

Challenges of docking in large, flexible and promiscuous binding sites.

Martin Kotev^a, Robert Soliva^b and Modesto Orozco^{a,c*}

^a *Joint BSC-IRB Program in Computational Biology. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona 08028, Spain*

^b *Joint BSC-IRB Program in Computational Biology. Barcelona Supercomputing Center. Barcelona 08031, Spain*

^c *Department of Biochemistry and Molecular Biology, University of Barcelona (UB), Barcelona 08029, Spain*

* Corresponding author. Modesto Orozco. modesto.orozco@irbbarcelona.org

Keywords: Induced-fit docking, PELE simulations, soluble epoxide hydrolase, benchmark

ABSTRACT

After decades of work, the correct determination of the binding mode of a small molecule into a target protein is still a challenging problem, whose difficulty depends on: i) the sizes of the binding site and the ligand; ii) the flexibility of both interacting partners, and iii) the differential solvation of bound and unbound partners. We have evaluated the performance of standard rigid(receptor)/flexible(ligand) docking approaches with respect to last-generation fully flexible docking methods to obtain reasonable poses in a very challenging case: soluble Epoxide Hydrolase (sEH), a flexible protein showing different binding sites. We found that full description of the flexibility of both protein and ligand and accurate description of solvation leads to significant improvement in the ability of docking to reproduce well known binding modes, and at the same time capture the intrinsic binding promiscuity of the protein.

1. Introduction

Accuracy in predicting how a small molecule binds to its target is an essential requirement for structure-based drug design (SBDD). The correct pose prediction of a small molecule in a binding site allows gaining insight on which chemical features are essential for achieving its biological activity,^{1,2} and using this information to improve the properties of the lead compound. In principle, the binding mode of ligands can be obtained by means of experimental techniques, but the time-restrictions in practical drug design projects are often incompatible with X-Ray and NMR time scales, forcing the use of computational approaches, particularly of docking algorithms.³

Current docking methods are typically based on a rigid receptor around which a ligand is rotated, translated and flexed until the best fit is found. Although docking has been proven to be an extremely useful tool,⁴ it yields suboptimal results in cross-docking experiments, i.e when one ligand is docked to the image of the receptor that is either unbound, or bound to a significantly different ligand.⁵ Bearing in mind that most macromolecular targets undergo some kind of induced-fit upon binding,⁶ the number of cases where standard docking procedures find problems is not negligible.

A simple strategy to model the flexibility of the macromolecular receptor is to submit it to a molecular dynamics (MD) simulation, from which a series of relevant conformations for rigid docking are derived.^{2,3} Although this might mitigate the negative impact of the rigid receptor approximation, plain MD does not usually provide sufficient protein conformational sampling, and there are not clear rules to extract a subset of representative configurations from the crude ensemble.⁷ Similar or even larger problems arise when protein models for docking are obtained from experimental ensembles.⁸ Alternative strategies to ensemble docking arise from the use of Induced-Fit Docking algorithms. One such approach is based on the iterative combination of rigid receptor docking plus protein structure prediction for certain areas of the active site.⁹ This method has proven to be especially relevant for detecting significantly large conformational transitions upon ligand binding, such as the kinase DFG-in/out conformational change.⁹ However, in general, solving the flexibility problem requires direct coupling between conformational exploration and binding, which in principle can be achieved by brute force molecular dynamics algorithms.¹⁰ Unfortunately, these calculations are still in the proof-of-concept stage requiring highly specialized computer resources which are rarely accessible to standard drug-design laboratories. Use of biased MD simulations can reduce the cost of calculations, but they are still too expensive and requires a previous knowledge of the expected binding mode,¹¹ something uncommon in the practical drug design scenario.

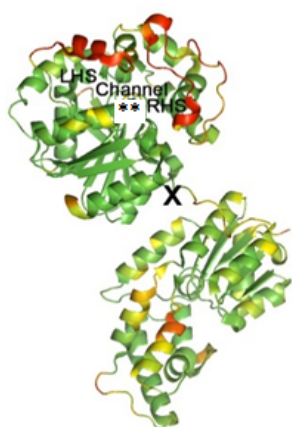
A few intermediate methods aiming to couple docking with conformational transitions, but escaping from the complexity of brute force MD simulations have been suggested. For example, in the essential dynamics molecular dynamics (ED/MD) approach the protein dynamics in the essential deformation space is coupled with sampling of ligand

degrees of freedom in the Cartesian space^{12,13} allowing a coupled sampling of protein and ligand flexibilities. Medusadock,¹⁴ models receptor and ligand flexibility simultaneously by using sets of discrete rotamers and RosettaLigand¹⁵ allows docking ligands to receptors by using a Monte Carlo minimization procedure in which the rigid body position and orientation of the small molecule and the protein side-chain conformations are optimized simultaneously.¹⁶ Finally, PELE (Protein Energy Landscape Exploration)¹⁷⁻²² implements a Metropolis Monte Carlo algorithm, where new trial configurations are produced for both ligand and protein while docking is performed. The program implements side chain prediction algorithms, specific ligand rotamer libraries, implicit solvent model,²³ and all-atom OPLS2005 force field²⁴ to compute receptor-ligand interactions.

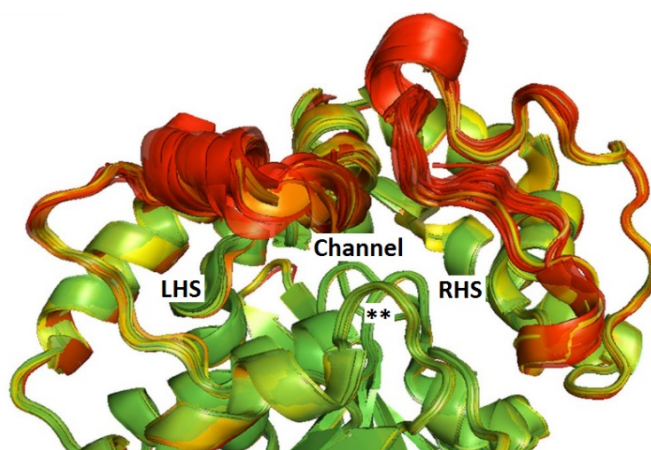
We explore here the performance of two state-of-the-art docking methods: GLIDE, based on the traditional paradigm of a rigid receptor and a flexible ligand, and PELE, a fully flexible approach. Of note, for the rigid receptor approximation, both standard docking against a series of experimental X-ray structures and docking against two MD-generated ensembles are tested. The benchmark system selected was Soluble Epoxide Hydrolase (sEH, Fig. 1a,b), an enzyme that metabolizes epoxyeicosanoic acids (EETs) generating products that have been linked to a variety of therapeutic areas, such as inflammation, pain and hypertension.²⁵ This protein was chosen, not only for its potential pharmacological interest, but because of its novelty (structural data on this protein were not available when GLIDE or PELE were developed) and its complexity, since it shows^{26,27} a complex binding site characterized by a huge hourglass-shaped binding cavity with the catalytic residues in a center channel separating a left-hand site (LHS) from a right-hand site (RHS), both able to bind ligands.^{26,27} The combination of high flexibility and complexity of the binding cavity (divided in 3 different sub-cavities) makes sEH an extremely challenging benchmark for docking approaches.

Figure 1.

a)



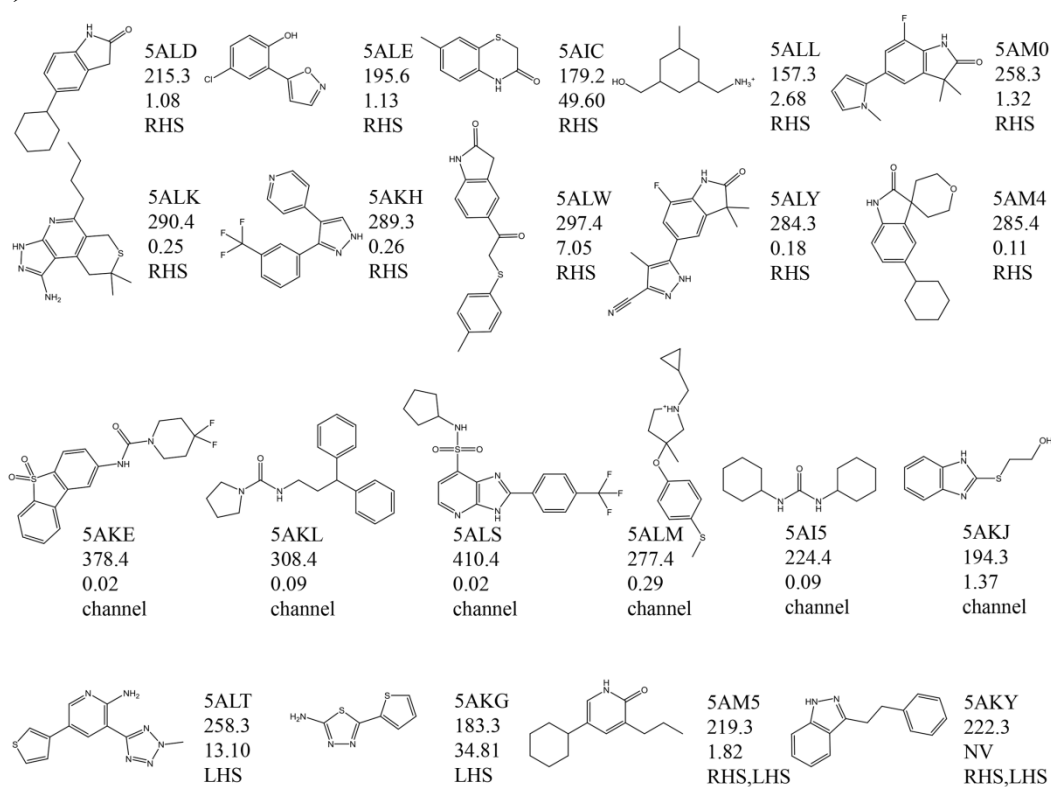
b)



Results presented here confirm the power of standard docking approaches (exemplified by GLIDE) but also their caveats. It is found that generating an ensemble of conformations via MD can in some instances improve the results as compared to using only an X-ray structure. However, for this challenging system it is found that representing both small molecule and protein flexibility simultaneously is crucial in order to have reasonable poses, to discriminate between alternative binding sites and to derive poses that can be useful for further lead optimization studies. In this sense, the use of fully flexible methods as PELE provides much improved results.

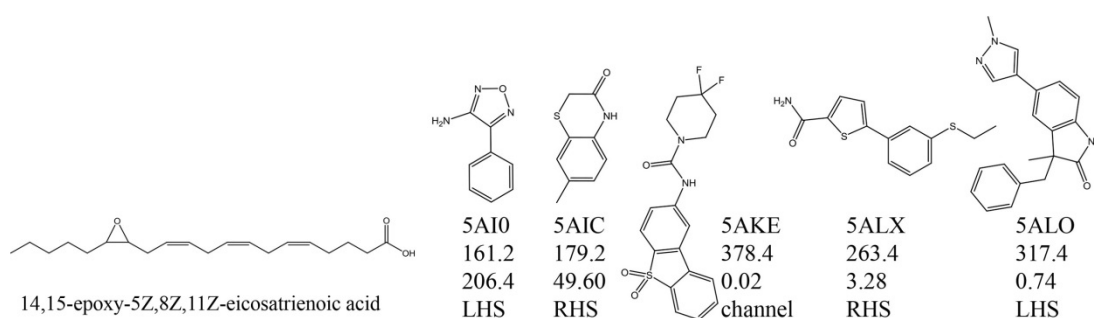
Figure 2.

a)

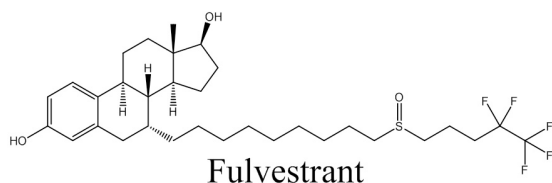


b)

c)



d)



2. Methods and computational details

More than 90 structures of human sEH are deposited in the PDB. A recently deposited one (PDB entry 5AHX) corresponds to the *apo* structure, while a variety of other crystals correspond to complexes with ligands of different sizes bound to different subcavities. For our docking experiment we selected 20 of these ligands based on diversity criteria (see Figure 2) and cross-docked them to 6 protein structures corresponding to the unbound *apo* structure (5AHX) and 5 bound states (see Figure 2) selected to cover the different binding modes of the protein (5AI0, 5AIC, 5AKE, 5ALY and 5ALO). Additionally, for the MD-ensemble approach, a series of protein conformations were generated by standard molecular dynamics simulations for the *apo* structure (5AHX) and also for the complex of sEH when bound to Fulvestrant (4J03), an anti-cancer agent that was serendipitously found to be a potent sEH inhibitor (see Figure 2). All the protein models were prepared for further calculations using the Protein Preparation Wizard (PPW) tool of Schrödinger software^{28,29} thus building any incomplete sidechains, optimizing the hydrogen bond network and predicting the protonation states of all His, Asp, Glu, Lys, and Arg sidechains. Solvent molecules were removed in all cases, as well as any other molecules that might be present from the crystallization buffers.

2.1. CMIP calculations

Classical molecular interaction potentials (CMIP using Na^+ and CH_4 as probes;³⁰) were used to define the recognition characteristics of sEH binding site. The electrostatic term in CMIP was obtained by solving the non-linear Poisson Boltzman equation³¹ (with standard protein and solvent dielectric definition, while the steric terms were accounted for by van der Waals formalism using standard AMBER parameters.³⁰)

2.2. Standard docking

Standard docking calculations (rigid target, flexible ligand) were performed using GLIDE, one of the most validated docking programs in the field.³²⁻³⁴ GLIDE grids were centered on catalytic triad residue D335, and extended to cover the whole binding site. Standard defaults for grid creation and ligand flexible SP (Standard Precision) docking were used.

2.3. MD-generated ensembles

In order to complete the comparison between fully flexible PELE simulations and rigid(receptor)/flexible(ligand) calculations with GLIDE against a series of X-ray structures, two MD generated ensembles were also tested. One of them was extracted from a 200 ns. MD trajectory of the apo structure (5AHX). The second was extracted from a 200 ns. MD trajectory of the holo structure of sEH when bound to Fluvestrant, which is an anti-cancer agent found to be a potent inhibitor of sEH and which, because of its big size (see Figure 2d), binds the enzyme occupying all LHS, channel and RHS subsites.

Molecular dynamics simulations of the two systems were performed with AMBER14³⁵ and the ff14SB force field. Explicitly solvated (TIP3P waters)³⁶ systems were simulated in the NPT ensemble in a truncated octahedron box. Sodium and chlorine ions were added to simulate physiological ionic strength and to bring the total charge to neutrality. The lid portion of sEH was removed in all simulations. SHAKE³⁷ was used to constrain chemical bonds, which allowed us to use an integration step of 2fs. For Fulvestrant, RESP charges were calculated at HF/6-31G** level together with GAFF³⁸ parameters. Total production runs of 200 ns were collected, with snapshots saved every 20 ps. All analyses from the simulations were carried out with CPPTRAJ³⁹. The clustering and extraction of snapshots for the simulations was performed by a hierarchical agglomerative algorithm and average-linkage method based on C α pairwise distances. After clustering, a representative set of 4 snapshots from the first 4 most populated clusters were taken for each of the two simulations and used for the ensemble rigid(receptor)/flexible(ligand) docking calculations.

2.4. PELE simulations

PPW-refined protein was manipulated to reduce its size by removing the lid region (Fig. 1a), which helped us to reduce the cost of calculations. The ligands under study were placed outside the protein at distances of around 20 – 25 Å RMSd away from the binding site, and an approximate distance of 3 – 5 Å to the loop K495-V500 (the selected position gives approximately equal access for entering and exploration of the LHS and RHS). For each simulation a sphere with center around the catalytic residue D335 and radius of 14 to 18 Å was used to limit the exploration to the approximate area of the sEH active site.

Following standard PELE use, the ligand originally placed at a random position is translated and rotated, and the protein is perturbed following the normal modes of lower frequency. In a further step, the complex is readapted, eliminating possible tensions produced in the perturbation step by performing a side chain prediction and full minimization. The resulting structure is accepted or rejected following the Metropolis criterion.

Ligand perturbation was switched between two modes of translation: a large one of 3.0 Å (when the ligand is far from the protein based on ligand SASA), and a small one of 0.5 Å (when the ligand is close to the protein). Along with the translation, the ligand was randomly rotated (rigid body rotation) in the range 0-15°. Backbone movements were performed by moving along the first six modes calculated from an elastic deformation model.⁴⁰ This was accomplished by performing an all-atom geometry optimization, where a constraint is added to each alpha carbon along a randomly chosen mode (with displacements of 1.0 Å). Finally a relaxation was accomplished with a 10° resolution side chain sampling for those residues having at least one atom within a 4 Å distance of any ligand atom (a 30° step was used for the ligand rotatable bonds).¹⁸ The final minimization included a weak constraint (1.0 kcal/mol/Å) every 10 alpha carbons, which help to avoid compaction of the protein due to the lack of explicit solvent (more details about PELE and setting up the simulations can be found at <https://pele.bsc.es>).

3. Results

3.1. Binding site analysis

As shown in Figure 1 most of the backbone flexibility in the sEH binding site is located in a few loops, namely those composed by amino acids N368-P379 (which make up part of the “ceiling” of the LHS), amino acids A411-M419 (which make up the whole of the ceiling of the RHS) and to a lesser extent, to the loop composed by K495-V500 (which is placed in the middle of the structure, lining a side of the central channel). These loop movements are responsible for most of the active site changes in volume and recognition. However, even with very slight backbone displacements, the subsites change in shape and recognition due also to a variety of side chain movements. When the X-ray structures of the complexes are superimposed, especially relevant side chain flexibility movements are seen for residues M310, W336, M339, Y343, I363, I375, Q384, M469 in the LHS and residues L408, L417, M419, V498 in the RHS. Of note, most of these residues are mainly big and hydrophobic, a feature that is line with the nature of the endogenous substrates of sEH (for an example, see Figure 2b), which are flexible carboxylic-acid bearing metabolites characterized by long aliphatic chains.

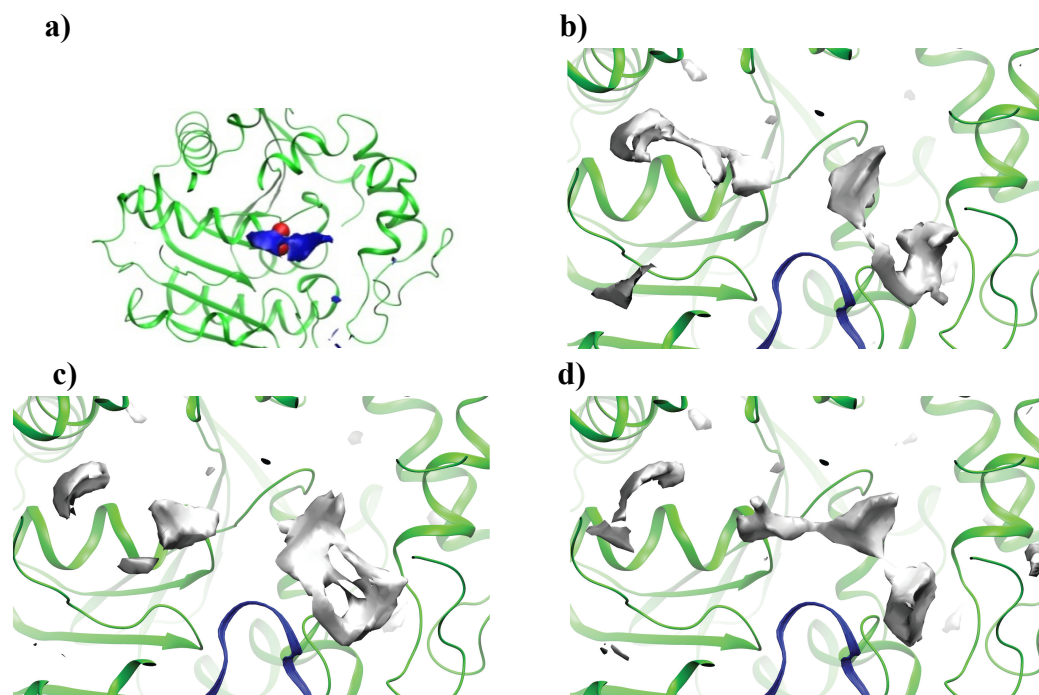
The mainly hydrophobic nature of the active site can be perceived in the CMIP profiles obtained for the +1 charged probe particle in the case of the apo structure of the enzyme. As can be seen in Figure 3a, most of the favorable interaction points are basically located in the central channel, where the catalytic triad (D335, Y383, and Y466) is found. This means that, as suspected, hydrophobic and van der Waals interactions will have a strong impact in the biological activity of the inhibitors binding either LHS or RHS. Of note, this analysis was replicated for a series of *holo* structures finding very close results in all cases.

In order to better perceive the induced fit effects of sEH in the different subsites, a series of CMIP calculations were carried out with the goal of comparing the three

binding pockets. The adaptation of the protein subsites to the presence of some of the inhibitors found in Figure 2 was tested. As a sample, three of them are discussed herein – 5ALT, 5ALY and 5ALS as representatives of LHS, RHS and Channel pockets, respectively. As seen on Figure 3, 5ALT and 5ALY are characterized by a lack of favorable vdW interactions (closed channel) in the catalytic triad region compared to 5ALS structure where a thin surface joins RHS with LHS is found. In 5ALT the RHS vdW surface is comparable in size to the LHS, the last is mainly formed by 4 hydrophobic and flexible sidechains of M310, M339, M469, I363 plus W336. CMIP calculations for the representative of the RHS complexes (5ALY, Fig. 3b) show better options for hydrophobic interactions in the right pocket. The main residue that increases the surface of RHS in 5ALY compared to 5ALT (and also changes frequently its position in the different crystals via induced-fit movements) is M419, while I363 and I375 make the difference between 5ALT and 5ALY in the LHS. The Channel representative – 5ALS (Fig. 3d) displays a clear shift and an opening movement of the entire K495-V500 loop and the formation of the channel “tube” together with a new hydrophobic pocket around F381.

In summary, this binding site analysis clearly underlines the plasticity of the each of the three subsites of sEH, which can adapt their conformation to the bound ligand by varying backbone and side chain conformations. The different subsite conformations are clearly incompatible for binding molecules other than the ones they were crystallized with. This picture translates, as could be expected, in a tough system for ligand pose prediction by rigid docking techniques (see below).

Figure 3.



3.2 Binding mode prediction by standard docking against X-ray structures

Both HTS and NMR-based screening experiments found an extremely high degree of hits for sEH, suggesting a large promiscuity probably related to its large and hydrophobic binding site. Very encouragingly, GLIDE is able to capture the “sticky” nature of the protein. Encouragingly, all the inhibitors considered (Figure 2a) appear as hits in the standard docking experiments. However, when going to a more detailed analysis less brilliant results emerge (see Table 1, Fig. 4) as across the 120 (20 ligands vs 6 protein structures) pairs calculated, only 19 top scoring poses are below 3.0 Å in RMSd (from the experimental binding pose), the number successes being reduced to 6 if the RMSd threshold is reduced to 2 Å).

Table 1

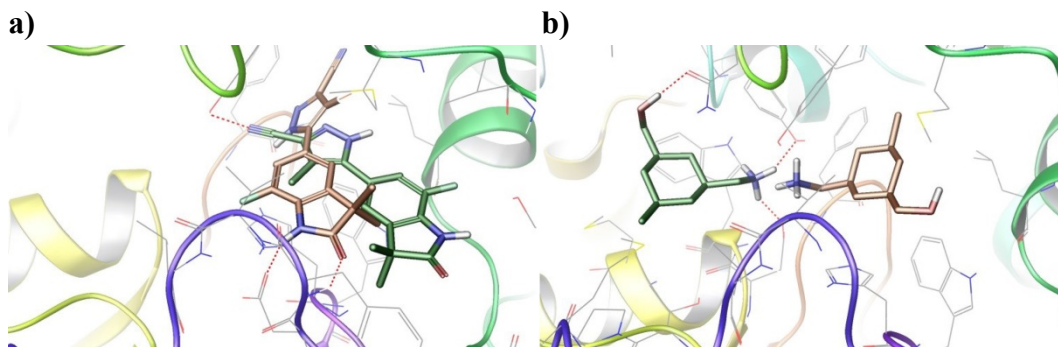
Glide SP cross-docking results for the set of 20 ligands (Figure 2a) against the 6 protein structures (Figure 2c). RMSd of the top scored pose / best RMSd amongst the top five scored poses are given when compared to the experimental pose. The ligand-protein pairs for the selected complexes chosen for subsequent PELE simulations are shown in bold.

Protein PDB entry	Ligand PDB entry									
	5AI5	5AIC	5AKE	5AKG	5AKH	5AKJ	5AKL	5AKY*	5ALD	5ALE
5AHX	1.5/1.4	3.1/3.1	2.9/2.8	15.3/5.7	1.5/1.5	7.2/7.2	10.4/10.1	2.1/2.1	2.5/2.5	1.0/1.0
5AI0	7.3/6.8	4.6/2.8	13.2/3.2	6.4/5.8	13.0/12.8	5.5/5.5	10.5/10.1	4.7/2.8	2.2/2.2	2.2/0.7
5ALX	7.6/7.1	11.1/3.0	10.9/10.9	6.3/5.7	12.2/12.2	8.6/5.9	9.9/9.8	2.7/2.3	11.8/5.7	11.3/1.0
5ALO	6.9/6.8	11.4/10.3	9.8/3.0	6.4/5.3	12.9/12.6	7.9/8.0	10.6/10.3	7.1/4.3	11.4/11.4	11.8/11.4
5AKE	0.8/0.4	9.5/5.0	2.5/2.4	7.0/5.3	13.5/11.2	6.3/1.2	1.4/1.4	3.7/3.7	3.0/3.0	5.7/5.7
5AIC	7.8/6.7	10.5/4.6	12.8/3.0	6.2/5.4	13.5/13.1	8.7/8.3	10.9/10.2	5.4/2.5	2.3/2.3	1.2/1.2

Protein PDB entry	Ligand PDB entry									
	5ALK	5ALL	5ALM	5ALS	5ALT	5ALW	5ALY	5AM0	5AM4	5AM5*
5AHX	13.1/12.8	3.3/3.2	6.4/1.5	7.5/7.5	14.8/14.6	15.3/5.5	6.5/3.9	5.7/4.4	4.1/2.1	2.0/2.0
5AI0	13.3/12.4	4.0/4.0	6.8/5.6	8.1/8.1	14.9/13.9	6.0/5.6	4.4/4.4	3.7/3.7	2.0/1.6	2.1/1.6
5ALX	12.9/12.3	9.9/9.8	5.1/5.4	8.0/7.9	14.8/3.3	15.5/13.5	15.3/13.8	12.8/12.6	14.2/14.2	6.7/5.5
5ALO	12.8/12.1	8.9/8.9	6.3/5.8	10.8/10.3	4.6/4.6	15.6/6.0	12.1/12.1	12.3/11.9	14.3/4.0	5.2/5.0
5AKE	12.3/12.7	4.2/4.2	4.9/3.8	2.1/2.1	14.7/6.1	8.7/4.5	12.2/12.3	12.0/12.0	13.3/12.9	6.1/6.0
5AIC	12.6/12.4	10.2/9.8	6.2/4.8	11.8/11.8	14.9/14.2	14.4/6.0	13.9/13.7	3.4/3.4	15.4/14.6	1.7/1.6

* — only the best result from both pockets LHS and RHS is shown

Figure 4.



The problem is not related to small errors in the scoring function, as the results do not improve dramatically if not only the 1st, but the five best scored poses are considered (see Table 1). In fact, for many ligands such as those appearing in 5AKG, 5AKJ, 5ALK, 5ALL or 5ALM (to name a few) it is impossible to find a reasonable binding mode when docked against any of the 6 structures of sEH considered. Likewise, GLIDE has severe problems finding correct binding modes for some protein conformations. For example, for 5ALO no ligand could be cross-docked correctly, for 5ALX only one is successfully cross-docked. Quite surprisingly, the *apo* structure seems to be among the best in accommodating more ligands (see 5AHX in Table 1), suggesting that when the protein adapts its structure to a given ligand it adopts a conformation that does recognize well many other compounds. This is confirmed by the fact that dual binding modes (for example 5AM5 and 5AKY) are not successfully recovered in GLIDE cross-docking experiments (see Table 1).

In summary, despite the general success of standard docking procedures (exemplified by GLIDE) to recognize potential ligands of sEH, binding poses are wrong in several cases, providing structural information which can mislead lead optimization procedures (see for instance Fig. 4). It seems that significant protein relaxation is coupled to the effective binding of ligands, which clearly reduces the success of standard docking procedures.

3.3 Binding mode prediction by standard docking against an MD-generated ensemble

As a further test of what can be expected from the rigid(receptor)/flexible(ligand) docking approach, it was decided to take a small ensemble of structures generated from MD simulations, to see if the results against X-ray structures could be improved by adding additional conformations of the protein. To this end, 2 MD simulations were carried out. One of them was that of the *apo* structure deposited in the pdb (5AHX). The second was the *holo* structure of sEH in complex with Fulvestrant (4J03).⁴¹ This complex was especially interesting as this is a big drug whose binding mode implies occupying all three subsites. From each of the 2 simulations, 4 representative snapshots

were used for docking the 20 compounds found in Figure 2a. The results are summarized in Table 2. Its analysis reveals a series of points.

The crystal structure 4J03, in line with the crystal structures featured in Table 1 above, seems to be able to recapitulate a small number of correct poses (only 3 ligands are placed under 3 Angstroms RMSd from their native poses as top scored). Encouragingly, when the ensemble of MD-derived structures (4J03_1 to 4J03_4) is taken instead of the single X-ray, the results improve for a few cases (see for instance 5AKE, 5AKL, 5ALL and 5ALS, where at least one MD-snapshot does indeed recapitulate the correct binding mode as top pose). However, it must be borne in mind that in some cases (for instance 5ALM) the ensemble seems to be worse than the X-ray structure.

A poorer performance is seen for the MD-generated ensemble of the apo structure (5AHX_1 to 5AHX_4). Not a single miss of its X-ray structure is recapitulated by the MD-ensemble, and discouragingly some of the ligands found to dock correctly on the X-ray are totally missed by the ensemble (see for instance 5AKE, 5AKH, 5ALE or 5AM5).

Table 2

Glide SP ensemble docking results for the set of 20 ligands (Figure 2a) against sEH structures derived from 2 MD simulations. 4 representative snapshots extracted from each of the MD trajectories of 5AHX and 4J03 were used (identified as 4J03_1 to 4J03_4 and 5AHX_1 to 5AHX_4). RMSd of the top scored pose / best RMSd amongst the top five scored poses are given when compared to the experimental pose.

Protein structure	Ligand PDB entry									
	5AI5	5AIC	5AKE	5AKG	5AKH	5AKJ	5AKL	5AKY*	5ALD	5ALE
4J03	1.4/0.6	4.7/4.7	3.3/3.3	6.6/6.3	5.7/4.8	2.0/1.3	7.5/1.5	6.1/1.8	9.8/7.1	5.0/5.0
4J03_1	1.6/1.6	10.3/8.9	9.0/9.0	7.9/7.7	13.1/11.7	1.2/1.1	10.3/9.5	5.2/5.2	11.2/10.3	5.0/5.0
4J03_2	1.4/0.9	9.1/9.1	8.9/8.8	13.5/9.9	14.6/12.4	1.5/1.0	8.0/7.9	5.4/5.4	12.7/10.2	5.0/5.0
4J03_3	1.5/1.4	8.7/7.7	2.6/2.6	14.1/7.1	13.6/11.1	2.5/1.3	1.00/0.7	4.9/4.8	12.0/5.8	4.8/4.8
4J03_4	1.4/1.4	11.8/3.1	2.8/2.8	6.6/6.5	14.4/6.2	6.7/6.1	12.2/11.2	6.1/6.0	7.6/4.1	7.3/2.9
5AHX	1.5/1.4	3.1/3.1	2.9/2.8	15.3/5.7	1.5/1.5	7.2/7.2	10.4/10.1	2.1/2.1	2.5/2.5	1.0/1.0
5AHX_1	6.6/6.4	11.1/10.1	3.4/3.3	6.8/6.3	13.6/11.4	6.7/6.7	11.8/10.6	7.0/2.9	10.6/10.4	10.8/10.4
5AHX_2	1.3/1.2	5.0/3.3	13.3/13.1	16.0/7.1	14.1/12.7	8.1/7.8	1.4/1.4	6.2/4.6	14.4/13.9	2.6/2.6
5AHX_3	2.3/2.0	14.9/10.6	10.0/10.0	13.8/6.3	15.6/13.5	5.2/5.2	11.2/10.8	6.3/4.7	14.3/12.2	6.0/6.0
5AHX_4	5.7/5.6	11.9/4.6	9.7/9.5	15.0/7.3	12.5/5.3	5.0/4.1	10.1/10.1	5.9/3.8	12.4/3.7	4.6/4.6

Protein structure	Ligand PDB entry									
	5ALK	5ALL	5ALM	5ALS	5ALT	5ALW	5ALY	5AM0	5AM4	5AM5*
4J03	15.1/14.9	9.9/1.0	2.2/1.3	8.4/2.8	13.1/12.0	8.2/5.7	7.0/7.0	14.5/12.0	3.2/3.2	7.3/4.8

4J03_1	14.4/11.3	10.6/7.8	8.2/4.2	11.5/6.5	6.6/6.6	8.3/8.3	12.3/10.6	12.8/12.4	12.5/12.1	6.0/5.7
4J03_2	10.6/9.7	2.3/2.3	7.8/5.8	2.1/2.1	7.1/6.3	8.3/7.8	12.1/10.6	11.8/11.2	12.5/12.2	6.6/6.6
4J03_3	13.3/13.1	7.4/7.4	8.8/2.9	2.3/2.3	14.3/10.8	13.5/6.8	12.2/7.2	11.6/11.6	14.2/14.0	6.1/5.6
4J03_4	16.9/14.1	1.1/0.9	9.6/9.3	9.9/9.9	12.2/12.0	4.7/4.7	5.2/4.6	4.7/3.6	3.8/3.5	5.9/5.5
5AHX	13.1/12.8	3.3/3.2	6.4/1.5	7.5/7.5	14.8/14.6	15.3/5.5	6.5/3.9	5.7/4.4	4.1/2.1	2.0/2.0
5AHX_1	11.1/11.1	9.8/8.6	5.3/5.3	9.1/3.5	5.9/5.9	13.8/13.3	15.1/14.5	12.3/11.4	13.2/13.1	4.9/4.9
5AHX_2	12.4/12.4	5.0/2.7	9.0/8.4	9.2/8.4	3.9/3.7	14.4/12.4	15.8/11.7	11.7/11.7	15.0/14.8	5.4/5.4
5AHX_3	14.5/13.1	11.3/9.8	8.4/8.1	13.6/12.5	12.3/6.6	14.8/12.0	15.4/15.4	14.3/14.0	15.4/15.2	5.9/4.8
5AHX_4	5.5/5.0	10.6/3.2	9.8/8.7	4.9/4.9	13.0/13.0	4.0/4.0	15.4/6.3	4.9/4.6	13.3/5.3	5.3/5.3

* – only the best result from both pockets LHS and RHS is shown

In summary, it is seen that ensemble docking against a series of MD-derived structures can allow recovering the correct pose for some ligands that could not be docked correctly against their X-ray. However, in the case of SEH, this is only true for the holo structure tested, but not the apo. Furthermore, in a significant number of cases, ensemble docking based on the holo structure still does not lead to correct, productive poses.

3.4 Binding mode prediction by flexible docking

A set of ligand/protein conformation pairs found in Table 1 were selected for PELE flexible docking calculations. These pairs (in **bold** in Table 1) were randomly selected among the “hard” cases for which standard docking fails to provide good poses, representing then a very difficult test set for PELE. Very remarkably, the results of the PELE simulation are strikingly good (Table 3 and Figure 5), since for all the cases the best scored pose is at less than 3 Å from the experimental one, and in more than 70% of the cases the RMSd between predicted (as the minimum energy structure) and experimental binding pose is less than 2 Å. This means that PELE provides then structural models of enough quality as to be useful in lead optimization procedures. Interestingly, in a few cases, the program finds one or two alternative solutions, distinct from the experimental one, which are very close in energy to the correct one, confirming the extreme promiscuity of this protein.

Table 3

PELE induced-fit docking results for the tough pairs highlighted in Table 1 RMSd of the top scoring poses (lowest interaction energy poses) from the ligand PDB entry native ones.

	Ligand PDB entry/ Protein PDB entry										
	5AI5/ 5ALX	5AIC/ 5AKE	5AKE/ 5ALX	5AKG/ 5AIC	5AKH/ 5AI0	5AKJ/ 5AHX	5AKL/ 5AHX	5AKY/ 5ALO (RHS)	5AKY/ 5ALO (LHS)	5ALD/ 5ALX	5ALE/ 5ALO
RMSd	1.7	0.7	2.6	1.9	1.4	1.6	1.6	2.7	2.1	1.4	1.3

	Ligand PDB entry/ Protein PDB entry										
	5ALK/ 5AHX	5ALL/ 5ALX	5ALM/ 5ALO	5ALS/ 5AI0	5ALT/ 5AHX	5ALW/ 5AI0	5ALY/ 5AI0	5AM0/ 5AHX	5AM4/ 5AKE	5AM5/ 5AKE (RHS)	5AM5/ 5AKE (LHS)
RMSd	0.9	0.8	2.9	2.4	1.6	1.9	0.5	0.8	0.8	1.5	2.5

The quality of PELE-derived poses allowed us to explore the flexibility pattern involved in binding to the different subcavities. Thus, binding to LHS subsite (an example is the binding of the 5ALT ligand to the apo structure (5AHX) requires a rearrangement of the structure via both backbone and sidechain movements which helps to define a good van der Waals recognition basin (around I363, M339, F381 and W336; see Figure 6a). Binding to the Channel requires rearrangements in Y383, Q384 and Y466 to define a good binding pocket, which allows the formation of crucial hydrogen bonds to the catalytic triad. Without these rearrangements the ligand 5AKL is placed by docking procedures more than 10 Å away from its actual binding mode in a different pocket (Table 1), while induced fit in PELE simulations allows to recover two symmetrically equivalent solutions regarding the two phenyl rings of the ligand (Fig. 5b and Fig. 6b). Even bigger conformational changes in the channel are required for the binding of the 5ALS ligand to the 5AI0 structure (compare Figure 5c and 6c) some of them already discussed in the CMIP analysis (see above). Binding of ligands to the RHS subsite is coupled also to some non-negligible conformational changes in the protein. For example, in the case of binding of 5ALK ligand to the *apo* structure, movements of M419 and V498 need to occur to define a successful binding point allowing a productive recognition of D496 and D335, since otherwise the ligand binds at sites that are located more than 13 Å away (see Table 1, Figure 5d and Figure 6d). Slightly different rearrangement (only V498) is required for a productive binding of the small ligand present in 5AIC (Figure 2a) to protein structure in 5AKE - without this change, different incorrect poses are recovered (Table 1, Figures 5e and 6e). Finally, dual RHS-LHS binders (that of compounds 5AM5 and 5AKY) require dual conformational changes in RHS and LHS (M469 in the LHS, and V498 in RHS), since otherwise poses generated are incorrect (Table 1). Fully flexible methods such as PELE allow us to capture properly the complex conformational choreography implicit to dual binding (see Figures 5f and 6f).

As a last note, it should also be mentioned that, in spite of the marked hydrophobic nature of the sEH active site, a few of the ligands partially interact with sEH via bridging water molecules. This is the case for instance of the 5AM5 and the 5ALT ligands (which bind in the LHS) and the 5ALE ligand (which binds in the RHS). As it is discussed above, PELE finds the correct binding mode even for these cases, although the program takes into account the role of solvent by an implicit solvation model, . An explanation for this behavior is that PELE, by way of exploring receptor and ligand degrees of freedom simultaneously, is able to correctly find the correct fit between protein and ligand in the very occluded active site of sEH, compensating for the absence

of certain water molecules. It might be expected that, in cases of more open binding sites, explicit waters should be added into the PELE calculations.

Figure 5.

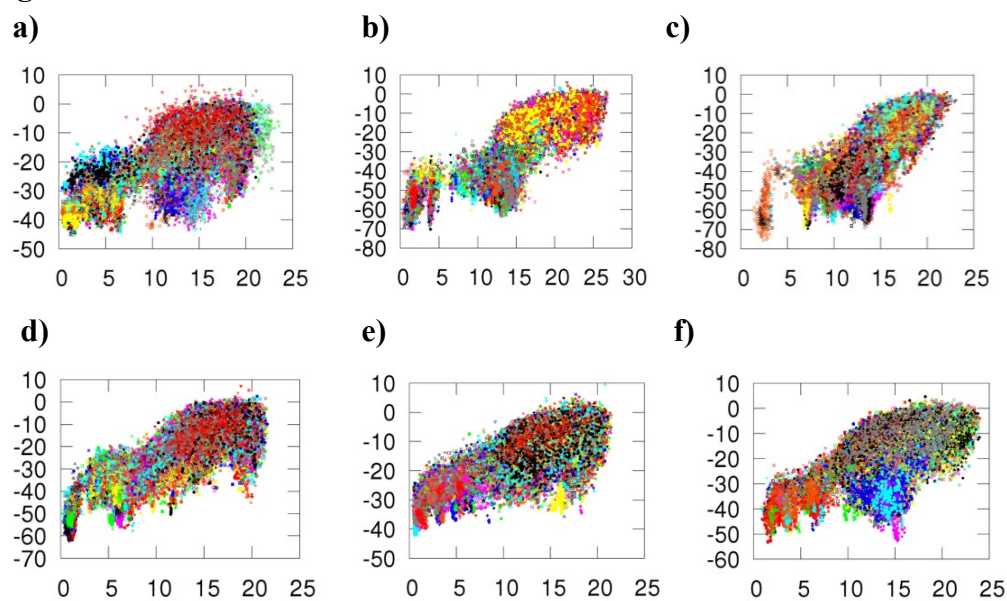
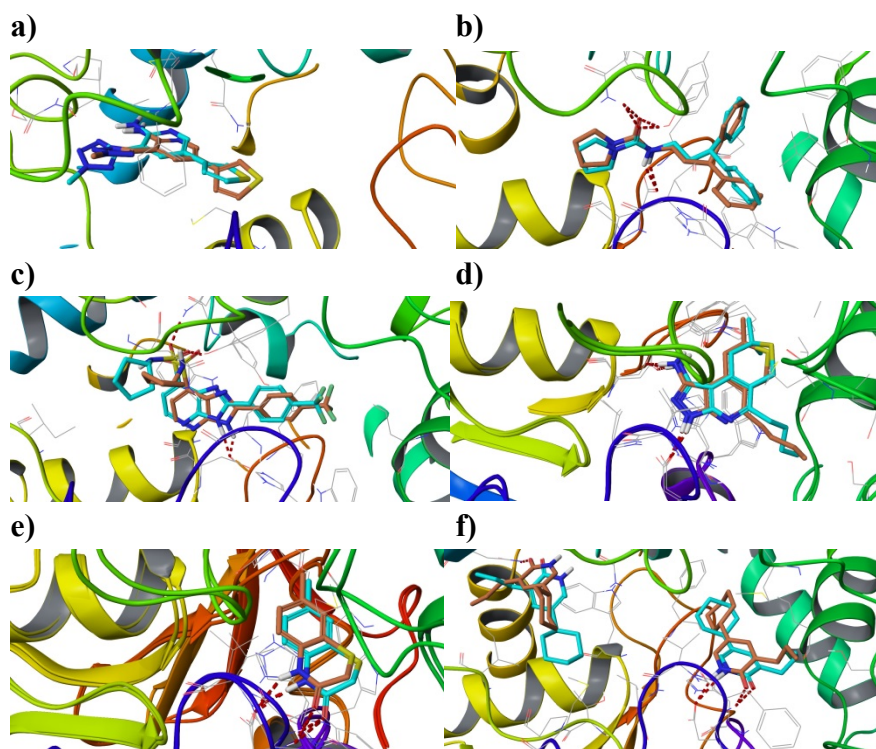


Figure 6.



4. Conclusions

Soluble Epoxide Hydrolase has a very malleable binding site which allows it to adapt to the presence of a wide range of varying chemical cores. Ligand-induced conformational movements lead to very significant changes in the recognition properties of the protein, visible in CMIP calculations, which makes the protein conformation over-specialized to recognize only similar ligands, failing to recognize others with a less similar structure or binding to alternative places of the large binding site. In this type of cases, standard docking approaches (rigid receptor) against an X-ray structure, which dominates the field of high throughput virtual screening, can be useful to retain potential binders but they are likely to produce reliable poses that can guide chemists to improve a lead compound only in a limited number of cases. This deficiency can be partially overcome by ensemble docking with a series of MD-derived structures, but only to a small extent and in case a holo structure is used. Very encouragingly, fully flexible docking methods such as PELE are not only able to successfully predict experimentally-determined binding in all the tested cases, but they are also capable of providing a fuzzy scenario of alternative binding modes which can provide unexpected clues to the design of completely new families of ligands.

5. Acknowledgements

The authors thank Nostrum Biodiscovery for providing us a copy of the PELE code and also Victor Guallar for comments and suggestions. M.K. was supported by a grant of the Fundació Marató TV3. Additional funding to this project comes from H2020 (BioExcel project and Human Brain Flagship project) and the Spanish MINECO (BIO2015/64802-R) funding agencies. IRB Barcelona is recipient of a Severo Ochoa Award of Excellence from MINECO (Government of Spain). MO is an ICREA Fellow researcher.

LEGENDS TO FIGURES

Figure 1. Structure of sEH. a) Catalytic (upper) and lid (lower) domains depicted with a B-factor ribbon (green-yellow-red spectra). The “x” denotes the site at which the lid region was cut to simplify the system. The ** sign is the approximate starting position for the ligands in the PELE simulations. b) Binding site of sEH, highlighting the LHS, RHS, and Channel subsites and B-factor ribbon marking the most flexible loops for 56 PDB entries. The color coding is from green (rigid) to red (flexible).

Figure 2. a) 2D Structures of the 20 compounds used for retrospective binding mode prediction, together with their PDB entries, MW, IC50 (μM) if available, and subsite location (LHS, RHS, Channel); b) The structure of 14,15-epoxy-5Z,8Z,11Z-eicosatrienoic acid c) 2D structures of the ligands whose protein conformation (together with the apo structure) was used for cross-docking the compounds found in a) – their PDB entries, MW, IC50 (μM) if available, and subsite location (LHS, RHS, Channel) are also shown; d) 2D structure of Fulvestrant, a potent inhibitor of sEH.

Figure 3. CMIP contours; a) CMIP surface for the +1 charged probe shown in blue for the apo structure 5AHX (contour for -7kcal/mol). The catalytic residue D335 (carboxyl moiety in red spheres) is shown behind; b)-d) Three generated contours (shown at -3kcal/mol) with a neutral charge probe shown as white surfaces for three representative structures of sEH: b) LHS subsite– 5ALT ; c) RHS subsite – 5ALY; d) Channel subsite– 5ALS. In dark blue the K495-V500 loop is shown.

Figure 4. Some examples of binding mode prediction obtained by the cross-docking calculations with GLIDE. In all cases, the experimental binding mode of the ligand is visualized in brown carbons and tubes whereas the predicted binding mode is visualized in green carbons and tubes: a) Complex 5ALY/5AI0; b) Complex 5ALL/5ALX.

Figure 5. PELE simulation profiles for a few selected ligand-protein complexes of Table 1. Each graph shows the RMSd distance in Å to the experimental ligand pose on the x-axis and the interaction energy in kcal/mol as predicted by PELE during the simulations is on the y-axis. Each color and shape in the graphs represents single PELE trajectory from a total of 128 in every complex simulation; a) 5ALT when simulated against 5AHX (5ALT/5AHX); b) 5AKL/5AHX; c) 5ALS/5AI0; d) 5ALK/5AHX; e) 5AIC/5AKE; f) 5AM5/5AKE.

Figure 6. Top poses predicted by PELE vs. X-ray crystal poses. PELE docked ligands are in light blue carbons (tubes representations), X-ray ligand structures are shown in brown carbon colors and tubes. H-bonds are shown in red dashed lines. a) 5ALT ligand simulated against the 5AHX structure (5ALT/5AHX); b) 5AKL/5AHX; c) 5ALS/5AI0; d) 5ALK/5AHX; e) 5AIC/5AKE; f) 5AM5/5AKE (both pockets).

References

1. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. *Br. J. Pharmacol.* **2011**, *162*, 1239.
2. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E.W. Jr. *Pharmacol. Rev.* **2013**, *66*, 334.
3. Kitchen, D. B.; Decornez, H.; Furr, J.R.; Bajorath, J. *Nat. Rev. Drug Discov.* **2004**, *3*, 935.
4. Good, A. C.; Oprea, T. I. *J. Comput. Aided Mol. Des.* **2008**, *22*, 169.
5. Laurie, A. T.; Jackson, R. M. *Curr. Protein Pept. Sci.* **2006**, *7*, 395.
6. Cozzini, P.; Kellogg, G.; Spyraakis, F.; Abraham, D.; Gabriele, C.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L.; Morris, G.; Orozco, M.; Perthinhez, T.; Rizzi, M.; Sotriffer, C. *J. Med. Chem.* **2008**, *51*, 6237.
7. Wong, C. F. *Expert Opin. Drug Discov.* **2015**, *10*, 1189.
8. Novoa, E. M.; Ribas de Pouplana, L.; Xavier Barril, X.; Orozco, M. *J. Chem. Theory Comput.* **2010**, *6*, 2547.
9. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534.
10. Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. *J. Am. Chem. Soc.* **2011**, *133*, 9181.
11. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. *Commun. ACM* **2008**, *51*, 91.
12. Carrillo, O.; Laughton, C. A.; Orozco, M. *J. Chem. Theory Comput.* **2012**, *8*, 792.
13. Chaudhuri, R.; Carrillo, O.; Laughton, C. A.; Orozco, M. *J. Chem. Theory Comput.* **2012**, *8*, 2204.
14. Ding, F.; Yin, S.; Dokholyan, N. V. *J. Chem Inf Model.* **2010**, *50*, 1623.
15. Davis, I. W.; Baker, D. *J. Mol. Biol.* **2009**, *385*, 381.
16. Mashiach, E.; Nussinov, R.; Wolfson, H. J. *Proteins* **2010**, *78*, 1503.
17. Borrelli, K. W.; Vitalis, A.; Alcantara, R.; Guallar, V. *J. Chem. Theory Comput.* **2005**, *1*, 1304.
18. Borrelli, K. W.; Cossins, B.; Guallar, V. *J. Comput. Chem.* **2010**, *31*, 1224.
19. Hosseini, A.; Alibès, A.; Noguera-Julian, M.; Gil, V.; Paredes, R.; Soliva, R.; Orozco, M.; Guallar, V. *J. Chem. Inf. Model.* **2016**, *56*, 915.
20. Cabeza de Vaca, I.; Lucas, M. F.; Guallar, V. *J. Chem. Theory Comput.* **2015**, *11*, 5598.
21. Takahashi, R.; Gil, V. A.; Guallar, V. *J. Chem. Theory Comput.* **2014**, *10*, 282.

22. Edman, K.; Hosseini, A.; Bjursell, M. K.; Aagaard, A.; Wissler, L.; Gunnarsson, A.; Kaminski, T.; Köhler, C.; Bäckström, S.; Jensen, T. J.; Cavallin, A.; Karlsson, U.; Nilsson, E.; Lecina, D.; Takahashi, R.; Grebner, C.; Geschwindner, S.; Lepistö, M.; Hogner, A. C.; Guallar V. *Structure* **2015**, *23*, 2280.
23. Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383.
24. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B.* **2001**, *105*, 6474.
25. Shen, H. C.; Hammock, B. D. *J. Med. Chem.* **2012**, *55*, 1789.
26. Öster, L.; Sofia Tapani, Yafeng Xue, Käck H. *Drug Discovery Today* **2015**, *20*, 1104.
27. Xue, Y.; Olsson, T.; Johansson, C. A.; Öster, L.; Beisel, H.-G.; Rohman, M.; Karis, D.; Bäckström, S. *ChemMedChem* **2016**, *11*, 495.
28. *Protein Preparation Wizard*, Schrödinger, LLC, New York, NY, 2016.
29. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221.
30. Gelpí, J. L.; Kalko, S. G.; Barril, X.; Cirera, J.; de La Cruz, X.; Luque, F. J.; Orozco, M. *Proteins* **2001**, *45*, 428.
31. Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187.
32. *Glide*, version 7.0, Schrödinger, LLC, New York, NY, 2016.
33. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J. Med. Chem.* **2004**, *47*, 1750.
34. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M.; Perry, J. K.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739.
35. Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham III, T.E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. AMBER 14, University of California, San Francisco, 2014.
36. Jorgensen, W. L., Chandrasekhar, J.; Madura, J.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
37. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
38. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D.A. *J. Comput. Chem.* **2004**, *25*, 1157.
35. 39. Roe, D. R.; Cheatham III, T. E. *J. Chem. Theory Comput.* **2013**, *9*, 3084.40.
- Bahar, I.; Atilgan, A. R.; Erman, B. *Fold. Des.* **1997**, *2*, 173.
41. Morisseau, C.; Pakhomova, S.; Hwang, S. H.; Newcomer, M. E.; Hammock, B. D. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 3818.