HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2019 July 23.

Published in final edited form as:

J Chem Inf Model. 2018 July 23; 58(7): 1426–1433. doi:10.1021/acs.jcim.8b00265.

MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations

Sarah E. Graham[†], **Noah Leja**[‡], and **Heather A. Carlson**[*,†,‡]

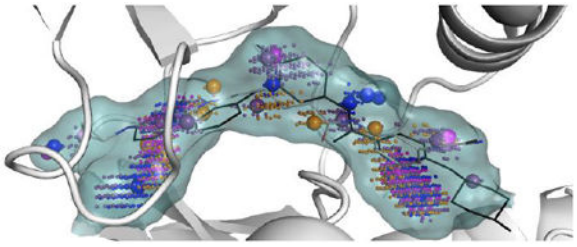
[†]Department of Biophysics, University of Michigan, 930 N. University Ave, Ann Arbor, MI, USA 48109-1055

[‡]Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, 428 Church St., Ann Arbor, MI, USA 48109-1065

Abstract

Mixed-solvent molecular dynamics (MixMD) is a cosolvent simulation technique for identifying binding hotspots and specific favorable interactions on a protein's surface. MixMD studies have the ability to identify these biologically relevant sites by examining the occupancy of the cosolvent over the course of the simulation. However, previous MixMD analysis required a great deal of manual inspection to identify relevant sites. To address this limitation, we have developed MixMD Probeview as a plugin for the freely available, open-source version of the molecular visualization program PyMOL. MixMD Probeview incorporates two analysis procedures: 1) to identify and rank whole binding sites and 2) to identify and rank local maxima for each probe type. These functionalities were validated using four common benchmark proteins, including two with both active and allosteric sites. In addition, three different cosolvent procedures were compared to examine the impact of including more than one cosolvent in the simulations. For all systems tested, MixMD Probeview successfully identified known active and allosteric sites based on the total occupancy of neutral probe molecules. As an easy-to-use PyMOL plugin, we expect that MixMD Probeview will facilitate identification and analysis of binding sites from cosolvent simulations performed on a wide range of systems.

Graphical Abstract



[*]**Corresponding Author:** Phone: 1-734-615-6841 carlsonh@umich.edu.

Supplemental Information

Starting conformation of cosolvent probes around DHFR (Figure S.1) and a comparison of local maxima for all solvent mixtures (Figures S.2–S.5). (PDF)

MixMD Probeview Python Script (txt)

MixMD Probeview User's Guide (PDF)

Introduction

First introduced in 2009[1], hotspot mapping with molecular dynamics (MD) simulations of small molecule probes and water is being increasingly applied towards the development of small molecule inhibitors. These cosolvent simulations provide two types of information. First, when many probes map a location, it identifies binding sites on the protein's surface, including ligand binding sites, protein-protein interaction sites, and other biologically relevant interactions. Secondly, the functional groups on the individual probes identify sites on the protein's surface that favor specific interactions, which can be used to inform structure-based drug design efforts. Several cosolvent simulation methods have been introduced, as recently reviewed.[2] While these methods all utilize mixtures of small molecule probes and water, they have a number of differences regarding the specific probes used, the protocol for simulation, and the method of identifying and ranking the results. For example, some cosolvent methods have focused on the use of a single probe molecule per simulation while others have multiple probes run simultaneously. The MixMD method developed by our group previously utilized a layered setup of a single probe type and water in a 5%/95% v/v probe to water ratio.[3] Introducing charged probes required a transition to ternary solvent mixtures to balance the number of positive and negative charges within the system.[4] Other simulation methods, including the SILCS method[5] from the MacKerell group and cosolvent simulations by Bakan et al.[6], have utilized 4–7 different types of probe molecules within the same simulation. Simulations containing multiple probe types clearly require fewer simulations than comparable methods that simulate each probe separately, but the extent to which this influences the predicted binding sites is unclear.

Traditionally, hotspots have been identified by overlapping density from multiple probe molecules.[4,7,8] In our MixMD method, the occupancy of probe molecules is determined by overlaying the protein and solvent system with a grid and counting the number of times a probe molecule occupies each region. The occupancy is then converted into "σ units", expressed as the number of standard deviations away from the mean occupancy. This allows for the maps to be viewed at different occupancy contour levels, in an analogous way to crystallographic electron density. The resulting maps are visualized in PyMOL to identify the highest occupied sites comprised of multiple probe types. These regions, or hotspots, are then ranked by maximal occupancy.[4] When applied to seven test systems, this method successfully identified known biologically relevant sites on the basis of maximal occupancy.[4] However, manually inspecting every probe map at multiple occupancy contour levels for every system is tedious and time-consuming, thereby limiting the number of systems that can be studied.

Other approaches have identified binding sites by converting the probe occupancies into theoretical binding affinities. In the SILCS method[5] and the method by Bakan et al.[6], the binding affinity at a specific grid point is calculated from the Boltzmann relationship:

$$\Delta G_i = -RT\ln\left(\frac{O_i}{O_{bulk}}\right) \quad (1)$$

where O_i is the occupancy at grid point *i*, O_{bulk} is the expected occupancy in bulk solvent, and T is the temperature. In the SILCS method, these energies are referred to as grid free energies, and they can be used to visualize predicted affinities on the surface of the protein or may be used to determine the theoretical binding affinity of a ligand having atoms at point i.[5] In the approach used by Bakan et al., distinct interaction sites are identified, and the lowest energy point, calculated from Equation 1, is selected to represent the site.[6] Nearby sites are merged and the energies are summed to yield theoretical affinities for each region. The affinities are then used to rank the "druggability" of each site. This approach was used successfully to identify known binding sites for five systems and to rank potential binding sites within each system by the maximum predicted affinity.[6]

While Equation 1 is straightforward to use, there are some inherent limitations in the calculation of binding affinities at the level of sub-atomic grid points using data from simulations of whole probe molecules. The binding affinity of a probe molecule is dependent on the contributions of every atom within the probe. For example, in the case of isopropyl alcohol, the hydroxyl group may be making hydrogen bonding interactions, while the methyl groups are making hydrophobic interactions. Partitioning the binding affinities calculated from the entire probe molecule's occupancy down to the grid point level neglects to consider these effects. Instead, we have focused on the analysis of overall occupancy of the probe molecules as a whole. Using a clustering method to identify separate regions on the protein's surface, we calculate the total occupancy of probe molecules for each site across all simulations. This identifies the regions that are highly occupied by multiple probe types across multiple simulations.

To facilitate application of our MixMD method, we have developed a plugin, which we call MixMD Probeview, for use with the freely available open-source version of PyMOL.[9] Requiring only a PDB-formatted file containing grid points and associated occupancies from a set of cosolvent simulations (easily obtained by post-processing of trajectories with AmberTools[10]), MixMD Probeview identifies binding sites composed of multiple probes as well as local maxima for individual probes. We have validated this method on four systems (including two with allosteric sites), using data taken from more than 2 μs of simulation time per system. Simulations were performed for multiple solvent setup procedures, including both solvents alone (ie. a single probe and water) and in several combinations (ie. 2 or more probe types and water). This allowed us to verify the ability of MixMD Probeview to identify binding sites for a range of systems and cosolvent procedures. Additionally, since simulations were completed for both individual probes and probes run in several different mixtures, we were able to compare the resulting probe occupancy and binding site prediction for different simulation methods. For each system and solvent mixture, the simulations were analyzed at two levels. The first being the ability to correctly predict and identify biologically relevant regions as highly ranked hotspots, and the second being the agreement in functional group mapping between individual and combined probe simulations.

Methods

Simulation Procedures

ABL kinase (PDB:3KFA)[11], Androgen receptor (AR, PDB:2AM9)[12], β-secretase (BACE, PDB:1W50)[13], and dihydrofolate reductase (DHFR, PDB:1DG8)[14] were selected as test systems. These proteins are commonly used benchmark systems and include systems with allosteric sites to provide a thorough test of MixMD Probeview's ability to predict binding sites. All ligands and water molecules in the crystal structures were removed, with the exception of the NADPH cofactor in DHFR which was retained and modeled using the parameters developed by Ryde.[15,16] Hydrogens were added and asparagine, glutamine, and histidine positions were optimized using MolProbity and the Protonate 3D tool in MOE.[17,18] Ionizable residues were assigned their default protonation state at pH 7. For each system, probes were run individually ("solo") or in one of two combined sets, given in Table 1. Portions of these simulations were completed previously by our group.[4] Solvent mixtures were chosen to minimize the need for two probes to compete for mapping the same type of interaction with the protein surface. For example, pyrimidine and imidazole are both aromatic probes and would be expected to occupy many of the same sites. For this reason, none of the probe mixtures include both pyrimidine and imidazole. In each case, a 5%/95% v/v ratio of probe molecules to TIP3P[19] water was maintained, with the 5% of probe molecules split evenly between probe types.

The simulations were initiated using a layered setup, with probe molecules placed around the protein, followed by a box of water to achieve the desired concentration. This setup was chosen to facilitate probe sampling at lower concentrations, consistent with previous development of the MixMD method.[3] The tleap module of AmberTools12 or 14[10, 20] was used for system setup, with the FF99SB[21] force field and previously developed solvent parameters[4, 22]. Probe molecules were distributed using tleap, without preferential placement in known binding sites. For example, setup of DHFR in each solvent type at the 5%/95% v/v concentration resulted in 197 pyrimidine/16,378 waters, 263 acetonitrile/14,472 waters, 176 methylammonium/176 acetate/20,879 waters, 182 isopropyl alcohol/14,663 waters, 165 N-metylacetamide/13,302 waters, and 240 imidazole/16,726 waters. Initial placement of these probe molecules relative to the active site is shown for DHFR as an example in Figure S.1 of the supplementary information. In addition, a different random number seed was used for each simulation so that initial velocities are set individually for each simulation. This helps to ensure that the initial system setup does not bias the results obtained. The systems were initially minimized, followed by heating to 300 K with restraints on the protein. The restraints were then gradually removed as the systems were equilibrated. For each system and solvent type, 10 simulations of 20 ns production time with a 2 fs timestep were completed with AMBER12 or 14.[10,20,23–25] Proper bulk solvent behavior over the course of the simulation was verified using radial distribution functions calculated using the cpptraj module in AmberTools14.[10] Following simulation, the last 10 ns of each trajectory were aligned, and the occupancy of the center of mass of each probe molecule was calculated on a 0.5 × 0.5 × 0.5 Å grid using an in-house modified version of the cpptraj module in AmberTools14.[10,26] The modification to cpptraj was necessary to allow for center-of-mass based occupancies to be calculated.

Analysis Procedures

Our PyMOL plugin, MixMD Probeview, was used for the analysis of the occupancy grids. The plugin and a detailed user guide are included in the supplementary information. The plugin consists of two analysis procedures: 1) to identify and rank whole binding sites and 2) to identify and rank maxima of each probe type. MixMD Probeview is written in Python and uses the scikit-learn package for clustering.[27] In order to identify whole binding sites on the protein's surface, the DBSCAN clustering algorithm was used. This algorithm is capable of identifying density connected regions of any shape or size and does not require a predefined cluster size or number of clusters.[28] DBSCAN clustering relies on three parameters: 1) a cutoff to determine which grid points to cluster, 2) epsilon (ε), the maximum distance by which two points can be separated and still be considered within the same cluster, and 3) the minimum number of points within an epsilon neighborhood for a point to be considered a core point. Clusters are created by grouping all points that are reachable within the epsilon distance and containing at least the minimum number of points. In practice, this allows for the automated identification of clusters of probe occupancy from either overlapping or adjacent grid points. The DBSCAN algorithm is particularly useful for identifying ligand binding sites because of its requirement for connected regions of density, thereby identifying sites that could be connected within the span of a few bond lengths. In the present study, grid points having greater than 10% of the maximum occupancy were used for clustering with a distance parameter of 3 Å. This is approximately the width of pyrimidine or twice the length of a carbon-carbon bond in ethane, and so would identify regions that could be connected within 1–2 bond lengths. The minimum number of points was set to 10 to remove small, sparse clusters from further analysis. Following clustering of the occupancy grid points, the resulting clusters can be ranked by either the maximum occupancy found in the cluster or the total sum of occupancy within the cluster.

While the DBSCAN clustering algorithm is suitable for identifying binding sites, it is not capable of differentiating groups of points whose edges are adjoining, as frequently happens in regions adjacent to local maxima. In order to identify and rank favorable probe binding sites for individual probe molecules, the Mean Shift clustering algorithm[29] was used. The Mean Shift algorithm was chosen as it is capable of identifying arbitrary shapes and sizes of clusters from data points with varying density in 3-D space, making it ideally suited to finding clusters corresponding to local maxima from cosolvent simulations. In the Mean Shift clustering procedure, the distribution of data is represented by a kernel density estimate with bandwidth parameter h. An iterative process is then applied to the data to identify a local density gradient followed by a shift of the center of the kernel until the gradient of the density is zero, and the peak in the data is identified.[29] This clustering process identifies the highest occupied region as the center, with lesser occupied regions surrounding this point grouped into the cluster based on the observed spatial distribution. Larger bandwidth values will generate fewer, larger clusters while a smaller bandwidth value will give a greater number of small clusters. The clusters can then be ranked by the maximum occupancy within the cluster.

As a comparison to other existing methods, binding site detection was also performed using the alternate, fast methods FTsite[7], Fpocket[30], and MOE siteview[17]. These analyses were

performed using the same initial starting crystal structures used for the setup of the MixMD simulations.

Results and Discussion

DBSCAN Clustering to Identify Binding Sites

Previous studies by our group established that biologically relevant sites could be identified based on maximum solvent occupancy in cosolvent simulations.[4] These simulations correctly identified the active and allosteric sites as being among the top ranked sites by occupancy. However, when solvent mixtures are used rather than single cosolvents some sites that would normally be ranked as having maximal occupancy may have intermediate occupancy values because of multiple, exchanging solvent molecules. Ranking by maximal occupancy in these cases would favor sites that bind a single probe type tightly rather than those sites which bind multiple probe types tightly. To account for this, we have moved to ranking based on total occupancy within a region. Occupancies were generated based on the center of mass of each probe molecule so that each probe would contribute equally when the total (summed) occupancies were calculated. The rankings shown in the following sections were generated using our PyMOL plugin with the occupancies of all neutral probe molecules. Previous MixMD studies have shown the ability to identify most active sites using the occupancies of neutral probe molecules.[4] Highly charged binding sites are typically not desirable pharmaceutical targets, and therefore we have focused on the identification of binding site regions using only neutral probes. Charged probe molecules were included in each set of simulations and yield additional insight into the binding preferences of each site that may be used along with the neutral probe results to identify specific favorable interactions.

ABL Kinase

Both active and allosteric ligands bind to ABL kinase, with varying specificity depending on ABL's conformational state. As shown in Figure 1, MixMD simulations identify both the active and allosteric sites as the top ranked sites for every solvent combination tested, though the ordering differed depending on the solvent set. Ligands that bind to the active site of the inactive, DFG-out form of ABL kinase (used to initiate the MixMD simulations) form interactions at the ATP binding site as well as the site that is occupied by phenylalanine in the DFG-in conformation.[11,31] These two sites are encompassed by the MixMD identified binding site, which shows two areas of density connecting over the activation loop. As shown in Figure 1, there is a patch of highly occupied probe density at both the left and right sides of the active-site ligand, corresponding to the ATP and phenylalanine positions, respectively. Summing over the clustered grid points identifies the active site as having the highest total occupancy for both the individual probe and solvent combination B simulations. In the case of the simulations of solvent combination A, the left and right portions of the active site are broken up into two clusters, as they are separated by slightly more than 3 Å. This results in the active site being ranked second, behind the allosteric site. Including the second cluster at the left side of the active site would have ranked the active site as the highest occupied cluster. Regardless, the top two sites clearly have a greater degree of occupancy than other sites, as seen in the boxplot in Figure 1.

The second site identified by the MixMD occupancy corresponds to the allosteric site of ABL kinase. Allosteric ligands bind in the myristate pocket, near the C-terminus. In the autoinhibited form of ABL kinase, the C-terminus adopts a bent conformation, allowing the SH-2 and SH-3 domains to close against the adjacent kinase face.[31] Ligands binding to this site can act to stabilize the autoinhibited form of ABL (eg. GNF-2, PDB:3K5V)[32], or may block bending of the helix to stabilize the active conformation (eg. DPH, PDB:3PYY)[33]. Both allosteric activators and inhibitors occupy the myristate binding site, shown in dark blue in Figure 1. Activators form additional interactions to the left of this site, which effectively blocks helix bending. These additional interactions are replicated in the MixMD simulations, and correspond to the small, light blue cluster to the left of the dark blue allosteric site.

Androgen Receptor

Androgen receptor (AR) is a soluble steroid-type protein that acts as an intracellular transcription factor.[34] AR is stimulated by androgens (e.g., testosterone and 5α-dihydrotestosterone) which bind to the active site and regulate gene expression for male sexual characteristics. Both agonists and antagonists of AR have been developed to treat conditions such as hypogonadism and prostate cancer.[34] As shown in Figure 2, ranking by total occupancy from MixMD simulations successfully identifies the active site as the top ranked site in all three solvent sets.

AR also contains two allosteric sites, as shown in Figure 2. Ligands binding to these sites alter the receptor's conformation, and subsequently, its ability to bind to steroid receptor coactivator 2–3(SRC2–3).[35] The inability to bind to SRC2–3 hinders the receptor's functionality, which ultimately diminishes the androgen response. These allosteric sites were identified in all three sets of simulations, but the ranking differed depending on the solvent set used. In the solo and solvent combination A simulations, the active site and two allosteric sites were the top three ranked sites. In the simulations of solvent combination B, the active site was ranked as number 1, but the two allosteric sites were ranked lower than one site that is a crystal packing interface. Comparing the distribution of occupancies among clusters, this discrepancy might be due to the smaller number of individual simulations for solvent combination B relative to the other solvent combinations. Averaging over a larger number of simulations might better distinguish functional binding sites from other easily desolvated sites on the protein's surface, as shown in the boxplot in Figure 2.

β-Secretase

BACE is responsible for cleavage of β-amyloid precursor protein.[37] The active site of BACE is a large cleft, containing a number of known subsites involved in ligand recognition.[38–40] Ligands do not have to make all of these interactions however, and effective ligands have been developed that bind within only a small region of the overall active site. For example, LY2811376 binds BACE with nanomolar affinity by engaging the catalytic aspartates and S1 and S3 subsites, and leads to decreased levels of Aβ in animals and humans.[41] MixMD simulations correctly identify these subsites, showing the highest levels of probe occupancy within the region occupied by LY2811376. As shown in Figure 3, MixMD identifies the active site cleft as the highest ranked site for every solvent set tested, though the spread of

the clusters differs. The cluster from the solo simulations spans the largest area, with probe occupancy extending across the binding cleft. In the solvent combination A and B simulations, a smaller region is mapped, but this smaller region corresponds to the portion of the active-site known to be targetable by small, high-affinity inhibitors.

Dihydrofolate Reductase

Dihydrofolate reductase (DHFR) is an enzyme that catalyzes the transformation of dihydrofolate to tetrahydrofolate, which is utilized for purine and thymidylate synthesis. Since DHFR is the sole source of tetrahydrofolate, DHFR is a common therapeutic target for many antibiotics, autoimmune disorders, and cancers.[42] As shown in Figure 4, MixMD correctly identifies the active site as the top-ranked site for every solvent mixture. All ligands binding within the active-site of DHFR occupy a T-shaped cleft, which is identified as the most-highly occupied site in our simulations. Some ligands extend beyond this core area to make additional interactions. For example, methotrexate contains two carboxylate groups that bind at the very edge of the active-site region. Identification of binding sites was based on neutral probe occupancy, so these sites are not visible in Figure 4, but are seen as local maxima of acetate (Figure S.4, supplementary information). This demonstrates the ability of MixMD to correctly identify the core active-site region as well as accessory sites that may be utilized by some ligands.

Comparing Local Maxima across Solvent Types

As demonstrated in the preceding sections, binding sites can be identified for any of the tested solvent mixtures by considering the total occupancy within a region as mapped by all of the neutral probes. In addition to binding site prediction, however, cosolvent simulations are also frequently used to identify specific interactions of individual probes for use in structure-based drug design. It is possible that solvents run in combination may compete with each other for binding, leading to fewer local sites being identified when solvent mixtures are used rather than solo cosolvent simulations. It is also possible that there may be cooperativity between probes, leading to additional local maxima in adjacent regions that cannot be observed in solo runs.

In order to compare the occupancies across the three sets of simulations, grid points were clustered using the mean shift algorithm implemented in MixMD Probeview to identify local maxima and surrounding points. Comparing the local maxima of each solvent within the active-site region shows differences for some systems between simulations done with each probe individually and those of combined solvent mixtures. For example, simulations of individual probes with ABL kinase identify local maxima for acetonitrile, imidazole, and isopropyl alcohol within the ATP binding portion (left side) of the active site (Figure 5). In solvent combination B, these three solvents are run in combination. In this case, acetonitrile and imidazole preferentially occupy this site over isopropyl alcohol. This result does not appear to be an artifact of system setup, as none of the simulations (either solo or combined) were initiated with these probe molecules directly in the active site. Moreover, the occupancies shown were generated by averaging over ten individual runs, each with different initial velocities set from a random number seed. Therefore, it appears that the differences in observed occupancies at this site are due to a preference for acetonitrile and

imidazole over isopropyl alcohol. While acetonitrile and imidazole still capture the tendency for hydrophobic and aromatic interactions within this region, hydrogen bonding information that may have been captured by isopropyl alcohol is lost. The observed preferential binding also has implications for calculating binding affinities based on probe occupancy. Most cosolvent methods use the Boltzmann relationship (Eq. 1) to calculate binding affinities based on the occupancy of probe molecules. In the event of preferential binding by some probes for a specific site, the non-favored probes would have artificially low occupancies relative to the expected distribution, leading to errors in the calculated binding affinities. Individual probe occupancy for every system is included in the supplementary information (Figures S.2–S.5).

Comparison with Alternate Methods for Binding-Site Detection

FTsite[7], Fpocket[30], and MOE siteview[17]were all sucessful in identifying the active site of each of the tested systems as the top-ranked site. These methods do not rely on molecular dynamics simulations, and so are much quicker than MixMD for generic binding-site detection. However, these methods were less capable of identifying and prioritizing allosteric sites of the tested systems. FTsite was not able to identify the allosteric sites on either ABL kinase or androgen receptor. MOE Siteview and Fpocket were able to identify the allosteric sites, but these sites were ranked below other regions that are not known binding sites. Rankings for each system are shown in Table 2. It is also important to note that the overall detail obtained from these binding site prediction methods differs. Fpocket provides polar and nonpolar interaction preferring regions, while the MixMD method can be used to identify interaction preferences of specific probe types.

Conclusions

MixMD Probeview successfully identified known active and allosteric sites based on total occupancy of all neutral probe solvents for all systems tested. For each system, the top-ranked site was either the active or allosteric site. In comparison with methods that do not rely on MD simulations, MixMD Probeview better prioritized known, allosteric binding sites over other sites that are not known to bind ligands. For systems having both active and allosteric sites, all of the additional known binding sites were ranked above the remaining sites, with the exception of one set of simulations for AR. As an easy-to-use plugin for the popular visualization software PyMOL, we expect that MixMD Probeview will facilitate identification of binding sites from cosolvent simulations performed on a wide range of systems.

In addition to Probeview’s ability to find regions containing multiple probe molecules, it automates identification and ranking of local maxima of each individual probe solvent by occupancy. Validation studies across both single and combined cosolvent mixtures allowed us to compare the differences in probe sampling across setup procedures. While the top-ranked sites identify the allosteric and active sites for every setup procedure tested, the solo probe simulations show the greatest separation between real binding sites and the rest of the protein surface. As shown in the boxplots in Figures 1–4, when a greater number of simulations are used for analysis, there is a greater separation in total occupancy between

known binding sites and less meaningful sites on the protein surface. However, the number of simulations that can be completed is limited by system size and computational resources. Researchers have frequently turned to combined solvent mixtures to reduce the overall number of simulations that must be completed, which appears to be an acceptable choice when the end goal is binding site identification. In regards to mapping all potential interactions within a binding site, the single probe simulations show the best ability to identify all potential interactions. When combined simulations are used, not all local maxima found in solo probe simulations are seen. This is due to other probe types binding more favorably and displacing the other potential probes. Therefore, when the goal of cosolvent simulations is to uncover all potential interactions within a binding site, using single probe solvents appears to be the most reliable choice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Chemical Computing Group for their generous donation of the MOE software used in this work. This work has been supported in part by the National Institutes of Health (R01 GM65372). The authors thank Katherine Guild for her assistance performing molecular dynamics simulations. Sarah Graham thanks the Rackham Graduate School for assistance in purchasing computing equipment. The authors are also grateful for computing cycles provided by Charles L. Brooks III at the University of Michigan.

References

(1). Seco J; Luque F; Barril X, Binding Site Detection and Druggability Index from First Principles. J. Med. Chem 2009, 52, 2363–2371. [PubMed: 19296650]

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

(3). Ung PMU; Ghanakota P; Graham SE; Lexa KW; Carlson HA, Identifying Binding Hot Spots on Protein Surfaces by Mixed-Solvent Molecular Dynamics: HIV-1 Protease as a Test Case. Biopolymers 2016, 105, 21–34. [PubMed: 26385317]

(4). Ghanakota P; Carlson HA, Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. J. Phys. Chem. B 2016, 120, 8685–95. [PubMed: 27258368]

(5). Raman E; Yu W; Lakkaraju S; Mackerell A, Inclusion of Multiple Fragment Types in the Site Identification by Ligand Competitive Saturation (SILCS) Approach. J. Chem. Inf. Model 2013, 53, 3384–3398. [PubMed: 24245913]

(6). Bakan A; Nevins N; Lakdawala A; Bahar I, Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. J. Chem. Theory Comput 2012, 8, 2435–2447. [PubMed: 22798729]

(7). Kozakov D; Grove LE; Hall DR; Bohnuud T; Mottarella SE; Luo L; Xia B; Beglov D; Vajda S, The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. Nat. Protoc 2015, 10, 733–755. [PubMed: 25855957]

(8). Mattos C; Bellamacina C; Peisach E; Pereira A; Vitkup D; Petsko G; Ringe D, Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. J. Mol. Biol 2006, 357, 1471–1482. [PubMed: 16488429]

(9). PyMOL 1.8.4.0, Schrodinger: New York, NY, 2016.

(10). Case DA; Babin V; Berryman JT; Betz RM; Cai Q; Cerutti DS; Cheatham TEI; Darden TA; Duke RE; Gohlke H; Goetz AW; Gusarov S; Homeyer N; Janowski P; Kaus J; Kolossváry I; Kovalenko A; Lee TS; LeGrand S; Luchko T; Luo R; Madej B; Merz KM; Paesani F; Roe DR; Roitberg A;

Sagui C; Salomon-Ferrer R; Seabra G; Simmerling CL; Smith W; Swails J; Walker RC; Wang J; Wolf RM; Wu X; Kollman PA AMBER 14, University of California: San Francisco, CA, 2014.

(11). Zhou T; Commodore L; Huang W-S; Wang Y; Sawyer T; Shakespeare W; Clackson T; Zhu X; Dalgarno D, Structural Analysis of DFG-in and DFG-out Dual Src-Abl Inhibitors Sharing a Common Vinyl Purine Template. Chem. Biol. Drug Des 2010, 75, 18–28. [PubMed: 19895503]

(12). 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]

(13). 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]

(14). Li R; Sirawaraporn R; Chitnumsub P; Sirawaraporn W; Wooden J; Athappilly F; Turley S; Hol WG, Three-Dimensional Structure of M *tuberculosis* Dihydrofolate Reductase Reveals Opportunities for the Design of Novel Tuberculosis Drugs. J. Mol. Biol 2000, 295, 307–323. [PubMed: 10623528]

(15). Holmberg N; Ryde U; Bulow L, Redesign of the Coenzyme Specificity in L-Lactate Dehydrogenase from *Bacillus stearothermophilus* Using Site-Directed Mutagenesis and Media Engineering. Protein Eng 1999, 12, 851–856. [PubMed: 10556245]

(16). Ryde U, Molecular Dynamics Simulations of Alcohol Dehydrogenase with a Four- or Five-Coordinate Catalytic Zinc Ion. Proteins: Struct., Funct., Bioinf 1995, 21, 40–56.

(17). Molecular Operating Environment (MOE) 2013.08, Chemical Computing Group Inc: Montreal, QC, Canada, 2013.

(18). Vincent BC; Arendall WB; Jeffrey JH; Daniel AK; Robert MI; Gary JK; Laura WM; Jane SR; David CR, MolProbity: All-Atom Structure Validation for Macromolecular Crystallography. Acta Crystallogr., Sect. D: Biol. Crystallogr 2010, 66, 12–21. [PubMed: 20057044]

(19). Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML, Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys 1983, 79, 926–935.

(20). Case DA; Darden TA; Cheatham TE; Simmerling CL; Wang J; Duke RE; Luo R; Walker RC; Zhang W; Merz KM; Roberts B; Hayik S; Roitberg A; Seabra G; Swails J; Goetz AW; Kollossvary I; Wong KF; Paesani F; Vanicek J; Wolf RM; Liu J; Wu X; Brozell SR; Steinbrecher T; Gohlke H; Cai Q; Ye X; Wang J; Hsieh MJ; Cui G; Roe DR; Mathews DH; Seetin MG; Salomon-Ferrer R; Sagui C; Babin V; Luchko T; Gusarov S; Kovalenko A; Kollman PA AMBER 12, University of California: San Francisco,CA, 2012.

(21). Hornak V; Abel R; Okur A; Strockbine B; Roitberg A; Simmerling C, Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. Proteins 2006, 65, 712–725. [PubMed: 16981200]

(22). 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]

(23). Case DA; Darden TA; Cheatham TE, III; Simmerling CL; Wang J; Duke RE; Luo R; Walker RC; Zhang W; Merz KM; Roberts B; Wang B; Hayik S; Roitberg A; Seabra G; Kolossváry I; Wong KF; Paesani F; Vanicek J; Liu J; Wu X; Brozell SR; Steinbrecher T; Gohlke H; Cai Q; Ye X; Wang J; Hsieh M-J; Cui G; Roe DR; Mathews DH; Seetin MG; Sagui C; Babin V; Luchko T; Gusarov S; Kovalenko A; Kollman PA AMBER 11, University of California: San Francisco, CA, 2010.

(24). Götz AW; Williamson MJ; Xu D; Poole D; Le Grand S; Walker RC, Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput 2012, 8, 1542–1555. [PubMed: 22582031]

(25). Salomon-Ferrer R; Götz AW; Poole D; Le Grand S; Walker RC, Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput 2013, 9, 3878–3888. [PubMed: 26592383]

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

(27). Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V; Vanderplas J; Passos A; Cournapeau D; Brucher M; Perrot M; Duchesnay E, Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res 2011, 12, 2825–2830.

(28). Ester M; Kriegel H-P; Sander J; Xu X A density-based algorithm for discovering clusters in large spatial databases with noise. *In* Proc. of 2nd International Conference on Knowledge Discovery and Data Mining, 1996, 96, 226–231.

(29). Comaniciu D; Meer P, Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans. Pattern Anal. Mach. Intell 2002, 24, 603–619.

(30). Le Guilloux V; Schmidtke P; Tuffery P, Fpocket: An Open Source Platform for Ligand Pocket Detection. BMC Bioinformatics 2009, 10. [PubMed: 19133123]

(31). Nagar B; Hantschel O; Young MA; Scheffzek K; Veach D; Bornmann W; Clarkson B; Superti-Furga G; Kuriyan J, Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase. Cell 2003, 112, 859–871. [PubMed: 12654251]

(32). Zhang J; Adrian FJ; Jahnke W; Cowan-Jacob SW; Li AG; Iacob RE; Sim T; Powers J; Dierks C; Sun F; Guo GR; Ding Q; Okram B; Choi Y; Wojciechowski A; Deng X; Liu G; Fendrich G; Strauss A; Vajpai N; Grzesiek S; Tuntland T; Liu Y; Bursulaya B; Azam M; Manley PW; Engen JR; Daley GQ; Warmuth M; Gray NS, Targeting Bcr-Abl by Combining Allosteric with ATP-Binding-Site Inhibitors. Nature 2010, 463, 501–506. [PubMed: 20072125]

(33). Yang J; Campobasso N; Biju MP; Fisher K; Pan XQ; Cottom J; Galbraith S; Ho T; Zhang H; Hong X; Ward P; Hofmann G; Siegfried B; Zappacosta F; Washio Y; Cao P; Qu J; Bertrand S; Wang DY; Head MS; Li H; Moores S; Lai Z; Johanson K; Burton G; Erickson-Miller C; Simpson G; Tummino P; Copeland RA; Oliff A, Discovery and Characterization of a Cell-Permeable, Small-Molecule c-Abl Kinase Activator that Binds to the Myristoyl Binding Site. Chem. Biol 2011, 18, 177–186. [PubMed: 21338916]

(34). Gao W; Bohl CE; Dalton JT, Chemistry and Structural Biology of Androgen Receptor. Chemical Rev 2005, 105, 3352–3370.

(35). 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]

(36). Nique F; Hebbe S; Peixoto C; Annoot D; Lefrancois JM; Duval E; Michoux L; Triballeau N; Lemoullec JM; Mollat P; Thauvin M; Prange T; Minet D; Clement-Lacroix P; Robin-Jagerschmidt C; Fleury D; Guedin D; Deprez P, Discovery of Diarylhydantoins as New Selective Androgen Receptor Modulators. J. Med. Chem 2012, 55, 8225–8235. [PubMed: 22897611]

(37). Lin X; Koelsch G; Wu S; Downs D; Dashti A; Tang J, Human Aspartic Protease Memapsin 2 Cleaves the β-Secretase Site of β-Amyloid Precursor Protein. Proc. Natl. Acad. Sci. U.S.A 2000, 97, 1456–1460. [PubMed: 10677483]

(38). Hong L; Koelsch G; Lin X; Wu S; Terzyan S; Ghosh AK; Zhang XC; Tang J, Structure of the Protease Domain of Memapsin 2 (β-Secretase) Complexed with Inhibitor. Science 2000, 290, 150–153. [PubMed: 11021803]

(39). Hong L; Turner RT, 3rd; Koelsch G; Shin D; Ghosh AK; Tang J, Crystal Structure of Memapsin 2 (Beta-Secretase) in Complex with an Inhibitor OM00–3. Biochemistry 2002, 41, 10963–10967. [PubMed: 12206667]

(40). Turner RT, 3rd; Hong L; Koelsch G; Ghosh AK; Tang J, Structural Locations and Functional Roles of New Subsites S5, S6, and S7 in Memapsin 2 (Beta-Secretase). Biochemistry 2005, 44, 105–112. [PubMed: 15628850]

(41). May PC; Dean RA; Lowe SL; Martenyi F; Sheehan SM; Boggs LN; Monk SA; Mathes BM; Mergott DJ; Watson BM; Stout SL; Timm DE; Smith LaBell E; Gonzales CR; Nakano M; Jhee SS; Yen M; Ereshefsky L; Lindstrom TD; Calligaro DO; Cocke PJ; Greg Hall D; Friedrich S; Citron M; Audia JE, Robust Central Reduction of Amyloid-β in Humans with an Orally Available, Non-Peptidic β-Secretase Inhibitor. J. Neurosci 2011, 31, 16507–16516. [PubMed: 22090477]

(42). Schnell JR; Dyson HJ; Wright PE, Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase. Annu. Rev. Biophys. Biomol. Struct 2004, 33, 119–140. [PubMed: 15139807]

(43). Wu YJ; Guernon J; Rajamani R; Toyn JH; Ahlijanian MK; Albright CF; Muckelbauer J; Chang C; Camac D; Macor JE; Thompson LA, Discovery of Furo[2,3-d][1,3]thiazinamines as Beta Amyloid Cleaving Enzyme-1 (BACE1) Inhibitors. Bioorg. Med. Chem. Lett 2016, 26, 5729–5731. [PubMed: 27816517]

(44). Ghosh AK; Reddy BS; Yen YC; Cardenas E; Rao KV; Downs D; Huang X; Tang J; Mesecar AD, Design of Potent and Highly Selective Inhibitors for Human beta-Secretase 2 (Memapsin 1), a Target for Type 2 Diabetes. Chemical Sci 2016, 7, 3117–3122.

(45). Lam T; Hilgers MT; Cunningham ML; Kwan BP; Nelson KJ; Brown-Driver V; Ong V; Trzoss M; Hough G; Shaw KJ; Finn J, Structure-Based Design of New Dihydrofolate Reductase Antibacterial Agents: 7-(benzimidazol-1-yl)-2,4-diaminoquinazolines. J. Med. Chem 2014, 57, 651–668. [PubMed: 24428639]

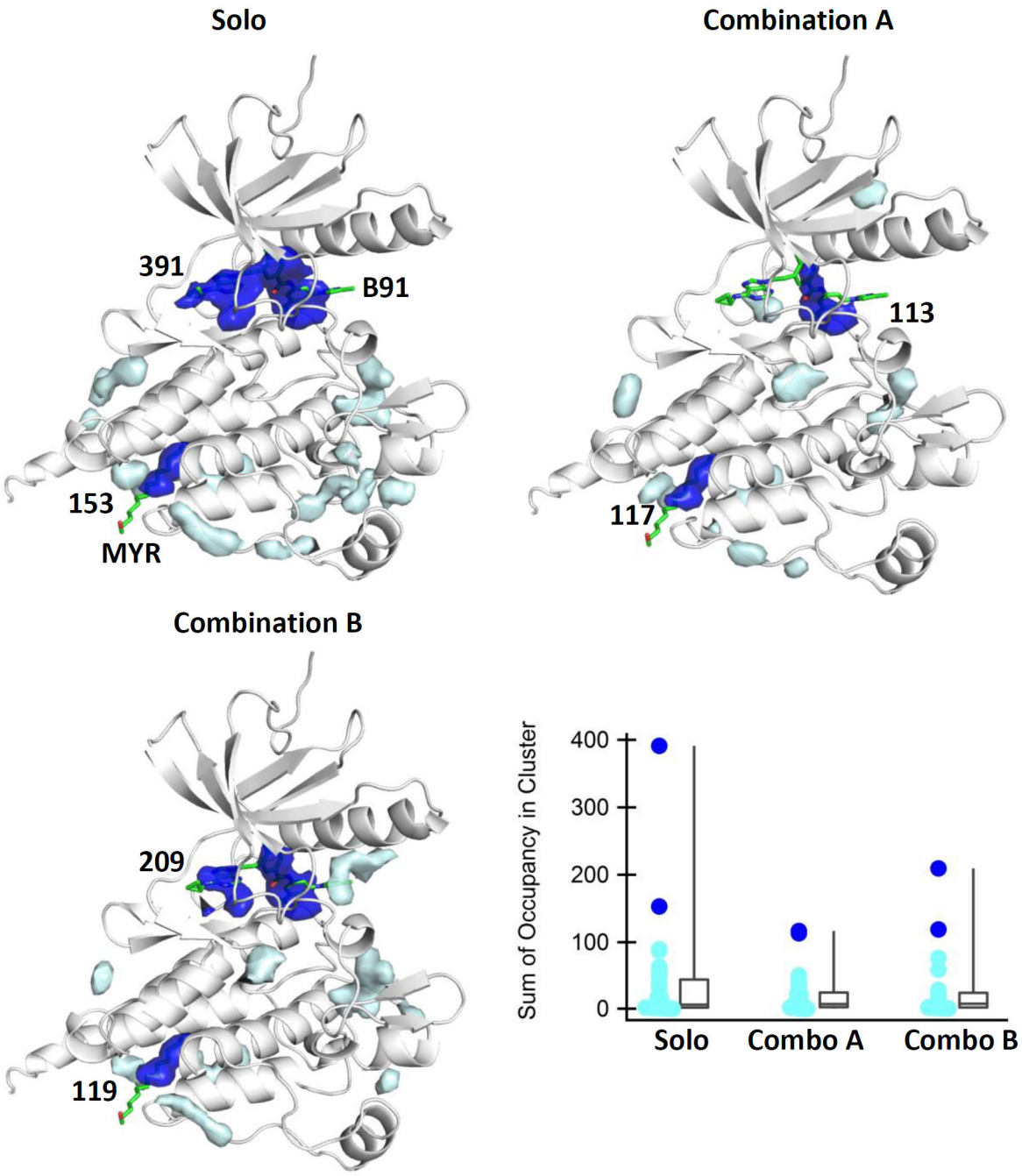


Figure 1:

Cluster ranking by total occupancy for ABL kinase. The active site ligand B91 (PDB: 3KFA)[11] and allosteric ligand (myristate, PDB:1OPJ)[31] are shown for reference. The top two sites for each solvent set are shown as dark blue clusters, with the total occupancy within these clusters given in bold. In every case, ranking by total occupancy identifies the active and allosteric sites as the highest ranked sites. The boxplot shows the distribution of total occupancies for each cluster and solvent set. The top two sites (corresponding to the

active and allosteric sites) are noticeably higher in occupancy than the remaining clusters (light blue).

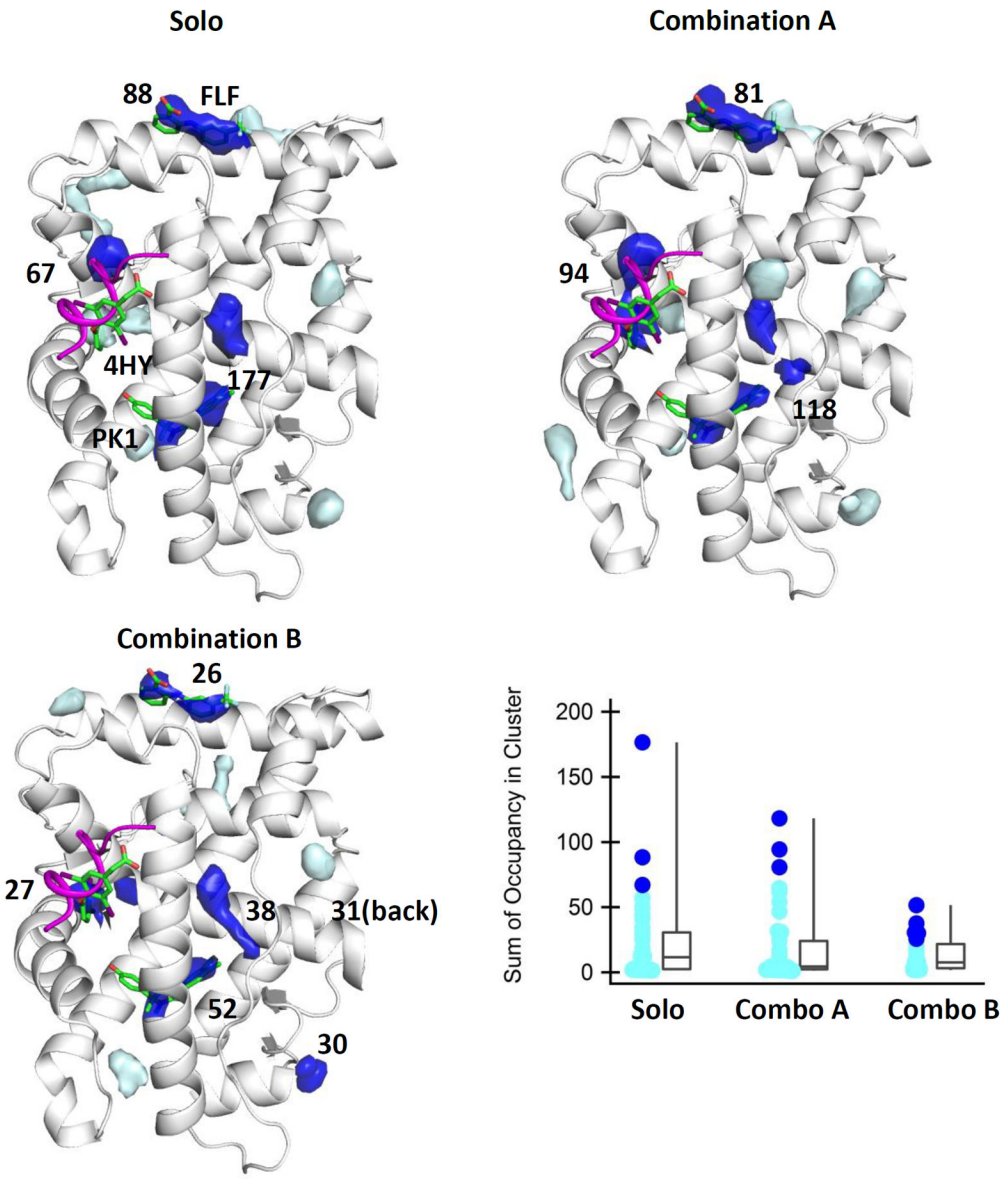


Figure 2:

Cluster ranking by total occupancy for the androgen receptor. The top ranked sites by occupancy are shown in dark blue, with the total occupancies for these clusters in bold. The remaining top ten clusters are shown in light blue. Active (PDB:3V4A, PK1)[36] and allosteric (PDB:2PIU,4HY and PDB:2PIX, FLF)[35] ligands are shown for reference. The SRC-2 coactivator peptide is shown in magenta (PDB:2QPY).[35]The active site is the top ranked site in all cases. In the solo and solvent combination A simulations, the two allosteric sites are the next highest ranked sites. However, in solvent combination B the total occupancies for the remaining sites are close together, making it difficult to discern the allosteric sites from ranking alone.

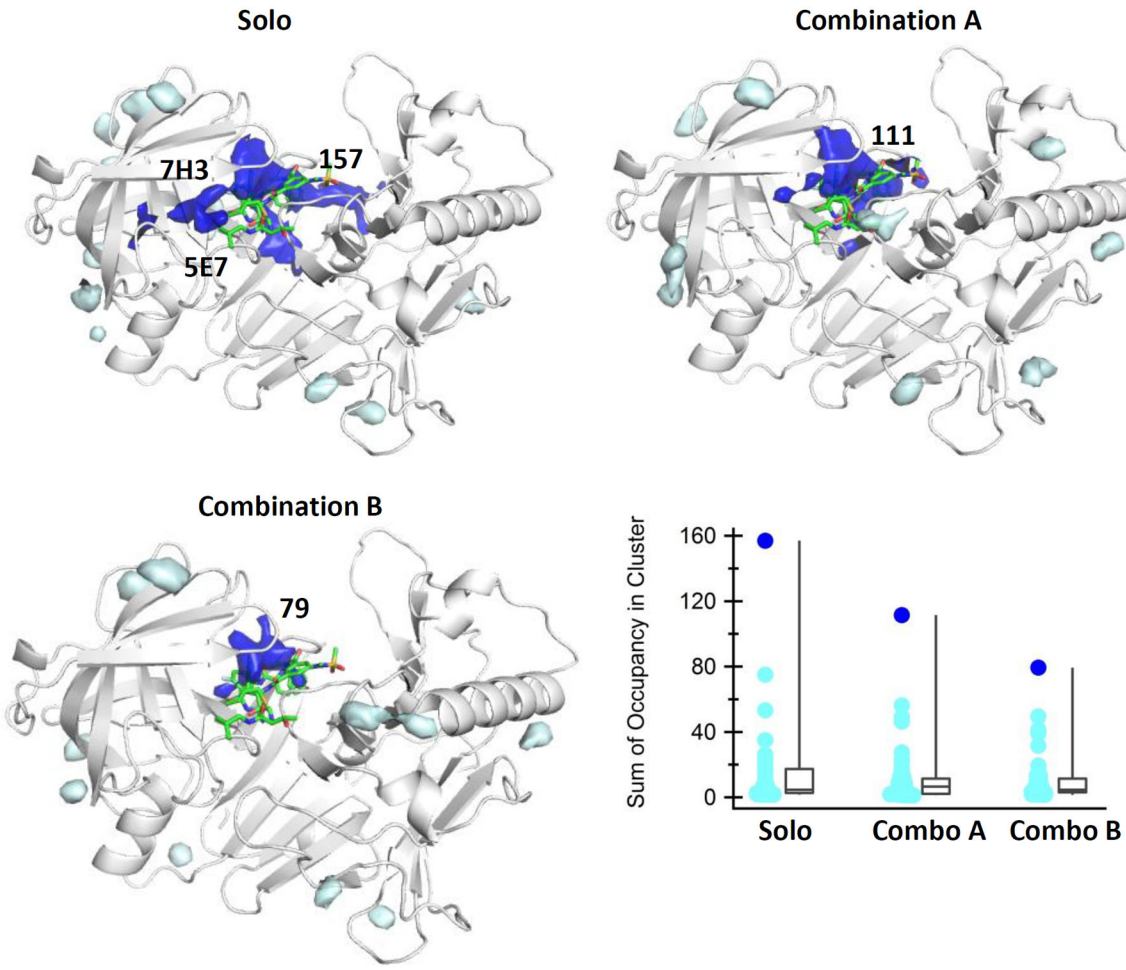


Figure 3:

BACE contains an extended binding cleft, with inhibitors 7H3 (PDB: 5TOL)[43] and 5E7 (PDB:5DQC)[44] shown for reference. In every case, MixMD correctly identifies the active site as the region with the highest total occupancy, shown in dark blue. The total occupancies of the top clusters are given in bold, with the remaining top ten clusters shown in light blue. The top cluster identified from solvent combinations A and B is smaller than that of the solo simulations, but overlaps with the subsites of BACE that have been targeted by small, high-affinity ligands.[41]

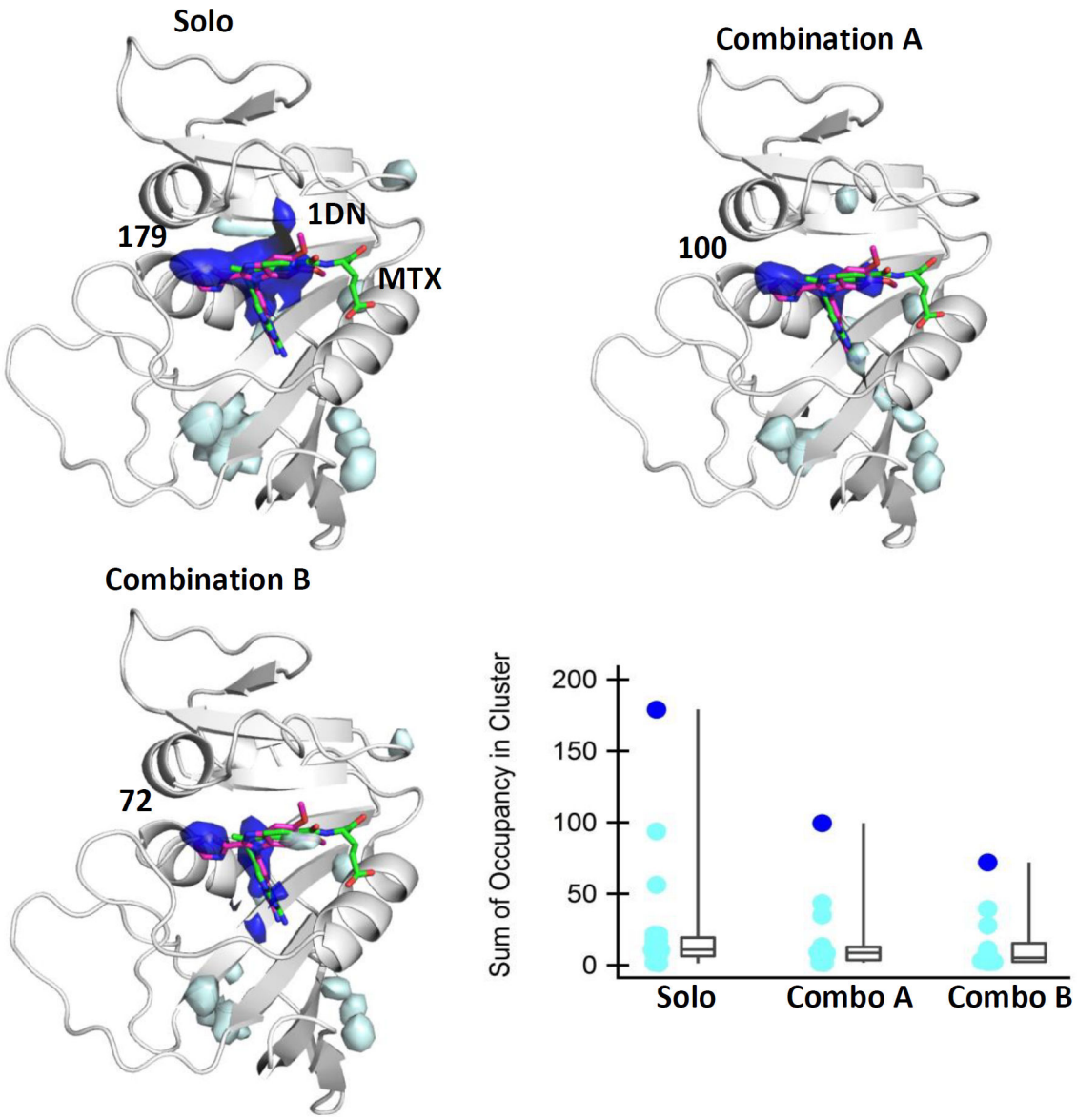


Figure 4.

The active site of DHFR is correctly identified as the top-ranked site (shown in dark blue) across all three sets of MixMD simulations. The total occupancy for the top sites is given in bold, with the remaining top ten clusters shown in light blue. Methotrexate and the ligand 1DN are shown for reference (PDB:1DF7, MTX and PDB:4LEK,1DN).[14, 45]

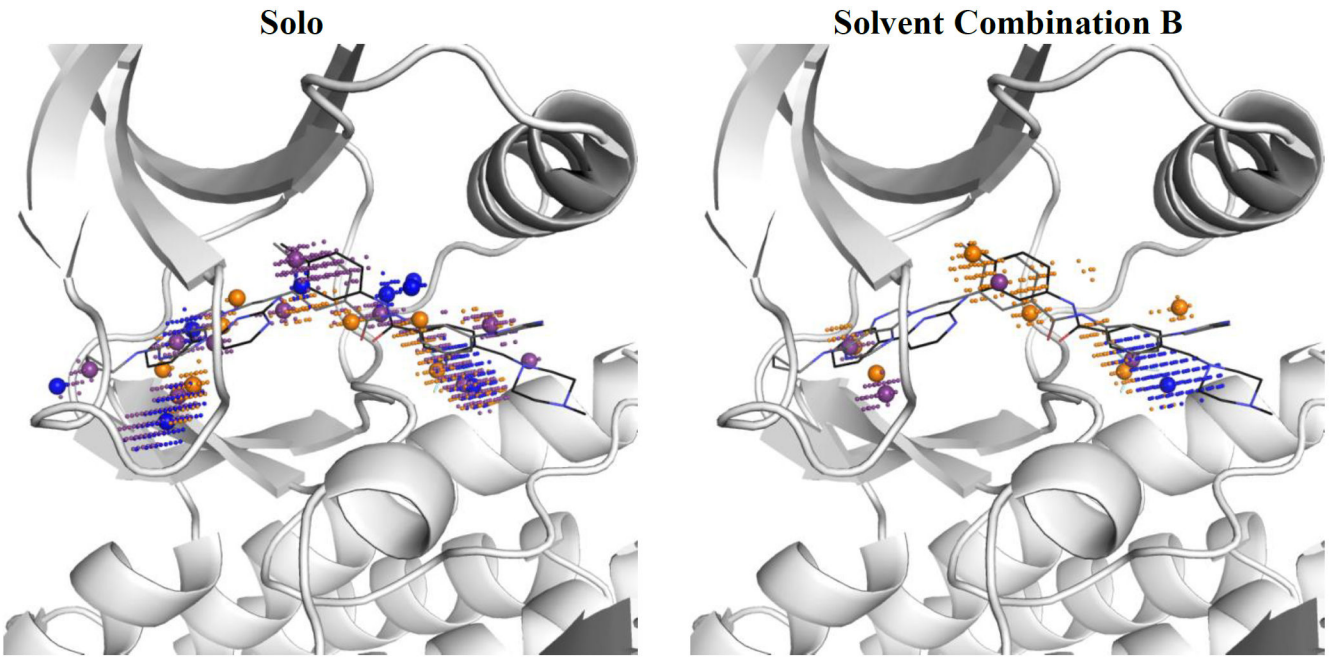


Figure 5:

Acetonitrile (orange), imidazole (purple), and isopropyl alcohol (blue) grid points with greater than 10% occupancy are shown for the active-site region of ABL kinase. Local maxima are shown as spheres, with surrounding grid points shown. Imatinib (PDB:1OPJ)[31] and B91 (PDB:3KFA)[11] are shown for reference. In the solo simulations, acetonitrile, imidazole, and isopropyl alcohol were each run individually. In the combined set B simulations, these three solvents were run in combination. Relative to the solo simulations, the occupancy in the combined simulations identifies fewer local maxima. For example, the isopropyl occupancy seen in the left portion of the ABL active site is absent in the combined solvent simulations, and it is replaced by imidazole and acetonitrile occupancy.

Table 1.

Probe mixtures used for each set of simulations. The solo probes were all run as a single probe in combination with water, except for methyalmmonium and acetate, which must be run together to achieve an overall neutral charge.

Solo	Combination A	Combination B
Acetonitrile (ACN)	Acetonitrile	Acetonitrile
Isopropyl Alcohol (IPA)	+ Isopropyl Alcohol	+ Isopropyl Alcohol
Imidazole (IMI)	Imidazole	+ Imidazole
N-methylacetamide (NMA)	+ N-methylacetamide	N-methylacetamide
Pyrimidine (PYR)	Pyrimidine	+ Pyrimidine
Methylammonium (MAI)	+ Methylammonium	+ Methylammonium
+ Acetate (ACT)	+ Acetate	+ Acetate

Table 2:

Identification and ranking of known active and allosteric sites by each method tested. Binding sites not found by a given method are abbreviated as NF. MixMD Probeview rankings are taken from the simulations of each probe individually.

	FTsite	Fpocket	MOE Siteview	MixMD Probeview
β-Secretase	1	1	1	1
Dihydrofolate Reductase	1	1	1	1
ABL Kinase: active site	1	1	1	1
ABL Kinase: allosteric site	NF	4	13	2
Androgen Receptor: active site	1	1	1	1
Androgen Receptor: allosteric sites	NF	5 & 7	5 & 10	2 & 3