Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics

Richard D. Smith[±], **Heather A. Carlson[±],[*]**
[±]Department of Medicinal Chemistry, University of Michigan, 428 Church Street, Ann Arbor, MI 48109-1056

Abstract

Protein dynamics play an important role in small molecule binding and can pose a significant challenge in the identification of potential binding sites. Cryptic binding sites have been defined as sites which require significant rearrangement of protein structure to become physically accessible to a ligand. MixMD is a computational protocol which maps the surface of the protein using molecular dynamics of the unbound protein solvated in a 5% box of probe molecules with explicit water. This method has successfully identified known active and allosteric sites which did not require reorganization. In this study, we apply the MixMD protocol to identify known cryptic sites of 12 proteins characterized by a wide range of conformational changes. Of these 12 proteins, three require reorganization of side chains, five require loop movements, and four require movement of more significant structures such as whole helices. In five cases, we find that standard MixMD simulations are able to map the cryptic binding sites with at least one probe type. In two cases (Guanylate Kinase and TIE-2), accelerated molecular dynamics, which increases sampling of torsional angles, was necessary to achieve mapping of portions of the cryptic binding site missed by standard MixMD. For more complex systems where movement of a helix or domain is necessary, MixMD was unable to map the binding site even with accelerated dynamics, possibly due to the limited timescale (100ns for individual simulations). In general, similar conformational dynamics are observed in water-only simulations and those with probe molecules. This could imply that the probes are not driving opening events but rather take advantage of mapping sites that spontaneously open as part of their inherent conformational behavior. Lastly, we show that docking to an ensemble of conformations from the standard MixMD simulations perform better than docking the apo crystal structure in 9 cases, and even better than half of the bound crystal structures. Poorer performance was seen in docking to ensembles of conformations from the accelerated MixMD simulations.
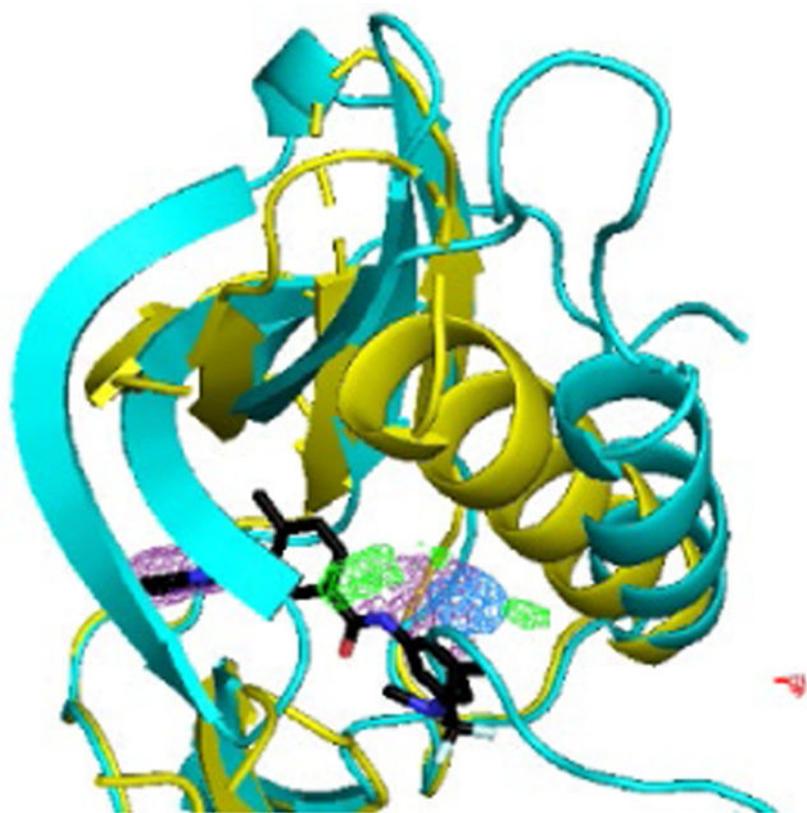
Graphical Abstarct

[*]Corresponding Author: carlsonh@umich.edu.

SUPPLEMENTAL INFORMATION

Data regarding the specific mapping of each probe in the standard and accelerated MD is given in Tables S1-S5. Binding site RMSDs, compared to the apo and bound structures is given in Figures S1-S24.

INTRODUCTION

Identifying and characterizing protein binding sites is a crucial first step to be able to predict and manipulate the binding characteristics of potential ligands.[1] Structural flexibility of protein targets has major consequences on molecular recognition, as the binding site(s) may sample many different conformational states.[2, 3] The ideal approach to identify binding sites in a prospective manner is to utilize the unbound structure of the protein. However, due to the flexible nature of proteins, using the unbound structure can present other difficulties.[4] Numerous static methods, both geometry-based and energy-based, have been unsuccessful in predicting sites with large movements.[5]

Cosolvent molecular dynamics (MD) techniques have been developed to map and predict favorable small-molecule "hotspots" for binding to proteins.[6-10] These techniques have the advantage of incorporating a fully flexible protein using MD and allowing for the competition of a small molecule with water at the protein surface. Several methods have been developed recently which have successfully utilized cosolvent MD techniques including: MixMD[11,12], SILCS[13], MDMix[6], as well as methods presented by Arcon *et al*.[7], Uehara and Tanaka[9], Kimura *et al*.[14], and Tan *et al*.[15]

The advantages and disadvantages of each method are highly dependent on probe types, solvent setup, and sampling. Mixed-solvent MD (MixMD) is our probe-mapping technique that has been extensively developed since its introduction in 2011, and it has demonstrated

utility in mapping both orthosteric and allosteric binding sites.[11, 16-19] The MixMD protocol accommodates full protein flexibility and preferentially locates hot spots along the surface of the protein.[11, 17] The method is based on using MD simulations of the protein in a mixture of water and an organic solvent (isopropanol, acetonitrile, pyrimidine, etc). Binding hot spots are identified by where the probe solvent has the highest occupancies on the protein surface. When the protein is allowed to be fully flexible during the simulation, the proper locations are found and random sites on the surface are reduced.[11]

The dynamic nature of proteins is a difficult challenge for traditional methods for structure-based drug design. In particular, working with cryptic sites presents a significantly demanding task. Cryptic sites are binding sites that are closed in the apo state and open through a structural rearrangement to bind ligands. The rearrangement can be as small as a side-chain rotation or as large as domain movement. Dynamic computational methods embrace protein flexibility[20] and are necessary to address cryptic sites. Computational techniques to identify cryptic or "transient" sites have used several methods to account for protein dynamics.[21, 22] TRAnsient Pockets in Proteins (TRAPP)[23] uses four techniques (tCONCORD[24], implicit solvent MD, L-RIP[25] and RIPlig[25]) for generating ensembles of protein structures to cluster and compare the RMSD to identify pockets on the protein that are transient, hence cryptic. The drawback of this method is that the center of the pocket must be known. The most recent method utilized Markov state models on MD simulations to identify "exposons" which successfully highlight the cryptic sites of two β-lactamases.[26] Gao *et al.* were able to generate bound conformations in long microsecond MD simulations using unbound starting structures for only 8 of the 39 systems they studied.[27] Oleinkovas *et al.* were unable to uncover cryptic sites for three systems using microsecond length MD simulations and therefore developed SWISH (Sampling Water Interfaces through Scaled Hamiltonians) to enhance sampling based on replica exchange MD.[28]

Cimermancic *et al.* identified a collection of proteins with cryptic sites for their binding-site prediction web server, Cryptosite.[4] A site was defined as being cryptic if it was not identifiable by FPocket[29] or ConCavity[30] using the unbound structure. This technique identified 93 distinct proteins for which a support vector machine model, based on the bound and unbound structures, was able to predict the cryptic binding sites in the unbound structures where as other programs had limited success due to the sites' cryptic nature.[4] The cryptic site definition used in the development of the Cryptosite dataset is based on the comparison of one bound and one unbound structure. This definition has received some criticism because not all unbound structures for these particular proteins present a cryptic binding site.[31] If the set of apo structures is expanded to include all unbound structures for the 93 Cryptosite proteins,[32] the solvent-mapping technique FTMap[33] detects hot spots within 5 Å of the proposed cryptic binding site in 88% of those proteins.[32] Additional inspection of the crystal structures showed that about half of the proteins had at least one unbound structure which could bind the ligand, without clashing with the protein.[32]

Overall, these cases serve as challenging proteins for cosolvent MD techniques, such as MixMD. Though cosolvent MD has been used on individual cases of cryptic pockets,[15] there are only two studies which provide a systematic examination across many proteins, those of Kimura *et al.*[14] and Schmidt *et al.*[34] Kimura *et al.* chose eight systems from the

Cryptosite study with varying degrees of pocket dynamics to investigate their ability to identify these cryptic binding sites with the aid of cosolvent MD simulations. They performed 100 ns MD runs on each system using Desmond and either 5% concentrations of isopropanol, resorcinol, or acetic acid as probes. They were able to map 6 of the 8 cryptic sites, with success defined as ranking the site in the top-three based on the free energy computed from probe occupancy.[14] Schmidt *et al.* utilized 10-500 ns cosolvent MD with 10% phenol on seven proteins to identify pocket cores which overlapped with the proposed cryptic sites in their chosen proteins.[34] They were successful in identifying these sites on six of the seven proteins.

Here, we aim to push MixMD to examine a range of conformational changes, particularly larger conformational changes. Accelerated MD[35] was necessary in cases where standard MD is unable to map the binding site, and it has shown utility in other cosolvent MD studies.[36] The main goal of this project is to successfully map the cryptic binding sites of 12 systems from the Cryptosite[4] dataset (using standard MixMD and/or accelerated MixMD). Unbound protein structures were used in our simulations to determine the ability of MixMD to induce conformational changes and allow mapping of cryptic sites. The accelerated MD method allows us to observe hidden sites that open upon large-scale motions of the protein backbones. We were able to visualize protein dynamics leading to the sampling of the cryptic site with small molecule probes in 7 of the 12 systems studied (Table 1).

METHODS

Choice of Protein Systems:

The 93 known cryptic sites from the Cryptosite[4] dataset were examined to identify proteins that would be amenable to MixMD. We required the functional unit of the protein to be monomeric, based on the biological unit described in the Protein Data Bank[37] for each structure. In most cases, we required the electron density map to be available and that the residues in the binding site be well-resolved and fit the experimental density. We focused on binary protein-ligand systems, so the unbound structure must not have cofactors/ions present in the binding site. This led to the selection of 12 systems: Ricin A-chain, Adipocyte Lipid Binding Protein (ALBP), Androgen Receptor (ligand binding domain), β-secretase (BACE-1), Guanylate Kinase (kinase domain), β-lactamase, TIE-2 (kinase domain), Protein Tyrosine Phosphatase 1B (PTP1B, allosteric site), Hepatocyte Growth Factor Receptor (c-Met, kinase domain), UDP-N-Acetylglucosamine Pyrophosphorylase (UAP1), Arylalkylamine N-acetyltransferase (AANAT), and Hsp90. The PDBIDs of the bound and unbound structures, as well as a brief description of the protein dynamics necessary to identify the cryptic binding site are listed in Table 1. An exception should be noted: PDBID 2am9 was chosen as the apo structure for the Androgen Receptor rather than 2ax9 (unbound structure from Cryptosite) because that structure was missing several important residues in the binding site. With our selection criteria, none of our systems overlap with those investigated in the study by Kimura *et al.*[14] However, the Schmidt *et al.*[34] paper appeared during the course of this project, and we happen to share two proteins in common with that study (Hsp90 and β-lactamase).

Setup of MixMD:

The unbound crystal structure of each protein was used in our simulations (See Table 1). The structures were stripped of all non-protein atoms. Hydrogen atoms were then added and histidine, asparagine, and glutamine tautomers were corrected by visual inspection using MOE[38] with guidance from MolProbity[39]. To neutralize the system, sufficient sodium or chloride ions were added using tleap of AmberTools 18[40]. Five different probes were used in our MixMD simulations including isopropanol (IPA), pyrimidine (PYR), acetonitrile (ACN), and a concurrent set of methylammonium (MAI) + acetate (ACT) ions, for a total of four solvation setups. Parameters for IPA, PYR, and ACN were based on those chosen by Lexa *et al.*[19] ACT and MAI parameters were developed by Ghanakota and Carlson.[41] The tleap module of AmberTools was used to generate the mixed-solvent boxes. First, a 7 Å shell of probe molecules was added around the protein. Second, TIP3P[42] water molecules were then added by the SolvateBox command, altering the box radius to obtain a ratio of ~5% of probe (or ~2.5% of each ACT and MAI as a mixture). Phenol was also used as a probe at 10% with RESP charges and the GAFF2 force field used as parameters for MD, analogous to recent mixed-solvent MD work by Schmidt *et al.*[34] However, across all 12 proteins, we were surprised to find that we could not map any binding sites with strong occupancies (35σ as defined below) using phenol; therefore, results are not presented in this work.

MD simulations.

Simulations were performed in Amber18[40] using FF14SB force field[43]. The SHAKE algorithm[44] was applied to restrain all bonds to hydrogen atoms, and a time step of 2 fs was used. Particle Mesh Ewald was used as implemented for the GPUs (pmemd.cuda)[45]. A 300K temperature was maintained using the Andersen Thermostat[46]. Solvent waters and probes were first minimized with 250 steps of steepest descent minimization, followed by 4750 steps of conjugate gradient keeping the protein restrained with a 500 kcal/mol-Å^2 weight. The whole system was then minimized with 250 steps of steepest descent minimization followed by 2250 steps of conjugate gradient. The temperature was gradually increased to 300 K in three 20 ps steps with a 10 kcal/mol-Å^2 restraint on the protein and an additional 270 ps of dynamics at 300 K using the same restraint on the protein. Restraints were removed from the side chains and gradually removed from the backbone over two 20 ps equilibration steps with backbone restraints of 5, and 1 kcal/mol-Å^2, and one 60 ps equilibration step with 0.1 kcal/mol-Å^2 backbone restraints. Finally, fully unrestrained dynamics were performed for 1.4 ns to allow for equilibration. The minimization, heating, and equilibration steps were performed using the sander module of Amber18. The pmemd.cuda module of Amber18 was then used to perform 100 ns of unrestrained, standard MD and accelerated MD for production. Ten independent simulations were carried out for each probe (PYR, ACN, IPA, phenol, and ACT+MAI) resulting in a production simulation time of 1 μs for each probe in both classical MD and accelerated MD. Both standard and accelerated MD were conducted for all probes with all 12 protein systems; however, the results of the accelerated MD are only discussed below where the standard MD was inadequate to find appropriate mapping. Data regarding the specific mapping of each probe in the standard and accelerated MD is given in the Supplemental Information, Tables S1-S5. Greater RMSD sampling is clearly seen in the accelerated MD. One 100 ns production run of each apo protein in water alone was simulated with both standard and accelerated MD

using the same equilibration procedure described above. This was to note the degree of sampling for the same time length without the presence of probe molecules.

Accelerated MixMD:

The accelerated MD method by McCammon and coworkers[35, 47] was chosen as it is fully implemented in Amber 18.[40] Accelerated MixMD is based on altering the potential energy landscape to accelerate the timescale in MD simulations and enhance torsion angle sampling. A bias potential was added to the true energy potential to enhance the escape rate from the potential wells and therefore accelerate and extend the time scale in MD simulations. The boost energy lies above the minimum of the potential energy surface. The calculation method from literature[35] was used, and the potential energy threshold boosts were calculated using:

$$\alpha P = N_{atoms} \times 0.2$$
$$E_{threshold}P = \text{Average Potential Energy} + \alpha P$$
$$\alpha D = 0.2 \times 3.5\,(N_{residues} + N_{probes})$$
$$E_{threshold}D = 3.5\,(N_{residues} + N_{probes}) + \text{Dihedral average energy}$$

Where α is the tuning factor that determines the depth of the modified energy potential well. P is the potential energy, D is potential for dihedral torsions, and E is the threshold energy (boost energy value). The boost factor measures the extent to which the simulation has been accelerated. Potential energy values and dihedral energy values were averaged over the 1.7 ns of unrestrained equilibration dynamics for each run.

Processing MixMD simulations.

The location of the probe atoms from the last 10 ns of the 10 runs were binned onto a 200 x 200 x 200 grid with 0.5 Å spacing, using the grid command from the cpptraj module from AmberTools[48]. The location of the probes appear as a mesh of occupancy density according to the formula $(x-\mu)/\sigma$ where x is the occupancy at the individual grid point, μ and σ are the mean and standard deviation of all grid points, respectively. The apo crystal structure was used to align and examine the maps of the corresponding probes and identify locations of maximal occupancy, and maps were contoured at a value of 35σ. The maps were visualized using the isomesh command in PyMOL[49]. Higher σ values for a given location indicates higher occupancy of the probe molecule at that location (phenol simulations showed no mapping at 35σ occupancies, though weak mapping was occasionally seen at lower occupancies). The location, maximum σ value, and rank of each potential probe's binding site was determined using the MixMD Probe View[50] plugin for PyMOL[49]. Binding-site RMSDs were calculated using only heavy atoms after backbone alignment to the unbound structure using cpptraj[48].

To determine if the maps had reached convergence, the location and σ-values of the probe maxima across 10-ns intervals from the trajectory were calculated (50-60 ns, 60-70 ns, 70-80 ns, 80-90 ns, and 90-100 ns). The average and standard deviation of the occupancies (in σ) for each mapped point, across the 5 different intervals, are reported in Supplemental Information, Tables S1-S5. A grid point was determined to be in the binding site if it was within 4 Å of the bound ligand and within 2 Å of the probe binding site determined by the

Probe View plugin. Analysis focused on the top-8 most highly occupied sites on the maps, but the mapped cryptic sites were primarily within the top-4 most occupied sites as we have observed in previous work.[16]

It is important to note that for each mapped hotspot there were probe maxima on the grids throughout the 50-100ns range (i.e., the hotspots were always there), but the occupancies of those maxima could shift between the 10-ns intervals. We found that the occupancies differ from system to system and probe to probe, with standard deviations ranging from 3.89 to over 24.76 σ (Tables S1-S5). The greatest fluctuation is seen in the charged maps. The simulations with large standard deviations indicate a local maximum was only occupied at greater than 35σ during some of the assessed intervals, including the 90-100 ns interval which labeled it a hotspot. It is not surprising some systems did not demonstrate the strong mapping until later in the simulation, given the complex conformational changes required for binding.

Docking ligands to the apo state, bound state, MixMD ensembles, and water-only MD ensembles.

All proteins were set up as described above. The ligand was extracted from the bound crystal structure and hydrogens were added using MOE[38], while inspecting for appropriate protonation and tautomerization.

MixMD ensembles were selected from the individual probe simulations. Ten representative structures for each of the 4 probe-mixture simulations were selected, giving a total of 40 structures used for each protein for either accelerated or standard MixMD. The snapshots from each of the 10 runs were clustered with cpptraj[48] using hierarchical agglomerative clustering based on the all-atom RMSD of the binding site defined by the apo residues within 4 Å of the ligand. Only snapshots from the last 10 ns of the 10 individual probe mixMD runs were used for clustering as those were the snapshots used to determine the mapping. The clustering parameter epsilon was set to 3 Å and 10 clusters were required. The centroid of each cluster was chosen as a representative. The same procedure was used to generate 10 clusters from the one water-only MD simulation; the only exception being that the entire trajectory was used.

We aimed to do the docking in a "prospective" manner to show how the MixMD data might be used for drug discovery, so the docking center of the binding sites were determined as the average of the clustered grid points determined with MixMD Probe View[50], combining maps from the four probe mixture simulations. The sites generated from Probe View were determined to be in the binding site if they were also on listed as top sites in Tables S1-S5. In some instances, clusters were combined if more than one of the clusters from MixMD Probe View mapped to the binding site. The conformation ensemble and docking center were determined separately for the standard and accelerated MixMD simulations.

GOLD v5.8[51] was used to perform docking to either the single apo structure, single bound structure, or ensemble of MixMD structures. The chemscore[52] scoring function was used, and a radius of 12 Å around the calculated site centers was used. Do_cavity was set to 0 and otherwise default parameters were used. The top pose was saved and the RMSD to the

bound conformation of the ligand was calculated using the rms_cur function in PyMOL.[49] We specifically did not look to identify the poses with the best RMSD because in a prospective application, only the top poses would be used.

RESULTS & DISCUSSION

In this study, we performed MixMD simulations on twelve proteins which contain cryptic binding sites. This set does not overlap with that of Kimura *et al.*[14] and was chosen to provide a challenging set to test the limits of MixMD. The degree of protein movement necessary to reveal the binding site varied from single side-chain movements in the case of Ricin and Adipocyte Lipid Binding Protein to larger domain movements in c-Met and PTP1B (see Figure 1). In general, we found good performance of standard MixMD to map cryptic binding sites, unless a helix was required to move (Table 1). In two cases, guanylate kinase and TIE-2, accelerated MixMD was needed to map the cryptic site. Specific details for mapping in each system are given farther below.

The goal of the MixMD protocol is to map binding hotspots using MD. Ideally, the simulations should sample the open conformation, uncovering the cryptic sites, during the analyzed time frame. When looking at the RMSD of the binding site residues for each system (Supplemental Information, Figures S1-S24), we observe some conformational sampling away from the unbound state that are more open, but the conformations do not necessarily fully replicate the bound structure. This is still enough dynamics in some probe-protein simulations to allow the smaller probes to sample the cryptic portions of binding sites. Binding site RMSDs approach 2 Å, compared to the bound structure in some runs for some probe-protein simulations. Comparisons to simulation of the protein in water alone (black line in Figures S1-S24) indicate the dynamics are the same magnitude whether the probes are present or not. This is slightly different than the findings of Schmidt *et al.*[34] Their protocol identified cryptic sites based on the differences between protein conformations from MD simulations in water alone compared to simulations using the mixed phenol-water solvent. It could be implied that the phenol probe may have induced conformational changes in their systems, though it is possible that the probes were opportunistic occupants of random reorganizations of the sites. Our comparison of MixMD to water-only MD would tend to point to our probes mapping inherent conformational states in our simulations.

To provide an example of how our data might be used in a "prospective" application, we used the MixMD and water-only MD simulations to provide binding-site conformations for docking calculations. We centered the docking site on the average of clustered points from sites determined with MixMD Probe View,[50] which overlapped with the protein in the apo state for half of our targets as one might expect for a cryptic site (Ricin, Androgen Receptor, BACE-1, TIE-2, AANAT, and PTP1B). Of course, half of the sites did not coincide with the protein, so some knowledge of the binding site was needed for some systems. This same weakness is seen in the study of Kimura *et al.*[14] but not in the Schmidt *et al.*[34] paper. However, it should be noted that all cosolvent methods like these suffer from the identification of spurious, extra sites on the surface of the proteins, which can complicate identifying the cryptic sites *a priori*.

We docked the known ligand (from the bound structure, see Table 1) into MD ensembles and compared the top pose to docking the ligand to the apo and bound states of each target. Table 2 shows that docking to the ensemble from the standard MixMD simulations out-performed the apo structure in 9 of 12 targets, but the accelerated MixMD out-performed the apo structure in 5 targets. Furthermore, the MixMD ensemble from standard MD performed better than the bound structure in 6 targets. It is encouraging that the sites obtained from MixMD for some targets would be on par with self-docking of the bound ligand to the bound crystal structure. The larger ligands of BACE-1 and AANAT did not successfully reproduce the bound structure for any of the protein conformations. For AANAT, β-Lactamase, UAP1, c-MET, and PTP1B the ligand was docked into the non-cryptic portion of the binding site identified from MixMD. Of course, the success of the docking to the structures was dependent upon the target, which is a known issue with docking protocols in general.[54-56] It cannot be ruled out that part of the success of docking to the MixMD conformations could have been the use of ensemble docking over using single crystal structures. It should be noted that for the standard MD simulations, the water-only ensembles performed comparably with the MixMD ensembles, but the MixMD ensembles were better than the water-only ensembles derived from accelerated MD.

The discussion below highlights the conformational sampling observed for each target and details the ability of MixMD to map the revealed binding sites. Each structure's known conformational change upon ligand binding can be seen in Figure 1. A summary of the mapping is also included in Table 1 and Supplemental Information Tables S1-S5. All probe maps in the figures are presented with a threshold of 35σ to represent the probe occupancy, and results for standard MD are presented unless noted otherwise. Analysis focused on the top-8 most highly occupied sites on the maps.

Ricin (A-chain):

The Ricin A-chain is responsible for the enzymatic depurination of a specific adenosine residue in the 28S rRNA.[57] Ricin only requires the movement of Tyr80 to accommodate an aromatic ring from the nucleotide base. In the case of the bound ligand in 1br6, this is mimicked by the pterin ring of a bound inhibitor.[58] The RMSD between Tyr80 in the unbound (PDBID: 1rtc[57]) and the bound (PDBID: 1br6[58]) is 2.96 Å, whereas all other binding site residues within 4 Å of the ligand have RMSDs less than 0.6 Å. This makes it the most amenable to MixMD as the rotation of one amino acid should occur during the MD simulation and allow for mapping by the probe molecules.

In our simulations with ACN and the ACT+MAI probes, Tyr80 does move out of the pocket. The location of the tyrosine in the average structure from the ACT+MAI simulations most closely resembles the bound conformation, allowing the two probes to enter and map the binding site when contoured at 35σ as seen in Figure 2. It is important to note that the MAI site is the top site for that probe, and ACT site is ranked second for the ACT probe (in terms of occupancy at 35σ). The other sites that are mapped at 35σ are all on or near exposed sites on the surface of the protein. The top ACT site has a phosphate buffer molecule bound in at least one Ricin structure (5u4l). Accelerated MD yields the same top site for IPA and ACN, which is occupied by crystal additive EDO in two Ricin structures (5j57 and 4ljr) and a

crystallographic ACT in a third structure (5j56). These are the third and fourth sites for IPA and ACN respectively with standard MD. Therefore, we have successfully mapped the cryptic binding site of Ricin and were able to find other lower-ranked sites where small molecules could bind. The biological implications of the other binding sites are unknown.

Adipocyte Lipid Binding Protein:

Adipocyte Lipid Binding Protein (ALBP) functions to solubilize intracellular hydrophobic lipids.[59, 60] The binding site is a deep hydrophobic cavity where the polar head group of the fatty acid ligand is bound at the bottom of the cavity by two arginine residues (R106 and R126) and a tyrosine (Y128).[59] Phenylalanine 57 (F57) has also been previously identified as a potential key residue for binding the fatty acid ligand due to its conformational change upon ligand binding[59]. The RMSD between the unbound (1alb)[60] and bound (1lic)[59] forms for F57 is 5.27 Å. However, the structure as a whole does not demonstrate significant dynamics in standard MD; the RMSD of all atoms between the bound and unbound structures is 1.16 Å. Large dynamic movements of the whole protein are not seen in our simulations, yet the probes were able to sample the binding site.

With standard MD, although we do not get significant movement of F57, we observe mapping of the hydrophobic core with both ACN and IPA (Figure 3A). These probes appear to be small enough to get into this deep hydrophobic pocket without the aid of accelerated MD. We would expect the ACT probe to map the acid moiety of the bound; however, this was not observed with standard MD. Accelerated MD was necessary to allow ACT to sample the bottom of this deep site. The ACT probe, with accelerated MD, bound near the two arginine residues which interact with the acid moiety of the bound ligand (Figure 3B). The simulation in this study has started with the unbound structure of the protein where F57 occupies the location of the hydrophobic tail of the fatty acid ligand (Figure 1). The bound structure has the phenyl ring of F57 exposed and oriented away from the binding site with the ligand in contact with Cβ of the residue indicating a large movement to allow for ligand binding.

Accelerated MD also demonstrates surprisingly different mapping with the uncharged probes. Here, the smaller uncharged probes, IPA and ACN, do not map the binding site as they did with standard MD, but PYR is able to map the binding site. Although the accelerated MD simulation with PYR did not map the tail where F57 was displaced, the loop appears to open some allowing probe molecules to map the interior of the binding site. In our simulations, F57 does not open to mimic the unbound structure; however, there is significant rotation of the phenyl ring. Therefore, we successfully mapped the cryptic site, although the complete opening of the phenylalanine residue was not observed.

Androgen Receptor:

The androgen receptor is a hormone receptor that controls gene expression and cell differentiation in humans.[61, 62] The cryptic site of interest is an allosteric site, not the primary orthosteric binding site of the protein.[61] Previous investigations into this receptor using the MixMD methodology with standard MD have demonstrated clear mapping of the allosteric site.[16] The site itself is not as difficult to map as most of the other systems because

it does not require backbone rearrangement to map the site, but it still can clearly be classified as cryptic due to the need for side-chain movement to make the site accessible. K720 and M734 move to accommodate the ligand with RMSDs between the unbound (2am9)[62] and bound (2piq)[61] conformations of 2.08 Å and 2.18 Å, respectively. These values are larger than the RMSD to the starting apo structure of the seven residues within 4 Å of the ligand (RMSD = 1.38 Å). The average RMSD over the last 10 ns of all standard MixMD simulations combined for K720 and M734 were 2.94 and 2.50 Å respectively, while the binding site RMSD was 1.94 Å. The average structure from the simulations shows K720 in the bound conformation, and M734 is closer to the bound conformation in three of the four probe simulations (Figure 4). Therefore, we see appropriate dynamics during the standard MixMD simulation for these two residues, allowing access to the binding site.

As expected, this site was mapped at 35σ after 100 ns by two of the five probes (PYR and IPA). Our mapping with standard MD reproduces the mapping seen by Ghanakota *et al*. Accelerated MixMD simulations were also able to map the active site and the other allosteric site, however with slightly decreased occupancies of the sites at 35σ. This is potentially due to the increase in the total potential energy applied during accelerated MD, which could cause the solvent molecules to spend less time and a grid point giving a lower occupancy value over the last 10 ns of simulation time analyzed for probe occupancy.

β-secretase (BACE-1):

BACE-1 is an aspartyl protease which functions to cleave the amyloid precursor protein (APP) into the amyloid β peptide, which misfolds and oligomerizes to form the main component of the plaques of patients with Alzheimer's disease.[63, 64] The peptide binding site is characterized by a large cleft which is covered by a flap region (residues 67-77) that is open in the unbound structure (1w50), but it closes and makes contacts with the peptide in the bound conformation.[63, 64] The RMSD between the unbound structure (1w50)[64] and the bound structure (3ixj)[63] is 1.46 Å and the RMSD of the flap residues is 4.37 Å. This binding site of BACE-1 is a shallow peptide-binding groove. In our simulations, the flap starts in the unbound, open state and closes in many of the simulations as can be seen in the average structures. The average RMSD of the flap residues for all snapshots in the last 10 ns used for the occupancy calculation was 3.80 Å and while the entire binding site RMSD was 2.63 Å. The open portion of the binding site is mapped by PYR and ACT. MAI maps the catalytic aspartate residues with the flap region closing in these simulations. It is not surprising that we are able to map this binding site, as it is an accessible binding site. However, it is interesting to note that in all simulations the flap closes on the binding site and the probes are able to map the formed binding site. Figure 5 displays the mapping of the binding site.

There has been some debate regarding the cryptic nature of this binding site. Vajda *et al*. noted that, upon inspection of the 52 unbound structures of BACE-1, there were two equal portions of those unbound structures. One portion represented a similar conformation to the bound-state, and the other portion was uniquely unbound in character.[31] Using FTMap[33], Beglov *et al*. found that the top hotspot for 1w50 was not in the cryptic binding site; however, there were several unbound structures identified where the site was mappable.[32] The mapping we observe is encouraging in that, even though the site may not be as cryptic

as initially thought, we observe that MixMD can provide dynamics that sample both the bound and unbound state and provide the ability to map a dynamic pocket.

Guanylate Kinase:

Guanylate kinase catalyzes the transfer of a phosphate from ATP to GMP forming GDP, assisting with the balance of GTP and GDP in the cell.[65, 66] This structure displays a large domain movement between the unbound (1ex6)[65] and bound (1gky)[66] structure. The GMP binding domain (residues 31-82) assembles upon GMP binding. The all-atom RMSD of the GMP binding domain between the unbound and bound structures is 8.21 Å. Although the domain demonstrates significant dynamics, especially with accelerated MD, it does not completely adopt the bound conformation during the simulation. The average RMSDs (standard MD/accelerated MD) of the GMP binding domain between the starting apo structure and the simulation snapshots used to calculate the occupancy are 2.61 Å/7.03 Å, 2.80 Å/7.77 Å , 2.50 Å/8.70 Å and 2.45 Å/7.75 Å for PYR, ACN, ACT+MAI, and IPA simulations, respectively.

No mapping of the cryptic portion of the binding site near the guanine ring is seen using standard MixMD; however, accelerated MixMD was able to provide enough conformational sampling. The accelerated MixMD simulations could map the cryptic GMP binding site as well as the ATP and Mg^{2+} binding sites with ACT, MAI, and PYR (Figure 6 A & B). The location of the phosphate is clearly mapped with the ACT probe as expected. This site is the second ranked site, which is initially mapped with high sigma values of 60 and 75σ. The top site corresponds to the ATP binding site, which is more exposed and requires less dynamics, whereas the cryptic site is the Guanylate binding site. Analysis of the PYR simulations displays mapping of the guanine of nucleotide at around 75σ. There is significant occupancy of MAI near the magnesium(II) binding site located near Asp98. Magnesium(II) was not present in either the unbound structure or with bound substrate (5GP), allowing the positively charged MAI ion to bind in this location. ACT was also able to map the ATP binding site (residues 9-16).

TIE-2 (Kinase Domain):

TIE-2 is a receptor tyrosine kinase involved in regulation of angiogenesis.[67, 68] The tyrosine kinase domain is crystallized and catalyzes the autophosphorylation of SH2 and SH3 domains utilizing ATP.[67, 68] This binding site is cryptic as there is a large domain movement which adapts to the peptide and nucleotide ligands (Figure 1). The all-atom RMSD of the residues within 4 Å of the bound ligand in pdb 2oo8[67] is 3.47 Å compared to the unbound structure (1fvr[68]). This domain movement includes an α-helix which closes on the bound ligand, the nucleotide binding loop (residues 833-836), and the activation loop (residues 980-983) moves out of the binding site, to allow for binding. In all probe simulations, on average, both the nucleotide binding loop and the activation loop move out of the binding site. The average RMSD (standard MD/accelerated MD) of the binding site to the initial starting structure for all snapshots analyzed from the MD run are 2.48/3.40 Å, 2.70/3.98 Å, 2.81/3.97 Å, and 2.46/3.71 Å for the PYR, ACN, ACT+MAI, and IPA simulations, respectively. The helix remains in the unbound conformation. Mapping near the nucleotide binding loop is only seen with standard MD (Figure 7A), while both sites are mapped using

accelerated dynamics (Figure 7B). The following paragraph discusses the accelerated MD results for TIE-2.

The dynamic movement of the two loops have allowed for mapping of the binding site in two of our probe simulations. The MAI probe maps near E972 of the helix, which does not close in our simulations. This MAI site overlaps perfectly with a morpholino group of the bound ligand in 2p4i. PYR could map precisely to the PYR moiety of the ligand which is near the nucleotide binding loop. This loop is disordered and unresolved in the bound crystal structure. The MAI and PYR sites are the top occupied sites for their respective simulations with occupancy > 75σ. At 35σ, there are two other PYR sites on the protein. The top site for IPA is also near the binding site where the activation loop has moved, and maps where D982 was located in the unbound structure. TIE-2 is a receptor tyrosine kinase which dimerizes upon binding of its substrate ligand, although these sites do not appear to be involved in dimerization. The asymmetric unit of the apo structure is a dimer and no mapping occurs for any probe at the dimer interface at 35σ.

Hsp90:

Hsp90 is a molecular chaperone involved in the folding and stabilization of proteins in the cell, and it has been the target of the development of cancer therapies.[69, 70] The ATP binding site is partially occluded in the unbound form by a disorganized portion of a helix (residues 106-109 of 2qfo), which is fully structured in the bound state (2wi7).[69] With standard MD we observe extensive mapping of the binding site with the uncharged probes, despite little rearrangement of the helix (Figure 8). The RMSD between the bound and unbound structures is 6.72 Å, whereas in all of our analyzed snapshots the average RMSD to the unbound structure is only 1.3 Å. It is important to note that no mapping of the occluded portion of the site is observed using accelerated MD, though we are unsure why. This indicates that even with slight movements of the protein, the probes are small enough to be able to sample the active site.

β-lactamase TEM-1 (M182T mutant):

Beta-lactamase TEM-1 is a protein which confers resistance to β-lactam containing antibiotics such as penicillin in bacteria.[71, 72] The bound structure (1pzo) contains two inhibitor molecules in the active site, with the CBT 301 site being cryptic.[71] TEM-1 β-lactamase has been a textbook case for cryptic sites and was used in one of the first attempts at identifying cryptic sites.[73, 74] These studies were successful at uncovering hidden allosteric sites by using Markov State Models of 81 microseconds of MD to observe clusters of structures for which residues outside of the binding site displayed correlated motions with binding site residues. They also identified the pocket with LIGSITE as the largest pocket available from these models. Those studies did not map the binding site but confirmed the large pocket and residues that were involved in the allosteric site. [73, 74] A second site was also confirmed in this study to interact with Alanine 232 and Alanine 249; however, no small molecules have been targeted to this site, and mutation to cysteine affects enzyme activity.[73] In our simulations, this site is mapped nearby with ACN and IPA with MixMD (Figure 9).

The challenge of mapping the cryptic (allosteric) site of this protein is due to occlusion by an alpha helix (residues 217-225, unbound (1jwp[72])-bound RMSD 4.71 Å). This helix must first move from its initial position during simulation to allow solvent probes to map the site. A second alpha helix must also move slightly to accommodate the ligand (residues 272-280, apo-bound RMSD 2.0 Å). This cryptic site is more of a challenge than the previously described sites since it requires manipulation of the entire peptide backbone to reveal the binding site as opposed to only a few side chains. Consistent with this assessment of difficulty, by 100 ns, simulations failed to map the cryptic sites at the location where the ligand binds at 35σ. This is most likely because the helix blocking the binding site had not moved dramatically from its original position to facilitate probe binding to the binding site (average RMSD to starting apo structure of residues 217-225: 1.77 Å, residues 272-280: 1.94 Å). There was some mapping with ACT near the edge of the binding site.

It seems that there is also no mapping of the cryptic site near the alpha helices with accelerated MixMD as well, but if the occupancy threshold is lowered to 20σ, sparse mapping of PYR begins to appear in the site obstructed by the helix. ACT also maps in the area occupied by ligands above the helix site at the lower 20σ occupancy. Investigation of the individual PYR simulations revealed that half of the runs successfully mapped the binding site with at least one or two rings, and one of the runs showed complete opening of the cryptic site and thorough mapping at the position of the bound ligand. Given that complete mapping occurred once and that mapping improved with more simulation time, it is plausible that we could achieve complete mapping of the cryptic site in every replicate given an even longer simulation.

c-Met:

c-Met is another receptor tyrosine kinase which has a large domain movement upon ligand binding.[75, 76] c-Met has a short alpha-helix (Residues 1222-1230) which is required to move out of the binding site for a ligand to bind. The helix is not structured in the ligand bound structure (3f82)[76]. The RMSDs between the apo (1r1w[75]) and holo (3f82[76]) binding site residues within 4 Å of the bound ligand in 3f82 is 3.09 Å. Neither standard nor accelerated mixMD were able to map the cryptic binding site. The accelerated mixMD simulations, which show the most dynamics, do not show significant conformational sampling with average binding-site RMSDs between the snapshots and the initial starting apo structure of 1.85 Å, 2.00 Å, 1.99 Å, and 2.01 Å for the PYR, ACN, ACT+MAI, and IPA runs, respectively. However, we do map several sites in the protein with ACN, ACT, and MAI. Most of these sites do appear to be pockets in the protein, but no crystal structures exist with a small molecule present in them. The pyridine ring of the inhibitor in 3f82 is mapped by PYR. This portion of the ligand is in part of the binding site where the helix dynamics is not necessary and the site is more exposed.

PTP1B:

Protein Tyrosine Phosphatase 1B (PTP1B) is a negative regulator of the insulin receptor through phosphorylation and has shown to be a promising target in the potential treatment of type II diabetes.[77] The allosteric site of PTP1B is a known cryptic site that requires some conformational changes to be exposed.[77, 78] Helix α7 at the allosteric site is involved in

interactions that keeps it in a conformation making it difficult for mapping of the allosteric site. Binding of inhibitors to this site cause disruption of these interactions and displacement of the helix from its position. Crystal structures that have a ligand at the allosteric site do not have the helix α7 resolved. The crystal structure used in our studies was void of any ligands at this site and included a resolved helix structure. This made it a challenge in our studies to examine whether the MixMD method will be able to induce conformational changes and cause the helix to move. The MixMD simulations did not show any mapping for the cryptic site caused by α7 helix near the allosteric site in this system.

AANAT:

AANAT catalyzes the Coenzyme A-dependent acetylation of serotonin as the first step in the synthesis of melatonin.[79] The bound structure contains a bisubstrate inhibitor which spans both the substrate and coenzyme binding sites.[80] In the unbound structure, a large portion of the helix of motif D is unstructured and occluding the coenzyme binding site. With standard MD this large loop does not move and the cryptic portion of the binding site is not mapped, but the non-cryptic portion of binding site is successfully mapped. No mapping of the site is seen in the cryptic portion of the binding site with accelerated MD likely because the structural rearrangement is simply not possible in the 100-ns simulation time window.

UAP1:

UDP-N-acetylglucosamine pyrophosphorylase (UAP1) from *Candida albicans* is involved in cell wall synthesis for this yeast strain. *C. albicans* is a fungal pathogen which generates infection in patients with weakened immune systems. It is known that this pathogen exhibits a high level of drug resistance, especially in drugs targeting cell wall synthesis.[81] Crystal structures have been reported for both the apo protein as well as the holo protein containing its natural substrate UDP-N-acetylglucosamine (UDP-GlcNAc) in its active site.[82] The bound structure has three loops which are closed on the reaction product, which is also a substrate for the reverse reaction, producing a induced-fit binding mechanism.[82] Even though the binding site is open in the unbound state, it remains open during the simulations, and we get slight mapping with ACN near the uridine binding site with standard MixMD.

CONCLUSION

We have employed the MixMD protocol, using both conventional as well as with accelerated MD simulations, for detecting cryptic binding sites of 12 proteins, with varying levels of success. MixMD was successful at mapping systems with a wide range of movement. We were able to map the binding site and observe appropriate conformational sampling for Ricin, Adipocyte Lipid Binding Protein (ALBP), Androgen Receptor (AR), BACE-1, Hsp90, Guanylate Kinase and TIE-2. The first three (Ricin, ALBP, and AR) were thought to be weak binding pockets by Beglov *et al.* as they only require single side-chain motions; however, very few bound ligands towards these proteins were investigated in that study. Our dynamics were also able allow for loop movements in the case of BACE-1 where the flaps are able to close on the binding site. Guanylate Kinase displayed movement of the GMP binding domain to allow mapping of the binding site. Additionally, for Guanylate Kinase, we were able to map the more exposed ATP with our charged probes (ACT and MAI). In

TIE-2, *two loops* were required to move, both the nucleotide binding loop and the activation loop. In our simulations, both loops vacated the binding site only when accelerated dynamics was performed.

The success of MixMD to map cryptic sites is comparable to the previous methods in that we were successful in mapping 7 different cryptic binding sites. The most comparable large scale studies from Gao *et al.*[27], Schmidt *et al.*[34] and from Kimura *et al.*[14], were able to identify 6, 7, and 8 binding sites, respectively. Direct comparisons between those methods and ours is a bit difficult because there is little overlap between our protein systems and those used in the studies above (only 2 of 39 systems from Gao *et al.*, 2 of 7 systems in Schmidt *et al.*, and none with Kimura *et al.*). We are confident that the reason the MixMD could not map the pockets that required larger motions was not due to lack of convergence because the backbone RMSDs were stable. It is possible longer simulations may allow for these motions, but given the lack of success using accelerated MixMD on these systems, it is unlikely.

In some cases, accelerated MixMD was required to map the cryptic binding site (Guanylate Kinase and TIE-2). We hypothesized that the decreased barrier of torsional rotation in accelerated MD would allow for more torsional sampling and better reveal the cryptic sites. This was true particularly in the use of the charged probes (Tables S4 and S5). In ALBP and Guanylate Kinase, accelerated dynamics was necessary to map the charged interactions in the binding site. Interestingly in some instances, mapping was not observed with accelerated dynamics yet could be seen with standard dynamics (acetate in ALBP and mapping of Hsp90). It would be difficult to speculate on the reason for this as many of these systems have complex dynamics, and the increased torsional energy may not allow for the appropriate sampling of the solvent in the binding site.

The cases where MixMD failed, all required whole helices to move to open the pockets. In the cases of PTP1B and c-Met, these were C-terminal helices. The c-MET termini were capped with an acetyl group as this is a domain from a larger protein. The terminus of the structure for PTP1B (2f6v) was not capped, as the terminal carboxylate was present in the structure, suggesting this was the end of the construct. It is possible that the dynamics of these helices may require the presence of the other protein domains. TEM-1 β-lactamase requires the movement of two helices; however, neither motion was observed in the simulations. It is possible that the time-scale for the MD was not long enough to observe helix movement or a different energy threshold may be necessary for accelerated dynamics to soften the torsional energy well more frequently and allow for increased large-scale motions. It is also possible that other probes could also be explored for these systems, such as ethanol, imidazole, or N-methyl acetamide. There was no consistent probe which mapped well across all of the systems, suggesting a variety of probes should be used with MixMD, unless first-hand knowledge of the chemical requirements of the desired site are known.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

REFERENCES

1. Ferreira L; dos Santos R; Oliva G; Andricopulo A, Molecular Docking and Structure-Based Drug Design Strategies. Molecules 2015, 20, 13384. [PubMed: 26205061]

2. Barril X; Morley SD, Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. J Med Chem 2005, 48, 4432–4443. [PubMed: 15974595]

3. Hall DR; Enyedy IJ, Computational solvent mapping in structure-based drug design. Future Med Chem 2015, 7, 337–353. [PubMed: 25826363]

4. Cimermancic P; Weinkam P; Rettenmaier TJ; Bichmann L; Keedy DA; Woldeyes RA; Schneidman-Duhovny D; Demerdash ON; Mitchell JC; Wells JA; Fraser JS; Sali A, CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. J Mol Biol 2016, 428, 709–719. [PubMed: 26854760]

5. Ghersi D; Sanchez R, Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures. J Struct Funct Genomics 2011, 12, 109–117. [PubMed: 21537951]

6. Alvarez-Garcia D; Barril X, Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. J Med Chem 2014, 57, 8530–8539. [PubMed: 25275946]

7. Arcon JP; Defelipe LA; Modenutti CP; Lopez ED; Alvarez-Garcia D; Barril X; Turjanski AG; Marti MA, Molecular Dynamics in Mixed Solvents Reveals Protein-Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. J Chem Inf Model 2017, 57, 846–863. [PubMed: 28318252]

8. Ghanakota P; Carlson HA, Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. J Med Chem 2016, 59, 10383–10399. [PubMed: 27486927]

9. Uehara S; Tanaka S, Cosolvent-Based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Druggable Protein Conformations. J Chem Inf Model 2017, 57, 742–756. [PubMed: 28388074]

10. Yu W; Lakkaraju SK; Raman EP; MacKerell AD Jr., Site-Identification by Ligand Competitive Saturation (SILCS) assisted pharmacophore modeling. J Comput Aided Mol Des 2014, 28, 491–507. [PubMed: 24610239]

11. Lexa KW; Carlson HA, Full protein flexibility is essential for proper hot-spot mapping. J Am Chem Soc 2011, 133, 200–202. [PubMed: 21158470]

12. Ghanakota P; van Vlijmen H; Sherman W; Beuming T, Large-Scale Validation of Mixed-Solvent Simulations to Assess Hotspots at Protein-Protein Interaction Interfaces. J Chem Inf Model 2018, 58, 784–793. [PubMed: 29617116]

13. Raman EP; Yu W; Guvench O; Mackerell AD, Reproducing crystal binding modes of ligand functional groups using Site-Identification by Ligand Competitive Saturation (SILCS) simulations. J Chem Inf Model 2011, 51, 877–896. [PubMed: 21456594]

14. Kimura SR; Hu HP; Ruvinsky AM; Sherman W; Favia AD, Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. J Chem Inf Model 2017, 57, 1388–1401. [PubMed: 28537745]

15. Tan YS; led P; Lang S; Stubbs CJ; Spring DR; Abell C; Best RB, Using Ligand-Mapping Simulations to Design a Ligand Selectively Targeting a Cryptic Surface Pocket of Polo-Like Kinase 1. Angew. Chem 2012, 51, 10078–10081. [PubMed: 22961729]

16. Ghanakota P; Carlson HA, Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. J Phys Chem B 2016, 120, 8685–8695. [PubMed: 27258368]

17. Lexa KW; Carlson HA, Protein flexibility in docking and surface mapping. Q Rev Biophys 2012, 45, 301–343. [PubMed: 22569329]

18. Lexa KW; Carlson HA, Improving protocols for protein mapping through proper comparison to crystallography data. J Chem Inf Model 2013, 53, 391–402. [PubMed: 23327200]

19. Lexa KW; Goh GB; Carlson HA, Parameter choice matters: validating probe parameters for use in mixed-solvent simulations. J Chem Inf Model 2014, 54, 2190–2199. [PubMed: 25058662]

20. Alvarez-Garcia D; Barril X, Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design. J Chem Theory Comput 2014, 10, 2608–2614. [PubMed: 26580781]

21. Amaro RE, Will the Real Cryptic Pocket Please Stand Out? Biophys J 2019, 116, 753–754. [PubMed: 30739726]

22. Kuzmanic A; Bowman GR; Juarez-Jimenez J; Michel J; Gervasio FL, Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. Acc Chem Res 2020, 53, 654–661. [PubMed: 32134250]

23. Stank A; Kokh DB; Horn M; Sizikova E; Neil R; Panecka J; Richter S; Wade RC, TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets. Nucleic Acids Res 2017, 45, W325–W330. [PubMed: 28431137]

24. Seeliger D; De Groot BL, tCONCOORD-GUI: visually supported conformational sampling of bioactive molecules. J Comput Chem 2009, 30, 1160–1166. [PubMed: 18942729]

25. Kokh DB; Czodrowski P; Rippmann F; Wade RC, Perturbation Approaches for Exploring Protein Binding Site Flexibility to Predict Transient Binding Pockets. J Chem Theory Comput 2016, 12, 4100–4113. [PubMed: 27399277]

26. Porter JR; Moeder KE; Sibbald CA; Zimmerman MI; Hart KM; Greenberg MJ; Bowman GR, Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. Biophys J 2019, 116, 818–830. [PubMed: 30744991]

27. Gao C; Desaphy J; Vieth M, Are induced fit protein conformational changes caused by ligand-binding predictable? A molecular dynamics investigation. J Comput Chem 2017, 38, 1229–1237. [PubMed: 28419481]

28. Oleinikovas V; Saladino G; Cossins BP; Gervasio FL, Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. J Am Chem Soc 2016, 138, 14257–14263. [PubMed: 27726386]

29. Le Guilloux V; Schmidtke P; Tuffery P, Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 2009, 10, 168. [PubMed: 19486540]

30. Capra JA; Laskowski RA; Thornton JM; Singh M; Funkhouser TA, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 2009, 5, e1000585. [PubMed: 19997483]

31. Vajda S; Beglov D; Wakefield AE; Egbert M; Whitty A, Cryptic binding sites on proteins: definition, detection, and druggability. Curr Opin Chem Biol 2018, 44, 1–8. [PubMed: 29800865]

32. Beglov D; Hall DR; Wakefield AE; Luo L; Allen KN; Kozakov D; Whitty A; Vajda S, Exploring the structural origins of cryptic sites on proteins. Proc Natl Acad Sci U S A 2018, 115, E3416–E3425. [PubMed: 29581267]

33. Ngan CH; Bohnuud T; Mottarella SE; Beglov D; Villar EA; Hall DR; Kozakov D; Vajda S, FTMAP: extended protein mapping with user-selected probe molecules. Nucleic Acids Res 2012, 40, W271–W275. [PubMed: 22589414]

34. Schmidt D; Boehm M; McClendon CL; Torella R; Gohlke H, Cosolvent-Enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-Based Drug Design. J Chem Theory Comput 2019, 15, 3331–3343. [PubMed: 30998331]

35. Hamelberg D; Mongan J; McCammon JA, Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 2004, 120, 11919–11929. [PubMed: 15268227]

36. Kalenkiewicz A; Grant BJ; Yang C-Y, Enrichment of Druggable Conformations from Apo Protein Structures Using Cosolvent-Accelerated Molecular Dynamics. Biology 2015, 4, 344–366. [PubMed: 25906084]

37. Berman HM; Westbrook J; Feng Z; Gilliland G; Bhat TN; Weissig H; Shindyalov IN; Bourne PE, The Protein Data Bank. Nucleic Acids Research 2000, 28, 235–242. [PubMed: 10592235]

38. MOE2016.08, 2016.8; Chemical Computing Group Inc.: 1010 Sherbrooke St. West, Suite #910, Montreal , QC, Canada, H3A 2R7.

39. Chen VB; Arendall WB; Headd JJ; Keedy DA; Immormino RM; Kapral GJ; Murray LW; Richardson JS; Richardson DC, *MolProbity* : all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D Biological Crystallography 2010, 66, 12–21. [PubMed: 20057044]

40. Case DA; Ben-Shalom IY; Brozell SR; Cerutti DS; Cheatham TEI; Cruzeiro VWD; Darden TA; Duke RE; Ghoreishi D; Gilson MK; Gohlke H; Goetz AW; Greene D; Harris R; Homeyer N; Izadi S; Kovalenko A; Kurtzman T; Lee TS; LeGrand S; Li P; Lin C; Liu J; Luchko T; Luo R; Mermelstein DJ; Merz KM; Miao Y; Monard G; Nguyen C; Nguyen H; Omelyan I; Onufriev A; Pan F; Qi R; Roe DR; Roitberg A; Sagui C; Schott-Verdugo S; Shen J; Simmerling CL; Smith J; Salomon-Ferrer R; Swails J; Walker RC; Wang J; Wei H; Wolf RM; Wu X; Xiao L; York DM; Kollman PA AMBER 2018, 2018.

41. Ghanakota P; Carlson HA, Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. The Journal of Physical Chemistry B 2016, 120, 8685–8695. [PubMed: 27258368]

42. Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML, Comparison of Simple Potential Functions for Simulating Liquid Water. Journal of Chemical Physics 1983, 79, 926–935.

43. Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 2015, 11, 3696–3713. [PubMed: 26574453]

44. Ryckaert JP; Ciccotti G; Berendsen HJC, Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. J Comput Phys 1977, 23, 327–341.

45. Gotz AW; Williamson MJ; Xu D; Poole D; Le Grand S; Walker RC, Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. Journal of Chemical Theory and Computation 2012, 8, 1542–1555. [PubMed: 22582031]

46. Andersen HC, Molecular-Dynamics Simulations at Constant Pressure and-or Temperature. Journal of Chemical Physics 1980, 72, 2384–2393.

47. Markwick PR; McCammon JA, Studying functional dynamics in bio-molecules using accelerated molecular dynamics. Phys Chem Chem Phys 2011, 13, 20053–20065. [PubMed: 22015376]

48. Roe DR; Cheatham TE 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J Chem Theory Comput 2013, 9, 3084–3095. [PubMed: 26583988]

49. Schrodinger L The PyMol Molecular Graphics System, 1.7.4.1.

50. Graham SE; Leja N; Carlson HA, MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations. J Chem Inf Model 2018, 58, 1426–1433. [PubMed: 29905479]

51. Jones G; Willett P; Glen RC; Leach AR; Taylor R, Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997, 267, 727–748. [PubMed: 9126849]

52. Verdonk ML; Cole JC; Hartshorn MJ; Murray CW; Taylor RD, Improved protein-ligand docking using GOLD. Proteins 2003, 52, 609–623. [PubMed: 12910460]

53. Iida S; Nakamura HK; Mashimo T; Fukunishi Y, Structural Fluctuations of Aromatic Residues in an Apo-Form Reveal Cryptic Binding Sites: Implications for Fragment-Based Drug Design. J Phys Chem B 2020, 124, 9977–9986. [PubMed: 33140952]

54. Carlson HA; Smith RD; Damm-Ganamet KL; Stuckey JA; Ahmed A; Convery MA; Somers DO; Kranz M; Elkins PA; Cui G; Peishoff CE; Lambert MH; Dunbar JB Jr., CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. J Chem Inf Model 2016, 56, 1063–1077. [PubMed: 27149958]

55. Smith RD; Dunbar JB Jr.; Ung PM; Esposito EX; Yang CY; Wang S; Carlson HA, CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions. J Chem Inf Model 2011, 51, 2115–2131. [PubMed: 21809884]

56. Su M; Yang Q; Du Y; Feng G; Liu Z; Li Y; Wang R, Comparative Assessment of Scoring Functions: The CASF-2016 Update. J Chem Inf Model 2019, 59, 895–913. [PubMed: 30481020]

57. Mlsna D; Monzingo AF; Katzin BJ; Ernst S; Robertus JD, Structure of recombinant ricin A chain at 2.3 A. Protein Sci 1993, 2, 429–435. [PubMed: 8453380]

58. Yan X; Hollis T; Svinth M; Day P; Monzingo AF; Milne GW; Robertus JD, Structure-based identification of a ricin inhibitor. J Mol Biol 1997, 266, 1043–1049. [PubMed: 9086280]

59. LaLonde JM; Bernlohr DA; Banaszak LJ, X-ray crystallographic structures of adipocyte lipid-binding protein complexed with palmitate and hexadecanesulfonic acid. Properties of cavity binding sites. Biochemistry 1994, 33, 4885–4895. [PubMed: 8161548]

60. Xu Z; Bernlohr DA; Banaszak LJ, Crystal structure of recombinant murine adipocyte lipid-binding protein. Biochemistry 1992, 31, 3484–3492. [PubMed: 1554730]

61. Estebanez-Perpina E; Arnold LA; Nguyen P; Rodrigues ED; Mar E; Bateman R; Pallai P; Shokat KM; Baxter JD; Guy RK; Webb P; Fletterick RJ, A surface on the androgen receptor that allosterically regulates coactivator binding. Proc Natl Acad Sci U S A 2007, 104, 16074–16079. [PubMed: 17911242]

62. Pereira de Jesus-Tran K; Cote PL; Cantin L; Blanchet J; Labrie F; Breton R, Comparison of crystal structures of human androgen receptor ligand-binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity. Protein Sci 2006, 15, 987–999. [PubMed: 16641486]

63. Bjorklund C; Oscarson S; Benkestock K; Borkakoti N; Jansson K; Lindberg J; Vrang L; Hallberg A; Rosenquist A; Samuelsson B, Design and synthesis of potent and selective BACE-1 inhibitors. J Med Chem 2010, 53, 1458–64. [PubMed: 20128595]

64. Patel S; Vuillard L; Cleasby A; Murray CW; Yon J, Apo and inhibitor complex structures of BACE (beta-secretase). J Mol Biol 2004, 343, 407–416. [PubMed: 15451669]

65. Blaszczyk J; Li Y; Yan H; Ji X, Crystal structure of unligated guanylate kinase from yeast reveals GMP-induced conformational changes. J Mol Biol 2001, 307, 247–257. [PubMed: 11243817]

66. Stehle T; Schulz GE, Refined structure of the complex between guanylate kinase and its substrate GMP at 2.0 A resolution. J Mol Biol 1992, 224, 1127–1141. [PubMed: 1314905]

67. Hodous BL; Geuns-Meyer SD; Hughes PE; Albrecht BK; Bellon S; Caenepeel S; Cee VJ; Chaffee SC; Emery M; Fretland J; Gallant P; Gu Y; Johnson RE; Kim JL; Long AM; Morrison M; Olivieri PR; Patel VF; Polverino A; Rose P; Wang L; Zhao H, Synthesis, structural analysis, and SAR studies of triazine derivatives as potent, selective Tie-2 inhibitors. Bioorg Med Chem Lett 2007, 17, 2886–2889. [PubMed: 17350837]

68. Shewchuk LM; Hassell AM; Ellis B; Holmes WD; Davis R; Horne EL; Kadwell SH; McKee DD; Moore JT, Structure of the Tie2 RTK domain: self-inhibition by the nucleotide binding loop, activation loop, and C-terminal tail. Structure 2000, 8, 1105–1113. [PubMed: 11080633]

69. Brough PA; Barril X; Borgognoni J; Chene P; Davies NG; Davis B; Drysdale MJ; Dymock B; Eccles SA; Garcia-Echeverria C; Fromont C; Hayes A; Hubbard RE; Jordan AM; Jensen MR; Massey A; Merrett A; Padfield A; Parsons R; Radimerski T; Raynaud FI; Robertson A; Roughley SD; Schoepfer J; Simmonite H; Sharp SY; Surgenor A; Valenti M; Walls S; Webb P; Wood M; Workman P; Wright L, Combining hit identification strategies: fragment-based and in silico approaches to orally active 2-aminothieno[2,3-d]pyrimidine inhibitors of the Hsp90 molecular chaperone. J Med Chem 2009, 52, 4794–809. [PubMed: 19610616]

70. Huth JR; Park C; Petros AM; Kunzer AR; Wendt MD; Wang X; Lynch CL; Mack JC; Swift KM; Judge RA; Chen J; Richardson PL; Jin S; Tahir SK; Matayoshi ED; Dorwin SA; Ladror US; Severin JM; Walter KA; Bartley DM; Fesik SW; Elmore SW; Hajduk PJ, Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. Chem Biol Drug Des 2007, 70, 1–12. [PubMed: 17630989]

71. Horn JR; Shoichet BK, Allosteric inhibition through core disruption. J Mol Biol 2004, 336, 1283–1291. [PubMed: 15037085]

72. Wang X; Minasov G; Shoichet BK, Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. J Mol Biol 2002, 320, 85–95. [PubMed: 12079336]

73. Bowman GR; Bolin ER; Hart KM; Maguire BC; Marqusee S, Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. Proc Natl Acad Sci U S A 2015, 112, 2734–2739. [PubMed: 25730859]

74. Bowman GR; Geissler PL, Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proc Natl Acad Sci U S A 2012, 109, 11681–11686. [PubMed: 22753506]

75. Schiering N; Knapp S; Marconi M; Flocco MM; Cui J; Perego R; Rusconi L; Cristiani C, Crystal structure of the tyrosine kinase domain of the hepatocyte growth factor receptor c-Met and its complex with the microbial alkaloid K-252a. Proc Natl Acad Sci U S A 2003, 100, 12654–12659. [PubMed: 14559966]

76. Schroeder GM; An Y; Cai ZW; Chen XT; Clark C; Cornelius LA; Dai J; Gullo-Brown J; Gupta A; Henley B; Hunt JT; Jeyaseelan R; Kamath A; Kim K; Lippy J; Lombardo LJ; Manne V; Oppenheimer S; Sack JS; Schmidt RJ; Shen G; Stefanski K; Tokarski JS; Trainor GL; Wautlet BS; Wei D; Williams DK; Zhang Y; Zhang Y; Fargnoli J; Borzilleri RM, Discovery of N-(4-(2-amino-3-chloropyridin-4-yloxy)-3-fluorophenyl)-4-ethoxy-1-(4-fluorophenyl)-2-oxo-1,2-dihydropyridine-3-carboxamide (BMS-777607), a selective and orally efficacious inhibitor of the Met kinase superfamily. J Med Chem 2009, 52, 1251–1254. [PubMed: 19260711]

77. Wiesmann C; Barr KJ; Kung J; Zhu J; Erlanson DA; Shen W; Fahr BJ; Zhong M; Taylor L; Randal M; McDowell RS; Hansen SK, Allosteric inhibition of protein tyrosine phosphatase 1B. Nat Struct Mol Biol 2004, 11, 730–737. [PubMed: 15258570]

78. Klopfenstein SR; Evdokimov AG; Colson AO; Fairweather NT; Neuman JJ; Maier MB; Gray JL; Gerwe GS; Stake GE; Howard BW; Farmer JA; Pokross ME; Downs TR; Kasibhatla B; Peters KG, 1,2,3,4-Tetrahydroisoquinolinyl sulfamic acids as phosphatase PTP1B inhibitors. Bioorg Med Chem Lett 2006, 16, 1574–1578. [PubMed: 16386905]

79. Hickman AB; Klein DC; Dyda F, Melatonin biosynthesis: the structure of serotonin N-acetyltransferase at 2.5 A resolution suggests a catalytic mechanism. Mol Cell 1999, 3, 23–32. [PubMed: 10024876]

80. Wolf E; De Angelis J; Khalil EM; Cole PA; Burley SK, X-ray crystallographic studies of serotonin N-acetyltransferase catalysis and inhibition. J Mol Biol 2002, 317, 215–24. [PubMed: 11902838]

81. Cannon RD; Lamping E; Holmes AR; Niimi K; Tanabe K; Niimi M; Monk BC, Candida albicans drug resistance another way to cope with stress. Microbiology 2007, 153, 3211–7. [PubMed: 17906120]

82. Maruyama D; Nishitani Y; Nonaka T; Kita A; Fukami TA; Mio T; Yamada-Okabe H; Yamada-Okabe T; Miki K, Crystal structure of uridine-diphospho-N-acetylglucosamine pyrophosphorylase from Candida albicans and catalytic reaction mechanism. J Biol Chem 2007, 282, 17221–30. [PubMed: 17392279]
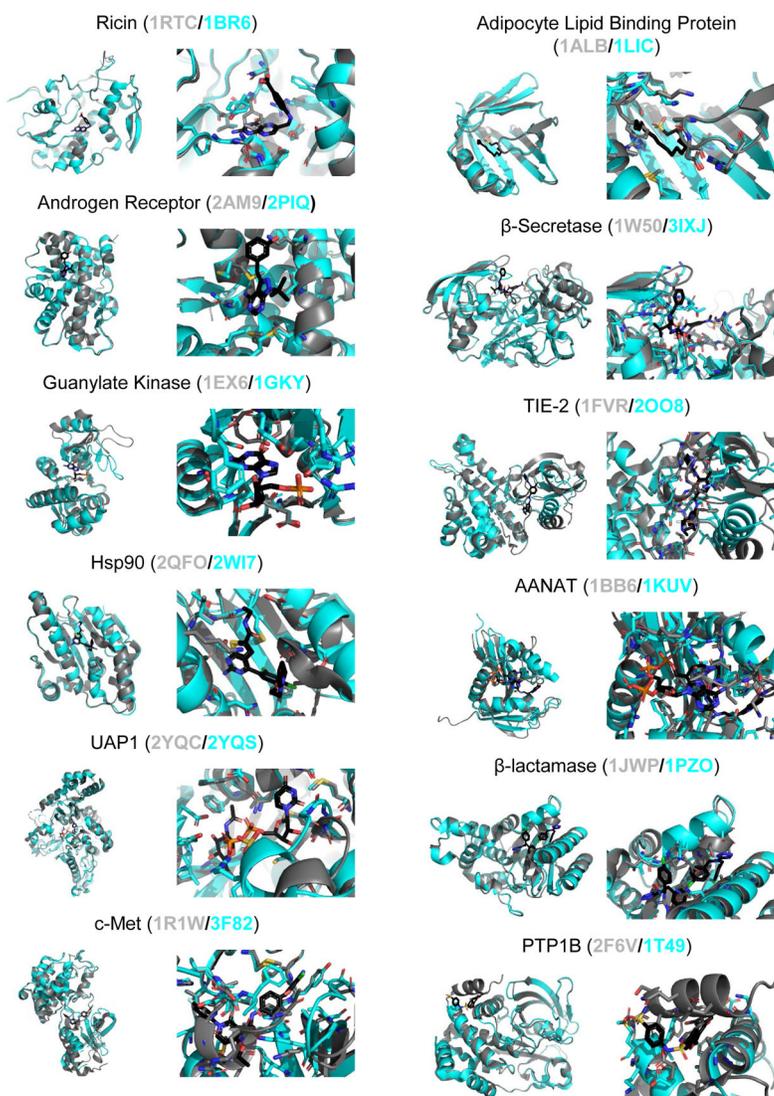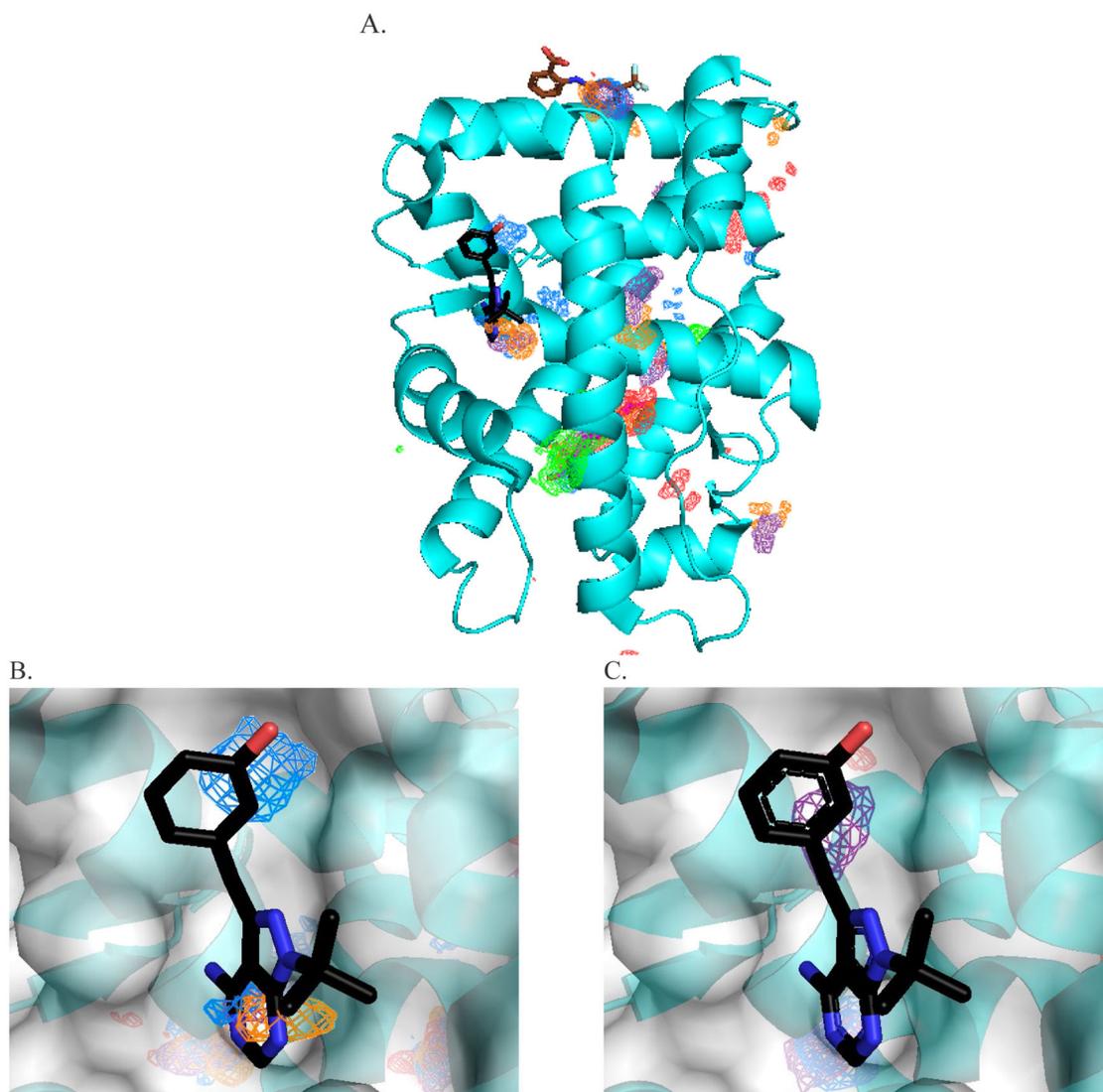
Figure 1.

Cryptic binding sites of the 12 proteins. The figures on the left for each protein show the overall structure of each protein, while the figures on the right display the dynamics within the active site. The dynamics are also described in Table 1. The unbound structure is shown in gray and the bound structure is in cyan. The ligand is drawn as sticks, and carbons are colored black.
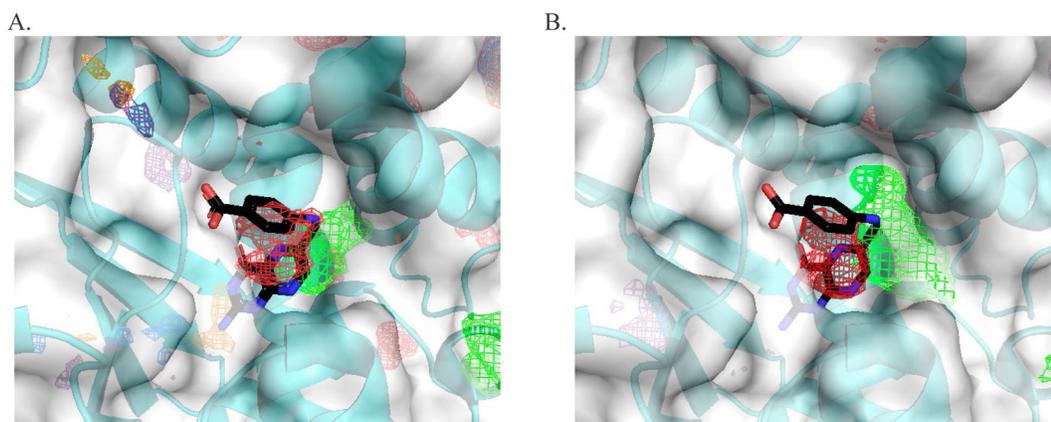
Figure 2.
Binding site of Ricin (A-chain) with standard (A) or accelerated (B) MixMD. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The surface and backbone of the bound structure (1br6) are in white and cyan, respectively. The ligand is drawn as sticks and carbons are colored black.
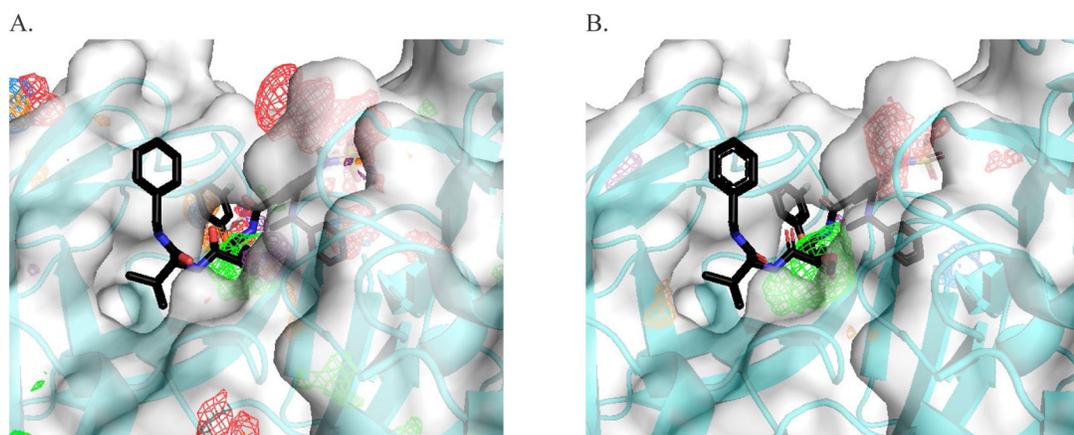
A.

B.
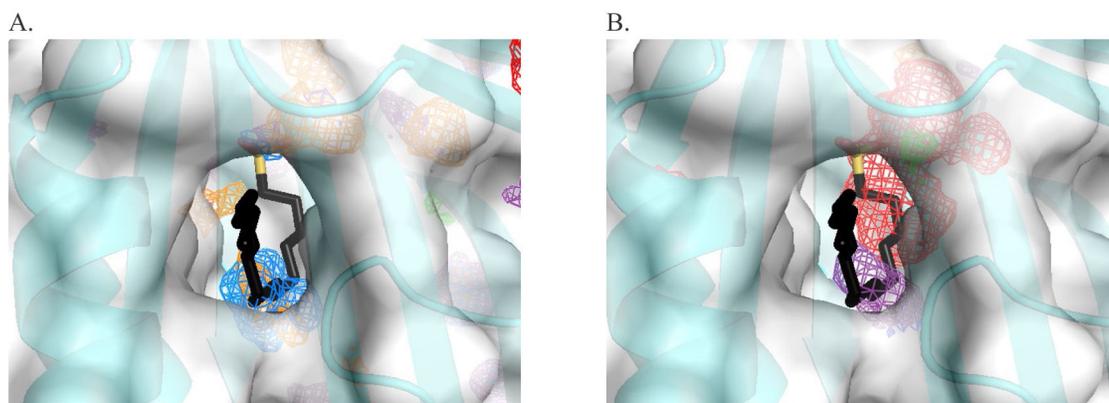
Figure 3.
Mapping of Adipocyte Lipid Binding Protein with standard (A) or accelerated (B) MixMD. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The bound structure (1lic) is shown with a cyan backbone and semitransparent, white surface.

A.

Figure 4.
A) Structure of Androgen Receptor with active site ligand (magenta (2am9)) and two allosteric sites (black (cryptic, 2piq) and brown (2pix)). Mapping of Androgen Receptor with standard (B) or accelerated (C) MixMD. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The backbone of the bound cryptic structure (2piq) is shown in cyan. B) and C) are close-ups of cryptic binding site with the protein surface shown in semitransparent white.

A.

B.

Figure 5.
Mapping of BACE-1 with standard (A) or accelerated (B) MixMD. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The bound structure (3ixj) is shown with a cyan backbone and white, semitransparent surface. The ligand is drawn as sticks and carbons are colored black.
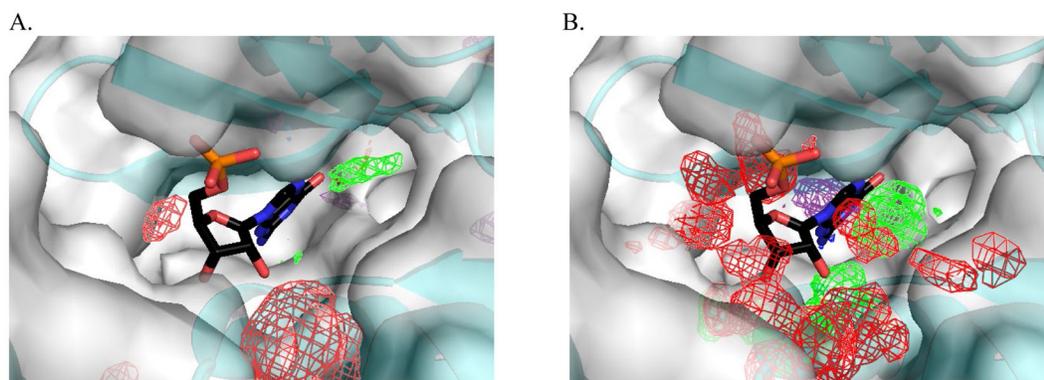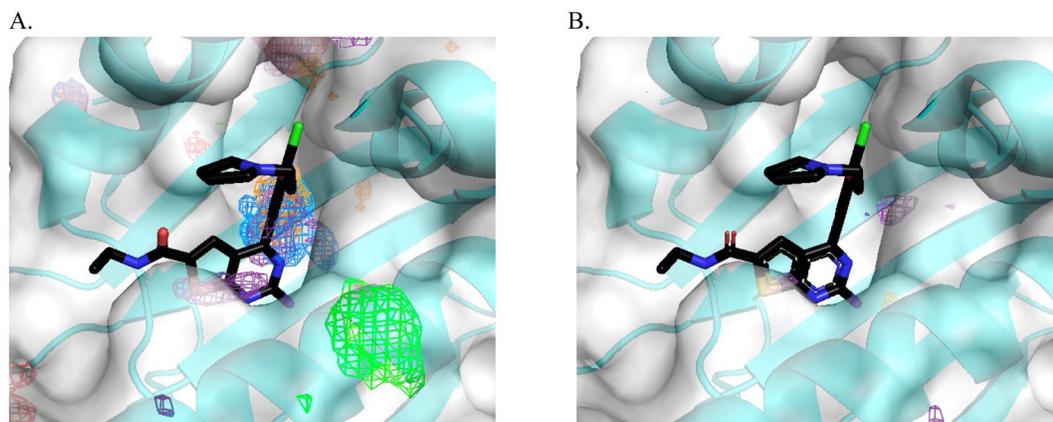
Figure 6.
Mapping of Guanylate Kinase with standard (A) or accelerated (B) MixMD. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The bound structure (1pzo) is shown with a cyan backbone and white surface. The ligand is drawn as sticks and carbons are colored black.

A. B.
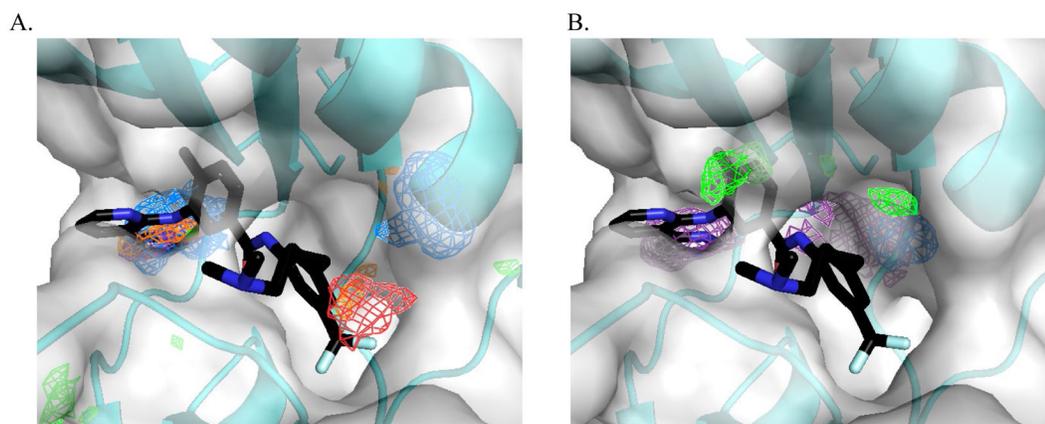
Figure 7.
Mapping of TIE-2 with standard (A) or accelerated (B) MixMD. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The backbone of the bound structure (2oo8) is in cyan, and its semitransparent surface is in white. The ligand is drawn in stick and carbons are colored black.
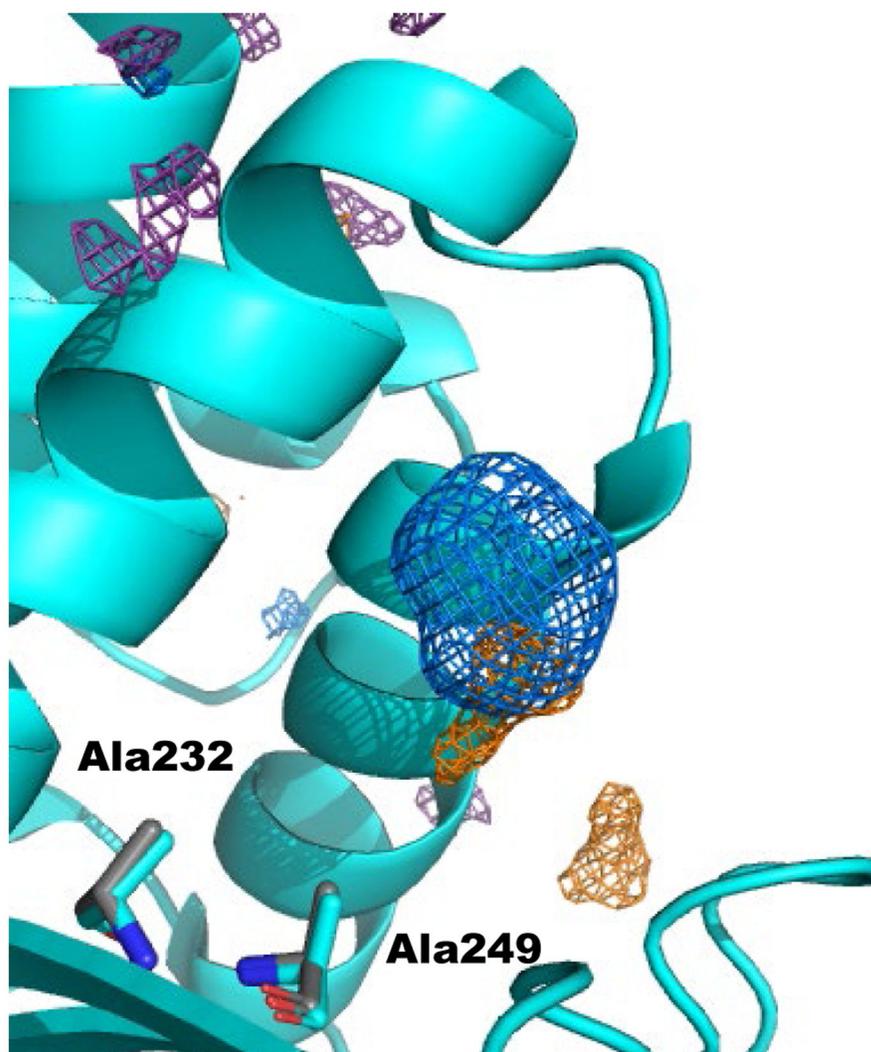
A.

B.

Figure 8.
Mapping of Hsp90 with standard (A) or accelerated (B) MixMD. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The backbone of the bound structure (2wi7) is shown in cyan, and its surface is white and semitransparent. The ligand is drawn in stick and carbons are colored black.

Figure 9.
Alternate binding site of β-lactamase. The occupancy of each probe is shown in mesh (PYR = purple, ACN = orange, IPA = blue, ACT = red, MAI = green) at 35σ. The backbone of the bound structure (1pxo) is shown in cyan.

Table 1.

Summary of mapping protein targets used, in order of increasing complexity of conformational changes. [a]

Protein Target	Apo PDB-ID	Bound PDB-ID	Dynamics	Mapped cryptic (stdMD)	Mapped cryptic (acMD)	Mapped other [b] (stdMD)	Mapped other [b] (acMD)
Ricin A-chain	1rtc	1br6	Y80 moves out of binding site.	Yes	Yes	—[c]	—
Adipocyte Lipid Binding Protein	1alb	1lic	F57 moves out of binding site.	Yes	Yes	—	—
Androgen Receptor	2am9 [d]	2piq	K720 and M734 move out of binding site	Yes	Yes	—	—
β-secretase (BACE-1)	1w50	3ixj	Flap (67-77) closes on binding site.	Yes	Yes	—	—
Guanylate Kinase	1ex6	1gky	GMP binding domain and activation loop move to form binding site.	No	Yes	Yes	Yes
TIE-2	1fvr	2oo8	Nucleotide binding loop and activation loop move out of binding site	No	Yes	Yes	Yes
Hsp90	2qf0	2wi7	Loop reorganization to helix, occludes binding site in unbound.	Yes	No	Yes	Yes
AANAT	1b6b	1kuv	Loop 1 forms helix H1 in bound, occludes binding site in unbound.	No	No	Yes	No
UAP1	2yqc	2yqs	Three loops close on ligand	No	No	Yes	Yes
β-lactamase	1jwp	1pzo	Helix (217-225) has to move.	No	No	Yes	Yes
c-Met	1r1w	3f82	Helix (1222-1230) has to move	No	No	Yes	Yes
PTP1B	2f6v	1t49	Helix σ7 has to move	No	No	Yes	Yes

[a] For details of the mapping of each probe with standard and accelerated MixMD, please see the Supplemental Information, Tables S1-S5.

[b] Some binding sites were large enough to have known binding hotspots beyond the cryptic portion.

[c] The em dash indicates cases where the whole binding site is cryptic; therefore, no other hotspots are available for mapping.

[d] 2am9 was used because the 2ax9 structure from the Cryptosite set was missing several key binding-site residues.

Table 2.

RMSD (Å) of docked ligand pose to the bound conformation.

Protein Target	Standard MD				Accelerated MD			
	apo[a]	bound[a]	Probe MD[b]	Water MD[c]	apo[a]	bound[a]	Probe MD[b]	Water MD[c]
Ricin A-chain	2.31	1.91	*1.76*	2.54	1.93	1.94	2.35	2.40
Adipocyte Lipid Binding Protein	3.17	2.48	*2.86*	3.05	2.86	2.36	3.10	2.63
Androgen Receptor	2.67	1.48	2.76	2.07	1.96	2.40	**2.07**	2.29
β-secretase (BACE-1)	4.88	5.81	**3.88**	3.81	4.79	5.41	**3.54**	5.33
Guanylate Kinase	1.55	1.42	1.77	0.93	1.39	0.93	1.81	2.38
TIE-2	2.18	3.96	**3.24**	2.51	3.08	1.30	*2.01*	3.96
Hsp90	1.66	3.40	**1.61**	1.36	3.26	3.40	**2.75**	2.99
AANAT	3.94	4.07	**3.38**	4.59	3.94[d]	4.07[d]	4.34[d]	4.74[d]
UAP1	2.28	2.42	**2.06**	3.14	3.24	3.12	3.36	3.79
β-lactamase	2.60	1.75	*2.56*	2.79	3.35	2.69	**1.93**	1.33
c-Met	3.46	2.52	*2.65*	2.14	2.69	1.52	*2.42*	3.25
PTP1B	2.49	1.62	*2.13*	2.23	2.99	3.45	3.06	2.82

[a] Note that the docking centers were identified by the data from the simulations, which leads to slight differences for docking to the apo and bound crystal structures in the standard MD vs accelerated MD columns.

[b] Italics is used to highlight when docking to the MixMD set out-performs the apo structure, and bold indicates the MixMD set out-performs the self-docking to the bound structure.

[c] The underline highlights where the water-only ensemble outperforms the MixMD ensemble.

[d] Site was not mapped with accelerated MixMD, so the site from the standard mixMD was used for docking.