Cosolvent-enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-based Drug Design

Denis Schmidt[§], Markus Boehm[$], Christopher L. McClendon[$], Rubben Torella[$], Holger Gohlke[§,#,*]

[§]Mathematisch-Naturwissenschaftliche Fakultät, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

[$]Medicinal Sciences, Pfizer Inc., Cambridge, Massachusetts, United States

[#]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, Jülich, Germany

Running title: Sampling and unbiased identification of cryptic pockets

Keywords: cryptic sites, transient pockets, molecular dynamics simulations, docking, protein, enhanced sampling, cosolvent

[*]Address: Universitätsstr. 1, 40225 Düsseldorf, Germany.
Phone: (+49) 211 81 13662; Fax: (+49) 211 81 13847
E-mail: gohlke@uni-duesseldorf.de.

1 Abstract

Modulating protein activity with small molecules binding to cryptic pockets offers great opportunities to overcome hurdles in drug design. Cryptic sites are atypical binding sites in proteins that are closed in the absence of a stabilizing ligand and are thus inherently difficult to identify. Many studies have proposed methods to predict cryptic sites. However, a general approach to prospectively sample open conformations of these sites and to identify cryptic pockets in an unbiased manner suitable for structure-based drug design remains elusive. Here, we describe an all-atom, explicit cosolvent, molecular dynamics (MD) simulations-based workflow to sample the open states of cryptic sites and identify opened pockets, in a manner that does not require *a priori* knowledge about these sites. Furthermore, the workflow relies on a target-independent parameterization that only distinguishes between binding pockets for peptides or small-molecules. We validated our approach on a diverse test set of seven proteins with crystallographically determined cryptic sites. The known cryptic sites were found among the three highest-ranked predicted cryptic sites, and an open site conformation was sampled and selected for most of the systems. Crystallographic ligand poses were well reproduced by docking into these identified open conformations for five of the systems. When the fully open state could not be reproduced, we were still able to predict the location of the cryptic site, or identify other cryptic sites that could be retrospectively validated with knowledge of the protein target. These characteristics render our approach valuable for investigating novel protein targets without any prior information.

2 Introduction

Proteins constitute by far the largest group of biological macromolecules whose function can be modulated by small molecules.[1] They commonly interact with their protein target via binding sites, concave clefts on the protein surface.[2] These binding sites are in most cases stable, even in the absence of a binding ligand.[3] In contrast, "cryptic" (or "transient") pockets require the presence of a ligand to open from a closed (*apo*) state.[4-6] Such cryptic sites have been recognized as valuable targets in drug design, particularly with regard to the discovery of allosteric modulators.[3, 5] In most cases, known cryptic sites have been identified serendipitously by means of X-ray crystallography in conjunction with screening or fragment tethering.[3] While these experimental approaches might be applicable to deliberately target yet unidentified cryptic sites of a protein of interest, they are typically cost and labor intensive. Thus, fast computational approaches that can *predict* potential cryptic sites starting from a given *apo* structure are greatly needed. A successful prediction requires two steps: First, efficient *sampling* of the closed *apo* structure is necessary to generate conformations of the cryptic sites in an open state. Second, accurate *identification* is required to select one or a few individual cryptic sites from the sampled conformational ensemble. For a computational approach to be a valuable prospective tool, the sampling and identification must be performed in an *unbiased* manner. Here, unbiased refers to the ability of the method to yield accurate results for a range of targets without the necessity of *a priori* target-specific information or parameterization.

Molecular dynamics (MD) simulations have proven to be successful in sampling cryptic sites.[4, 7-9] Inspired by the application of mixed-solvent MD simulations to identify hot spots and binding sites,[10-12] the incorporation of cosolvents, small probe molecules added to the solvent, was found to improve the sampling of the open state.[11-15] Different agents have been proposed as cosolvents, including (halogenated or hydroxylated) benzene and isopropanol.[11-13, 15, 16] More recently, mixed-solvent MDs were combined with enhanced MD techniques, namely accelerated MDs and Hamiltonian Replica Exchange simulations.[14, 16] While the sampling could be improved, the introduction of further simulation parameters makes deriving a general, broadly applicable simulation protocol challenging. Kokh *et al.* proposed non-equilibrium MD protocols to sample cryptic sites with a focus on slow protein motions.[17] These protocols are based on local perturbations of the protein structure. The authors applied identical simulation protocols to four protein systems, to our knowledge the broadest study where results for a consistent simulation protocol were reported. While shown to be successful in revealing the flexible elements of the binding sites of these four systems, the described protocols required *a priori* knowledge about the presence of a cryptic site.

Unlike the sampling, the identification of cryptic sites, with or without prior information about the cryptic site, has been addressed much less frequently. In a case study on β-lactamase, Bowman and colleagues used clustering, pocket identification, mutual information, and, later, Markov state models to analyze very extensive MD simulations.[8, 9] They were able to predict more than 50 potential cryptic sites on a single protein. However, other studies claimed that only few of those might actually be genuine cryptic sites.[6, 18] Although Bowman *et al.* later substantiated the existence of and communication with the active site for two of the predicted pockets,[8] this discussion stresses the experimental difficulty to validate cryptic sites with respect to their existence, let alone (functional) relevance, and highlights the necessity of predictions with high specificity. Kokh and coworkers developed the tool TRAPP[19] to monitor the dynamics of a given binding site and detect transient pockets. Yet again, TRAPP requires the *a priori* definition of the binding site.[19] Kimura and coworkers adapted a mixed-solvent protocol for hotspot detection to identify cryptic sites.[15] They showed partial or full opening of cryptic sites in eight systems and the existence of highly favorable hotspots therein. However, different cosolvents were used for the different test systems, and, in a few cases, the merging of hotspots was done manually during the hotspot clustering procedure. Therefore, it is unlikely that a common protocol was applied for all eight test systems. While all above methods are based on the identification of cryptic sites from a conformational ensemble, Cimermancic *et al.* developed an interesting alternative.[20] By training support vector machines on residue features, they were able to map potential cryptic site locations onto protein structures. Yet, although certain residue features are optionally derived from MD simulations in this approach, the cryptic sites' open conformations were not explicitly sampled and, thus, are not available for subsequent drug design applications. Hence, many studies have investigated the sampling of cryptic sites, and few have studied the combined sampling and identification of cryptic sites. However, a method that applies a general protocol to prospectively sample and identify cryptic sites suitable for structure-based drug design in an unbiased manner has remained elusive. Again, unbiased refers to a method that samples cryptic site conformations *and* identifies cryptic site locations without *a priori* information on the cryptic site or the necessity of a target-dependent parameterization.

In the present study, we developed a method to sample and prospectively identify cryptic sites that does not rely on *a priori* knowledge about these sites. First, we evaluated different cosolvents to identify the best choice to sample cryptic sites. Second, using this single cosolvent, we generated MD ensembles for all proteins in our test set. Third, we devised a post-processing workflow that specifically identifies cryptic sites from MD ensembles and yields representative structures of these pockets in the open state. Finally, we validated these representative structures by testing whether a known crystallographic binder of the cryptic site could be docked into the sampled pocket conformations.

Our method builds upon all-atom, explicit cosolvent MD simulations. Cosolvent simulations have been frequently used to identify binding sites and hotspots therein (see ref. [21] for a review), and the application of cosolvents to specifically enhance the opening of cryptic sites has been described.[11-16] However, no single, generally applicable MD protocol has emerged so far. Thus, we systematically test different concentrations of multiple cosolvents to identify the, on average, best cosolvent condition to sample cryptic sites. With this cosolvent, multiple simulations are run for each protein in our test set. From the resulting MD ensembles, our method identifies regions on the protein with a higher likelihood to form pockets than in simulations conducted in water only. The use of MD simulations in water as a reference state reduces the number of false positive pockets. Such identified regions, or "pocket cores", are then used to screen for corresponding structures in the MD ensembles, which are clustered to identify representatives as starting points for structure-based drug design applications. The workflow is validated on a set of seven protein systems for which cryptic sites, associated with diverse conformational changes, are crystallographically known. This validation not only assesses the possibility to sample and identify the cryptic sites in question without using *a priori* information about these pockets, but also includes the re-docking of the crystallographic ligands into the representatives extracted from the MD ensembles.

3 Methods

Structure Preparation

The *apo* and *holo* crystal structures of Aldose Reductase (PDB IDs 1x96/4prr), Bcl-xL (PDB IDs 1r2d/4ehr), β-lactamase (PDB IDs 1jwp/1pzo), HSP90 (PDB IDs 1yer/1uyd), Interleukin-2 (PDB IDs 1m47/1m48), MDM2 (PDB IDs 1z1m/3jzk), and p38α (PDB IDs 1wbs/3hvc) were retrieved from the Protein Data Bank[22] (Figure 1). For PDB ID 1z1m, an NMR structure, only the first model was used, and hydrogens were removed before the following preparation step. *Apo* structures were prepared for MD simulations using the Protein Preparation Wizard[23, 24] of the Schrödinger suite (release 2017-2). Crystallized water molecules and additives were removed. Bond orders were assigned, hydrogens and disulfide bonds were added, and protonation states and conformational flips of histidine, asparagine, and glutamine side chains were assigned. N- and C-termini and chain breaks were capped using *N*-methyl amide (NME) and acetyl (ACE) capping groups, respectively. Missing atoms in residue side chains (which only occurred in Interleukin-2) were modelled using Prime[25-27].

Molecular Dynamics Simulations (General Information)

The Amber16 suite of programs[28] was used for all simulations. Unless otherwise mentioned, settings for MD simulations were taken from ref. [29]. The ff14SB force field[30] was used for protein residues and the GAFF2 force field[31] for small organic molecules. For NPT simulations, the Berendsen barostat with a pressure relaxation time of 1.0 ps was used.

Preparation of Solvent Boxes

Solvent boxes with defined cosolvent concentrations were prepared for simulations of the test systems. For this, TIP3P water molecules were placed in a box of 20^3 Å^3 (ethanol) or 30^3 Å^3 (isopropanol, phenol) using PackMol[32] (version 16.344) together with a number of organic solvent molecules matching the desired volume concentration (5%, 10%, 15%, or 20%). The force field parameters for ethanol (ETA) and isopropanol (IPA) were taken from ref. [33]. For phenol (IPH), atomic partial charges were fitted to reproduce electrostatic potentials calculated by *ab initio* methods using the RESP procedure as described in ref. [29]. The force field parameters were assigned by the *parmchk2* module of the Amber suite in agreement with the GAFF2 force field. The resulting boxes were minimized and equilibrated using the *sander* module. The boxes were minimized using the steepest descent and conjugate gradient methods for 5000 steps each. For equilibration, the systems were simulated at 300 K for 100 ps using canonical ensemble (NVT) conditions and subsequently for 50 ps using isothermal-isobaric (NPT) conditions to adjust the density. The *skinnb* parameter for particle mesh Ewald (PME) summation was set from 2 to 1 Å. This was required to avoid the radius of direct Ewald summation for nonbonded contributions (PME *cutoff* + *skinnb*) exceeding the box size.

Water/Cosolvent/Protein Molecular Dynamics Simulations

The *tleap* module was used to solvate the prepared protein structures with the prepared solvent boxes. The protein structures were placed in a truncated octahedron with a margin of at least 12 Å and neutralized using Na^+ or Cl^- ions, as required. All simulations were run using the GPU accelerated version of PMEMD[34]. Initially, the systems were minimized using the steepest descent and conjugate gradient methods for 5000 steps each. The solute was restrained to the initial coordinates using harmonic restraints. This minimization cycle was repeated three times, using force constants for the restraints of 25 kcal mol^{-1} $Å^{-2}$, 5 kcal mol^{-1} $Å^{-2}$, and zero, respectively. The systems were heated in two steps of 50 ps each, both using a time step of 1 fs and the Langevin thermostat set to a collision frequency of 2.0 ps^{-1}. First, the systems were heated to 100 K using NVT conditions, then to 300 K using NPT conditions. Densities were allowed to adapt during additional 200 ps of NPT simulations. During heating and density adaptation, the solute was restrained as during minimization, using a force constant of 5 kcal mol^{-1} $Å^{-2}$. The positional restraints were gradually reduced over 80 ps after density adaptation. The systems were then simulated for 200 ps in the NVT ensemble, which were excluded from analysis. For the evaluation of (co-)solvent effects, each such prepared system was simulated for an additional 500 ns. Snapshots were stored every 20 ps. For the identification of cryptic sites (only for water and IPH10), a total of ten independent replicas were generated by randomizing the velocities during the heating stage.

Pocket Identification

We used PocketAnalyzer^{PCA} 1.3[35] to detect pockets for all snapshots from the MD simulations. PocketAnalyzer^{PCA} is a grid-based pocket detection method, i.e. each identified pocket is represented by a set of grid points forming a negative imprint (henceforth referred to as a pocket grid *P*). It uses three parameters for pocket detection, the degree of buriedness (*dob*), the minimum number of neighbors (*mnb*), and the maximum cluster size (*mcs*). To evaluate the effect of cosolvents on the opening of cryptic sites, these parameters were set to 10, 8, and 100 (*dob*, *mnb*, and *mcs*) for test systems that have a peptidic native ligand (Bcl-xL, Interleukin-2, MDM2) binding to the cryptic site and to 12, 8, and 100 (*dob*, *mnb*, and *mcs*) for test systems binding a small molecule in the cryptic site (ALR, β-lactamase, HSP90, p38α). The smaller *dob* parameters allows PocketAnalyzer^{PCA} to identify even shallow pockets, which are typical for protein-protein interfaces[36, 37]. A grid spacing of 0.8 Å was used consistently.

Evaluation of (Co-)Solvent Effects On Cryptic Site Sampling

To evaluate the effect of different cosolvents and their concentrations on the formation of cryptic sites, pocket grids identified during MD simulations by PocketAnalyzer^{PCA} were compared with reference grids, representing the locations of known cryptic pockets. For each test system, the reference grids were defined around those ligand atoms that bind to a cryptic site, as follows. Each retrieved *holo* structure was aligned to the corresponding *apo* structure using Chimera 1.11[38], and the ligand that binds to the cryptic site was selected. Typically, only parts of the ligand bind to the actual cryptic site, while the rest binds to a continuously open (non-cryptic) pocket of the binding site or extend into the solvent. To truly focus on the cryptic site, ligand atoms within a radius of 2 Å around the atoms of the *apo* structure, i.e., ligand atoms that would cause a steric clash in the *apo* structure, were identified. Finally, a grid was built within a 2.5 Å radius around each of those atoms (henceforth referred to as the reference grid), matching the orientation, localization, and spacing of the pocket grids generated by PocketAnalyzer^{PCA}. Note that the reference grids slightly exceed the actually accessible cryptic pockets in the *holo* structure because they penetrate the van der Waals volume of the receptors by approximately 0.8 Å (due to the 2.5 Å radius). Without this "grace volume", i.e., when using the actual pocket grids identified in the *holo* structures, the detection of pocket opening during the MD trajectories is highly sensitive to the side chain conformations in the *holo* structure. The overlap between the reference grid and a pocket grid is calculated according to eq. (1).

$$\boldsymbol{overlap(P, R)} = \sum_{\boldsymbol{g \in P}} \boldsymbol{\delta(g, R)}\,, \boldsymbol{with\ \delta} = \begin{cases} \boldsymbol{1, if\ g \in R} \\ \boldsymbol{0, otherwise} \end{cases} \tag{1}$$

Here, *g* denotes a grid point in the pocket grid *P*, *R* denotes the reference grid, and *δ* is a delta function. The overlap(P, R) is zero for a pocket grid *P* that does not overlap with the cryptic site, and its maximum is $|R|$. The "accessible cryptic volume" (ACV) in an MD snapshot, i.e., the magnitude of opening of the cryptic site, is quantified using eq. (2).

$$\boldsymbol{ACV_i} = \frac{\sum_j \boldsymbol{overlap(P_{i,j}, R)}}{\sum_j \boldsymbol{overlap(P_{holo,j}, R)}} \tag{2}$$

where $P_{i,j}$ and $P_{holo,j}$ denote the pocket grids for the *j*th pocket in the *i*th MD snapshot and the *holo* structure, respectively. The denominator scales the *ACV* to the accessible volume in the *holo* structure, which is required since the reference grid exceeds the actual cryptic site (*vide supra*). *ACV* = 1 implies that the volumes of the cryptic site in the respective MD snapshot and in the *holo* structure are equal in terms of the number of grid points, while minor side-chain rearrangements are possible. The *i*th MD snapshot was considered in an *open* state if $ACV_i \geq 1$, i.e., the volume of the cryptic site in the respective MD snapshot is as large or larger than in the *holo* structure (Figure 2).

Identification of Cryptic Sites Using Pocket Cores

To identify those regions on the protein surface where pockets are frequently present, the pocket grids identified throughout one or more trajectories were combined to a summary grid *S*. A summary grid includes all grid points that are in any pocket of any snapshot of that combined MD ensemble (eq. 3):

$$S = \{g | \forall\, i, j: g \in P_{i,j}\} \quad (3)$$

where $P_{i,j}$ is the pocket grid for the *j*th pocket in the *i*th MD snapshot. Every grid point of the summary grid has an associated occupancy, Occ_g, which is the relative number of snapshots in the ensemble in which this grid point has been observed (eq. 4).

$$\boldsymbol{Occ_g = \frac{\sum_{i,j} \delta(g, P_{i,j})}{N}, with\ \delta = \begin{cases} 1, if\ g \in P_{i,j} \\ 0, otherwise \end{cases}} \quad (4)$$

δ is a delta function, and N is the total number of snapshots of a given simulation or set of simulations. As each grid point can only be in one pocket grid j for each snapshot i, the limit of the sum in the numerator is N, and the upper limit of Occ_g is one. In turn, $Occ_g = 1$ indicates that *g* is part of the volume of a pocket in every snapshot. Groups of grid points with high occupancy indicate recurring pockets over an ensemble (Figure 3A). Due to side-chain fluctuations, the occupancies of grid points within a pocket will generally increase from the periphery to the center (or "core") (Figure 3C, left panel). To identify cryptic pockets, difference grids were computed by subtracting the occupancy values of summary grids of the simulations with and without cosolvent:

$$\boldsymbol{D_{1-2} = \{g | g \in S_1 \cup S_2\, ;\ Occ_g = Occ_{g,S1} - Occ_{g,S2} \wedge \forall g \notin S_{\{1,2\}}: Occ_{g,S\{1,2\}} = 0\}} \quad (5)$$

with $Occ_{g,S\{1,2\}}$ denoting the occupancy of a grid point in a summary grid for simulations with or without cosolvent, respectively.

Phenol (IPH) was found to be the best cosolvent to identify cryptic sites at a volume concentration of 10% (IPH10, see Results section for the evaluation of cosolvent compositions, Figure 2). Consequently, "pocket cores" were identified from the difference grid $D_{IPH10-WAT}$. Pocket cores are subsets of grid points in such a difference grid with positive occupancies, indicating regions where pockets form more frequently in the presence of phenol as cosolvent, i.e., they constitute potential cryptic sites. The definition of a lower bound occupancy threshold to identify such regions did not appear straightforward to us, as the tendency of pocket opening caused by a cosolvent likely depends on that cosolvent, other simulation conditions including the extent of convergence, and the intrinsic probability for opening of the cryptic site itself. Instead, we considered pocket cores as sets of grid points with a locally increased occupancy compared to the surrounding grid points.

To provide an unambiguous way for its identification, a pocket core was defined as an isolated group of connected and highly occupied grid points, as follows. First, the difference grid is converted to a graph, where each grid point is a node, and edges indicate a direct spatial neighborhood of two nodes. Neighbors are defined as adjacent grid points along x-, y-, and z-directions and all possible diagonals, resulting in, at most, 26 neighbors per node. Second, nodes are gradually removed from the graph in the order of increasing occupancy, i.e., the graph is "depleted" of low-occupancy nodes (Figure 3B). For this, the lower occupancy threshold is increased in steps of 0.001. Upon the removal of nodes (and their respective edges), the initial graph disconnects into subgraphs. Third, after each removal step, nodes with less than four neighbors are iteratively removed, and connected components are identified in the resulting graph. The removal of nodes based on the number of neighbors accelerates the separation of connected components and smoothens the resulting pocket cores. A connected component is then considered a pocket core if it fulfills additional constraints: I) The number of nodes in a connected component (i.e., grid points in a pocket core) must be between 120 and 300 nodes (about 60 to 150 Å^3, for the used grid spacing of 0.8 Å), which has been empirically identified as best range for the used test systems. Note that these thresholds are defined for the pocket cores and the actual pockets will be mostly larger (see section "Pocket selection"). II) A connected component must be > 2.5 Å away from a "floppy end". A floppy end is defined as a stretch of five residues before a capping group, except for those residues that are in a secondary structure element other than *turn* or *bend* for more than 70% of the simulation time; the secondary structure propensity for each residue was calculated using the *cpptraj* module[39] of the Amber suite. This additional constraint was implemented because such floppy ends tend to generate false positive pockets in our experience. Identified connected components were mapped back onto the grid, ranked according to their average occupancy value, and reported as pocket cores. Unless otherwise noted, all above described steps were implemented in-house using Python programs (see Listing S1).

Pocket Selection, Clustering, and Docking

The similarity between a pocket core and a pocket grid *P* is calculated based on the number of overlapping grid points. As the identified pocket cores, by design, are smaller than the largest pocket grids found for a certain binding site, the asymmetric Tversky index[40] was used as similarity measure. The Tversky parameters α and β were set to 0.75 and 0.25, respectively, which favors pocket grids that are a superset of the pocket core. Hence, pockets larger than the pocket core are preferred. Pocket grids with a Tversky index > 0.6 were selected from the MD ensemble, together with their protein structures. All steps were implemented in-house using Python programs (see Listing S2).

Subsequently, the structures selected for a given pocket core were clustered based on the heavy atom RMSD of the binding site residues. The binding site residues were defined as those residues with at least four atoms within 5 Å of the pocket core. Structures were clustered using the DBSCAN algorithm as implemented in *cpptraj* with $\varepsilon = 1$ Å and the *minpoints* parameter set to 4. Clusters with less than ~0.1% of the sampled conformations were excluded from further analysis, which corresponds to 300 members in our setup.

For docking, the cluster representatives and the structures with the largest pocket volume (with respect to the pocket grid) were used. Docking grids were calculated using Schrodinger's Glide[41-43] software (release 2018-1). The binding site was defined by the above selected binding site residues using an ASL (Atom Specification Language) expression. The reference crystal structures were prepared using the Protein Preparation Wizard. The ligands binding to the cryptic sites were extracted and docked using Glide with SP scoring. Twenty ligand poses were generated and optimized using the built-in post-docking optimization routines, and the heavy atom RMSD to the input pose was calculated. These settings are comparable to those of a prospective docking application. The docking pose with the lowest RMSD to the input structure was considered the best pose to assess to what extent the crystallized pose could be reproduced. We have not considered the docking score to identify the best pose for the following reasons. The docking trials described where used as a means to validate whether the automatically sampled and selected pockets are i) open and ii) resemble the open state in such a way that a ligand can bind at all. In contrast, the docking score is influenced by the performance of the scoring function for the specific protein and ligand. In fact, even when re-docking the ligands into their respective crystal structures, only two cases were observed in which the best pose by score and by RMSD were identical (Table S1). Furthermore, only for half of the docking runs, the best pose by score showed an RMSD less than 2 Å (Table S1).

4 Results and Discussion

Evaluation Dataset

Seven proteins for which crystallographic structures have shown the opportunity to form cryptic sites were selected from literature as test systems[7-9, 15, 19]. At least two crystal structures were available for all test systems: one in the *apo* state, with the cryptic site being closed, and one in the *holo* state, with the cryptic site being open. The test systems include Aldose reductase (ALR), Interleukin-2 (IL-2), Mouse double minute 2 homolog (MDM2), Mitogen-activated protein kinase p38α, β-lactamase (β-Lac), B-cell lymphoma-extra large (Bcl-xL), and Heat shock protein 90 (HSP90) (Figure 1). They differ with respect to the conformational change they undergo upon the close-to-open transition of the cryptic site. These conformational changes vary from being dominated by single side chain movements such as Leu300 in ALR[19, 44] or Phe42 in IL-2[19, 45], breathing-like motions[46] of the pocket as in MDM2[47] and p38α[19, 48], major backbone and side-chain rearrangements in β-Lac[49], transitions of secondary structure elements (HSP90[19, 50]), or the combination of multiple conformational changes throughout the binding interface (Bcl-xL[13, 51, 52]). All proteins are well recognized test systems and have been tested in different studies on cryptic sites before.[7-9, 15, 19] Hence, we consider our dataset relevant, diverse, and challenging. Especially for HSP90 and β-Lac, the sampling of the cryptic site using MD simulations has been acknowledged as difficult.[16, 17] With seven proteins, our evaluation dataset is also amongst the largest ones that have been used for cryptic site prediction. Kimura and coworkers used eight structures, which were selected from a dataset compiled by Cimermancic *et al.*[20] Except for p38α, however, we are not aware of comparable studies on the MD simulations-based sampling of cryptic sites for these eight structures. A comparison of the two datasets is thus difficult. Furthermore, the authors used different solvents for the identification of cryptic sites for different proteins, where in this study, several cosolvents were evaluated but the pocket identification was based on a single cosolvent.

Evaluation of Cosolvent Compositions for Sampling Cryptic Pockets

In order to allow prospective applications, we intended to identify the overall best cosolvent condition to foster the opening of cryptic sites during MD simulations. The test systems were simulated in the presence of varying volume concentrations (5%, 10%, 15%, and 20%) of the cosolvents ethanol (ETA), isopropanol (IPA) and phenol (IPH). A volume concentration of 5% has been proposed in other mixed-solvent studies[15, 53, 54] but we presumed that higher concentrations of cosolvent will have a larger effect on the formation of cryptic sites. The three selected cosolvents were used because they are neutral and have an amphipathic character, and should thus be able to bind to and stabilize hydrophobic cryptic sites. Furthermore, they mimic side chains of amino acids and should be able to displace those during simulations. The choice of isopropanol and phenol is further corroborated by a recent study[15], indicating that isopropanol and resorcinol (benzene-1,3-diol) can induce pocket opening.

For each solvent, the fraction of frames in which the cryptic site is open was assessed (Figure 2A) by measuring the *accessible cryptic volume* (ACV, eq. 2). The ACV is rigorously defined, and we argue that this metric is a more system-independent and direct assessment of the opening of a cryptic site than those proposed by other authors, such as the set of residues lining the binding site, projections of inter-residue distances around the binding site, or selected backbone dihedrals[15-17]. Not unexpectedly, our results indicate that the test systems are differently susceptible to the opening of cryptic sites: In IL-2 and MDM2, the cryptic sites open readily in the presence of cosolvents, whereas they are less frequently observable for other proteins in the test set, especially β-Lac and HSP90, which is in line with the large conformational changes required and the challenging sampling process.[16, 17]

In simulations considering water only, the known cryptic sites open to full extent (ACV ≥ 1, i.e., the observed volume is equal to or larger than the volume of the *holo* pocket) only in < 1.5% of the combined simulation time. In contrast, organic molecules as cosolvents increase the fraction of frames with open cryptic sites. At concentrations of 15% or above, however, we observed unfolding events in several MD trajectories. Hence, we discourage the use of such high concentrations of cosolvents. Averaged over the seven test systems, cryptic pockets open significantly ($p < 0.05$) more frequently (> 35%) upon the addition of 10% phenol (henceforth denoted as IPH10). Note that, as our assessment of cryptic site formation is rather strict, low rates for β-Lac and Bcl-xL may conceal that cryptic sites open at least partially.

In ALR, where the conformational change upon pocket opening is highly localized, this can be clearly seen as sudden increase in ACV. The pocket opens after about 100 ns and remains open for the rest of the simulation time (Figure 2B). Interestingly, this opening was not observed for all concentrations of IPH (cf. ALR-IPH15 in Figure 2), which indicates the necessity to perform multiple, independent MD simulations. We observed a flip of the Φ backbone dihedral of Leu300 upon pocket opening (Figure S1), which is consistent with the opening mechanism described by other authors using cosolvent-free MD simulations at elevated temperatures, and structural changes observed in crystal structures.[55] This confirms that the observed opening of the pocket is not an artifact of our simulations, but resembles the expected mechanism. A selected snapshot from the ALR simulation (Figure 2C, highlighted in purple) furthermore shows that the overall structure remains intact after more than 400 ns of simulation time. The sampled conformation closely resembles the crystal structure with Leu300 flipped out of the cryptic site, and one of the phenol molecules mimics the position of the nitrofuran moiety of the crystal ligand. To conclude, based on our systematic evaluation across seven relevant and diverse test systems, we consider IPH10, i.e., a volume concentration of 10% phenol, as the best cosolvent composition to sample cryptic pockets for prospective identification, and thus applied it throughout the remaining simulations.

Unbiased Identification of Cryptic Sites

To prospectively identify cryptic sites, ten independent trajectories were simulated for each test system with and without the addition of 10% phenol as cosolvent. The radial distribution functions for oxygens of water (Figure S2A) are as expected for TIP3P water, and those for C1 atoms of phenol are below 1.5 for distances > 10 Å in all test systems (Figure S2B), which indicates the absence of large phenol aggregates or a phase separation under the chosen simulation conditions. The RMS average correlation (RAC)[56] was calculated for the combined trajectories to monitor convergence of the simulations (Figure S3). The RAC drops to about 1 Å and 0.5 Å for the simulations with and without cosolvent, respectively, after a lag time of about 2 μs for the combined simulations. The RAC curves show a steady decrease from this point onwards, indicating the absence of large conformational changes. The principal component analysis of heavy atom coordinates of pocket-lining atoms indicates that the sampled space covers the conformations of the crystal structures very well, except for HSP90 (Figures S4-S10; see also next section for discussion). Furthermore, visual inspection revealed an opening of the cryptic pockets on all test systems; only partial opening was observed for the most challenging test systems, Bcl-xL, β-Lac, and HSP90 (Table 1). For an unbiased identification, pocket grids (see Methods section for definition) were calculated throughout all simulations using PocketAnalyzer^{PCA}. Summary (eq. 3) and difference grids (eq. 5) (Figure 3A) were calculated accordingly. Difference grids are a means to focus the identification on sites that are more frequently open in the presence of phenol as cosolvent than in pure water. Subsequently, pocket cores were identified from the difference grids (Figure 3B, C). Finally, they are ranked according to their average occupancy (eq. 4, 5), which is approximately the fraction of simulation time that a pocket was more frequently open at this location when the system was simulated with cosolvent plus water than when simulated with water alone.

The largest number of pocket cores (seven) was identified in p38α (Figure 4A), which is also the largest protein in the test set (> 350 amino acids). Even with the large number of pockets identified, the known cryptic site was matched by the pocket core on the first rank (Table 1). Additional cryptic pocket cores were mainly found on the surface close to helix αC and between helices αD, αE and αF (Figure 4A, Table 1). Encouragingly, these additional pockets have also been described in another computational study on the dynamics of p38α.[57] Furthermore, two of these pockets are either part of a known protein-protein interface (docking groove/CD/ED pocket) or match the position of a fragment in a CDK2 crystal structure.[57] In contrast to p38α, only two pocket cores were found for ALR and MDM2 (Table 1); again, the known cryptic sites were successfully matched by the best ranked pocket cores in both cases.

Overall, for all test systems except β-Lac (discussed further below), pocket cores matching the location of the known cryptic sites were identified within the three top ranked pockets without tailoring settings for the different targets, with half of the identified cryptic cores ranked at the top. We consider this result a significant advancement over existing methods for cryptic pocket sampling and identification, as no *a priori* knowledge of pocket location, target-dependent preferable cosolvent or other parameters for pocket identification were used, in contrast to previously described methods.[15, 19]. The two parameter sets used for PocketAnalyzer^{PCA} reflect fundamental differences in pockets binding small molecules *versus* peptides, which in future studies allows one to tune cryptic pocket identification based on the type of ligand desired to bind at this location, consistent with previous structural observations comparing how proteins bind small molecules and other proteins or peptides differently at the same site[37].

Structural Assessment of Sampled and Identified Cryptic Pockets

In ALR, MDM2, and p38α, the known cryptic site was correctly matched by the highest ranking pocket core. The sampled conformations resembled the conformations observed in the crystal structures according to visual inspection, and were further validated by PCA of the pocket-lining atoms (Figures S4-S10) as well as by docking experiments (see next section).

In IL-2, the identified pocket core matches the cryptic site in the IL-2Rα binding site, which is blocked by Phe42 in the *apo* structure (Figure 5). The distribution of the χ_1 angle of Phe42, which has been described as a key metric for pocket opening[58], revealed a shift from about -60° (corresponding to the *apo* structure) to about -180° (corresponding to the *holo* structure) (Figure S11). Interestingly, the ratio of the height of the peaks at -180° and -60°, respectively, is about 1:2 (open:closed) in our simulations without cosolvent. This ratio is higher than in other MD[59] or mixed-solvent MD studies[54] of IL-2, which indicates that the cryptic site opens, although to a smaller extent, even in MD simulations without cosolvent. This explains the low average occupancy, which causes the pocket core to be ranked only third. Next to the IL-2Rα binding site, a second cryptic site has been described in IL-2,[60] which shows coupling to the first one.[59] The difference grid $D_{\text{IPH10-WAT}}$ clearly showed that also for this site the open conformation is sampled in the cosolvent simulations (Figure S12). However, this site is not considered a pocket core due to its proximity to the loop between helices B and C,[45] which is not resolved in the starting structure (PDB 1m47), and hence, this site is excluded (see section "Methods" for details on the identification of pocket cores). Thus, not identifying this second site does not reflect a weakness of our approach but a limitation of the starting structure.

In Bcl-xL, two pocket cores were identified matching the crystallographic ligand (Figure 5). The first one matches the trimethylsilyl moiety of the ligand and requires the displacement of Arg100 and Tyr195. Interestingly, the latter residue is only shifted in the used reference structure (Figure 5) but is completely displaced in other crystal structures (such as PDB ID 3sp7), thereby creating a larger void. Encouragingly, both conformations are sampled and selected using our approach. The second pocket core is located between helices α3 and α4, around Leu108. This binding interface is highly flexible and has been crystallized in different conformations (compare PDB IDs 4ehr, 3sp7, and 3zln). The different crystal structures have a shift of α3 in common, which increases the accessible volume around Leu108. Additionally, Tyr101 and Phe105 are displaced compared to the *apo* state. The latter transition especially involves changes of the secondary structure of Phe105 and adjacent residues (Figure 5). In our simulations, we observe the shift of α3, the displacement of Tyr101, and a shift, but not the full displacement including changes in the secondary structure, of Phe105. Furthermore, these transitions occur independently rather than in a concerted manner during our simulations.

Only a partial opening of the cryptic site is observed in our simulations for β-Lac and HSP90. In both cases, we identified a hydrogen bonding network in the *apo* structure that remains intact during our simulations, but is broken in the open state of the cryptic site in the known *holo* structure. We assume that this hydrogen bonding network locks the structures in the closed state. Our simulation conditions were selected to stabilize exposed hydrophobic patches in water rather than to disrupt electrostatic interactions. As a consequence, very long or enhanced-sampling simulations might be more suitable to sample the full exposure of such locked cryptic sites.[9, 17] Consequently, none of the pocket cores identified in β-Lac matched the expected cryptic site. However, another pocket core, ranked second by occupancy, is located in the vicinity of Ala232 (Figure 4B), which was not open in the *holo* structure used herein. The existence of this additional cryptic site has been proposed before by Bowman and coworkers.[8] Notably, our simulations are much shorter than the aggregated 81 µs used in that study. In a more recent crystal structure (PDB ID 5hw5) a Xenon atom binds at the entrance of the cryptic site identified by this pocket core (Figure 4B). The use of Xenon has been described as a method for the identification of hydrophobic pockets in crystallography[61] and, hence, indicates the existence of a flexible hydrophobic pocket. Therefore, despite the fact that the expected cryptic site could not be identified, our pocket cores revealed additional pockets, one of which has been experimentally validated.

The cryptic site of HSP90, which is a subpocket of the existing binding site, was not sampled in its completely open configuration, which becomes also evident from the PCA (Figure S7). Accordingly, it could not be located by a pocket core. However, our approach identified a pocket core, ranked second, matching the *existing* binding site. As the pocket core identification is based on the difference between cosolvent-containing and cosolvent-free simulations and the binding site in question is observable in the *apo* structure, this finding indicates that the binding site closes in cosolvent-free simulations. Interestingly, the shape of this pocket core overlaps with Phe138, indicating the existence of a cryptic subpocket around this residue. A closer inspection of the trajectories revealed three possible rotamer states of Phe138 (Figure S13A). When simulated in water alone, the existing pocket closes due to a side chain rotation of Phe138 (Figure S13A, B). Additionally, the binding site slightly contracts by an inwards movement of Phe138 and the following residues. In contrast, the rotamer observed in the crystal structure (open and closed) is the dominant species in the simulation with cosolvent (Figure S13D). Hence, in this case, the cosolvent stabilizes the existing pocket and prevents it from closing. A third rotamer is observed in IPH10, in which Phe138 is rotated away from the binding site, which gives access to a new cryptic subpocket between Phe138 and the β-sheet (Figure S13C), as indicated by the pocket core. Lastly, the highest ranked pocket core in HSP90 is located underneath the loop formed by the C-terminal end of helix α3 and the subsequent residues to Phe138, indicating a high flexibility and partial detachment of this loop (Figure S14). This lid and its structural dynamics have been described to play an important role in the activation cycle of HSP90.[62, 63]

In summary, we were able to generate fully or partially open conformations of the cryptic sites in all our test systems using a single simulation protocol. We consider the definition of distinct cryptic sites and the sampling of explicit binding site conformations important for the subsequent exploitation of newly identified pockets. Other methods for cryptic site prediction, such as CryptoSite[20], do not yield this (Figure S15), but are therefore computationally less demanding. In those cases where the cryptic site was formed only partially, the transition into the fully open state would require breaking hydrogen bond networks and/or changes in the secondary structure, which can require microsecond-long simulations. This is in agreement with other studies, where the transition from the folded to the unfolded state of helix α3 in HSP90 could not be observed in equilibrium MD simulations,[17] and the opening of the cryptic site in β-Lac was only observable in many microsecond-long simulations.[8, 16]

Automated Selection of Representative Protein Structures and Docking

After the identification of pocket cores, representative structures of the cryptic pockets were automatically selected by our workflow. We validated these structures with respect to their ability to bind the crystal structure ligand in the open cryptic site by computational docking. For the correctly identified cryptic site, the binding site was defined by residues within a fixed distance around the pocket cores, sampled conformations were clustered on these residues, and clusters with less than ~0.01% of the sampled conformations were excluded to focus on relevant conformations (see section "Methods"). Our approach yielded a single cluster for all test systems, except for ALR (two clusters) and p38α (four clusters) (Table 2). For each cluster, the cluster representative and structure with the largest pocket (with respect to the number of grid points of the pocket grid) were used for docking. Accordingly, at most eight structures (p38α), but predominantly only two structures, were selected for docking in a consistent manner.

The crystallized ligands could be re-docked into ALR and MDM2 with a minimum RMSD of 0.8 and 1.1 Å, respectively (Table 2), which is considerably better than the commonly accepted threshold of 2 Å for re-docking,[64] i.e., docking a ligand into its native receptor. The docking into ALR works remarkably well, despite the observation that His110, which forms a hydrogen bond with the carboxylic acid of the ligand in the crystal structure, flips during the simulations due to the absence of a binding partner and, thus, cannot form this interaction in the docked pose. In IL-2 and p38α, the crystallographic ligand pose was reproduced with an RMSD of 2.3 and 2.7 Å, respectively. Such RMSD values are still considered good for cross-docking experiments, i.e., docking a ligand into a non-native *holo* structure.[65] Particularly for p38α, the cryptic site is highly buried. As a consequence, it is not accessible in the closed state (RMSD > 13 Å, Table 2), while in the native structure the ligand pose can be well reproduced (best re-docking RMSD = 0.2 Å, Table S1). In that respect, the docking results clearly indicate the opening of the cryptic site. For HSP90 and Bcl-xL, the crystallographic ligand pose could not be approximated closer than an RMSD of 3.6 Å. This is likely due to the incomplete sampling of the open state of the cryptic site as already discussed. However, the re-docking of the ligands into their native structure resulted in equally large RMSD values (Table S1). In Bcl-xL, docking is likely challenged by the large size of the ligand containing 47 heteroatoms and 12 rotatable bonds. Encouragingly, a visual inspection shows that the docking pose resembles the overall binding mode well and, in particular, mimics the binding to the cryptic subpocket opened by Phe105 (Figure S16). It is important to stress that without first sampling and identifying the cryptic pockets, in none of the six test systems a satisfying docking solution was found, with the best ligand pose having an RMSD of 3.0 Å compared to the crystal structure (Table 2).

In conclusion, docking into only a few representative protein structures, selected in an automated way from the identified cryptic pocket core matching the cryptic site, enabled us to reproduce the crystallized ligand poses for five out of the six systems for which the cryptic site could be identified. This validates that our MD protocol is able to sample conformations resembling the open state, starting from the closed state crystal structure. We emphasize that a single protocol was used for all test systems, not only to sample and identify pocket cores, but also to define the binding site, cluster and select structures, and re-dock the ligands. In contrast to previous studies,[13, 15-17] no attempts were made to manually select the best cryptic pocket conformation from the MD ensemble, which highlights that our workflow for cryptic site identification and selection can extract relevant binding site conformations from the conformational ensemble.

5 Concluding Remarks

We have developed a new approach to sample and identify cryptic sites on protein structures in a manner that does not require prior information on the system, for use in structure-based drug design on novel protein targets. We systematically tested different cosolvents in equilibrium MD simulations and identified a volume concentration of 10% phenol (IPH10) as the best cosolvent composition to foster the opening of cryptic sites, while keeping the unfolding of proteins at a low probability. From the simulation data, we identified pockets by a grid-based approach and located regions where pockets are more frequently observable in the presence of cosolvent than in pure water. We refer to such regions as pocket cores. For the majority of test cases, one of the top-ranking pocket cores matched the location of the cryptic site. Hence, these pockets could have been identified in a prospective manner. Subsequently, we devised a workflow to select only a few (mostly two) representative structures for the pocket cores from the MD ensemble, and re-docked the known ligands that bind to the cryptic site. The crystallographic ligand poses could be well reproduced for five out of six test systems for which a pocket core was correctly identified.

The number of cryptic sites proposed by our approach is small (at most seven), which we consider an appropriate number of pockets for subsequent structure-based design approaches. In contrast, other methods have proposed up to 50 cryptic pockets for a single protein.[9] Several cryptic sites that we detected in addition to the crystallographic known pockets (p38α, β-Lac, HSP90, and IL-2) could be rationalized retrospectively based on findings from other studies. This demonstrates that our protocol is specific enough to yield only a small number of proposed cryptic sites, while being sufficiently sensitive to identify relevant ones nonetheless.

Our approach is challenged by sampling cryptic sites where the opening process involves time scales that are demanding for current unbiased MD simulations. This is true for β-Lac, HSP90, and Bcl-xL. However, even for these challenging targets, high-ranking pocket cores were identified in two cases that indicated the location of the cryptic sites. Other studies have succeeded in sampling such difficult cryptic sites, in part by using enhanced sampling techniques.[9, 16, 17] The proposed methods, however, require the *a priori* knowledge about the cryptic site location or the use of system-specific parameters. Still, integrating enhanced sampling techniques might further improve the sampling of cryptic sites by our method. Finding a consistent simulation protocol applicable to a wide range of targets, while keeping the specificity high, might be challenging, however.

In conclusion, our method to sample and identify cryptic sites does not rely on *a priori* knowledge about these sites and uses a target-independent parameterization, which renders it valuable for investigating novel protein targets and for *de novo* hit identification in small molecule ligand discovery.

6 Acknowledgements

We thank C. Pfleger and S.M.A. Hermans for fruitful discussions and critically reading the manuscript. Computational support and infrastructure was provided by the “Zentrum für Informations- und Medientechnologie” (ZIM) at Heinrich Heine University Düsseldorf. We are grateful for computing time provided by the John von Neumann Institute for Computing (NIC) to HG on the supercomputer JURECA at Jülich Supercomputing Centre (JSC) (user ID: HKF7).

7 Associated Content

Supplementary figures show the distribution of the Leu300 Φ backbone dihedral of ALR as a function of the simulation time for the first simulation with 10% phenol as cosolvent (Figure S1), the radial distribution function for C1 atoms of phenol, calculated for the simulations with 10% phenol as cosolvent (Figure S2), RMS average correlation (RAC) of the combined trajectories for the simulations without cosolvent and with 10% phenol as cosolvent (Figure S3), PC analyses of the sampled spaces for simulations in IPH10 (Figures S4-S10), the histogram of χ_1 dihedral for Phe42 in IL-2 for simulations without and with 10% phenol as cosolvent (Figure S11), sampling of the second cryptic site in IL-2 (Figure S12), conformations of Phe138 in HSP90 (Figure S13), the dynamics of the lid region of HSP90 in IPH10 (Figure S14), the results of an orthogonal approach of cryptic site identification (CryptoSite, Figure S15), a comparison of docked and crystallized configurations of the ligand in Bxl-xL (Figure S16), the comprehensive docking results for docking into the *apo*, *holo*, and the sampled conformations (Table S1) and the Python routines referred to in the methods (Listing S1 and S2).

The Supporting Information is available free of charge on the ACS Publications website.

8 Funding

This study was supported by funds from Pfizer, Inc., and, in part, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 270650915 (Research Training Group GRK 2158, TP4a to HG). Financial support by DFG for funds (INST 208/704-1 FUGG) to purchase the hybrid computer cluster used in this study is gratefully acknowledged.

9 Notes

The authors declare the following competing interest(s): M. Boehm, C. L. McClendon, and R. Torella are employees of Pfizer, Inc.

10 References

1. Hopkins, A. L.; Groom, C. R., The druggable genome. *Nat. Rev. Drug Discov.* **2002,** *1* (9), 727-730.

2. Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M., Protein clefts in molecular recognition and function. *Protein Sci.* **1996,** *5* (12), 2438-2452.

3. Hardy, J. A.; Wells, J. A., Searching for new allosteric sites in enzymes. *Curr. Opin. Struct. Biol.* **2004,** *14* (6), 706-715.

4. Frembgen-Kesner, T.; Elcock, A. H., Computational sampling of a cryptic drug binding site in a protein receptor: Explicit solvent molecular dynamics and inhibitor docking to p38 MAP kinase. *J. Mol. Biol.* **2006,** *359* (1), 202-214.

5. Lu, S.; Ji, M.; Ni, D.; Zhang, J., Discovery of hidden allosteric sites as novel targets for allosteric drug design. *Drug Discovery Today* **2018,** *23* (2), 359-365.

6. Vajda, S.; Beglov, D.; Wakefield, A. E.; Egbert, M.; Whitty, A., Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.* **2018,** *44*, 1-8.

7. Eyrisch, S.; Helms, V., Transient pockets on protein surfaces involved in protein−protein interaction. *J. Med. Chem.* **2007,** *50* (15), 3457-3464.

8. Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S., Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2015,** *112* (9), 2734-2739.

9. Bowman, G. R.; Geissler, P. L., Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl. Acad. Sci. U.S.A.* **2012,** *109* (29), 11681-11686.

10. Seco, J.; Luque, F. J.; Barril, X., Binding site detection and druggability index from first principles. *J. Med. Chem.* **2009,** *52* (8), 2363-2371.

11. Basse, N.; Kaar, J. L.; Settanni, G.; Joerger, A. C.; Rutherford, T. J.; Fersht, A. R., Toward the rational design of p53-stabilizing drugs: Probing the surface of the oncogenic Y220C mutant. *Chem. Biol.* **2010,** *17* (1), 46-56.

12. Tan, Y. S.; Śledź, P.; Lang, S.; Stubbs, C. J.; Spring, D. R.; Abell, C.; Best, R. B., Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of Polo-Like kinase 1. *Angew. Chem. Int. Ed.* **2012,** *51* (40), 10078-10081.

13. Tan, Y. S.; Spring, D. R.; Abell, C.; Verma, C., The use of chlorobenzene as a probe molecule in molecular dynamics simulations. *J. Chem. Inf. Model.* **2014,** *54* (7), 1821-1827.

14. Kalenkiewicz, A.; Grant, B.; Yang, C.-Y., Enrichment of druggable conformations from apo protein structures using cosolvent-accelerated molecular dynamics. *Biology* **2015,** *4* (2), 344-366.

15. Kimura, S. R.; Hu, H. P.; Ruvinsky, A. M.; Sherman, W.; Favia, A. D., Deciphering cryptic binding sites on proteins by mixed-solvent molecular dynamics. *J. Chem. Inf. Model.* **2017,** *57* (6), 1388-1401.

16. Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L., Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *J. Am. Chem. Soc.* **2016,** *138*, 14257-14263.

17. Kokh, D. B.; Czodrowski, P.; Rippmann, F.; Wade, R. C., Perturbation approaches for exploring protein binding site flexibility to predict transient binding pockets. *J. Chem. Theory Comput.* **2016,** *12*, 4100-4113.

18. Beglov, D.; Hall, D.; Wakefield, A.; Luo, L.; Allen, K.; Kozakov, D.; Whitty, A.; Vajda, S., Exploring the structural origins of cryptic sites on proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2018,** *115* (15), E3416-E3425.

19. Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C., TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins. *J. Chem. Inf. Model.* **2013,** *53* (5), 1235-1252.

20. Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N. A.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A., CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* **2016,** *428* (4), 709-719.

21. Ghanakota, P.; Carlson, H. A., Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J. Med. Chem.* **2016,** *59* (23), 10383-10399.

22. Berman, H. M.; Westbrook, J.; Zukang, F.; Gillliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28*, 235-242.

23. *Schrödinger Release 2016-2: Protein Preparation Wizard*. Schrödinger, LLC: New York, NY, 2016.

24. Sastry, M. G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013,** *27*, 221-234.

25. *Schrödinger Release 2016-2: Prime*. Schrödinger, LLC: New York, NY, 2016.

26. Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B., On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002,** *320*, 597-608.

27. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A., A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.* **2004,** *55*, 351-367.

28. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005,** *26*, 1668-1688.

29. Pfleger, C.; Minges, A.; Boehm, M.; McClendon, C. L.; Torella, R.; Gohlke, H., Ensemble- and rigidity theory-based perturbation approach to analyze dynamic allostery. *J. Chem. Theory Comput.* **2017,** *13*, 6343-6357.

30. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the accuracy of protein side-chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015,** *11*, 3696-3713.

31. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25*, 1157-1174.

32. Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M., PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009,** *30*, 2157-2164.

33. 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.

34. Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C., Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013,** *9*, 3878-3888.

35. Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K., Pocket-space maps to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.* **2011,** *51*, 2666-2679.

36. Metz, A.; Pfleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K.-H.; Gohlke, H., Hot spots and transient pockets: Predicting the determinants of small-molecule binding to a protein–protein interface. *J. Chem. Inf. Model.* **2012,** *52* (1), 120-133.

37. Wells, J. A.; McClendon, C. L., Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007,** *450* (7172), 1001-1009.

38. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004,** *25*, 1605-1612.

39. Roe, D. R.; Cheatham, T. E., PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **2013,** *9* (7), 3084-3095.

40. Tversky, A., Features of similarity. *Psychol. Rev.* **1977,** *84*, 327-352.

41. *Schrödinger Release 2016-2: Glide*. Schrödinger, LLC: New York, NY, 2016.

42. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004,** *47*, 1750-1759.

43. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein−ligand complexes. *J. Med. Chem.* **2006,** *49*, 6177-6196.

44. Sotriffer, C. A.; Krämer, O.; Klebe, G., Probing flexibility and "induced-fit" phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations. *Proteins: Struct., Funct., Genet.* **2004,** *56* (1), 52-66.

45. Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C., Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl. Acad. Sci. U.S.A.* **2003,** *100*, 1603-1608.

46. Stank, A.; Kokh, D. B.; Fuller, J. C.; Wade, R. C., Protein binding pocket dynamics. *Acc. Chem. Res.* **2016,** *49*, 809-815.

47. Uhrinova, S.; Uhrin, D.; Powers, H.; Watt, K.; Zheleva, D.; Fischer, P.; McInnes, C.; Barlow, P. N., Structure of free MDM2 N-terminal domain reveals conformational adjustments that accompany p53-binding. *J. Mol. Biol.* **2005,** *350* (3), 587-598.

48. Diskin, R.; Engelberg, D.; Livnah, O., A novel lipid binding site formed by the MAP kinase insert in p38α. *J. Mol. Biol.* **2008,** *375* (1), 70-79.

49. Horn, J. R.; Shoichet, B. K., Allosteric inhibition through core disruption. *J. Mol. Biol.* **2004,** *336* (5), 1283-1291.

50. Wright, L.; Barril, X.; Dymock, B.; Sheridan, L.; Surgenor, A.; Beswick, M.; Drysdale, M.; Collier, A.; Massey, A.; Davies, N.; Fink, A.; Fromont, C.; Aherne, W.; Boxall, K.; Sharp, S.; Workman, P.; Hubbard, R. E., Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem. Biol.* **2004,** *11* (6), 775-785.

51. Liu, X.; Dai, S.; Zhu, Y.; Marrack, P.; Kappler, J. W., The structure of a Bcl-xL/Bim fragment complex. *Immunity* **2003,** *19* (3), 341-352.

52. Schroeder, G. M.; Wei, D.; Banfi, P.; Cai, Z.-W.; Lippy, J.; Menichincheri, M.; Modugno, M.; Naglich, J.; Penhallow, B.; Perez, H. L.; Sack, J.; Schmidt, R. J.; Tebben, A.; Yan, C.; Zhang, L.; Galvani, A.; Lombardo, L. J.; Borzilleri, R. M., Pyrazole and pyrimidine phenylacylsulfonamides as dual Bcl-2/Bcl-xL antagonists. *Bioorg. Med. Chem. Lett.* **2012,** *22* (12), 3951-3956.

53. Graham, S. E.; Leja, N.; Carlson, H. A., MixMD Probeview: Robust binding site Prediction from cosolvent simulations. *J. Chem. Inf. Model.* **2018,** *58* (7), 1426-1433.

54. 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.

55. Rechlin, C.; Scheer, F.; Terwesten, F.; Wulsdorf, T.; Pol, E.; Fridh, V.; Toth, P.; Diederich, W. E.; Heine, A.; Klebe, G., Price for opening the transient specificity pocket in human Aldose Reductase upon ligand binding: Structural, thermodynamic, kinetic, and computational analysis. *ACS Chem. Biol.* **2017,** *12* (5), 1397-1415.

56. Galindo-Murillo, R.; Roe, D. R.; Cheatham, T. E., Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta, Gen. Subj.* **2015,** *1850* (5), 1041-1058.

57. Gomez-Gutierrez, P.; Rubio-Martinez, J.; Perez, J. J., Identification of potential small-molecule binding pockets in p38α MAP kinase. *J. Chem. Inf. Model.* **2017,** *57* (10), 2566-2574.
58. Thanos, C. D.; Randal, M.; Wells, J. A., Potent small-molecule binding to a dynamic hot spot on IL-2. *J. Am. Chem. Soc.* **2003,** *125*, 15280-15281.
59. McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P., Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput.* **2009,** *5*, 2486-2502.
60. Hyde, J.; Braisted, A. C.; Randal, M.; Arkin, M. R., Discovery and characterization of cooperative ligand binding in the adaptive region of Interleukin-2. *Biochemistry* **2003,** *42*, 6475-6483.
61. Prangé, T.; Schiltz, M.; Pernot, L.; Colloc'h, N.; Longhi, S.; Bourguet, W.; Fourme, R., Exploring hydrophobic sites in proteins with xenon or krypton. *Proteins: Struct., Funct., Genet.* **1998,** *30*, 61-73.
62. Krukenberg, K. A.; Street, T. O.; Lavery, L. A.; Agard, D. A., Conformational dynamics of the molecular chaperone Hsp90. *Q. Rev. Biophys.* **2011,** *44*, 229-255.
63. Zhang, H.; Zhou, C.; Chen, W.; Xu, Y.; Shi, Y.; Wen, Y.; Zhang, N., A dynamic view of ATP-coupled functioning cycle of Hsp90 N-terminal domain. *Sci. Rep.* **2015,** *5* (1), 9542-9542.
64. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000,** *295* (2), 337-356.
65. Cavasotto, C. N.; Abagyan, R. A., Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004,** *337* (1), 209-225.

11 Figure Captions

Figure 1. Overlay of *apo* and *holo* structures of the seven protein systems (Aldose Reductase (ALR), Interleukin-2, MDM2, p38α, β-lactamase, HSP90, Bcl-xL) used in this study. Proteins are shown in cartoon representation in gray. Residues around the cryptic site are highlighted by color (blue: *apo* structure, orange: *holo* structure). The ligand binding to the cryptic binding site is shown in stick representation. The spectrum of conformational changes upon formation of the cryptic site ranges from single or few side chain rearrangements (ALR, Interleukin-2) to small shifts of secondary structure elements (MDM2, p38α, β-lactamase) to formation of secondary structure elements (HSP90) and large-scale deformation of the binding site (Bcl-xL).

Figure 2. Cosolvent mixture screens identify the average best sampling conditions. (A) Barplot quantifying the effect of the used solvent combinations (columns) on the opening of cryptic sites for the seven test systems (rows). Colors of the bars represent the solvents (blue: water, green: ETA, red: IPA, purple: IPH) and the concentrations (light shade: 5% to dark shade: 20%). The height of a bar indicates the fraction of frames of a single trajectory with an open cryptic site (accessible cryptic volume (eq. 2) $ACV \geq 1$). Triangles mark simulations where partial unfolding was observed, usually starting by the segregation of α-helices from the rest of the protein. The test systems are varyingly susceptible to the opening of their cryptic sites. On average over the seven test systems (last row), IPH10 has a significantly stronger effect than water and the highest effect of all solvents tested. (B) ACV of cryptic site in Aldose Reductase simulated in IPH10 as a function of the simulation time. The sampled cryptic site volume is scaled to the volume of the cryptic site in the reference crystal structure. After about 100 ns, the cryptic site opens and remains open for the rest of the simulation. (C) Snapshot of the simulation of Aldose Reductase in IPH10 ($t \approx 426$ ns, highlighted in purple) in comparison to its crystal structure. Orientation of ALR as in Figure 1. The cryptic site is open and the binding site conformation closely resembles the one of the crystal structure. The IPH probe molecule matches the position of the nitrofuran moiety of the crystallized ligand.

Figure 3. Workflow for the identification of cryptic sites. (A) Summary grids (eq. 3) are calculated for the simulations with (S_{IPH10}) and without (S_{WAT}) IPH as cosolvent, depicted by a purple and blue protein, respectively. Probabilities to find a pocket at each grid point (occupancies, eq. 4) are depicted by color, ranging from white (low) to red (high). The difference of these grids, $\boldsymbol{D_{IPH10-WAT}}$ (eq. 5), indicates regions in which pockets occur more frequently in the presence of cosolvent. Pocket cores are calculated from the difference grid as

depicted in panel B. Snapshots are selected from the cosolvent simulation if their pockets match the identified pocket cores. Lastly, the extracted snapshots are clustered based on the RMSD of the binding site residues. (B) Pocket cores are calculated from the difference grids by successively removing grid points with increasing occupancy ("depletion") and identifying connected groups of grid points subject to additional constraints, such as size (see section "Methods" for details). (C) Results of the workflow applied to ALR. Protein structures are represented as in Figure 1, unless otherwise noted. Left panel: Summary grid for cosolvent simulations (S_{IPH10}), shown with the closed-state crystal structure of ALR. For clarity, only grid points in the vicinity to the ligand in the reference crystal structure (PDB ID 4prr, not displayed) are shown. Grid points are shown as spheres. The occupancy of each grid point is indicated by color and sphere scale from zero (white/small) to 0.8 (red/large). The occupancy of grid points decreases from the core of the pocket to the outside. Middle panel: Pocket core calculated from the difference grid. Grid points that are part of the pocket core are shown as spheres with semi-transparent surface. Right panel: *Holo* structure of ALR (carbon atoms in orange) with bound ligand compared to the cluster representative of the highest occupied cluster (carbon atoms in gray). Only residues in the binding site, as defined by distance to the pocket core, are shown in stick representation and labeled.

Figure 4. Pocket cores are identified for selected test systems. (A) Closed-state structure of p38α shown as gray ribbon and transparent surface with the N- and C-terminal domains oriented to the top and bottom, respectively. The crystallized ATP-competitive inhibitor is shown in stick representation and highlighted by a blue arrow. The identified cryptic cores are shown as groups of spheres. The average occupancy of each pocket is indicated by color, and ranges from zero (white) to 0.35 (red), and the sphere size, which ranges from small to large. The cryptic core with the highest occupancy (red) matches the cryptic binding site. The ligand binding to the cryptic site is shown in orange as semi-transparent surface. Its position was determined by alignment with the open-state crystal structure. Notably, no pocket core was identified in the ATP-binding site, due to the reference simulations in water. (B) Crystal structure of β-Lac (gray ribbon) with bound Xe atoms (yellow spheres) (PDB ID 5hw5). The position of Xenon atom Xe302 marks the entrance of a new cryptic site identified by a pocket core. The pocket core is represented by spheres with semi-transparent surface (color scale as in panel A). Xe303 binds to the expected cryptic site. The ligand binding to the cryptic site is shown as an orange, semi-transparent surface. Other highlighted residues and secondary structure elements are discussed in the main text.

Figure 5. Correctly identified cryptic pockets. Protein representations as in Figure 1. Pocket cores matching the known cryptic sites are indicated by semi-transparent surfaces. The average occupancy of the pocket cores is indicated by surface color from zero (white) to 0.5 (red). Highlighted residues are discussed in the main text. Residue F105 in Bcl-xL is indicated by "*" and labeled outside the protein structure for clarity.

12 Tables

Table 1: Results of cryptic site identification.

Target	Observed opening of cryptic site	Number of cryptic cores identified	Rank of cryptic pocket[a]
ALR	Yes	2	1
β-Lac	Partially	4	---
Bcl-xL	Partially	3	(2/3)[b]
IL-2	Yes	5	3
MDM2	Yes	2	1
p38α	Yes	7	1
HSP90	Partially	4	(2)[a]

[a] Ranked according to the average occupancy of a pocket core.

[b] Brackets indicate results that are based on a partial opening of the binding site.

Table 2: Results of docking experiments to validated selected cryptic site conformations.

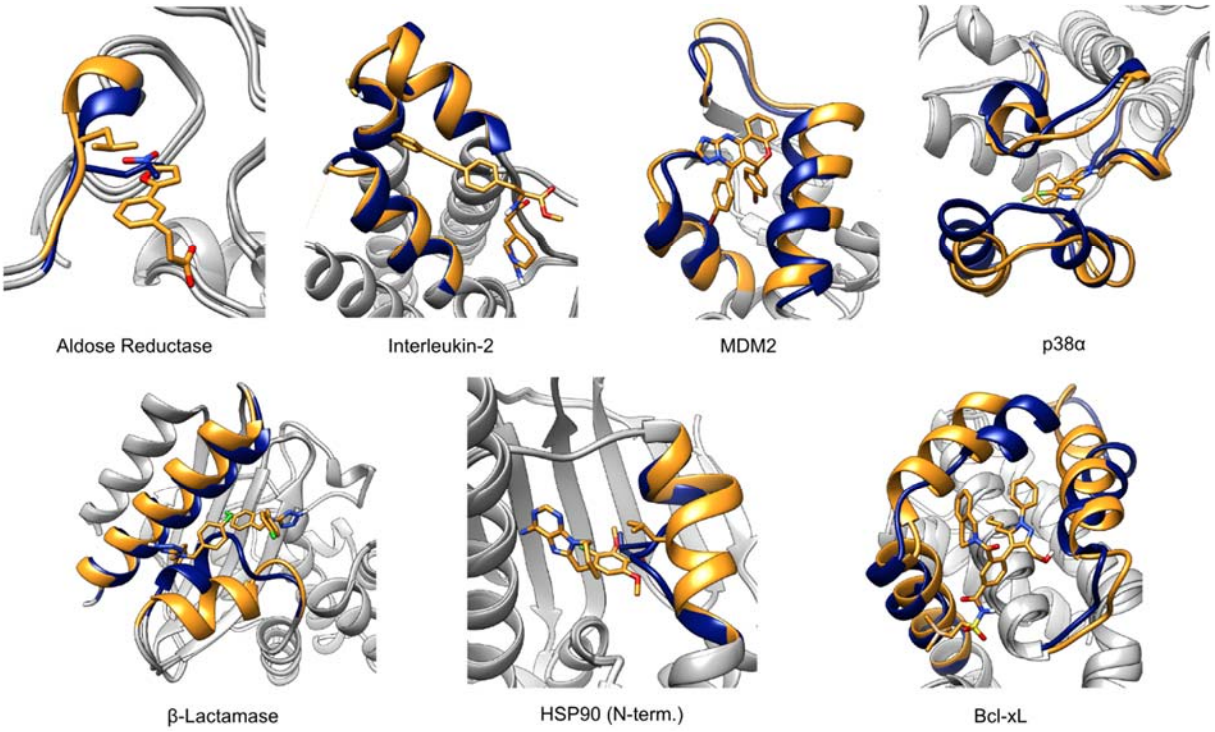
Target	Number of clusters	RMSD - simulation[a]	RMSD - crystal structure[a]
ALR	2	0.8	4.5
Bcl-xL[b]	1	3.6	4.4
IL-2	1	2.3	3.0
MDM2	1	1.1	5.4
p38α	4	2.7	13.3
HSP90	1	3.6	4.9

[a] Best RMSD of docking pose for docking into structures selected from MD simulations or the crystal structure.

[b] Only the third pocket core, located around Leu108, was used for comparability.

13 Figures

Figure 1



Aldose Reductase

Interleukin-2

MDM2

p38α

β-Lactamase

HSP90 (N-term.)

Bcl-xL

Figure 2

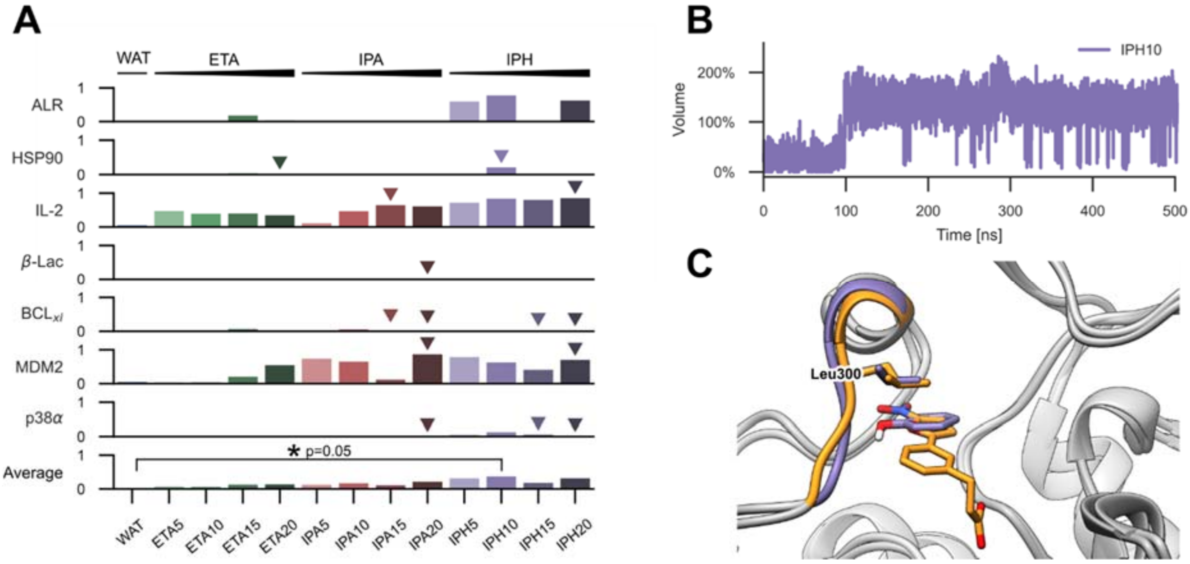

A
WAT
ETA
IPA
IPH
ALR
HSP90
IL-2
β-Lac
BCL$_{xl}$
MDM2
p38α
Average
1
0
* p=0.05
WAT
ETA5
ETA10
ETA15
ETA20
IPA5
IPA10
IPA15
IPA20
IPH5
IPH10
IPH15
IPH20
B
IPH10
Volume
200%
100%
0%
0
100
200
300
400
500
Time [ns]
C
Leu300

Figure 3

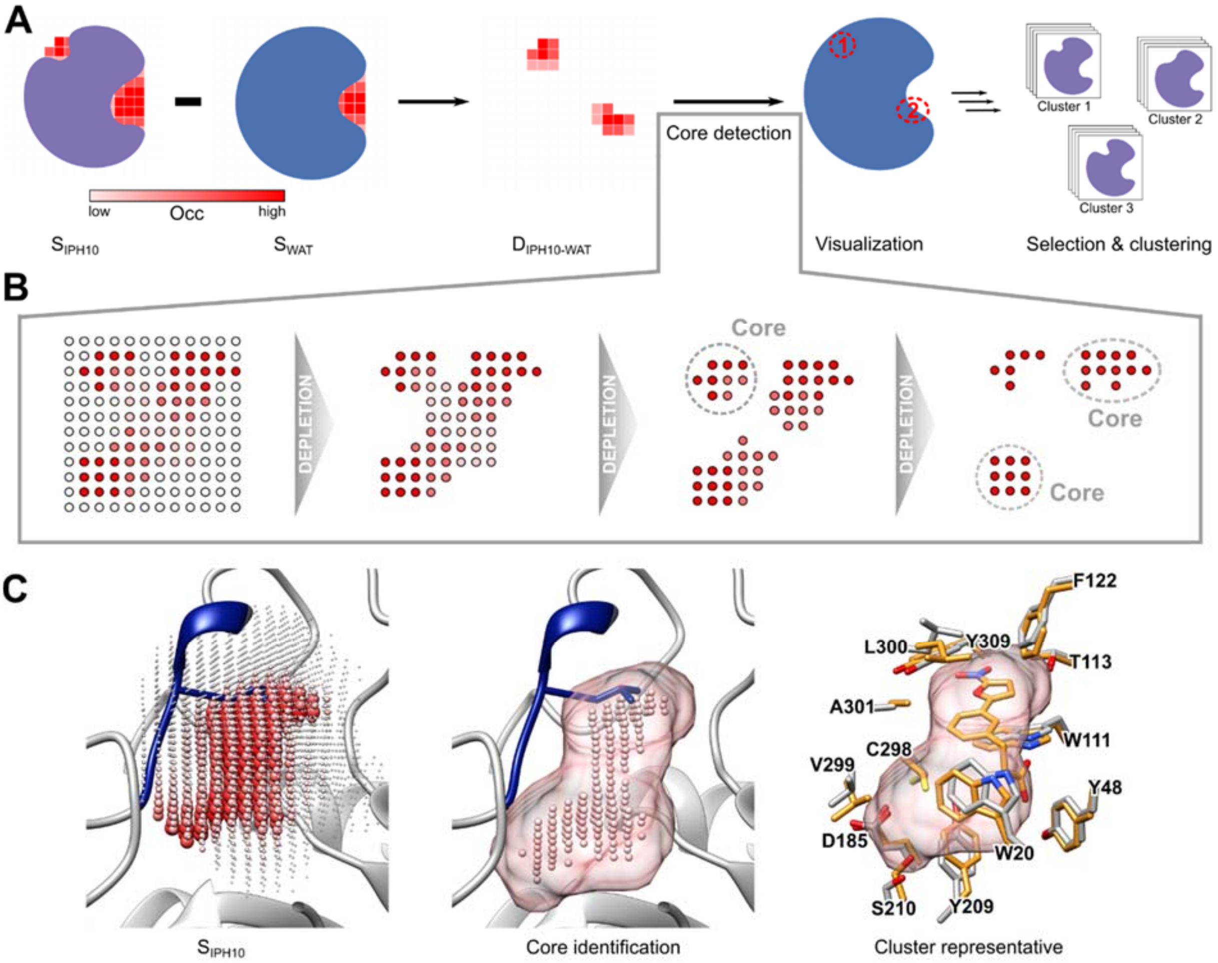

A
low
Occ
high
S_{IPH10}
S_{WAT}
$D_{IPH10-WAT}$
Core detection
1
2
Visualization
Cluster 1
Cluster 2
Cluster 3
Selection & clustering
B
DEPLETION
DEPLETION
Core
DEPLETION
Core
Core
C
F122
L300
Y309
T113
A301
W111
C298
V299
Y48
D185
W20
S210
Y209
S_{IPH10}
Core identification
Cluster representative

Figure 4

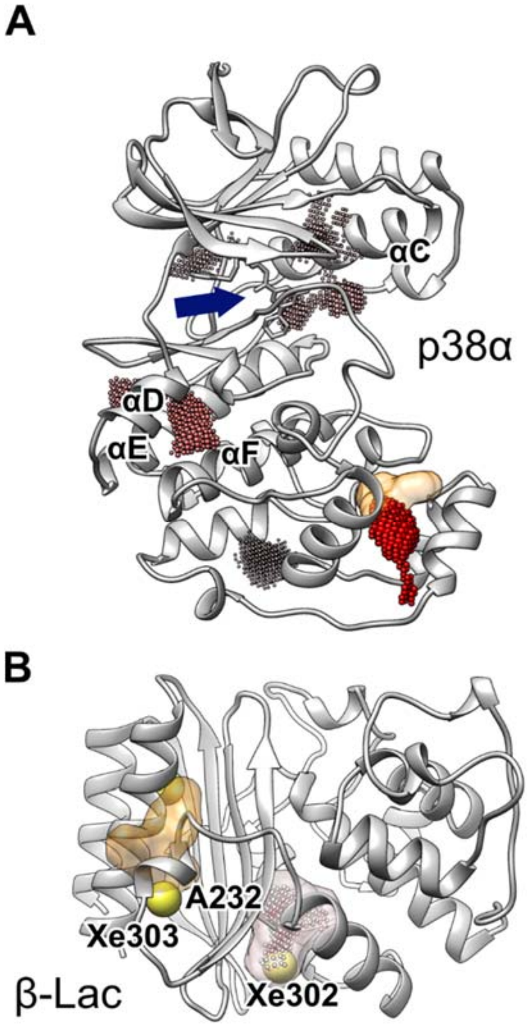
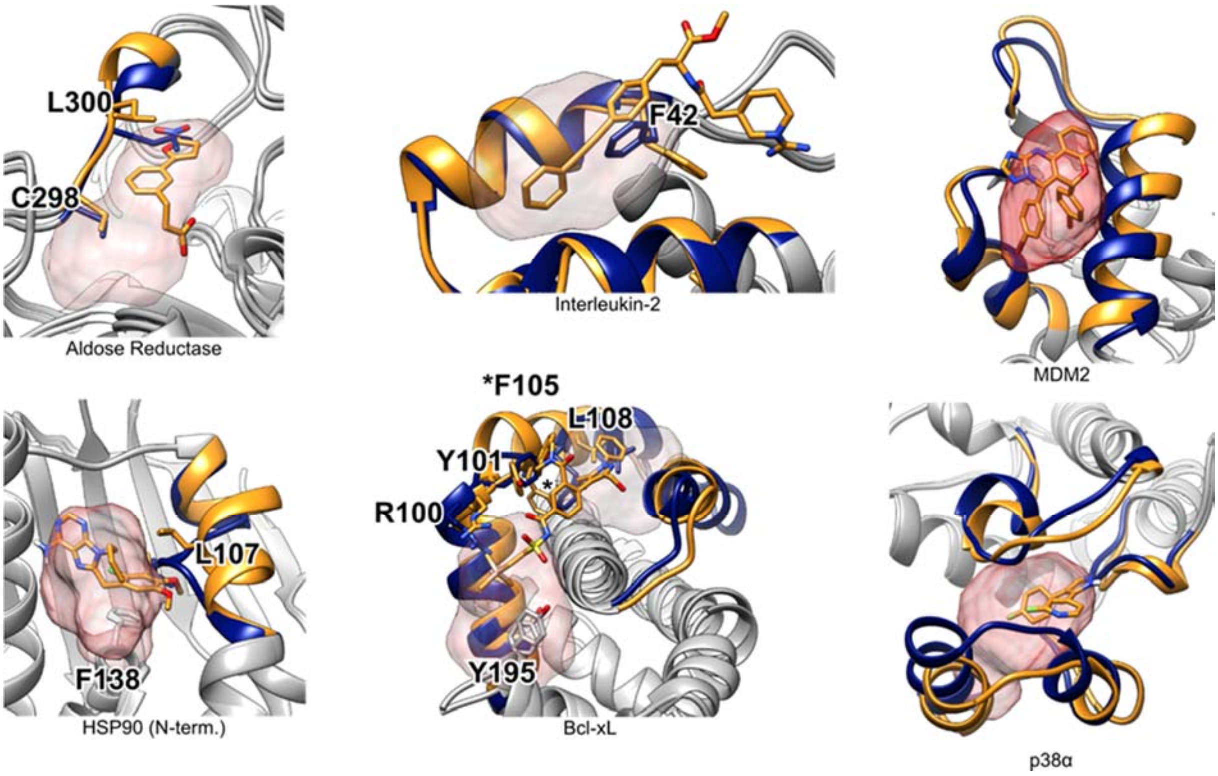


Figure 5


L300
C298
Aldose Reductase
F42
Interleukin-2
MDM2
L107
F138
HSP90 (N-term.)
*F105
L108
Y101
R100
*
Y195
Bcl-xL
p38α

TOC graphic