Published in final edited form as:
J Chem Inf Model. 2011 April 25; 51(4): 877–896. doi:10.1021/ci100462t.

Reproducing Crystal Binding Modes of Ligand Functional Groups using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations

E. Prabhu Raman[1], **Wenbo Yu**[1], **Olgun Guvench**[2], and **Alexander D. MacKerell Jr.**[1,*]

[1] Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, 20 Penn Street HSF II, Baltimore MD 21201

[2] Department of Pharmaceutical Sciences, University of New England College of Pharmacy, 716 Stevens Ave, Portland ME 04103

Abstract

The applicability of a computational method, Site Identification by Ligand Competitive Saturation (SILCS), to identify regions on a protein surface with which different types of functional groups on low-molecular weight inhibitors interact is demonstrated. The method involves molecular dynamics (MD) simulations of a protein in an aqueous solution of chemically diverse small molecules from which probability distributions of fragments types, termed FragMaps, are obtained. In the present application, SILCS simulations are performed with an aqueous solution of 1 M benzene and propane to map the affinity pattern of the protein for aromatic and aliphatic functional groups. In addition, water hydrogen and oxygen atoms serve as probes for hydrogen bond donor and acceptor affinity, respectively. The method is tested using a set of 7 proteins for which crystal structures of complexes with several high affinity inhibitors are known. Good agreement is obtained between FragMaps and the positions of chemically similar functional groups in inhibitors as observed in the X-ray crystallographic structures. Quantitative capabilities of the SILCS approach are demonstrated by converting FragMaps to free energies, termed Grid Free Energies (GFE), and showing correlation between the GFE values and experimental binding affinities. For proteins for which ligand decoy sets are available, GFE values are shown to typically score the crystal conformation and conformations similar to it more favorable than decoys. Additionally, SILCS is tested for its ability to capture the subtle differences in ligand affinity across homologous proteins, information which may be of utility towards specificity-guided drug design. Taken together, our results show that SILCS can recapitulate the known location of functional groups of bound inhibitors for a number of proteins, suggesting that the method may be of utility for rational drug design.

INTRODUCTION

Structure-based drug design (SBDD), which involves the discovery and optimization of small molecules that bind to a target with known 3D structure, is of central relevance to medicinal chemistry. Computational methods have important roles to play in SBDD from early to late phases in a typical drug discovery project[1] resulting in time and cost savings.

[*] Corresponding author: Alexander D. MacKerell, Jr., Room 629, HSF II, 20 Penn Street, Baltimore, MD 21201, Tel: 410-706-7442, Fax: 410-706-5017, alex@outerbanks.umaryland.edu.

Supporting Information Available

Supporting information contains the Figures and Tables referred to in the paper. Chemical structures, annotation and crystal parameters relevant to all ligands analyzed are also presented. This information is available free of charge via the Internet at http://pubs.acs.org/.

Docking followed by scoring of millions of compounds may be used to identify novel lead compounds in the early stages of SBDD campaigns. However, due to approximations used in the scoring functions as required for computational feasibility, a high population of false positives (or false hits) is typically identified. More importantly, only a limited range of chemical space is represented in the seemingly large libraries used in virtual and high-throughput screenings (HTS).[2] On the other end, there exist highly accurate, albeit computationally expensive methods such as free energy perturbation (FEP) that have been used in SBDD.[3, 4] Since such calculations require an already known binding mode of a lead compound, they are of more utility in later stages of SBDD such as lead optimization.

Fragment-based drug discovery (FBDD) is a promising method for identifying novel high affinity small molecule binders.[5] The basic idea behind FBDD is to identify small molecule fragments that bind to specific sites on the protein and the subsequent optimization or linking of the fragments to create a higher affinity binder. In order to detect fragment binding in FBDD, NMR spectroscopy or X-ray crystallography are among the experimental approaches usually employed. [2] Atomic-detail information from these methods has been successfully combined with fragment evolution yielding lead candidates in numerous cases.[6, 7] However, such methods require significant time and resources and are limited by the requirement that fragment affinities be in the millimolar range or better[8], thereby limiting the use of simpler and smaller fragments in the discovery process. Another problem associated with X-ray crystallography is overemphasis on a single conformation as opposed to a conformational distribution.

Using a theoretical model, Hann et. al.[9] predicted that less complex (or small) ligands are likely to have multiple binding sites on a protein. This is due to their small size such that the chance of "mismatch" interactions that would decrease binding is less leading to non-unique binding sites. While weak binding affinity and the resulting non-unique binding modes may be experimentally problematic to detect, this is not the case in *in-silico* methods.[10–17] SILCS (Site Identification by Ligand Competitive Saturation) is such an *in-silico* method[18] that involves molecular dynamics (MD) simulation of the target protein in the presence of an aqueous solution of organic solutes to obtain their probability of binding to different sites on the protein. In this study, we test the ability of the SILCS method to capture the location of known functional group binding sites on a protein surface based on experimental binding modes of larger, more complex ligands. Data from the SILCS method is based on 3D probability distributions of the molecular fragments, called FragMaps, on the protein surface. In the present study we show that a minimalistic choice of fragments – benzene, propane and water – still captures to a reasonable extent the crystal binding modes of larger molecules of high affinity. We then extend the SILCS method to allow for quantitation of ligand relative affinities by converting FragMaps to free energies, termed Grid Free Energies (GFE), based on simple inverse-Boltzmann weighting of the FragMap and the bulk solution probabilities. Using this metric it is shown that the SILCS method can be sensitive to differences in structures seen across homologous proteins and that quantitative predictions based on the SILCS method correlate with experimental binding affinities. Finally, the limitations of the method in its current implementation and potential for improvement are discussed.

METHODS

Protein set selection

Since one of the goals of the study is to validate the generality of the SILCS methodology with regard to identifying binding sites of ligand functional groups on diverse protein surfaces, seven proteins from five families were chosen. Additionally, each of the chosen proteins has between two to seven crystal structures of its complexes with different ligands.

Effort was also made to select proteins for which experimental binding affinity data exist. Proteins were chosen from several sources in the literature, many of which were accessed from the Ligand-Protein Database (LPDB)[19]. These proteins, along with their PDB identifiers, resolution of the crystals and additional information is presented in Table 1. Additionally, for the protein-ligand complex crystal structures used in this study, Table S9, Supporting Information, provides the PDB ID, Ligand ID, Chain and the SMILES string for the ligands and Table S10 provides the resolution and other crystallographic parameters of those crystal structures.

SILCS MD simulations

All simulations were performed using the CHARMM molecular simulation program[20] the CHARMM protein force field[21] with CMAP backbone correction[22], and the TIP3P water model[23]. Where available, the crystal structure of the apo (unliganded) form of the protein was obtained from the PDB[24]. For cases for which the apo form was not available, protein coordinates from one of the protein-ligand complex crystal structures were used following deletion of the crystallographic ligand. For HIV-protease, SILCS simulations were performed starting with both forms of the protein. Crystal water molecules were retained, as were any structurally important ions. The Reduce software[25] was used to place missing hydrogen positions and to choose optimal Asn and Gln sidechain amide and His sidechain ring orientations. Benzene and propane molecules ("fragments") were placed on a square grid by randomly choosing between the two molecules at each grid point. For each protein, ten such grids were generated with grid spacing selected to yield a concentration of ~ 1M benzene and ~ 1M propane when combined with a box of water molecules at the experimental density of water. Ten protein+fragments+water systems were generated by overlaying the crystal coordinates of the protein and water molecules with each of the ten different solutions and removing all fragment and water molecules with any atom within 2 Å of any protein atom. The net system charge was made neutral by replacing random water molecules with the appropriate number of sodium or chloride ions.

The system in the presence of periodic boundary conditions[26] was minimized for 500 steps with the steepest descent algorithm[27] while employing harmonic positional restraints having a force constant of 1 kcal*mol^{-1}Å^{-2} per atomic mass unit on protein non-hydrogen atoms. The "leap frog" version of the Verlet integrator[26] with a time step of 2 fs was used for molecular dynamics (MD) simulations. Water geometries and bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[28] Long-range electrostatic interactions were handled with the particle-mesh Ewald method[29] with a real space cutoff of 8 Å, a switching function[30] was applied to the Lennard-Jones interactions in the range of 5 to 8 Å, and a long-range isotropic correction[26] was applied to the pressure for Lennard-Jones interactions beyond the 8 Å cutoff length. Following minimization, with the same positional restraints, the system was heated to 298 K by periodic reassignment of velocities[31] over 20 ps, followed by an equilibration for 20 ps using velocity reassignment. In the production simulations that followed, the positional restraints were removed and replaced by very weak restraints on only the protein backbone Cα atoms with a force constant of 0.01 kcal*mol^{-1}A^{-2} per atomic mass unit to prevent the rotation of the protein in the simulation box. Each of the ten systems were simulated for 20 ns at 298 K and 1 atm pressure with a Nosé-Hoover thermostat[32, 33] and the Langevin piston barostat[34], thus yielding a cumulative sampling time of 200 ns for each SILCS simulation. Snapshots were output every 10 ps as more frequent saving of snapshots did not change the results (not shown).

As detailed previously[18], to prevent the aggregation of hydrophobic fragments and to promote faster convergence, a repulsive interaction energy term was introduced only between benzene:benzene, propane:propane and benzene:propane molecular pairs. This was achieved by adding an additional massless particle to the center of mass of benzene and to the central carbon of propane molecules. Each such particle does not interact with any other atoms in the system but with other particles on the hydrophobic molecules through the Lennard-Jones (LJ) force field term[26] with parameters ($\varepsilon = -0.01$ kcal/mol; $R_{min} = 12.0$ Å). Combined with the same switching function and cutoffs used for other LJ interactions, this leads to a purely repulsive energy vs. distance profile for hydrophobic molecular pairs[18]. Since the repulsive interaction energy approaches zero as the separation distance increases to the cutoff length of 8Å, the additional term is not expected to perturb small molecule:protein interactions in different sites on the protein surface.

FragMap construction

3D probability distributions of the fragment atoms at various sites called "FragMaps" were constructed for four atom types – benzene carbons, propane carbons, water hydrogens and water oxygens – from the SILCS simulation trajectories using the following protocol. Atoms from snapshots output every 10 ps from the last 5 ns of each SILCS simulation trajectory were binned into 1 Å X 1 Å X 1 Å cubic volume elements (voxels) of a grid spanning the entire system, and the voxel occupancy for each FragMap atom type was thus calculated. For benzene and propane, carbon atoms were binned if they were closer than 5 Å from the protein surface while for water hydrogens and oxygens a 2.5 Å cutoff was used. For the purpose of normalization, a bulk voxel occupancy for each fragment type was calculated. To this end, simulations of rectangular boxes of size 70 Å X 67 Å X 60 Å composed only of benzene+propane+water were performed using the same protocol as the simulations with protein. 10 such boxes were simulated for 5 ns each to yield a cumulative sampling time of 50 ns. Fragment and water atoms were binned into the 3D grid and the average voxel occupancy was calculated for each FragMap atom type over the entire volume. To check for any change in voxel occupancies with change in box size, another round of benzene +propane+water simulations were performed with a system size of 62 Å X 53 Å X 48 Å. There was no significant difference in voxel occupancies between the two differently-sized systems and averages over these two systems were used to compute the bulk voxel occupancy. The voxel occupancies computed in the presence of the protein were divided by the value in bulk to obtain a normalized probability. Normalized FragMaps were converted to free energies via a Boltzmann-based transform of the normalized probability to yield a "Grid Free Energy (GFE)" for each fragment type *f*, for each voxel with coordinates *x,y,z*:

$$GFE^{f}_{x,y,z} = \min\left\{-RT\log_e \frac{voxel\ occupancy^{f}_{x,y,z}}{\left\langle bulk\ voxel\ occupancy^{f}_{x,y,z}\right\rangle}, 0\right\} \tag{1}$$

where *f* corresponds to one of the four FragMap types - benzene, propane, hydrogen bond donor or acceptor. The GFE values were allowed to assume a maximum of zero. This was done in order to avoid unphysically high LGFE scores (see Eqn. 2 below for definition) resulting from ligand atoms positioned beyond the cutoff of 2.5/5 Å with respect to the protein surface. Unless indicated otherwise, all visualization of benzene and propane FragMaps were made with a GFE cutoff of −1.2 kcal/mol and hydrogen bond donor/ acceptor FragMaps with a cutoff of −0.5 kcal/mol.

Analysis of decoy sets and scoring scheme

To establish the ability of FragMaps to decipher the correct binding modes of ligands, a collection of ligand-decoy conformations was used as a validation set. For each protein-ligand complex tabulated in the LPDB, there exist two CHARMM pre-computed decoys sets: (i) surface-distributed decoys and (ii) binding-site decoys, as described in Roche et al.[19] In short, surface-distributed decoys were generated by translating the ligand onto each point on a non-overlapping spherical grid built around the protein, which extends to 5 Å beyond the maximum extent of the protein surface. Random rotations followed by a short minimization were used to generate multiple conformations on each grid point. Clustering was then performed based on both root mean square deviation (RMSD) and distance of the center of mass of the decoy with respect to the crystal conformation to yield 50 diversely-distributed decoys on the protein surface. Binding site decoys were generated using the replica method in CHARMM to simultaneously propagate 25 copies of the bound conformation of a ligand by Langevin dynamics with restraints applied to bias the copies away from the experimental bound position. At the end of 1000 simulation steps, the conformations were minimized resulting in a near-continuum of decoy positions close to the experimental bound pose of the ligand in the binding site.

To quantify the overlap of FragMaps with an arbitrary ligand conformation, the ligand coordinates were first transformed based on the alignment of the protein conformation in the protein-decoy structure to the protein conformation with which the SILCS simulations were initiated. An alignment of structures was also required when comparing the overlap between ligands and FragMaps of two different homologous proteins. For cases where a protein-ligand co-crystal structure for one of the homologs did not exist (eg. ATI-D with trypsin and FXI-2 with Factor Xa), the position of the ligand was assigned based on alignment of the protein conformation of the homologue lacking a co-crystal structure with the ligand with the homologue with known protein-ligand conformation based on optimal alignment of backbone Cα atoms and the FragMaps were then analyzed in relation to the transformed protein coordinates.

Atoms in the ligands were classified as aromatic carbon, aliphatic carbon (sp^3 methyl or methylene), hydrogen bond donor (hydrogen bonded to a nitrogen or oxygen) and hydrogen bond acceptor atoms (oxygen atoms), consistent with the four FragMap types. It should be noted that in the present study only oxygen atoms were assumed to act as acceptors; an assumption that will be tested in future studies. All ligands analyzed in this study are shown in Figure S8 of the Supporting information. The protonation state of the ligands was assigned as it was listed in the LPDB. For ligands obtained from other sources, MOE[35] was used to assign the protontation state consistent with pH 7. Each classified atom of a ligand with coordinates $(x_i.y_i,z_i)$ was assigned a score equal to the Grid Free Energy value of the corresponding FragMap type f, $GFE^{f}_{x_i,y_i,z_i}$ of the voxel it occupies. A sum over each of the four classes of atoms separately yields the GFE score for the four FragMap types (e.g. adding the benzene FragMap $GFE^{benzene}_{x,y,z}$ values corresponding to the position of aromatic carbon atoms of the aligned ligand conformation yields the aromatic GFE score for that ligand). A sum of benzene, propane, hydrogen bond acceptor and donor GFE scores for a ligand yields the Ligand Grid Free Energy (LGFE) score:

$$LGFE=\sum_{\substack{aromatic\\ atoms\ i}} GFE^{benzene}_{x_i,y_i,z_i}+\sum_{\substack{aliphatic\\ atoms\ i}} GFE^{propane}_{x_i,y_i,z_i}+\sum_{\substack{H-donor\\ atoms\ i}} GFE^{H-donor}_{x_i,y_i,z_i}+\sum_{\substack{H-acceptor\\ atoms\ i}} GFE^{H-acceptor}_{x_i,y_i,z_i} \tag{2}$$

Protein preparation

For α-thrombin, eight loop residues and two terminal residues from the heavy chain were missing from the apo structure (PDB 3D49) and their positions were constructed using the MODELLER package[36] with a ligand-bound (holo) structure (PDB 1BMM) as a template. Since the loop residues in 1BMM structure were not in close contact with the active site, they are unlikely to be differently structured in the apo form. The crystal structure of another holo form of the protein (PDB 1AE8) was used to build 5 and 3 missing residues from the N and C terminal, respectively, of the light chain of α-thrombin. Again, this region being distal to the binding pocket is unlikely to be differently structured in the apo form. 100 models were generated and were clustered based on the RMSD of modeled residues using the program NMRCLUST[37]. The most representative structure of the largest cluster was selected and a 1 ns MD simulation with GBMV implicit solvent model[38] was performed with all atoms of the protein other than those being modeled constrained. Additionally, in the SILCS simulations, the weak restraint of 0.01 kcal*mol^{-1}A^{-2} per atomic mass unit applied to Cα atoms was *not* applied to the modeled residues.

RESULTS

The SILCS methodology is applied in the present work to 7 proteins for which crystallographic data of satisfactory resolution of protein-ligand complexes are available (Table 1). SILCS MD simulations were initiated with the apo or/and holo forms depending on the availability of crystallographic structure coordinates in the PDB;[24] ligands were deleted from the holo structures prior to the MD simulations. Probabilities of fragment atoms calculated for each voxel, called FragMaps, were computed from the last 5 ns of the ten independent 20 ns SILCS simulations for each protein. A $-RT\log_e$ transform of the relative grid probabilities was done to convert FragMaps probabilities into free energies, referred to as the grid free energy (GFE) representation (Equation 1). The $\mathrm{GFE}_{x,y,z}$ value quantifies the per-atom free energy of binding for each atom type to each voxel in the simulation system. Since the desolvation penalty is included implicitly in the GFE values, it justifies their use to quantify the contribution of each type of atom to the overall free energy of binding of a ligand to the protein.

Fragments used in the present study include benzene, propane and water, as used in our previous study[18]. It has been noted that about 75% of bioactive molecules contain aromatic rings[39] and 40% of drug-like molecules contain a benzene ring.[40] This motivated the inclusion of benzene. Propane, another hydrophobic fragment was included based on its ability to penetrate smaller protein pockets and to identify protein pockets with affinities for aliphatic groups that are ubiquitous in known drug-like molecules. The inclusion of both benzene and propane, which may both be considered hydrophobic, allows for different regions of the protein that favor one class over the other to be identified. Water can both donate and accept hydrogen bonds and served as a probe for chemical groups with those functionalities. Further motivation for the choice of these fragments is their small size. All fragments are small enough to have no dihedral rotational degrees of freedom involving non-hydrogen atoms as well to maximize their diffusion constants to facilitate sampling over the protein surface. In addition, to achieve convergence in a computationally feasible timescale, a minimal set of representative fragments was chosen so as to maximize their individual concentrations, which maximizes their binding probability[18].

FragMap convergence

A major concern in the SILCS method is the required duration of the simulations to assure adequate convergence of the FragMaps. In our previous study[18], convergence focused on the change in FragMaps with increased sampling. While that criteria was considered in the present study, we also judged the quality of convergence based on the “predictive power” of the FragMaps. A preliminary investigation revealed a higher predictive power with FragMaps constructed from the last 5-ns segments of each trajectory, where the predictive power is judged by the degree of overlap of FragMaps with the position of the relevant ligand functional groups in protein-ligand crystal conformations. Using trypsin as a test case, we investigated the change in the conformational distribution of fragments as a function of time. Figure 1a shows the radial distribution function, $g(r)$ of benzene molecules with respect to the center of trypsin computed using the first and last 5-ns segments of the 20-ns SILCS trajectories. The distribution obtained from the last 5-ns segment shows a slightly higher population of fragments with low *r* values, indicating deeper penetration of fragments into protein pockets. This observation is explained based on the time required for changes in protein conformation and/or the solvation shell of the protein to occur that are required to accommodate fragment penetration. To check if the difference seen in the $g(r)$ profile is significant for the purpose of identifying key binding interactions, we performed the following analysis. The ten 20-ns trajectories were divided into two sets of five trajectories each, and two benzene FragMaps were constructed from the 0–5ns segments and two from the 15–20ns segments. Figure 1b shows the two 0–5ns benzene FragMaps overlaid on trypsin and Figure 1c shows the same for the 15–20ns dataset. The two sets of FragMaps generated from the 0–5ns dataset concur on the prediction of most benzene binding sites, however only one set predicts the relatively buried ligand binding site namely, specificity S1-pocket (Figure 1b, arrow). This site is predicted by both sets of trajectories by FragMaps generated from the 15–20ns dataset. Since these maps were constructed from trajectories that have different initial positions of fragments in the simulation box, the similar prediction implies convergence to a satisfactory extent rather than redundant unconverged results. These observations indicate that convergence in the mapping of pockets with higher degrees of burial requires longer simulation time than that of surface exposed sites. This conjecture was further tested using α-thrombin and HIV-protease, where the degree of burial of the inhibitor binding sites is higher than it is in trypsin. Figures S1 and S2 show that FragMaps constructed from later segments of SILCS trajectories overlap better with inhibitor functional groups in the crystal binding modes. We also observed that LGFE scores constructed from later parts of the simulations better distinguish the crystal conformation from decoy ligand conformations (Figure S3). These results, especially with HIV-protease, demonstrate that even for significantly buried pockets, the SILCS method identifies the key interactions necessary for ligand binding when adequate sampling is preformed (see below).

Additional evaluation of the convergence properties was performed by taking the difference of the raw voxel occupancies of FragMaps from the 15–20ns segment of trypsin SILCS simulations 1 to 5 and 6 to 10. For fully converged results, such a “difference map” should have a value of zero for all voxels. Figure S4 of supplementary information shows the histogram of occupancy differences from the “difference maps” to be centered around zero as expected for random errors. Importantly, the difference map values approach zero near the cutoff values used for FragMap visualization, indicating the significance of the cutoff values of GFE=−1.2 kcal/mol for benzene and propane FragMaps and GFE=−0.5 kcal/mol for hydrogen bond donor and acceptor FragMaps.

SILCS validation by recapitulation of crystallographic inhibitor-protein complexes

In the following analyses, a set of representative ligands was chosen for each protein and the overlap of the different FragMap types with positions of chemical groups in the crystallographic binding mode of the full ligand is discussed. This approach tested the ability of FragMaps to identify key interactions between the protein and low MW ligands. Quantification of the extent of FragMap-ligand overlap is done by first aligning protein structure from the ligand-protein crystal structure with the structure of the protein used to initiate the SILCS simulations, thereby placing the ligand on the same 3D grid on which the FragMaps were computed. The atoms in the ligand are placed in the following classifications: aromatic carbon, aliphatic carbon, hydrogen bond donor or hydrogen bond acceptor, corresponding to the four FragMap types. Each such classified atom is assigned a score equal to the GFE of the corresponding FragMap type of the voxel it occupies, with the sum over the GFE scores of appropriate atoms in a ligand yielding the ligand GFE score (LGFE, Equation 2). In the following analysis we present the LGFE scores and also partition it into GFE scores associated with the four different atom types to show their individual contributions (Eq. 2). A comparison of the LGFE score of the crystal conformation and decoys is made to show the predictive power of the SILCS FragMap-ligand overlap. The ligand-protein database (LPDB)[19] contains two kinds of decoy conformation sets: surface distributed and binding site decoys, both of which are used as validation populations. A significant fraction of the binding site decoys tend to have conformations that are very similar to the crystal (i.e. < 1 Å RMSD). Therefore, we show the RMSD of the conformations along with GFE scores.

Trypsin

Bovine trypsin is a serine protease for which co-crystal structures of the protein with several small molecule inhibitors exist[19]. The crystal structure of the apo form of trypsin (PDB 1S0Q) was used to initiate the SILCS simulations with the resulting FragMaps compared with 4 trypsin-ligand co-crystal structures. Figure 2a–d shows the primary specificity S1-pocket of the protein with FragMaps overlaid on the crystal binding modes of the ligands in crystal structures 3PTB, 1TNH, 1TNI and 1TPP respectively, which we refer to as TI-A, TI-B, TI-C and TI-D. In all four cases the benzene FragMap overlaps with the crystal orientation of the aromatic rings, with the extent of overlap being larger in TI-A, TI-B and TI-D. The aromatic ring in TI-C is positioned approximately perpendicular to the plane formed by the benzene FragMap density resulting in decreased overlap of ligand atoms with the benzene FragMap. Interestingly, TI-C also has the poorest binding affinity among the four as denoted in Figure 2 (3G, bottom right of each panel). The presence of propane FragMap in the same region suggests that this site has an affinity for aliphatic groups, which is consistent with the fact that S1-pocket binds lysine and arginine residues of natural peptide substrates of trypsin.[41] The neighborhood of negatively charged Asp189, which makes favorable electrostatic interactions with substrate Lys/Arg residue sidechains, shows the presence of hydrogen-bond donor FragMaps. For all four inhibitors, the ammonium/amidinium hydrogens show overlap with the hydrogen-bond donor FragMap (Figure 2, blue arrows). The negatively charged Asp189 at the bottom of the pocket and the surrounding hydrophobic residues create an environment favorable for binding of aliphatic/aromatic groups with hydrogen bond donors at their end[41]. The overlap shown in Figure 2 demonstrates that FragMaps capture these two thermodynamically important interactions. The benzamidine group in TI-D binds to the S1-pocket in the same orientation as benzamidine alone (TI-A). The additional acid and alcohol groups in TI-D are in the vicinity of a hydrogen-bond acceptor FragMap, consistent with the better affinity of that compound as compared to TI-A. The LGFE values as displayed on the bottom left of each panel in Figure 2 correctly predict the least and the most favorable ligands.

To quantify the utility of the overlaps shown in Figure 2 the LGFE scores of crystal conformations of the four inhibitors were compared to values obtained for the surface distributed and binding site decoys. TI-A, TI-B and TI-C do not contain any hydrogen bond acceptors, whereas a carboxylate and an alcohol are present in TI-D. Figure 3 shows the GFE scores computed over aromatic, aliphatic, hydrogen-bond donor and acceptor atoms separately, as well as the total sum (LGFE score) for the four inhibitor crystal conformations (red lines) and the respective surface distributed decoys (black histograms). For TI-A, TI-B and TI-D ligands, the GFE scores for aromatic and H-donor atoms (except TI-A) are generally lower than most decoys (Figure 3, column 1 and 3). The aliphatic score for the crystal conformation is consistently more unfavorable than that for most decoys. The lack of overlap of the methylene groups in all inhibitors with propane FragMaps is also evident in Figure 2 and indicates that the aliphatic moieties act as linkers between the benzene ring and the positively charged group interacting with Asp189. Nevertheless, primarily due to the favorable aromatic and to some extent, H-donor GFE values, the LGFE score for the crystal conformation is more favorable than that for almost all decoys for TI-A, TI-B and TI-D. Notably, low (i.e. favorable) LGFE scores for the crystal conformations are attributed to the simultaneous low values across all categories. In other words, a decoy conformation may have a lower GFE score than the crystal conformation for one class of atoms, but when the contribution of all categories is considered, the crystallographic conformation tends to have the lower (more favorable) LGFE. The poor overlap observed for TI-C is quantified by high (unfavorable) GFE scores across all categories for the crystal conformation. Therefore, the LGFE score does not distinguish the crystal conformation from the decoys. This may be related to the relatively high binding free energy of this ligand. In addition, it should be noted that the B factors for ligand in the TI-C crystal structure (1TNI)[42] are much higher than that of other inhibitors indicating that the location of the ligand may be poorly defined or that the ligand has additional flexibility in the pocket, a feature not accounted for in our approach of using the single crystal conformation. This potential limitation is addressed in the discussion below.

The ability of FragMaps to distinguish the native conformations from decoys within the binding site was next evaluated. Figure 4 shows the GFE scores for crystal (red line) and decoy (black filled squares) orientations as a function of RMSD with respect to the crystal conformation. For aromatic, hydrogen-bond donor and acceptor atoms, the general trend is that the GFE score increases with increasing RMSD as expected. For these FragMap types, the lowest GFE value is almost always assumed by a conformation which has an RMSD < 2 Å. Again, TI-C with the most unfavorable experimental affinity is an exception showing poor correlation of GFE scores with RMSD. For all inhibitors, the aliphatic GFE score correlates poorly with RMSD, presumably for the same reason as suggested above for surface-distributed decoys. Most importantly, the LGFE for TI-A, TI-B and TI-D show (Figure 4, right column) that only conformations very similar to the crystal (with RMSD < 2 Å) have LGFE scores close to that of the crystal conformation. Thus the LGFE score distinguishes conformations within the binding site that are similar to the crystal orientation against the ones that are different.

α-Thrombin

Human α-thrombin (AT), another serine protease, is a target for anti-coagulation drugs and has therefore been well investigated. Crystal structures of several high-affinity inhibitors in complex with this protein are available in the PDB. The crystal structure of the apo form (PDB 3D49) was used to seed all SILCS simulations after removing the non-covalently bound peptide, hirudin, that binds distal to the ligand-binding pocket. Missing residues from the protein structure were constructed using the MODELLER package [36] as detailed in Methods.

Baum et al.[43] thermodynamically characterized several structurally-similar high-affinity inhibitors ofα-thrombin and obtained the X-ray crystal structures of the complexes. Since both the experimental affinities and binding modes are available, this series of inhibitors was selected to illustrate the capability of SILCS FragMaps in lead optimization. Figure 5a–c shows the overlay of the crystal orientation of three α-thrombin inhibitors ATI-A (PDB 2ZGX), ATI-B (PDB 2ZDA) and ATI-C (PDB 2ZO3), which have increasing binding affinity going from ATI-A to ATI-B to ATI-C. The benzene and propane FragMaps are able to identify the three binding pockets S1, S2 and S3 of the inhibitors (Figure 5). Hydrogen-bond donor FragMaps are located close to the positively charged amidinium group of benzamidine at the base of the S1 pocket, although they are occluded from the view in Figure 5. Both propane and benzene densities are observed to coincide with the benzene ring of the benzamidine group in the three inhibitors. Similarly, the ligand ATI-D (Figure 5d) contains an aliphatic chain terminating with an ammonium group, which places methylene groups in the region located by propane FragMaps. The position of the amide of the peptide linker between S1 and S2 binding regions coincides with the hydrogen bond donor FragMap as shown in Figure 5a by the blue arrow. The S2 pocket shows both propane and benzene FragMaps, which is in agreement with the crystal-binding mode of the aliphatic proline group. Finally, the S3 pocket also shows a mixed density of benzene and propane FragMaps, indicating this as a hydrophobic pocket. The structural differences between the three inhibitors characterized by Baum et al. occur only in the region that binds to the S3 pocket. Consequently, their binding modes in the S1 and S2 pockets are the same. FragMaps in the S3 site show two regions of hydrophobic affinity, one of which is occupied by benzene and the other by propane FragMaps. ATI-B replaces the terminal S3 methyl with a phenyl group. Figure 5b shows that the phenyl group occupies one of the two hydrophobic regions in the S3 site (purple arrow) contributing to the −1.43 kcal/mol[43] more favorable experimentally measured binding free energy ΔG^0. ATI-C adds an additional benzene group which occupies the second hydrophobic site in the S3 pocket (Figure 5c, green arrow) resulting in a −0.57 kcal/mol [43] decrease in ΔG^0. The LGFE values for the three inhibitors indicated in Figure 5a, b and c capture the trend of increasing affinity. This case study suggests that SILCS FragMaps can provide accurate structural information that allows for improved understanding of changes in ligand-protein affinities, information that is useful to guide lead optimization. Baum et al.[43] note that Glu217 undergoes a conformational change to accommodate the binding of ATI-C. We show in Figure S11 of supporting information, two conformations of Glu217 obtained from the SILCS simulation trajectories that overlap with the two crystallographic positions of the same residue in complex with ATI-C (PDB 2ZO3), thus demonstrating the utility of including protein flexibility in affinity mapping.

To further quantify the ability of FragMaps to identify key protein-ligand interactions, we used three other known inhibitors of α-thrombin namely ATI-D (PDB 1AE8), ATI-E (PDB 1BMN) and ATI-F (PDB 1D4P) shown in Figures 5d, e and f, respectively, which have decoy conformations available[19]. Following the same strategy used for the trypsin inhibitors, the GFE scores for the individual fragment types and for the full ligands were calculated and are shown in Figure 6. The GFE scores of the aromatic and aliphatic atoms are significantly more favorable than that for the decoys. The hydrogen-bond donor GFE values are weakly negative and score favorably only in the case of ATI-F. The low absolute values of hydrogen bond donor and acceptor GFE values are explained in the discussion section. The hydrogen bond acceptor scores do not distinguish the crystal orientations from decoys. This is analogous to the aliphatic carbons in the trypsin inhibitors, indicating that not all functional groups interact with high affinity with the protein, rather they may play the role of linkers. For example, two of the four oxygen atoms in ATI-D face away from the protein surface and thus are unlikely to form enthalpically favorable interactions with the protein. More importantly, the LGFE scores for the three inhibitors distinguish the crystal from decoy conformations very well as seen in the right column of Figure 6. Figure 7 shows the dependence of GFE scores with RMSD for the three inhibitors in the context of the binding site decoys. The aromatic, aliphatic and to some extent, the hydrogen bond donor GFE scores tend to distinguish the conformations with low RMSD. However the hydrogen-bond acceptor scores do not show this behavior, consistent with the surface-distributed decoys. The right column of Figure 7 shows that the ligand GFE scores distinguish well the structures similar to the crystal conformation from those that are significantly different.

HIV protease

HIV protease (HP) has a number of high affinity inhibitors for which protein-ligand complex crystal structures have been solved. LPDB contains nearly 50 such crystal structures reflecting the immense interest in this target for drug design. As the binding pocket is significantly secluded from solution, mapping its affinity pattern in the context of SILCS poses a difficult problem, as the exchange of molecules from the active site with the surrounding solvent must occur during the MD simulations. Further adding to the challenge is that ligand binding causes significant conformational change in the vicinity of the active site[44] (Figure 8e). Both an apo form (PDB 2HB4) and a holo form (PDB 1G2K, with the ligand, HPI-A, removed) were used to seed two series of SILCS simulations. FragMaps calculated from the last 5 ns of the 10 simulations for the apo and holo forms of the protein are displayed in Figure 8a and b, respectively, with a slightly higher GFE cutoff of −1.0 kcal/mol for benzene and propane than the values of −1.2 kcal/mol used for the other proteins. Inhibitors from 5 high-resolution protein-ligand crystal structures, namely HPI-B to F (PDB 1B6K, 1D4L, 1HBV, 1HPV, 1HVI), were analyzed. Common to the inhibitors are 4 distinct hydrophobic groups, which are denoted by numbers in Figure 8a, b, as well as a fifth site common to two of the inhibitors. The flaps covering the active site are flexible on a timescale > 100μs.[44, 45] Therefore, opening and closing events have a very low probability of occurring on the 20-ns time scale of the present simulations such that only local conformational changes in the apo (open) and holo (closed) states will occur. In both cases some sampling of the four hydrophobic sites is occurring as evidenced by the aromatic and/ or aliphatic FragMaps indicated by the arrows in Figures 8a and b. However, as expected the FragMap overlap with the hydrophobic groups is not as optimal as for the holo protein. In the following analysis we discuss only the FragMaps of the holo protein, though results from the apo form are included in Fig. S5 of the supporting information.

An important characteristic feature of the HIV-protease binding pocket is the presence of a tetra-coordinated structural water molecule that accepts two hydrogen bonds from the amides of flap residues ILE50 and ILE50′ (of both subunits) and donates two to the oxygen atoms in the carbonyl and sulfonamide groups of inhibitors[46, 47] and has been indicated to induce the fit of the flaps to the inhibitor[48]. Figure 8c shows the PDB co-crystal structure of HPI-C overlaid with FragMaps and the bridging water molecule (Wat301), which is conserved across all co-crystal structures analyzed. Hydrogen-bond acceptor FragMaps capture the position of Wat301 accurately. Lowering the grid free energy cutoff value from −0.5 to −1.0 kcal/mol still retains this density region, indicating it to have very high affinity in this site, as has been observed previously.[49] Another important interaction of the HPI-C ligand is the hydroxyl group hydrogen bonding to the catalytic ASP25 and ASP25′ residues of the two subunits at the bottom of HP binding pocket. Figure 8d shows the hydroxyl oxygen of the inhibitor and the overlap of the likely position of the alcohol hydrogen with hydrogen-bond donor FragMaps in the vicinity of ASP25 residues of both subunits. Lam et al.[48] sought to replace the hydrogen bonding pattern both near the conserved tetra-coordinated Wat301 and near the catalytic site with a seven-membered cyclic urea resulting in one of the highest affinity inhibitors out of 53 listed in the LPDB (−14.23 kcal/mol), which we refer to as HPI-G. The geometry of the inhibitor from the co-crystal structure (PDB 1DMP) overlaid with FragMaps is shown in Figure 8f. The four aromatic hydrophobic groups of the inhibitor show good overlap with benzene/propane FragMaps. The cyclic urea oxygen, which surrogates for Wat301, overlaps with the hydrogen-bond acceptor map (red arrow). The position of hydroxyl oxygens at the catalytic site is consistent with the presence of hydrogen-bond donor FragMaps (blue arrow).

LGFE scores were calculated for the seven HIV-protease inhibitors studies. The data were then presented as a correlation plot of the calculated relative LGFE scores and relative experimental free energies of binding (Figure S5 in supporting information). Results show that the ligand GFE scores computed both from apo and holo FragMaps are able to capture most, but not all trends in the experimental relative binding free energies. Thus, SILCS FragMaps were able to capture the thermodynamically important interactions and can point to regions on the protein surface that have high affinity for different chemical functionalities, information that can be valuable in scaffold design and lead optimization.

Similar to the previous two examples, GFE scores were calculated for two ligands and their decoy conformations[19] (HPI-A, C with PDB IDs 1G2K, 1D4L, respectively) for both surface-distributed and binding-site decoys. Figure S6 in supplementary information shows that the GFE values of the crystal conformation for all FragMap types are significantly more favorable than that for decoys, such that the LGFE score distinguishes well the crystal from decoy conformations; the difference between the LGFE score of the crystal conformation and that of the most favorable decoy conformation is > 10 kcal/mol in both cases. This is because, unlike trypsin, HIV protease inhibitors are large such that random placement of the ligand on the protein surface leads to the loss of several favorable contacts causing an unfavorable LGFE score. Thus, the case of HIV protease provides a very easy distinction of the crystal conformation from surface decoys by the FragMaps. Similarly, Figure S7 shows that the ligand GFE scores of the binding-site decoys increase nearly monotonically with RMSD thus distinguishing crystal-like conformations from ones that are significantly different.

FK506 binding protein (FKBP12)

FKBP12 is a cytosolic enzyme that catalyzes cis-trans isomerization of prolyl amide bonds and is known for its function in T-cell activation.[50] To test the ability of FragMaps to identify key binding interactions, the synthetic ligand from PDB 1FKG and two natural product ligands, FK506 and rapamycin, were analyzed. Due to the large contact interfaces between the natural product ligands and the protein, this case provides an opportunity to test our methodology's predictability of interactions involving large molecules. The protein conformation used to initiate the SILCS simulations was the one in complex with one of the synthetic inhibitors[51] (PDB 1FKG). Figure 9 shows the crystal structure of FKBP12 overlaid with FragMaps and the three ligands (PDB IDs 1FKG, 1FKJ, 1FKL which we refer to as FI-A, FI-B and FI-C, respectively). FragMaps identify three hydrophobic regions in the binding site marked S1, S2 and S3 in Figure 9a. S1 is one of the regions on the protein surface with the highest affinity for propane (minimum GFE value of site S1 < −2.4 kcal/mol) and is occupied by a cyclic aliphatic group in all cases. S2 is another site with a high aliphatic affinity and is occupied by an alkyl group in all three inhibitors. For both S1 and S2, the average propane FragMap GFE value is much more favorable than for benzene, which is consistent with the presence of aliphatic moieties interacting with those sites. The third hydrophobic site, S3, is occupied by a phenyl group in the 1FKG inhibitor (Figure 9a) and a cyclic aliphatic group in FK506 and rapamycin (Figure 9b, c), and therefore is consistent with the presence of both aliphatic and aromatic FragMaps. It is known that the removal of the benzene group from FI-A, which interacts with Ile56 and Tyr82 in S3, causes a 3 to 6 fold loss of affinity.[51] The phenyl group in FI-A that faces away from the protein surface (Figure 9a) shows no overlap with any FragMap. It has been argued that this group contributes to binding affinity more through intramolecular interactions and consequently enables the ligand to assume a favorable conformation for binding FKBP12, than through any favorable interactions with the protein itself[51]. The hydrogen-bond acceptor FragMap captures one conserved hydrogen bond acceptor interaction between a carbonyl oxygen common to all inhibitors and the Ile56 backbone amide (shown by the red arrow in Figure 9a). Figure 9c shows that a fourth hydrophobic site S4 is occupied by a methyl group in rapamycin. Thus, FragMaps identify the three main hydrophobic interactions that are shared by the ligands described above, as well as a fourth hydrophobic interaction unique to rapamycin. Furthermore, consistent with the structures of the three inhibitors, site S1 has a marked preference for aliphatic hydrophobic groups vs. aromatic hydrophobic groups as evidenced by the propane FragMap density being much higher than that of benzene.

Bacterial nicotinic acid mononucleotide adenylyltransferase (NadD)

The protein NadD from the HxGH-motif containing nucleotidyl transferase superfamily has been recognized as a promising new target for developing novel antibiotics against drug resistant bacterial infections.[52] Recently, several small molecule inhibitors were identified for *Bacillus anthracis* NadD and their X-ray co-crystal structures with the protein were solved.[53, 54] Though the inhibitor compounds are all located at the "handshake" dimer interface in the crystal structures, NadD inhibition occurs in its monomeric form in solution and oligomerization caused by crystal packing is not expected to change the complex structure significantly.[53] Thus, the crystal conformation of the monomeric apo-form of NadD (PDB 3DV2) was used to seed SILCS simulations.

Figure 10 shows the overlay of the crystal orientation of four NadD ligands and the FragMaps. The protein surface from one of the four co-crystals (PDB 3MMX) is used instead of the apo form in the figure to facilitate visualization due to changes in sidechain conformations of residues in the binding site. Together, benzene and propane FragMaps identify three hydrophobic pockets, which we refer to as P1, P2 and P3, which are used by the four inhibitors to make contact with the protein (purple and green arrows in Figure 10a–d). Overlap of the benzene FragMap in the P1 site is seen with the crystal orientations of the aromatic bicyclic ring of NI-A (PDB 3HFJ), the anthracene of NI-B (PDB 3MLA), the central phenyl of NI-C (PDB 3MLB) and the naphthalene of NI-D (PDB 3MMX) as shown in Figure 10a, b, c and d, respectively. In site P2, overlap between benzene the FragMap is seen with the crystal orientation of the fluorophenyl group of NI-A as shown in Figure 10a. Additionally, the presence of propane density at this position suggests that both aliphatic and aromatic carbons can be accommodated. This is supported by the overlaps of propane FragMaps and the crystal orientation of the aliphatic chain between two amide groups in NI-B and NI-C (Figure 10b and c, respectively). The benzene FragMap also identifies an aromatic binding mode in site P3 where overlap occurs with the chlorophenyl moiety of NI-B. Compared to NI-A, the additional overlap with FragMaps of NI-B is consistent with the experimental observations that NI-B has a lower experimental K_i value than NI-A. Consistent with experimental IC_{50} values is the fact that inhibitor compound NI-D, which has the smallest overlaps with FragMaps among the four inhibitors (Figure 10d), shows the lowest inhibitory effects.[54]

Interestingly, beyond the NadD inhibitors, FragMaps also identified binding modes of some small compounds that exist in the experimental environment and are related to inhibition. Figure 10b shows that the hydrogen bond acceptor FragMap overlaps with the oxygen in a formate molecule found to mediate the interaction between the protein and NI-B, which replaces an interaction between a carboxylate group of the natural substrate and the protein. [54] This information was used to design NI-D for which an overlap between the hydrogen bond acceptor FragMap and carboxylate group is observed (red arrow in the left side of Figure 10d). Moreover, in the NI-D-protein complex, there is a citric acid molecule near the binding pocket. As Figure 10d shows, overlaps between hydrogen bond acceptor FragMaps and several oxygen atoms of citric acid are present (red arrows on the right side of Figure 10d). Indeed, this position corresponds to the ATP phosphate binding site in the substrate bound crystal structure and thus supplies additional information for design of new inhibitors. Thus, SILCS FragMaps capture important binding sites, different subsets of which are exploited by different ligands to form favorable interaction with the protein.

Ribonuclease A

Bovine pancreatic ribonuclease A (RNaseA) has several co-crystal structures of the protein in complex with inhibitors. The crystal structure of the apo form (PDB 1JVT) was used to seed the SILCS simulations. Figure 11 shows the overlay of the crystal orientation of the

RNaseA inhibitors with the FragMaps. There are three important binding subsites on the protein surface with which the inhibitors make contact. This includes B1 and B2 subsites, which interact with aromatic rings and a P1 subsite, which interacts with the phosphoryl groups of a bound substrate. Figures 11a–e show the overlap of FragMaps with a representative set of inhibitors selected from Dechene et. al. [55] Inhibitors 5′-ADP (RIA; PDB 1O0H, Figure 11a) and 2′,5′-ADP (RI-B; PDB 1O0O, Figure 11b) were both designed to bind to the P1-B2 region of the active site.[56] There are overlaps between the hydrogen bond acceptor FragMap and phosphate oxygens for both inhibitors at the P1 subsite (red arrows) whereas optimal overlap between benzene FragMap and adenine group at the B2 subsite is only found for inhibitor RI-A, as shown by the purple arrow in Figure 11a. This finding is consistent with the lower K_i value of RI-A at 1.2 μM compared to that of RI-B at 8 μM, a trend that is captured, albeit only qualitatively, by the more favorable LGFE score of −11.8 kcal/mol for RI-A compared to −1.9 kcal/mol for RI-B. Different FragMap overlap patterns are also found for inhibitors U-2′-p (RI-C; PDB 1O0M) and U-3′-p (RI-D; PDB 1O0N) shown in Figure 11c and d, respectively, which were both bound to the B1-P1 region[56] Overlap between benzene FragMaps and uracil groups as well as between hydrogen bond acceptor FragMaps and uracil oxygen atoms at the B1 subsite are present for both inhibitors as shown by purple and red arrows, respectively, in the B1 subsite. The 2′ and 3′ phosphate groups in RI-C and RI-D, respectively, make contact with two different sites both of which show hydrogen bond acceptor densities shown by red arrows in the P1 subsite. The 10 fold higher experimental binding affinity of RI-C with respect to RI-D is not captured by the LGFE scores ($\mathrm{LGFE}_{\mathrm{RI\text{-}C}} = -6.4$ kcal/mol, $\mathrm{LGFE}_{\mathrm{RI\text{-}D}} = -7.1$ kcal/mol). This can be explained by two significant limitations inherent in the approach used to quantify overlap of atoms with FragMaps. First is the usage of a single crystal conformation of the ligand as opposed to an ensemble, as discussed below. Second is the simplistic treatment of all oxygen atoms as one class of hydrogen bond acceptors. As explained in the discussion section, use of a conformational distribution of ligands and more elaborate functional group classification schemes have the potential to overcome these limitations.

Finally, inhibitor pdUppA-3′-p (RI-E, PDB 1QHC, Figure 11e) can be treated as a combination of inhibitor RI-A and RI-D for which all types of overlaps between FragMaps and crystal orientations of inhibitor functional groups discussed above are evident. Benefit from all these overlaps as identified by FragMaps results in the highest affinity inhibitor RI-E among the five analyzed here..[57] LGFE score of RI-E correctly predicts it as the most favorable binder (Figure S5). This indicates that optimizing interactions based on overlap with FragMaps may help in the design of inhibitors with improved affinities.

SILCS FragMaps capture the affinity differences between homologous proteins

While the above results validate the capability of the SILCS methodology to recapitulate the location of binding sites on protein surfaces as well as rank ligands based on binding affinities in several cases, it would be desirable if the method were able to identify dissimilarities in binding patterns in homologous proteins. Such a capability would allow for the method to be used to direct ligand modifications to improve selectivity as well as affinity. In this section, this potential is examined by comparing FragMaps and GFE scores for the serine proteases α-thrombin, trypsin and Factor Xa.

α-thrombin has the same fold as trypsin as well as significant structural similarity of the S1-pocket. Benzamidine is known to bind both proteins in the S1-pocket (PDB IDs 3PTB and 1DWB)[19]. The LGFE of benzamidine in complex with trypsin is slightly more favorable at −9.3 kcal/mol than with α-thrombin at −7.5 kcal/mol. This captures the experimental trend in binding affinity, which was measured as −6.46 and −3.98 kcal/mol towards trypsin and α-thrombin, respectively.[19] However, there exist subtle differences in the region surrounding the binding pocket. Prominently, the “60-loop” present in α-thrombin, which surrounds the

S2 site, is absent in trypsin. To understand the impact of this difference and other subtle sequence differences, including in the vicinity of the S3 pocket, on the affinity pattern of the proteins FragMaps were analyzed and used to calculate LGFE scores for an additional ligand common to the proteins. De Simone et al.[58] obtained the crystal structure of EOC-D-Phe-Pro-Abh (referred to as ATI-D) in complex with α-thrombin and measured its selectivity to α-thrombin versus trypsin and found it to be very selective to α-thrombin (selectivity was quantified by K_4 ratios as described in Ref [58] as > 18,000). Figure 12a and b shows the crystal-binding mode of ATI-D overlaid with FragMaps obtained for α-thrombin and for apo trypsin, respectively. The FragMaps in the S1 pocket of both proteins are similar, consistent with the affinity of benzamidine for both proteins; however, the aliphatic/aromatic densities observed in the S2 and S3 pockets of α-thrombin are shifted away from the pocket in the case of trypsin. This leads to decreased overlap with the FragMaps in the prospective binding mode of ATI-D to trypsin. The change in the affinity pattern of the protein around the binding pocket captured by the FragMaps is thus the likely cause of the differential affinity of ATI-D for trypsin. To quantify these differences the LGFE score of the inhibitors were calculated and are displayed on the lower left corner of each panel in Figure 12. A significantly more favorable LGFE for α-thrombin suggests that the FragMaps predict correctly the higher affinity of this inhibitor as compared to trypsin.

Factor Xa (FX) is another serine protease sharing the same fold as α-thrombin and trypsin. The absence of the “60-loop” from Factor Xa makes it much more structurally similar to trypsin. Consequently, there exist inhibitors that have high affinity for trypsin and Factor Xa but not to α-thrombin. Due to the availability of X-ray co-crystal structures of Factor Xa inhibitors and experimental binding affinities to both Factor Xa and trypsin[59], we chose the Factor Xa/trypsin pair as a more strict test of selectivity. The crystal structure of a ligand-Factor Xa complex (PDB 1MQ5) was used to seed the SILCS simulations. Maignan et al.[59] synthesized several high affinity inhibitors (K_i=0.7 to 22nM) of Factor Xa, of which only one (referred to as FXI-1) showed reasonable affinity (K_i=69nM) to trypsin. Figure 12c and d show the similar crystal binding orientations of the inhibitor to Factor Xa and trypsin, respectively (PDB 1EZQ and 1F0U). FXI-1, as well as other inhibitors, bind in an extended orientation with two main interactions with the S1 pocket and the hydrophobic S4 pocket[59]. The FragMaps correctly predict this by hydrophobic densities in the S1 and S4 pockets and a hydrogen bond donor density at the bottom of the S1 pocket (Figure 12d; difficult to visualize in Figure 12c). FragMaps also predict a third hydrophobic region (marked as S2′). It has been shown that replacement of the ester group in FXI-1 that occupies this site with a terminal methyl causes a decrease in binding affinity[59, 60]. Figure 12c and d show similar levels of overlap of FXI-1 with Factor Xa and trypsin, which is consistent with the fact that it has high affinity to both proteins. LGFE scores of the ligand for Factor Xa and trypsin are both highly favorable and hence consistent with the high affinity the inhibitor has for both proteins. However, the LGFE scores do not correctly predict the relative binding affinities

Similar analysis was performed on another inhibitor (FXI-2) that binds both Factor Xa and trypsin (Figure 12e and f). FXI-2 could only be co-crystallized with trypsin [9] and therefore, Figure 12e shows the inhibitor conformation obtained by alignment of the protein structures as described in Methods. For this inhibitor, the GFE scores calculated with the FragMaps of the two proteins do not correlate with the inhibition constants. Figure 12e shows that the inhibitor overlaps with the benzene/propane densities in the S4 pocket; however, most of the overlap is made by the atoms of sulphonyl group and the ring sulphur atom. (Maignan et al. note that the sulfonamide makes only hydrophobic interactions with the protein and not electrostatic). One possible reason for the LGFE score not capturing the relative binding affinities of FXI-2 to Factor Xa and trypsin could be that the benzene molecule used in SILCS simulations may not capture the binding pattern of a sulfonamide group and consequently is not able to predict the probability of such a group in the S4 pocket. A more diverse choice of fragments in SILCS simulations could help alleviate this problem and efforts towards this are actively underway. Another reason could be the use of static crystal structures to compute overlap with FragMaps, as discussed below. Nevertheless, the first example discussed in this section showed that FragMaps are capable of identifying changes in affinity pattern of the protein across homologs that have significant differences. For homologs with very similar affinity patterns, the scoring scheme in its current implementation does not distinguish ligands that have high affinity to both proteins.

DISCUSSION

Presented is a validation study of the SILCS approach in the context of its ability to recapitulate the location of functional groups in ligands bound to a diverse collection of proteins. Results consistently show that the experimentally determined positions of functional groups of ligands coincide with SILCS FragMaps of similar chemical type. This is shown qualitatively by visualizing the overlap of the FragMaps and the ligands and quantitatively by using the GFE scores. Two initial conclusions emerge from this result. First, MD simulations of a protein in a 1 M aqueous solution of two organic solutes are able to identify thermodynamically important protein-ligand interaction sites. Second, despite being simplistic, the fragments used are able to identify binding regions of modified analogs of the fragments (e.g. benzene vs. benzamidine) or incorporated into larger ligands (e.g. ligands of α-thrombin or HIV protease). Additional observations from the present study will be discussed below.

The extent of agreement of the SILCS FragMaps with the experimental data is primarily associated with three main assumptions used in our study. One is the assumption of additivity of fragment binding free energies that is inherent to fragment-based methods.[8] In the present work we introduce additivity on a per-atom basis through GFE scoring. Further, the use of only four FragMap types leads to atoms in diverse chemical functional groups being treated as identical in the scoring scheme. For example, the water oxygen atom is used as a probe for any oxygen atom present in the context of an ether, alcohol, carbonyl, carboxylate or phosphate group. The same atom type in different chemical functional groups is likely to have different properties, such as the desolvation penalty, and, consequently, different binding free energy contributions. Additionally, selected atoms in any given ligand are not assigned to FragMap categories resulting in some contributions being unaccounted for in GFE scoring (for example the aryl-halogen in the trypsin ligand TI-B or the sulphonamide in FXI-2). While the SILCS methodology may inherently require the atom-based additivity assumption, the use of only four FragMap types is a specific choice that will be addressed in future studies.

Possibly the most significant limitation in the present study is the use of a single experimental crystal conformation of the inhibitors. This impacts the qualitative comparison of the FragMaps with the bound inhibitors as well as the quantitative calculation of the GFE scores. In general, this approximation inherently contrasts one of the significant advantages of the SILCS approach, the explicit inclusion of conformational distributions of both the protein and the fragments being studied. To overcome this limitation, future studies will involve the use of conformational distributions of ligand-protein complexes obtained from MD simulations. It is anticipated that the use of probability distributions of the ligands will allow for more representative comparisons with the FragMaps as well as significantly improving the accuracy of the calculated GFE scores. Given these assumptions the level of agreement between the experimental binding orientations and relative binding affinities of the ligands with the SILCS simulations speaks to the robustness of the methodology. Improved agreement with experimental data is anticipated in future studies designed to overcome these assumptions.

The limitations concerning the use of a single crystal conformation contributes to the FragMaps not necessarily recapitulating the location of all functional groups in the ligands. As stated above, the use of distributions of the ligands obtained from MD simulations would allow for more rigorous comparisons, though other contributions exist. Many of the comparisons involve different ligand-protein structures, such that alignment of the proteins followed by placement of the ligand on the structure used to initiate the SILCS simulations represents a significant approximation, which is exacerbated by local conformational changes in the protein associated with the different ligands. This leads to the slight offset of the FragMaps from the crystal atomic positions of functional groups in some large inhibitors (e.g. HIV protease-inhibitor shown in Figure 8f). Again, MD simulations of the ligand-protein complexes may partially overcome this issue. Differences will also be present due to the intrinsic affinity of the protein for the ligand not being additive in that each functional group does not make ideal contributions to binding. This is due to spatial constraints associated with the ligand structure and due to certain moieties acting as linkers rather than directly contributing to binding. Interestingly, this was observed for aliphatic moieties in the trypsin inhibitors TI-B, TI-C and TI-D (Figure 2), as evidenced by the GFE score for the crystal orientation as compared to the surface (Figure 3) and binding-site (Figure 4) decoys. In contrast, with the α-thrombin inhibitors ATI-D, ATI-E and ATI-F hydrogen bond acceptors, associated with amide and ester moieties acting as linkers, do not make favorable contributions as compared to the surface decoys (Figure 6), which is consistent with the acceptor oxygens pointing away from the protein towards the solvent (Figure 5). Such variable contributions of different functional groups as a function of system/ligand is consistent with previous observations based on crystallography that individual fragments do not assume the same binding modes as occurs in full ligands.[61] Being able to differentiate between functional groups contributing to binding versus those acting as linkers/scaffolds is anticipated to significantly facilitate compound optimization.

Conversion of the FragMaps based on probability distributions to free energies allows for ligand free energies of binding to be estimated via summation of the GFE terms. Both the contribution of individual moieties on the ligands may be evaluated as well as the overall binding of inhibitors to be quantitatively ranked. This was shown using decoy conformations included in the LPDB in combination with GFE scores (Figures 3, 4, 6 and 7). While some of the individual GFE scores for the four FragMap types did not distinguish crystal from decoy conformations, the ligand GFE score in the majority of cases had the most favorable value for the crystal conformation. Overall, the ligand GFE (LGFE) scores for all the compounds studied were high negative values for tightly binding ligands. This is due to the GFE being based on an equilibrium between bound and solvated states thereby inherently taking the desolvation penalty of both the fragments and protein into account as well as direct fragment-protein interactions and other (eg. entropic) contributions to binding.

It is also notable that the GFE scoring scheme shows the potential to be able to rank ligands, yielding reasonable correlation with experimental affinities in some cases. Figure S5 in the supporting information shows the correlation of relative LGFE scores with the relative experimental binding free energies. The ranking of ligands by LGFE score agrees with experimental data for a majority of ligands of trypsin, HIV-protease,α-thrombin and RNaseA, though exceptions exist. The observed discrepancies are likely due to the limitations highlighted above and efforts to overcome these are underway.

The individual GFE contributions of the four types of FragMaps indicate the importance of various types of interactions to ligand binding. In ligands that have a significant number of aromatic/aliphatic atoms, the GFE score for the crystal conformation across these atom types is favorably negative and these atoms tend to make the most contribution to LGFE. This indicates the importance of hydrophobic interactions in driving ligand binding. On the other hand, while H-bond donor and acceptor GFE scores distinguish crystal from decoys reasonably well in many cases, their values tend to be close to zero, which can be attributed to the desolvation penalty that needs to be paid for hydrogen bonding with the protein to occur. Thus, the discriminatory ability of hydrogen bonding interactions is based on their ability to form adequately favorable interactions with the protein to overcome desolvation, though overall these contributions do not appear to drive ligand binding.

An important feature of the methodology is the use of SILCS simulations initiated from a single protein conformation, typically the apo form in the present work, to quantitatively assess the binding of different inhibitors. This occurs due to the inclusion of protein flexibility in the SILCS simulations. While small harmonic restraints were included on the protein Cα atoms to restrain the overall rotation and translation of the protein in the present study, they do not significantly damp protein motions. This allows for the identification of the buried benzene binding site in trypsin (Figure 1) as well as SILCS simulations of the HIV protease initiated from both apo and holo forms to both recapitulate the binding modes of a number of inhibitors. Such capabilities require adequate sampling in the simulations (20 ns presently) and the extent of convergence will be dictated by the magnitude of the conformational change occurring in the protein upon ligand binding. This is indicated by the apo vs. holo HIV protease FragMaps also showing significant differences and emphasizes how care must be taken to assure adequate convergence in different systems.

Finally, computational feasibility is an important advantage of the SILCS method. On average, it took about 10 days for a single 20-ns trajectory running on a 2X4-core node of a commodity-computing cluster. Since each of the 10 trajectories is independent, they can be run in parallel. Accordingly. computational time is not a limiting factor considering the months to years of time that is spent in finding a lead molecule or in its optimization in a typical drug discovery project. Furthermore, an additional advantage of the FragMaps is their applicability to any ligand of the protein under study.

SUMMARY

We applied Site Identification by Ligand Competitive Saturation (SILCS) to a test set of 7 proteins for which multiple protein-ligand co-crystal structures exist. The method involved MD simulations of a protein immersed in a solution of organic solutes. We chose benzene and propane as the organic fragments to probe the binding pattern of the proteins for aromatic and aliphatic functional groups, respectively, and water served as a probe for hydrogen bond donors and acceptors. The simulation data was processed to yield probability distributions of atoms in the fragments used, referred to as FragMaps. FragMaps may be used in the context of normalized probability distributions or converted to a grid free energy representation, which provides the per-atom free energy contribution for ligand binding as was done in the present study. Visualization of FragMaps with protein-ligand crystal structures showed that FragMaps correctly predict the binding locations of functional groups in ligands. The significance of the overlap between FragMaps and the relevant atoms of the functional groups in ligands was quantified by GFE scores. Importantly, GFE values are true free energies, since, per the definition of free energy, they are the logarithm of the ratio of the probablities of observing a given probe atom either adjacent to the protein surface or in bulk solution. GFEs therefore incorporate both enthalpic and entropic contributions, and because the MD simulations are performed in explicit solvent and include full protein sidechain mobility, also account for molecular solvation effects and protein flexibility. The GFE score metric was successful in distinguishing crystal from decoy conformations thus confirming quantitatively the overlap shown in the visualizations. For several ligands for which data exists, we correlated the ligand GFE scores with experimentally measured affinities. However, due to the simplifying assumptions in our current approach, most notably the use of single crystal structures and the simple classification of ligand atoms, GFE scores did not rank correctly the ligands of all proteins. In addition, subject to certain limitations it is demonstrated that FragMaps can be sensitive to differences in structure between homologous proteins, information that can be used for specificity guided drug design. This capability in combination with the ability of SILCS to identify potential interactions of fragments with a protein that are likely to impart thermodynamic stability to a protein-ligand complex suggests the potential for the use of the SILCS method in structure based drug design.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants CA120215 and CA107331 and the Samuel Waxman Cancer Research Foundation. The authors acknowledge computer time and resources from the Computer Aided Drug Design (CADD) Center at the University of Maryland, Baltimore.

References

1. Jorgensen WL. The many roles of computation in drug discovery. Science. 2004; 303:1813–1818. [PubMed: 15031495]
2. Congreve M, Chessari G, Tisi D, Woodhead AJ. Recent developments in fragment-based drug discovery. J Med Chem. 2008; 51:3661–3680. [PubMed: 18457385]
3. Kollman PA. Free energy calculations: Applications to chemical and biochemical phenomena. Chem Rev. 1993; 93:2395–2417.
4. Guvench O, MacKerell AD Jr. Computational evaluation of protein-small molecule binding. Curr Opin Struct Biol. 2009; 19:56–61. [PubMed: 19162472]
5. Erlanson DA, McDowell RS, O'Brien T. Fragment-based drug discovery. J Med Chem. 2004; 47:3463–3482. [PubMed: 15214773]
6. de Kloe GE, Bailey D, Leurs R, de Esch IJ. Transforming fragments into candidates: small becomes big in medicinal chemistry. Drug Discov Today. 2009; 14:630–646. [PubMed: 19443265]
7. Murray CW, Blundell TL. Structural biology in fragment-based drug design. Curr Opin Struct Biol. 2010; 20:497–507. [PubMed: 20471246]
8. Rees DC, Congreve M, Murray CW, Carr R. Fragment-based lead discovery. Nat Rev Drug Discov. 2004; 3:660–672. [PubMed: 15286733]
9. Hann MM, Leach AR, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci. 2001; 41:856–864. [PubMed: 11410068]
10. Miranker A, Karplus M. Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins. 1991; 11:29–34. [PubMed: 1961699]
11. Ben-Shimon A, Eisenstein M. Computational mapping of anchoring spots on protein surfaces. J Mol Biol. 2010; 402:259–277. [PubMed: 20643147]
12. Seco J, Luque FJ, Barril X. Binding site detection and druggability index from first principles. J Med Chem. 2009; 52:2363–2371. [PubMed: 19296650]
13. Yang CY, Wang S. Computational Analysis of Protein Hotspots. Med Chem Lett. 2010; 1:125–129.
14. Silberstein M, Dennis S, Brown L, Kortvelyesi T, Clodfelter K, Vajda S. Identification of substrate binding sites in enzymes by computational solvent mapping. J Mol Biol. 2003; 332:1095–1113. [PubMed: 14499612]
15. Landon MR, Lancia DR Jr, Yu J, Thiel SC, Vajda S. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. J Med Chem. 2007; 50:1231–1240. [PubMed: 17305325]

16. Dey F, Caflisch A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. J Chem Inf Model. 2008; 48:679–690. [PubMed: 18307332]

17. Lexa KW, Carlson HA. Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping. J Am Chem Soc. 2011; 133:200–202.

18. Guvench O, MacKerell AD Jr. Computational fragment-based binding site identification by ligand competitive saturation. PLoS Comput Biol. 2009; 5:e1000435. [PubMed: 19593374]

19. Roche O, Kiyama R, Brooks CL 3rd. Ligand-protein database: linking protein-ligand complex structures to binding data. J Med Chem. 2001; 44:3592–3598. [PubMed: 11606123]

20. Brooks BR, Brooks CL III, MacKerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The biomolecular simulation program. J Comput Chem. 2009; 30:1545–1614. [PubMed: 19444816]

21. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling dynamics studies of proteins. J Phys Chem B. 1998; 102:3586–3616.

22. MacKerell AD Jr, Feig M, Brooks CL III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem. 2004; 25:1400–1415. [PubMed: 15185334]

23. Durell SR, Brooks BR, Ben-Naim A. Solvent-induced forces between two hydrophilic groups. J Phys Chem. 1994; 98:2198–2202.

24. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977; 112:535–542. [PubMed: 875032]

25. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol. 1999; 285:1735–1747. [PubMed: 9917408]

26. Allen, MP.; Tildesley, DJ. Computer Simulation of Liquids. Oxford University Press; Oxford: 1987. p. 1-383.

27. Levitt M, Lifson S. Refinement of protein conformations using a macromolecular energy minimization procedure. J Mol Biol. 1969; 46:269–279. [PubMed: 5360040]

28. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys. 1977; 23:327–341.

29. Darden T, York D, Pedersen L. Particle mesh Ewald: an N•log(N) method for Ewald sums in large systems. J Chem Phys. 1993; 98:10089–10092.

30. Steinbach PJ, Brooks BR. New spherical-cutoff methods for long-range forces in macromolecular simulation. J Comput Chem. 1994; 15:667–683.

31. Andersen HC. Molecular dynamics simulations at constant pressure and/or temperature. J Chem Phys. 1980; 72:2384–2393.

32. Nosé S. A molecular dynamics method for simulations in the canonical ensemble. Mol Phys. 1984; 52:255–268.

33. Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A. 1985; 31:1695–1697. [PubMed: 9895674]

34. Feller SE, Zhang YH, Pastor RW, Brooks BR. Constant pressure molecular dynamics simulation: the Langevin piston method. J Chem Phys. 1995; 103:4613–4621.

35. MOE, Chemical Computing Group. 2009. p. 10

36. Eswar, N.; Webb, B.; Marti-Renom, MA.; Madhusudhan, MS.; Eramian, D.; Shen, MY.; Pieper, U.; Sali, A. Curr Protoc Bioinformatics. Vol. 15. John Wiley & Sons. Inc; 2006. Comparative protein structure modeling using Modeller; p. 5.6.1-5.6.30.

37. Kelley LA, Gardner SP, Sutcliffe MJ. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. Protein Eng. 1996; 9:1063–1065. [PubMed: 8961360]

38. Chocholousova J, Feig M. Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. J Comput Chem. 2006; 27:719–729. [PubMed: 16518883]

39. Ertl P, Jelfs S, Muhlbacher J, Schuffenhauer A, Selzer P. Quest for the rings. In silico exploration of ring universe to identify novel bioactive heteroaromatic scaffolds. J Med Chem. 2006; 49:4568–4573. [PubMed: 16854061]

40. Kolb P, Caflisch A. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. J Med Chem. 2006; 49:7384–7392. [PubMed: 17149868]

41. Leiros HK, Brandsdal BO, Andersen OA, Os V, Leiros I, Helland R, Otlewski J, Willassen NP, Smalas AO. Trypsin specificity as elucidated by LIE calculations, X-ray structures, and association constant measurements. Protein Sci. 2004; 13:1056–1070. [PubMed: 15044735]

42. Kurinov IV, Harrison RW. Prediction of new serine proteinase inhibitors. Nat Struct Biol. 1994; 1:735–743. [PubMed: 7634078]

43. Baum B, Muley L, Heine A, Smolinski M, Hangauer D, Klebe G. Think twice: understanding the high potency of bis(phenyl)methane inhibitors of thrombin. J Mol Biol. 2009; 391:552–564. [PubMed: 19520086]

44. Hornak V, Okur A, Rizzo RC, Simmerling C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. Proc Natl Acad Sci USA. 2006; 103:915–920. [PubMed: 16418268]

45. Ishima R, Torchia DA. Protein dynamics from NMR. Nat Struct Biol. 2000; 7:740–743. [PubMed: 10966641]

46. Ala PJ, DeLoskey RJ, Huston EE, Jadhav PK, Lam PY, Eyermann CJ, Hodge CN, Schadt MC, Lewandowski FA, Weber PC, McCabe DD, Duke JL, Chang CH. Molecular recognition of cyclic urea HIV-1 protease inhibitors. J Biol Chem. 1998; 273:12325–12331. [PubMed: 9575185]

47. Suresh CH, Vargheese AM, Vijayalakshmi KP, Mohan N, Koga N. Role of structural water molecule in HIV protease-inhibitor complexes: a QM/MM study. J Comput Chem. 2008; 29:1840–1849. [PubMed: 18351589]

48. Lam PY, Jadhav PK, Eyermann CJ, Hodge CN, Ru Y, Bacheler LT, Meek JL, Otto MJ, Rayner MM, Wong YN, et al. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. Science. 1994; 263:380–384. [PubMed: 8278812]

49. Lu YP, Yang CY, Wang SM. Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes. J Am Chem Soc. 2006; 128:11830–11839. [PubMed: 16953623]

50. Harding MW, Galat A, Uehling DE, Schreiber SL. A receptor for the immunosuppressant FK506 is a cis-trans peptidyl-prolyl isomerase. Nature. 1989; 341:758–760. [PubMed: 2477715]

51. Holt DA, Luengo JI, Yamashita DS, Oh HJ, Konialian AL, Yen HK, Rozamus LW, Brandt M, Bossard MJ, Levy MA, Eggleston DS, Liang J, Schultz LW, Stout TJ, Clardy J. Design, Synthesis, and Kinetic Evaluation of High-Affinity Fkbp Ligands and the X-Ray Crystal-Structures of Their Complexes with Fkbp12. J Am Chem Soc. 1993; 115:9925–9938.

52. Gerdes SY, Scholle MD, D'Souza M, Bernal A, Baev MV, Farrell M, Kurnasov OV, Daugherty MD, Mseeh F, Polanuyer BM, Campbell JW, Anantha S, Shatalin KY, Chowdhury SA, Fonstein MY, Osterman AL. From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. J Bacteriol. 2002; 184:4555–4572. [PubMed: 12142426]

53. Sorci L, Pan Y, Eyobo Y, Rodionova I, Huang N, Kurnasov O, Zhong S, MacKerell AD Jr, Zhang H, Osterman AL. Targeting NAD biosynthesis in bacterial pathogens: Structure-based development of inhibitors of nicotinate mononucleotide adenylyltransferase NadD. Chem Biol. 2009; 16:849–861. [PubMed: 19716475]

54. Huang N, Kolhatkar R, Eyobo Y, Sorci L, Rodionova I, Osterman AL, Mackerell AD Jr, Zhang H. Complexes of bacterial nicotinate mononucleotide adenylyltransferase with inhibitors: implication for structure-based drug design and improvement. J Med Chem. 2010; 53:5229–5239. [PubMed: 20578699]

55. Dechene M, Wink G, Smith M, Swartz P, Mattos C. Multiple solvent crystal structures of ribonuclease A: an assessment of the method. Proteins. 2009; 76:861–881. [PubMed: 19291738]

56. Leonidas DD, Chavali GB, Oikonomakos NG, Chrysina ED, Kosmopoulou MN, Vlassi M, Frankling C, Acharya KR. High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors. Protein Sci. 2003; 12:2559–2574. [PubMed: 14573867]

57. Leonidas DD, Shapiro R, Irons LI, Russo N, Acharya KR. Toward rational design of ribonuclease inhibitors: high-resolution crystal structure of a ribonuclease A complex with a potent 3′,5′-pyrophosphate-linked dinucleotide inhibitor. Biochemistry. 1999; 38:10287–10297. [PubMed: 10441122]

58. De Simone G, Balliano G, Milla P, Gallina C, Giordano C, Tarricone C, Rizzi M, Bolognesi M, Ascenzi P. Human alpha-thrombin inhibition by the highly selective compounds N-ethoxycarbonyl-D-Phe-Pro-alpha-azaLys p-nitrophenyl ester and N-carbobenzoxy-Pro-alpha-azaLys p-nitrophenyl ester: a kinetic, thermodynamic and X-ray crystallographic study. J Mol Biol. 1997; 269:558–569. [PubMed: 9217260]

59. Maignan S, Guilloteau JP, Pouzieux S, Choi-Sledeski YM, Becker MR, Klein SI, Ewing WR, Pauls HW, Spada AP, Mikol V. Crystal structures of human factor Xa complexed with potent inhibitors. J Med Chem. 2000; 43:3226–3232. [PubMed: 10966741]

60. Klein SI, Czekaj M, Gardner CJ, Guertin KR, Cheney DL, Spada AP, Bolton SA, Brown K, Colussi D, Heran CL, Morgan SR, Leadley RJ, Dunwiddie CT, Perrone MH, Chu V. Identification and initial structure-activity relationships of a novel class of nonpeptide inhibitors of blood coagulation factor Xa. J Med Chem. 1998; 41:437–450. [PubMed: 9484495]

61. Babaoglu K, Shoichet BK. Deconstructing fragment-based inhibitor discovery. Nat Chem Biol. 2006; 2:720–723. [PubMed: 17072304]

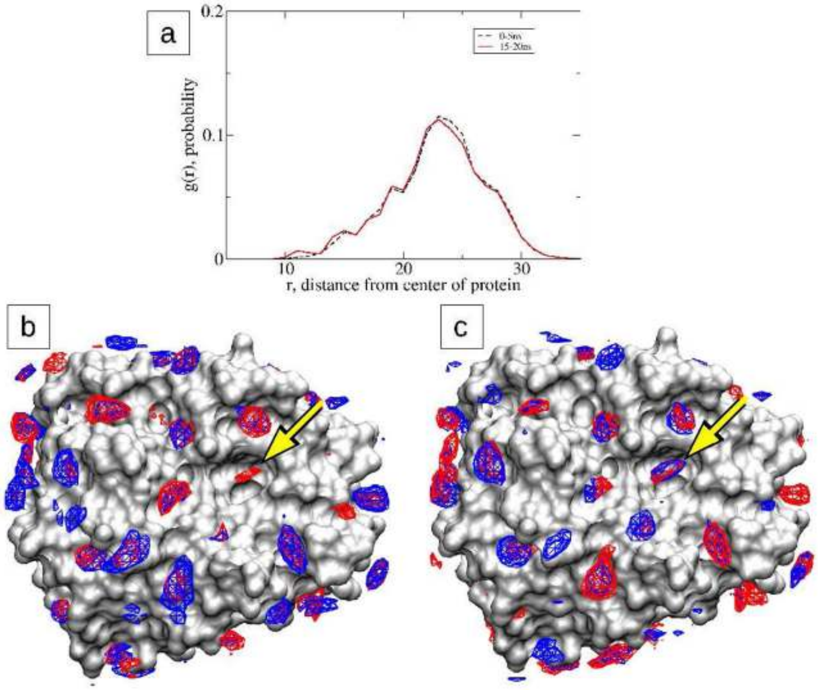

Figure 1.

Analysis of benzene fragments from SILCS simulations of trypsin. **a**) radial distribution function *g*(*r*), of the center of mass of benzene molecules from 0–5ns and 15–20 ns segments of all trajectories. Overlay of two benzene FragMaps constructed by dividing the **b**) 0–5ns and **c**) 15–20ns segments of trajectories into two groups shown as red and blue wireframe representations on the crystal structure of apo-trypsin. Most benzene binding sites are predicted simultaneously by both sets of trajectories indicating convergence. An arrow shows the relatively buried S1-specificity pocket.

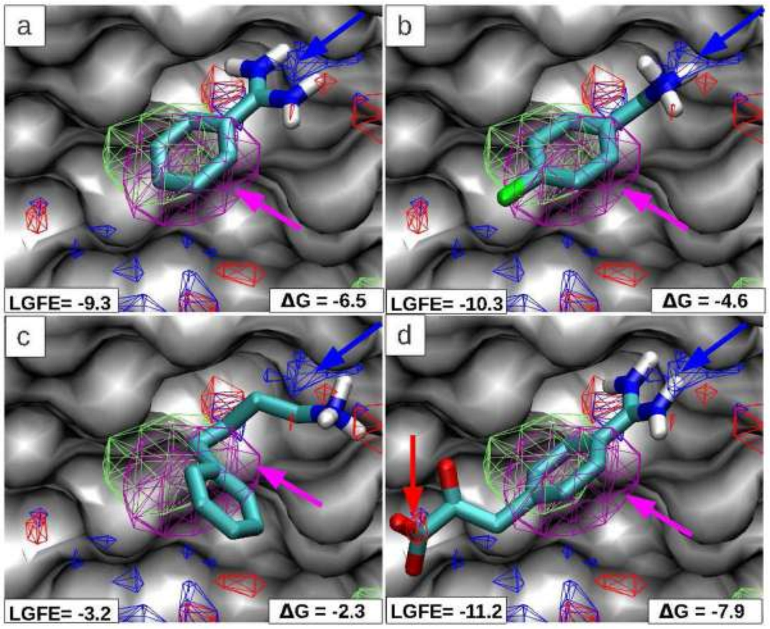

Figure 2.

S1-pocket of trypsin is shown with benzene (purple), propane (green), hydrogen bond donor (blue) and acceptor (red) FragMaps. Four inhibitors are overlaid (a) TI-A, (b) TI-B, (c) TI-C and (d) TI-D. The position of FragMaps that overlap with inhibitor atoms are indicated by arrows colored same as the corresponding FragMap wireframes. Only polar hydrogen atoms of the inhibitors are shown for clarity and the alcohol hydrogen of 1TPP inhibitor is not shown due to the uncertainty in its position. Protein atoms occluding the view of the inhibitor-FragMap overlap are removed from the visualization. The units of ligand grid free energy score (LGFE) and experimental binding affinities (from LPDB[19]) are kcal/mol.

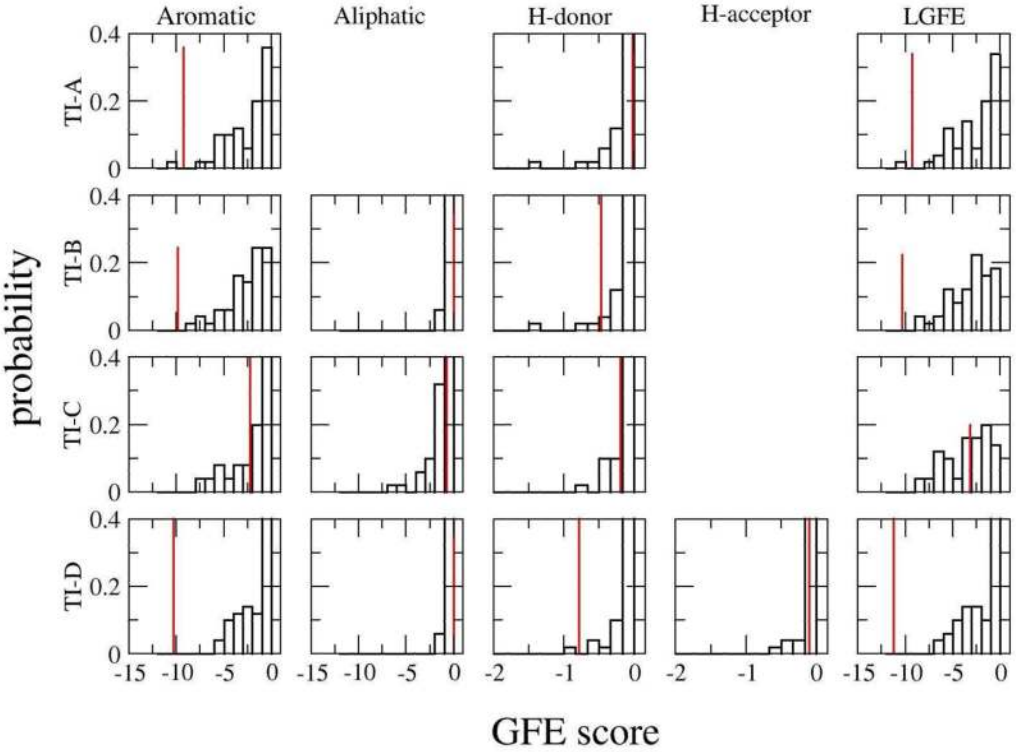

Figure 3.

Grid free energy scores computed for the crystal (red line) and surface distributed decoy (black histograms) conformations over aromatic, aliphatic, hydrogen-bond donor and hydrogen-bond acceptor atoms for four inhibitors of trypsin (marked on the left). Scores over each category of atoms are shown in the first four columns and the right column shows the sum over all as the ligand grid free energy (LGFE) score. H-bond acceptor GFE scores are not shown for TI-A, TI-B and TI-C ligands as they do not contain a hydrogen bond acceptor atom. Aliphatic GFE score is not shown for TI-A, as it does not contain methylene/ methyl groups.

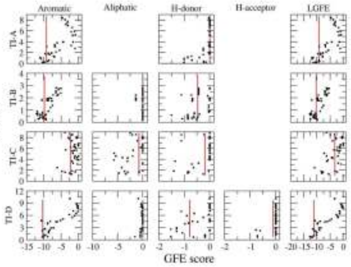

Figure 4.

Grid free energy scores computed for the crystal (red line) and binding site decoy conformations[19] (black squares) as a function of RMSD with respect to crystal conformation, over aromatic, aliphatic, hydrogen-bond donor and hydrogen-bond acceptor atoms for four inhibitors of trypsin (marked on the left). Scores over each category of atoms are shown separately and the right column shows the sum over all as the ligand grid free energy (LGFE) score.

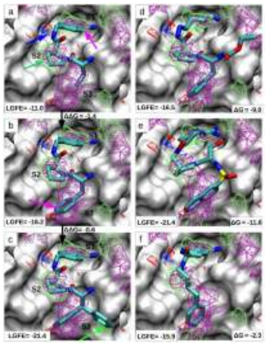

Figure 5.

α-thrombin FragMaps overlaid on the crystal binding mode of ligands (a) ATI-A (b) ATI-B and (c) ATI-C, (d) ATI-D, (e) ATI-E and (f) ATI-F. Benzene, propane, hydrogen bond donor and acceptor FragMaps are displayed as purple, green, blue and red wireframe representations, respectively. Arrows of the same color point to areas of overlap between the FragMaps and ligand atoms. Experimental ΔΔG values for ATI-A, B, C are from ITC experiments of Baum et. al.[43] Experimental ΔG of ATI-D, E, F are as listed in LPDB[19]. The units of ΔΔG, ΔG and computed LGFE are kcal/mol.

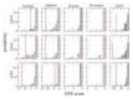

Figure 6.
Grid free energy score computed for the crystal (red line) and surface distributed decoys (black histograms) over aromatic, aliphatic, hydrogen-bond donor and hydrogen-bond acceptor atoms for three inhibitors of α-thrombin (marked on left of each row). Sums over each category of atoms are shown in the first four columns and the right column shows the sum over all as the ligand grid free energy (LGFE) score.

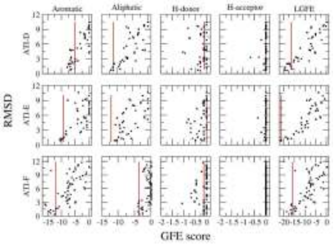

Figure 7.

Grid free energy score computed for the crystal (red line) and binding site decoys[19] (black squares) as a function of RMSD with respect to crystal conformation, over aromatic, aliphatic, hydrogen-bond donor and hydrogen-bond acceptor atoms for three inhibitors of α-thrombin (marked on the left of each row). Scores over each category of atoms are shown separately and the right column shows the global sum over all categories as the ligand grid free energy (LGFE) score.

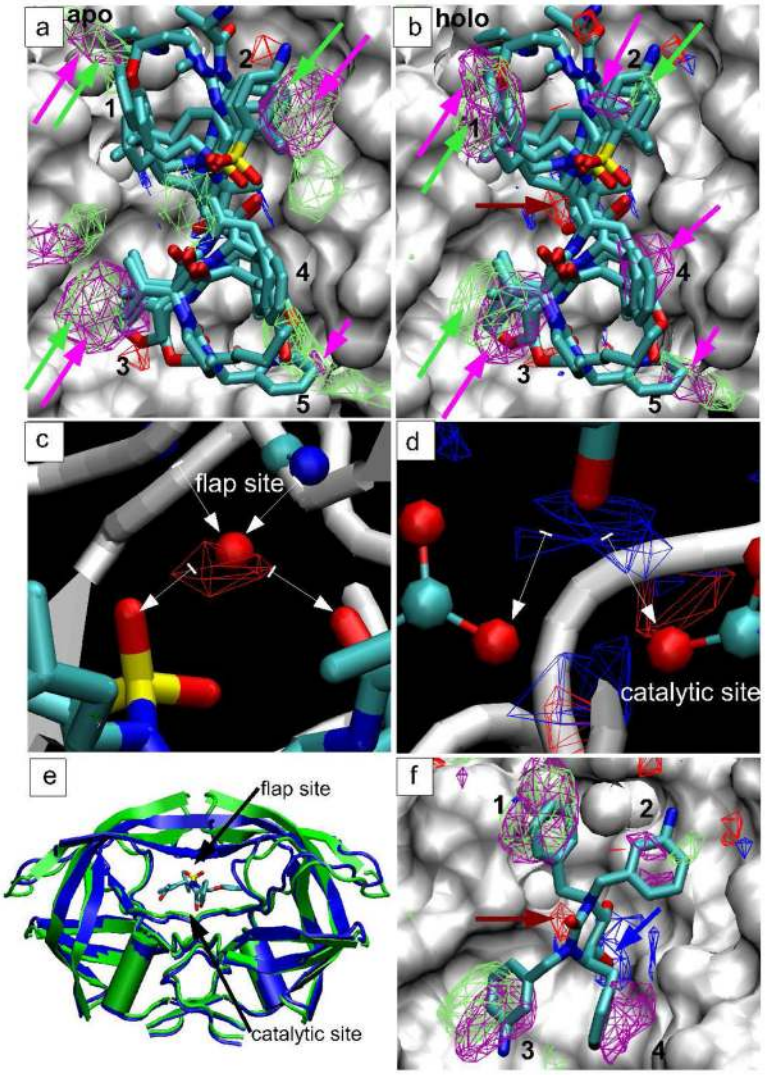

Figure 8.

(**a**) apo (PDB 2HB4) and **b**) holo (PDB 1G2K) forms of HIV-protease with benzene (purple), propane (green), hydrogen bond donor (blue) and acceptor (red) FragMaps, and with 5 inhibitors (HPI-B to F) overlaid to display the canonical binding mode. Arrows display the 5 conserved hydrophobic affinity regions. **c**) The flap site in holo protein (PDB 1D4L) showing the overlap of conserved water molecule Wat301 and water oxygen FragMap. **d**) Catalytic site of holo protein showing the overlap of the potential position of alcohol hydrogen of HPI-C and hydrogen-bond donor FragMap. Arrows denote the direction of hydrogen bonds. **e**) apo (green) and holo (blue) forms of the protein with inhibitor HPI-A in the active site. **f**) Overlay of FragMaps with HPI-G (PDB 1DMP) inhibitor. Red and blue arrows show overlap with hydrogen bond acceptor and donor atoms, respectively. A GFE cutoff of −1.0 kcal/mol was used for benzene and propane FragMaps for optimal visualization.

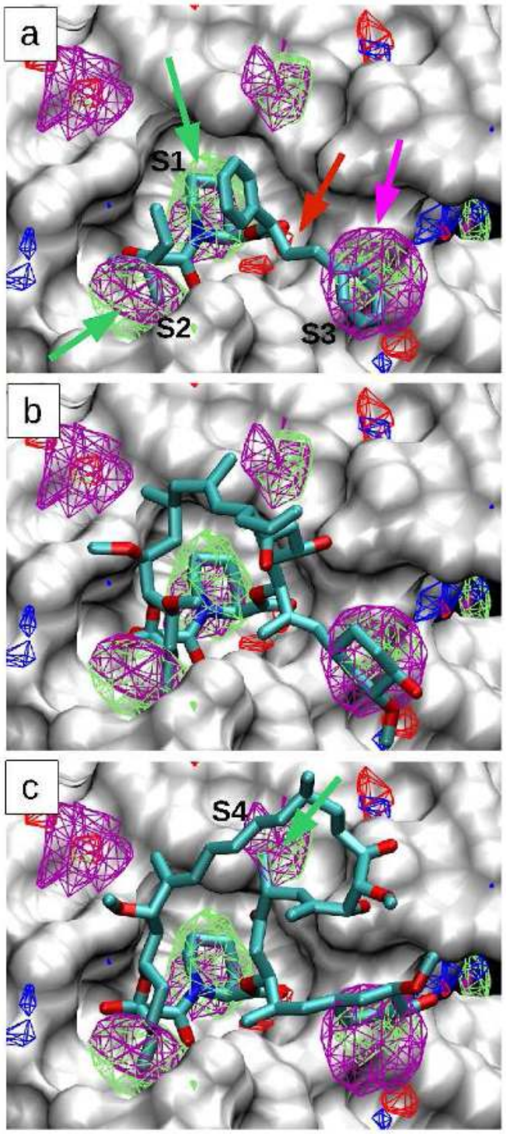

Figure 9.

FKBP12 in complex with a) a synthetic inhibitor (PDB 1FKG), b) FK-506 and c) rapamycin. Benzene, propane, hydrogen bond donor and acceptor FragMaps are displayed as purple, green, blue and red wireframe representations. Arrows of the same color point to areas of overlap between the FragMaps and ligand functional groups.

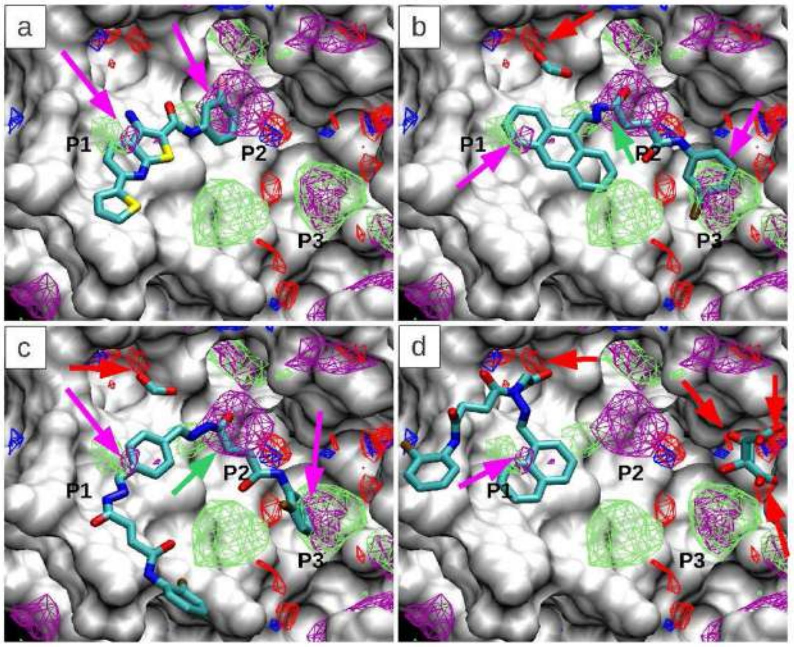

Figure 10.
NadD in complex with four inhibitors a) NI-A, b) NI-B and c) NI-C and d) NI-D. Benzene, propane, hydrogen bond donor and acceptor FragMaps are displayed as purple, green, blue and red wireframe representations. Arrows of the same color point to areas of overlap between the FragMaps and ligand functional groups.

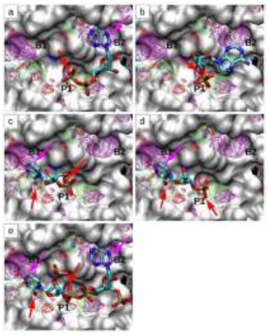

Figure 11.

RNaseA in complex with 5 inhibitors a) RI-A, b) RI-B, c) RI-C, d) RI-D and e) RI-E. Benzene, propane, hydrogen bond donor and acceptor FragMaps are displayed as purple, green, blue and red wireframe representations. Arrows of the same color point to areas of overlap between the FragMaps and ligand functional groups.

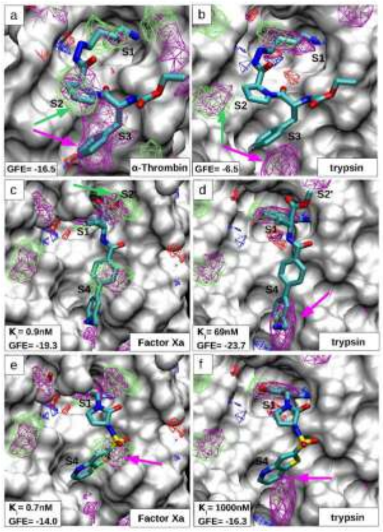

Figure 12.

Comparison of FragMaps between homologs of the serine protease family. Benzene, propane, hydrogen bond donor and acceptor FragMaps are displayed as purple, green, blue and red wireframe representations. The α-thrombin inhibitor ATI-D is overlaid with FragMaps and protein structures of (a) α-thrombin and (b) trypsin. Conformation of inhibitor FXI-1 in complex with factor Xa (PDB 1EZQ) and trypsin (1F0U) is overlaid with the corresponding FragMaps and protein structures in (c) and (d), respectively. Conformation of FXI-2 (from trypsin co-crystal PDB 1FOT) overlaid with FragMaps and protein structure of factor Xa (trypsin) in e(f). Experimental K $_i$ values are from Ref[59]. LGFE values are in units of kcal/mol. The "60-loop" of α-thrombin has been omitted from (a) to facilitate visualization of the ligand binding pocket.

Table 1

Proteins used in this study

Protein	Family	apo/holo	PDB ID of simulated structure (resolution in Å)	Number of inhibitor crystal structures analyzed	Decoy analysis
Trypsin	Serine protease	apo	1S0Q (1.04)	4	Y
α-thrombin	Serine protease	apo	3D49 (1.50)	6	Y
HIV protease	Aspartic protease	apo, holo	apo – 2HB4 (2.15) holo – 1G2K (1.95)	7	Y
FKBP12	Isomerase	holo	1FKG (2.00)	3	N
Factor Xa	Serine protease	holo	1MQ5 (2.10)	2	N
NadD	Transferase	apo	3DV2 (2.30)	4	N
Ribonuclease A	Ribonuclease	apo	1JVT (2.05)	5	N