mol Des 18:189–208
28. Clark M, Cramer RD, Van Opdenbosch N (1989) Validation of the general purpose tripos 5.2 force field. J Comput Chem 10:982–1012. doi:10.1002/jcc.540100804
29. Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2016.
30. LigPrep, version 2.3, Schrödinger, LLC, New York, NY, 2009.
31. Case DA, Babin V, Berryman JT, et al (2014) AMBER 14. University of California, San Francisco.
32. Bayly CI, McKay D, Truchon J-F (2011) An informal AMBER small molecule force field: parm@Frosst
33. Kroemer RT (2003) Molecular modelling probes: docking and scoring. Biochem Soc Trans 31(5):980–984. doi:10.1042/BST0310980
34. Yusuf D, Davis AM, Kleywegt GJ, Schmitt S (2008) An alternative method for the evaluation of docking performance: RSR vs RMSD. J Chem Inf Model 48:1411–1422. doi:10.1021/ci800084x
35. Warren GL, Do TD, Kelley BP et al (2012) Essential considerations for using protein-ligand structures in drug discovery. Drug Discov Today 17:1270–1281. doi:10.1016/j.drudis.2012.06.011
36. Cozzini P, Kellogg GE, Spyrakis F et al (2009) Target flexibility: an emerging consideration in drug discovery and design. J Med Chem 51:804–828. doi:10.1021/jm800562d.Target
37. Spyrakis F, BidonChanal A, Barril X, Luque FJ (2011) Protein flexibility and ligand recognition: challenges for molecular modeling. Curr Top Med Chem 11:192–210. doi:10.2174/156802611794863571
38. Barril X, Fradera X (2006) Incorporating protein flexibility into docking and structure-based drug design. Expert Opin Drug Discov 1:335–349. doi:10.1517/17460441.1.4.335
39. Cheng LS, Amaro RE, Xu D et al (2008) Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. J Med Chem 51:3878–3894. doi:10.1021/jm8001197
40. Abagyan R, Rueda M, Bottegoni G (2010) Recipes for the selection of experimental protein conformations for virtual screening. J Chem Inf Model 50:186–193. doi:10.1021/ci9003943
41. Campbell AJ, Lamb ML, Joseph-McCarthy D (2014) Ensemble-based docking using biased molecular dynamics. J Chem Inf Model 54:2127–2138. doi:10.1021/ci400729j
42. Birch L, Murray CW, Hartshorn MJ et al (2002) Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. J Comput Aided Mol Des 16:855–869. doi:10.1023/A:1023844626572
43. Barril X (2014) Ligand discovery: Docking points. Nat Chem 6:560–561. doi:10.1038/nchem.1986
44. Fischer M, Coleman RG, Fraser JS, Shoichet BK (2014) Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. Nat Chem 6:575–583. doi:10.1038/nchem.1954
45. Álvarez-García D, Barril X (2014) Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. J Med Chem, 57(20):8530–8539. doi:10.1021/jm5010418
46. Bento AP, Gaulton A, Hersey A et al (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42:1083–1090. doi:10.1093/nar/gkt1031
47. Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235
48. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20:318–331. doi:10.1016/j.drudis.2014.10.012
49. Wale N (2011) Machine learning in drug discovery and development. Drug Dev Res 72:112–119. doi:10.1002/ddr.20407