JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

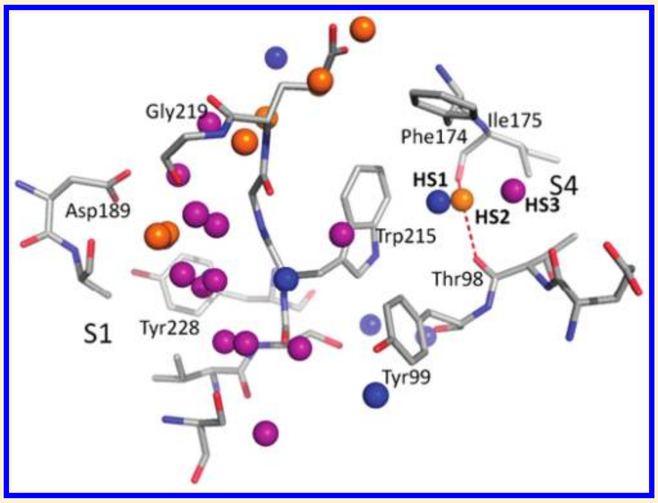
Protein Pharmacophore Selection Using Hydration-Site Analysis

Bingjie Hu[†] and Markus A. Lill*[,†]

[†]Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University,
575 Stadium Mall Drive, West Lafayette, Indiana 47906, United States

Ⓢ *Supporting Information*

ABSTRACT: Virtual screening using pharmacophore models is an efficient method to identify potential lead compounds for target proteins. Pharmacophore models based on protein structures are advantageous because a priori knowledge of active ligands is not required and the models are not biased by the chemical space of previously identified actives. However, in order to capture most potential interactions between all potentially binding ligands and the protein, the size of the pharmacophore model, i.e. number of pharmacophore elements, is typically quite large and therefore reduces the efficiency of pharmacophore based screening. We have developed a new method to select important pharmacophore elements using hydration-site information. The basic premise is that ligand functional groups that replace water molecules in the *apo* protein contribute strongly to the overall binding affinity of the ligand, due to the additional free energy gained from releasing the water molecule into the bulk solvent. We computed the free energy of water released from the binding site for each hydration site using thermodynamic analysis of molecular dynamics (MD) simulations. Pharmacophores which are colocalized with hydration sites with estimated favorable contributions to the free energy of binding are selected to generate a reduced pharmacophore model. We constructed reduced pharmacophore models for three protein systems and demonstrated good enrichment quality combined with high efficiency. The reduction in pharmacophore model size reduces the required screening time by a factor of 200−500 compared to using all protein pharmacophore elements. We also describe a training process using a small set of known actives to reliably select the optimal set of criteria for pharmacophore selection for each protein system.



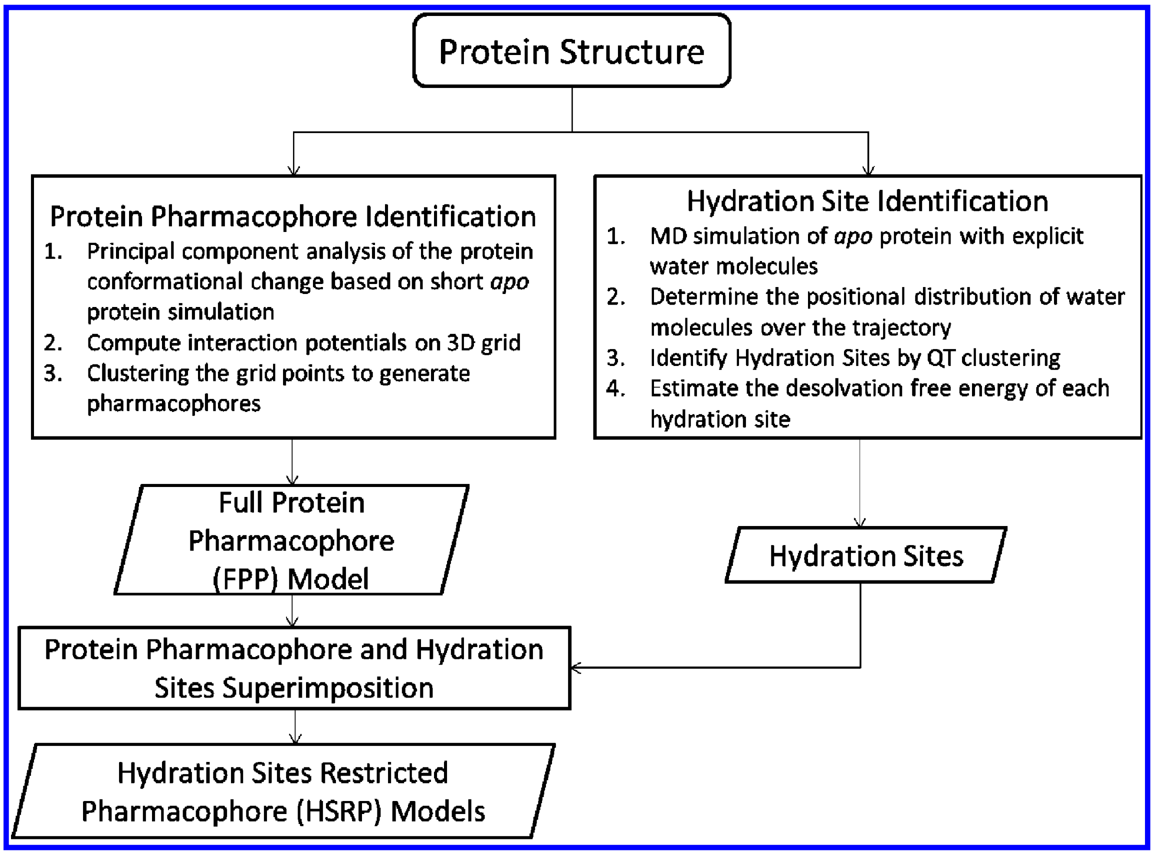
■ INTRODUCTION

Pharmacophore models are widely used to screen large chemical data sets to identify potential actives for a specific target protein. These models are typically derived from an analysis of several known actives. In addition to manual definition of pharmacophores by experienced medicinal chemists, a number of methods[1−5] have emerged to deduce structural features common to biologically active ligands and predicted to be important for the biological activity. If explicit information about the 3D structure of the binding site of the target protein is known, these data can be used to guide the development of the pharmacophore model. In the program LigandScout,[6] for example, interactions between protein and ligand in an experimentally determined protein−ligand structure are utilized to guide the pharmacophore selection process. Pharmacophores that are not present in the particular protein−ligand structure but could be important for the binding of structurally different ligands, however, may be neglected in the resulting pharmacophore model. Alternatively, the binding site of the target protein can also be utilized for pharmacophore perception without the inclusion of ligand information.[7] An interaction map between probes characterizing potential ligand features and binding site residues is translated into potential locations of pharmacophore elements. These potential pharmacophores are not restricted to previously identified active ligands. However, the size of the pharmacophore model, i.e. the number of pharmacophore elements, is typically quite large in order to include most of the potential interactions between all potentially binding ligands and the protein. Therefore clustering methods are applied to reduce the number of pharmacophores in the resulting model.

In the present study, we hypothesize that the number of pharmacophores derived from a protein structure can be reduced using free energy of desolvation estimates, resulting in pharmacophore models with good enrichment quality and increased efficiency compared to models containing all possible pharmacophore elements from the binding site. The underlying assumption is that water molecules bound to the ligand-free protein are released into bulk solvent when a ligand binds to this particular moiety. If the displaced water molecule gains free energy upon release, this energy contributes to the binding free energy of the ligand. Friesner and co-workers have used this concept in the displaced-solvent functional (DSF) method to predict the

Received: December 22, 2011
Published: March 8, 2012

Scheme 1. Overall Scheme for Generating Hydration-Site Restricted Pharmacophore Models


Protein Structure
Protein Pharmacophore Identification
1. Principal component analysis of the protein conformational change based on short *apo* protein simulation
2. Compute interaction potentials on 3D grid
3. Clustering the grid points to generate pharmacophores
Hydration Site Identification
1. MD simulation of *apo* protein with explicit water molecules
2. Determine the positional distribution of water molecules over the trajectory
3. Identify Hydration Sites by QT clustering
4. Estimate the desolvation free energy of each hydration site
Full Protein Pharmacophore (FPP) Model
Hydration Sites
Protein Pharmacophore and Hydration Sites Superimposition
Hydration Sites Restricted Pharmacophore (HSRP) Models

contribution of solvent displacement by a bound ligand to the overall free energy of binding.[8,9]

In this paper, we developed a related concept in the context of pharmacophore selection. The basic premise is that ligand functional groups which replace water molecules bound to the ligand-free protein contribute strongly to the overall binding affinity of the ligand. This is due to the additional free energy gained from releasing the water molecule into the bulk solvent.

The pharmacophore modeling process starts with the identification of potential protein pharmacophores based on an analysis of accessible hydrogen bonds and hydrophobic cavities in the binding site. Then, hydration sites defining the potential location of the water molecules are identified in the binding site of the protein. The free energy of water released from each hydration site is computed using thermodynamic analysis of molecular dynamics (MD) simulations. Finally, overlap between hydration sites and the potential protein pharmacophores is determined. Only pharmacophores displaying overlap with hydration sites with estimated favorable contribution to the overall free energy of binding are selected for the final reduced pharmacophore model. We tested various sizes of pharmacophore models and different overlap criteria on three protein systems. Finally, we developed a training process using a small set of known actives to reliably select the optimal criterion for pharmacophore selection in each protein system.

MATERIAL AND METHODS

Data Set and Target Proteins Preparation. As a proof-of-concept study, we first applied our method to factor Xa (fXa). Three different protein–ligand complexes of fXa (PDB code: 1F0S,[10] 1MQ6,[11] 1NFU[12]) were used to study the influence of different starting structures on the construction of the pharmacophore models. We also tested our method on HIV protease-1 (HIVPR, PDB code: 1AJV[13]) and *Pneumocystis carinii* dihydrofolate reductase (pcDHFR, PDB code: 1DAJ[14]). For each protein structure, the side-chain conformations of ASN, GLN, and HIS, and tautomers and protonation states of HIS were adjusted using the Reduce program.[15] Hydrogen atoms were added to the structures using the tleap module of AMBER10.[16]

The dictionary of useful decoys (DUD)[17] data set was used to perform virtual screening studies. For each active in the DUD data set, there is a corresponding set of decoys with similar physical properties but dissimilar topology. The data set we used was composed of 146 actives and 5717 decoys for fXa, 62 actives and 2020 decoys for HIVPR, and 102 actives and 2094 decoys for pcDHFR. For each ligand, multiple conformations were generated using Openeye Omega[18] with the energy window of 15 kcal/mol and 1000 conformations in maximum. Duplicate conformers were removed using a 0.2 Å RMSD cutoff for ligands with zero to three rotatable bonds, a 0.3 Å cutoff for ligands with four to six rotatable bonds, and a 0.4 Å cutoff for all ligands with more than six rotatable bonds.

Our in-house program *clusterconformer* was used to define the pharmacophore elements for each ligand conformation. Hydrogen-bond pharmacophores are placed at the position of potential donor and acceptor groups of the ligand. Ligand atoms (excluding hydrogen atoms) are defined to be hydrophobic if they are not hydrogen-bond donor or acceptor or directly bonded to a ligand's donor or acceptor atom. The hydrophobic atoms from each ligand conformation were clustered using hierarchical clustering with a minimum distance between cluster centers of 2.0 Å. Clustering is performed to reduce the number of hydrophobic ligand pharmacophores.

This significantly reduces the cost of clique detection and consequently increases the efficiency of pharmacophore-based screening.

Overview of Procedure To Generate Hydration-Site Restricted Pharmacophore Models. Scheme 1 displays the overall procedure to generate hydration-site restricted pharmacophore models. Alternative protein conformations are generated based on principal component analysis of a short MD simulation. The hydrogen-bond and hydrophobic potentials averaged over the various protein conformations are projected on a 3D grid, and clustering over the grid points is used to generate protein pharmacophores.

Hydration sites were defined based on the positional occupancy of water molecules on a 3D grid computed over a MD trajectory of the *apo* protein. The spatial overlap between hydration sites and protein pharmacophores is used to define hydration-site restricted pharmacophore models with reduced number of pharmacophore elements. The details of this procedure will be discussed in the following sections.

Protein Pharmacophore Identification. We included small-scale protein flexibility during the process of identifying protein pharmacophores in order to account for the binding of diverse ligands.[19−22] The amplitude of the protein conformational change was estimated by using principal component analysis of the covariance matrix derived from a 50 ps *apo* simulation of each protein using our in-house MD program.[23] Simulations are performed with the Amber02 force field using a water cap of 25 Å around the center of the binding site and temperature coupling utilizing a Berendsen thermostat.[24] A group-based cutoff of 10 Å was chosen for all nonbonded interactions. The coordinates of residues in the binding site were translated by the first principal component in both directions. The principal components on a short MD trajectory allow only for the modeling of small-scale protein flexibility. Thus, the size of the protein's conformational change is limited to approximately 1−2 Å RMSD. A 3D grid with 0.4 Å spacing between grid points was placed in the binding site for each of the three protein structures (minimized X-ray structure and two structures derived by coordinate translation following first principal component). The interaction potential for hydrogen-bonding and hydrophobic ligand atoms placed at individual grid points was computed using a continuous form of the ChemScore[25,26] scoring function. In detail, the interaction potential for a hydrogen-bond donating atom i on a grid point j was computed by

$$V_j^{donor} = V_{j,dist}^{donor} \cdot V_{j,angle}^{donor} \tag{1a}$$

with

$$V_{j,dist}^{donor} = \begin{cases} 0, & r_{ij} > 0.7\ \text{Å} \\ -0.5 - 0.5\cdot\cos\left(\dfrac{\pi}{0.5\ \text{Å}}\cdot(r_{ij} - 0.2\ \text{Å})\right), \\ \quad 0.2\ \text{Å} < r_{ij} < 0.7\ \text{Å} \\ -1, & r_{ij} < 0.2\ \text{Å} \end{cases} \tag{1b}$$

and

$$V_{j,angle}^{donor} = \begin{cases} 0, & \cos\varphi < \cos(88°) \\ 0.5 + 0.5\cdot\cos\left(\dfrac{\pi}{\cos(88°) - \cos(27°)}\cdot \right. \\ \quad \left. (\cos\varphi - \cos(27°))\right), \\ \quad \cos(88°) < \cos\varphi < \cos(27°) \\ 1, & \cos\varphi > \cos(27°) \end{cases} \tag{1c}$$

with $r_{ij} = |r - 1.85\ \text{Å}|$ and r was the distance between atom i and grid point j. The angle φ was defined by the angle between lone pair, acceptor of protein, and grid point. The same functional form was used for hydrogen-bond acceptors with φ defined by the angle between protein's donor hydrogen atom, donor heavy atom, and grid point. The hydrophobic potential was computed by

$$V_j^{hydrophobic} = \begin{cases} 0, & r > r_{ij}^{vdW} + 1.7\ \text{Å} \\ -0.5 - 0.5\cdot\cos\left(\dfrac{\pi}{1.4\ \text{Å}}\cdot(r - r_{ij}^{vdW} - 0.3\ \text{Å})\right), \\ \quad r_{ij}^{vdW} + 0.3\ \text{Å} < r < r_{ij}^{vdW} + 1.7\ \text{Å} \\ -1, & r < r_{ij}^{vdW} + 0.3\ \text{Å} \end{cases} \tag{2}$$

r_{ij}^{vdW} is the sum of van der Waals radii of protein atom and grid point. The grid point can be considered to represent a potential binding site of a hydrophobic ligand atom. Thus, the van der Waals radius of a carbon atom is assumed for the grid point as carbon atoms are most frequently engaged in hydrophobic contacts between protein and ligand. The distance and angle threshold values in eqs 1 and 2 were adjusted to reproduce the overall form of the original ChemScore scoring function. The modified function, however, provides continuous derivatives of the potential with respect to the coordinates.

For each protein donor or acceptor group one pharmacophore element was defined with the center of the pharmacophore computed by

$$c = \sum_i x_i \cdot \varepsilon_i \tag{3}$$

The sum is over all grid points i with favorable interaction potential from the particular protein donor and acceptor based on eq 1 where x_i and ε_i are the coordinates and absolute interaction potential of each grid point, respectively.

The hydrophobic pharmacophore elements were defined by k-means clustering of all grid points with negative hydrophobic scores. The number of clusters, k, was adjusted until the minimum distance between a cluster center i and any other cluster center was on average smaller than 2 Å. The type of protein pharmacophore was defined by the type of ligand atom that would produce favorable binding to the protein atoms associated with the pharmacophore. For example, a protein donor group could favorably interact with a ligand acceptor. The protein pharmacophore associated with that interaction is defined to be an acceptor protein pharmacophore.

Hydration Site Identification. A molecular dynamic (MD) simulation with the binding site filled with explicit water molecules was performed for each target system. The ligand in each protein–ligand complex was removed, but the crystallographic water molecules were kept. The protein was then solvated in an octahedron SPC[27] water box with a minimum of 10 Å between any protein atom to the edge of the box. Chlorine and sodium ions were then added to neutralize the systems.

We performed the MD simulations using GROMACS[28] with AMBER03 force field. Each system was energy minimized for 5000 steps using steepest descent algorithm to relieve the steric clashes within the system. The water molecules of the systems were then equilibrated for 250 ps with periodic boundary conditions in all three dimensions and with all protein heavy atoms harmonically restrained (spring constants of 1000 kJ mol^{-1} nm^{-2}). The Nose-Hoover thermostat[29,30] was used for temperature coupling at 300 K, and the Parrinello-Rahman[31] approach was used for pressure coupling at 1 bar. The electrostatic interactions were calculated exactly for atom pairs within 10 Å and by Particle Mesh Ewald[32,33] method for pairs beyond this cutoff. The Lennard-Jones interactions were truncated at 14 Å. Another 1 ns equilibration was performed under the same settings with the protein unrestrained. Finally a 10 ns production run was performed and the coordinates were saved every 10 ps to generate 1000 frames for subsequent analysis.

Using the 1000 snapshots generated from the MD simulation, the hydration sites were identified. First, the protein binding site was defined as a box surrounding its original ligand plus 3 Å in each dimension. A 3D grid was placed over the binding site using a grid spacing of 0.25 Å. In each snapshot, the positions of all the waters' oxygen atoms in the binding site were determined. A Gaussian distribution function centered on the oxygen atom centroid was distributed onto the 3D grid (Figure 1). To keep consistent with the definition of the 1 Å radius hydration sites described below, we used 0.33 Å as the standard deviation of this Gaussian function such that the Gaussian distribution covers 99.7% of the water occupancy within a 1 Å (three times the standard deviation) radius sphere. The distribution function was averaged over the MD trajectory, and pronounced peaks in this averaged function represent tightly binding water molecules which maintain their position throughout the MD simulation.

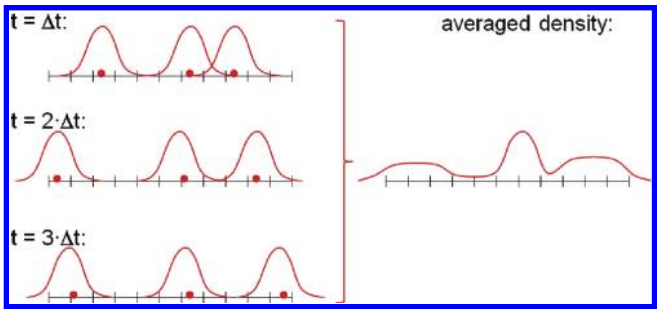


Figure 1. 1D example for determining water distribution function. Oxygen atoms (red spots) are located on grid, and probability of occupancy is mapped using a Gaussian distribution function (left side). The distribution function is averaged over many snapshots (in this example: three) of an MD simulation (right side). Tight binding water molecules display pronounced peaks.

For defining tight binding water molecules, water molecules were assigned to the position of the probability peaks using the quality threshold (QT) clustering algorithm: For each grid point all other grid points that are within 1 Å radius sphere are identified. The sphere that has the maximum occupancy (summation of the probabilities over all grid points in that sphere) was selected as first hydration site, and all grid points contained in this sphere were removed from subsequent QT clustering steps. This clustering process was repeated until the occupancy in an identified hydration site becomes less than twice the expected occupancy of a 1 Å radius sphere in bulk solvent. The latter was determined by analyzing the pseudohydration sites in a MD simulation of bulk solvent (described in the next section). A pseudohydration site was defined as a randomly selected 1 Å radius sphere in the bulk solvent. The same Gaussian distribution functions were used to compute the occupancy probability of each grid point. The occupancy of a pseudohydration site was thus a simple summation of the probabilities on the grid points inside the defined sphere. Water molecules from the MD trajectory were assigned to each hydration site if its oxygen position is within the hydration site sphere. The 1 Å radius sphere, which has been used in previous hydration site studies,[8,9] ensures there is at most one water molecule in each hydration site per MD snapshot.

Free Energies of Desolvation of Hydration Sites. The desolvation energy of each hydration site was determined by analyzing the enthalpy and entropy change of the water molecules inside a hydration site with respect to bulk solvent

$$\Delta G_{hs} = \Delta H_{hs} - T\Delta S_{hs} \tag{4}$$

where ΔH_{hs} and ΔS_{hs} are the enthalpic and entropic change of transferring a water molecule from the bulk solvent into the hydration site of the protein cavity. The change of the pressure-volume work associated with a volume change can be neglected.[34] Thus the enthalpic change can be estimated by the change of the interaction energies

$$\Delta H_{hs} \approx \Delta E_{hs} = E_{hs} - E_{bulk} \tag{5}$$

where E_{hs} is the interaction energy of a water molecule in the hydration site with the surrounding protein and water atoms. It was determined based on the average sum of van der Waals and electrostatic interactions between each water molecule inside a given hydration site with the protein and all the other water molecules. In detail, water molecules inside each hydration site were recorded for each frame of the MD trajectory. Each recorded water molecule was defined as an energy group, and its van der Waals and electrostatic interactions with the surrounding environment were extracted using the g_energy analysis tool in the GROMACS package. E_{bulk} is the interaction energy of a water molecule with its surrounding environment in the bulk solvent. To calculate E_{bulk}, a 10 ns MD simulation of a water box with 3272 explicit SPC[27] water molecules was performed following the same procedures and parameters as the solvated protein simulation described above. Averaging the interaction energies of 2058 water molecules with their surrounding environment over the 10 ns MD trajectory leads to an estimated E_{bulk} of −18.18 kcal/mol (standard error: 0.002 kcal/mol), a value comparable to that documented by Friesner and co-workers using the TIP4P water model.[9]

Assuming no change in the momenta part of the partition function upon transferring a water molecule from the bulk solvent into the protein cavity, ΔS_{hs} can be estimated by[35]

$$\Delta S_{hs} = R\ln\left(\frac{C^\circ}{8\pi^2}\right) - R\int p_{ext}(q)\ln p_{ext}(q)dq \tag{6}$$

where C° is the concentration of pure water (1 molecule/29.9 Å^3), R is the gas constant, and $p_{ext}(q)$ is the external mode probability density function (PDF) of the water molecules' translational and rotational motions during the molecular dynamics simulation. It should be noted that higher-order correlations between water molecules in the binding site[36] are neglected in this approach.

To estimate $p_{ext}(q)$ for each hydration site, we analyzed the translational and rotational motions of the water molecules in that hydration site using a method adapted from McCammon and co-workers.[35] For each hydration site, the translational degrees of freedom of water molecules in this site were defined by the fluctuation of the position of its center oxygen in the protein coordinate system. The Euler angles representing the spatial orientation of the water molecules in reference to the Cartesian coordinate system were used to calculate the rotational degrees of freedom. In detail, the rotated system (X, Y, Z) for quantifying the rotation of a water molecule was defined based on its H_1-O-H_2 plane: the unit vector in the direction of $O-H_1$ defines X, the unit vector orthogonal to X in the H_1-O-H_2 plane defines Y, and the unit vector orthogonal to the H_1-O-H_2 plane defines Z. The Euler angles were then computed based on this rotated system. Two 3 × 3 zero-mean covariance matrices were constructed for the translational and rotational motions respectively assuming decoupled translational and rotational motions. One 6 × 6 zero-mean covariance matrix was also constructed assuming the translational and rotational motions are coupled. Principal components analysis was performed by diagonalizing the zero-mean covariance matrices, and the original coordinates from the 1000 snapshots were projected onto each of the principle component dimensions. A histogram was constructed for each principle component dimension with 70 bins to allow $p_{ext}(q)$ to be calculated by normalizing the histogram. The configurational entropy of each dimension was then numerically integrated using the composite Simpson's rule. The overall configurational entropy is then summed over all the principal component dimensions. No significant difference of the estimated configurational entropy was observed between using the two 3 × 3 matrices and the one 6 × 6 matrix by performing the paired t tests at the significance level of 0.01. Therefore, the results discussed in this paper all utilize the two 3 × 3 matrices. The estimated configurational entropy using 6 × 6 matrix is reported in the Supporting Information S1.

Construction of Hydration-Site-Restricted Pharmacophore (HSRP) Models. To test the hypothesis that the number of pharmacophore elements derived from a protein structure can be reduced using free energy of desolvation estimates to increase efficiency combined with good enrichment quality of the pharmacophore models, we used the hydration site information to construct a set of hydration-site-restricted pharmacophore (HSRP) models for virtual screening. Those hydration sites with positive ΔG_{hs} (posHS) should be energetically rewarding if replaced by the ligand; therefore, we constructed HSRP models based on rewarding hydration sites with positive ΔG_{hs}. To accomplish this, we superimposed the precomputed protein pharmacophore elements onto the identified hydration sites using the original crystal structure as a common reference frame. The pharmacophore elements that were within 1.0 Å (radius of the hydration site) to any of the hydration sites centers were selected to construct the posHS HSRP models. Because the protein pharmacophore elements were generated based on three protein conformations whereas the hydration sites were identified by analyzing the MD trajectory, it might be too restrictive to simply use the radius of the hydration site as a selection criterion. Therefore we also used 1.5 Å and 2.0 Å distance cutoffs to construct the HSRP models.

In addition to using an absolute cutoff (i.e., with positive ΔG_{hs}) in defining the rewarding hydration sites, we also selected the hydration sites based on their free energy rankings. We first ranked all the identified hydration sites in a descending order based on their estimated free energies. Then we used top 25% (T0.25), top 50% (T0.50), top 75% (T0.75), or all of the hydration sites (allHS) to select the pharmacophore elements following the same methods and distance cutoffs described above.

Virtual Screening. Virtual screening was performed using our in-house program, *Hydro-Pharm*. The program enumerates all possible matches of protein and ligand pharmacophores using a modified Bron-Kerbosch clique detection algorithm,[37,38] then translates these matches into binding poses, and selects the optimally ranked ligand poses using a pharmacophore-based scoring function adapted from LigandScout.[6] First, the length of the edge between each pair of ligand pharmacophores was determined. The edge lengths were also determined for the protein pharmacophore pairs. All ligand pharmacophore edges that match the protein pharmacophore edges based on physicochemical properties (hydrophobicity, hydrogen bond donor/acceptor) of their vertices and edge lengths were identified. A 1.0 Å tolerance was allowed in the matching of the edge lengths. This matching can be represented by a graph in which each node represents a matching ligand-protein pharmacophore pair. The clique detection algorithm then identifies all the completely connected subgraphs from this graph. The Kabsch algorithm[39] was then used to position the ligand into the protein binding site based on the matching pharmacophore elements.

The ligands that were positioned inside of the protein binding site were scored using the "pharmacophore fit" scoring function adapted from LigandScout.[6] The scoring function is based on the matching between the ligand and protein pharmacophore features. The feature matches are defined as follows: a ligand pharmacophore with donor/acceptor properties matches with a donor/acceptor protein pharmacophore (see the definition of protein pharmacophores above) if the distance between the ligand and protein pharmacophores is within 1.0 Å and the angle formed by the donor-hydrogen vector and acceptor-lone pair vector is within 45°; a ligand pharmacophore with hydrophobic property matches with a hydrophobic protein pharmacophore if the distance is within 1.5 Å.

The scoring function is defined as

$$S = c^{*}N_{MFP} + S_{RMSD} \tag{7}$$

where c is the weighting factor for the number of matched feature pairs with the default value of 10. N_{MFP} is the number of geometrically matched feature pairs. S_{RMSD} is the root-mean square deviation (RMSD) distance score of the matched feature pair defined as

$$S_{RMSD} = 9 - 3^{*}\min(RMSD_{FP}, 3) \tag{8}$$

where $RMSD_{FP}$ is the RMSD of the matched feature pair distances.

Measures of Virtual Screening Success. To analyze the virtual screening results, the ligands for each protein system were ranked based on their pharmacophore fit scores. The Receiver Operating Characteristic (ROC) curve displaying the fraction of ranked actives (true positive rate) at a given fraction

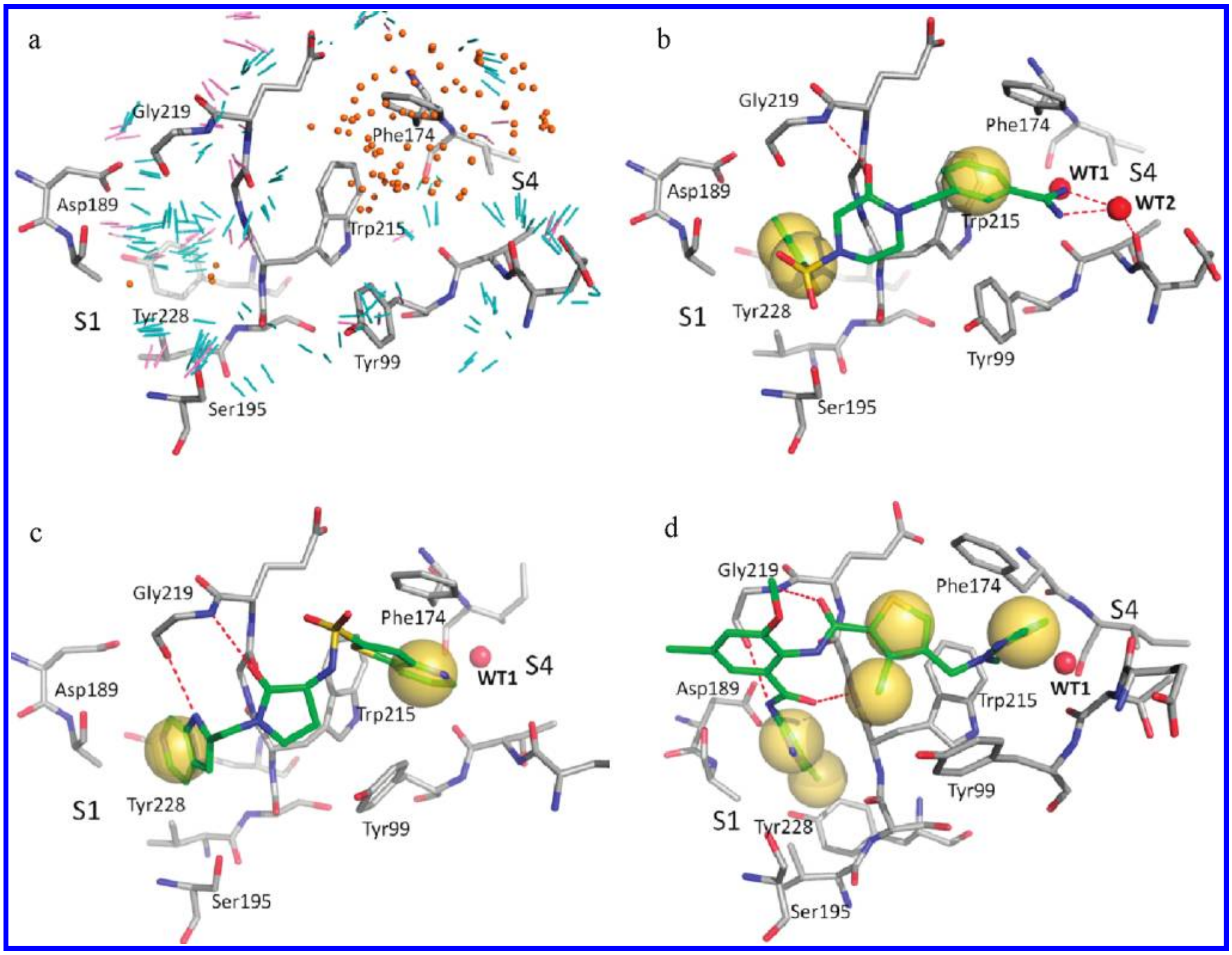


Figure 2. Protein pharmacophores identified in the active site of fXa:1NFU (a) compared with the interaction features between fXa and its cocrystallized ligands (b: RRP in 1NFU, c: PR2 in 1F0S, d: XLD in 1MQ6). The coordinates for binding site residues shown as gray sticks represent the minimized crystal structure of fXa. a: The pharmacophore elements representing potential ligand functional groups are coded as follows: orange dots as hydrophobic groups, cyan lines as hydrogen bond donor groups, pink lines as hydrogen bond acceptor groups. b-d: Red dash, hydrogen bond interaction; Yellow sphere, hydrophobic interaction. A conserved water molecule (WT1) observed in all three crystal structures and a mediating water molecule in 1NFU (WT2) are shown as red spheres.

of ranked decoys (false positive rate) was plotted for each virtual screening run. The enrichment factor, defined as

$$EF = \frac{True\ positive\ rate}{False\ postive\ rate} \tag{9}$$

was calculated at 1% (EF1), 10% (EF10), and 20% (EF20) of ranked decoys. The area-under-the-curve (AUC) was also calculated for each ROC curve and used to assess the enrichment quality of the pharmacophore models.

Model Selection by Training on Enrichment of Small Subsets. As described in the construction of the HSRP models, we selected the protein pharmacophores using different energy cutoffs for the free energy of desolvation of a hydration site and different cutoffs for the maximum distance between hydration sites and protein pharmacophores. To reliably select the optimal HSRP models for different protein systems, we performed a training process on small subsets of the DUD data set for each protein system. Subsets with 5 (sub5), 10 (sub10), and 20 (sub20) active ligands were randomly selected from the DUD data set of each protein system. Based on the proportion of actives to decoys in the full DUD data set, a corresponding number of decoys were randomly selected for the sub5, sub10, and sub20 sets. This random selection was repeated 20 times for each subset. The virtual screening was then performed on these subsets using different HSRP models.

■ RESULTS AND DISCUSSION

Identification of Protein Pharmacophore Elements. We first generated the protein pharmacophores for the three protein systems and analyzed how the pharmacophores collocate with residues known to be involved in protein–ligand interactions. The pharmacophore elements identified in the 1NFU structure of fXa are shown in Figure 2a. The fXa binding site can be defined by the S1 and S4 subpockets. The anionic S1 pocket contains several hydrogen bond donor/acceptor pharmacophore elements, many of which are contributed by residues Asp189, Ser195, and Tyr228, which line the S1 pocket. Several hydrophobic features are also found surrounding the phenol ring of Tyr228 in the S1 pocket. The S4 pocket is surrounded by the aromatic residues Tyr99, Phe174, and Trp215. A number of hydrophobic pharmacophore elements are identified surrounding the benzene ring of

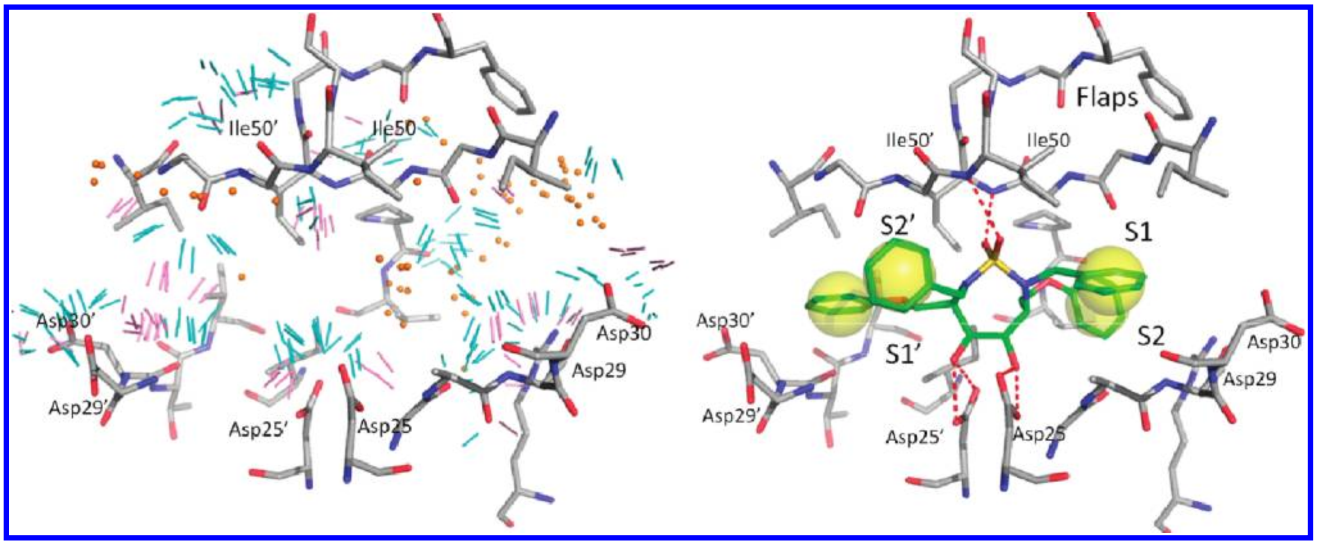


Figure 3. Protein pharmacophores identified in the active site of HIVPR (left) and the interaction between the cyclic sulfamide HIVPR inhibitor 1AJV:NMB and the binding site residues. The binding site residues are shown as gray sticks. The pharmacophore features are color-coded as described in Figure 2.

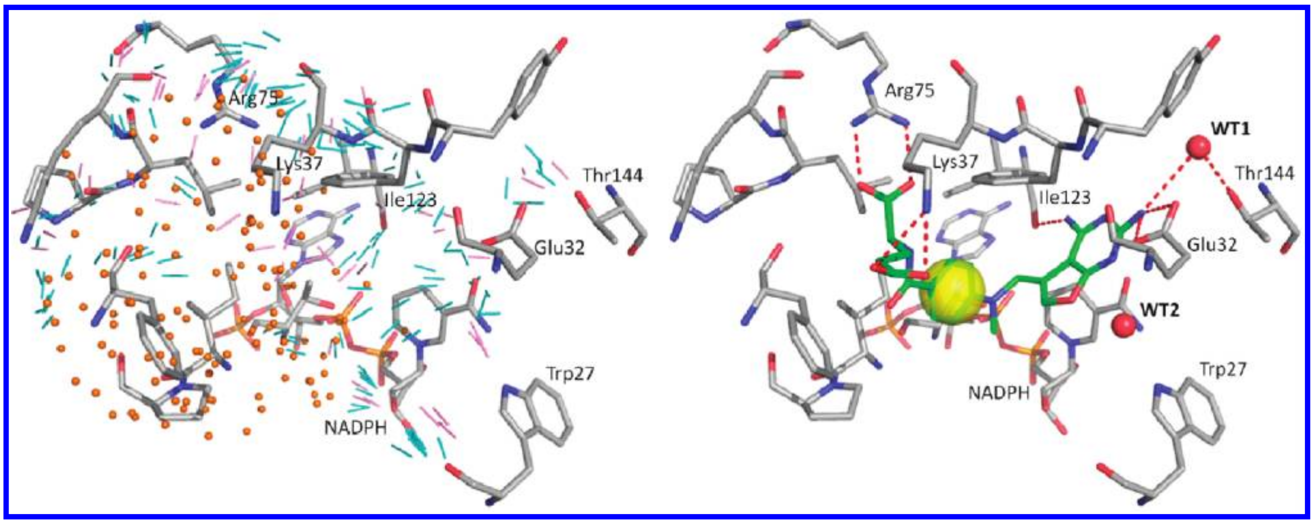


Figure 4. Protein pharmacophores identified in the active site of pcDHFR (left) and the interaction between the classical antifolate MTXO and binding site residues of 1DAJ (right). The binding site residues and the cofactor NADPH are shown as gray sticks. The pharmacophore features are color-coded as described in Figure 2.

Phe174 and the indole group of Trp215. This agrees with the experimental results that the S4 pocket has a high affinity for hydrophobic groups.[40,41] Although these pharmacophore elements are identified in the 1NFU X-ray structure and cover the key interactions with its cocrystallized ketopiperazine inhibitor 1NFU:RRP (Figure 2b), they also encompass key interactions between the sulfonamidopyrrolidinone inhibitor 1F0S:PR2 (Figure 2c) and the nonamidine inhibitor 1MQ6:XLD (Figure 2d) with fXa.

The binding site of HIVPR contains eight charged residues (Asp25/25′, Asp29/29′, Asp30/30′, Arg8/8′).[42] These charged residues contribute the majority of hydrogen bond donor/acceptor groups to the pharmacophore model (Figure 3 left). Hydrophobic features are identified surrounding the flaps and the S1/S1′, S2/S2′ subsites of HIVPR. These pharmacophore elements also contain key interaction features between the ligand and the binding site residues (Figure 3 right).

The binding site of pcDHFR contains many hydrophobic residues (Figure 4 left). Therefore a large number of hydrophobic pharmacophore elements are found in the binding site of pcDHFR. Hydrogen bond donor/acceptor pharmacophore elements were found near the polar side chain of residues Lys37, Arg79, Glu32, and Thr144 (Figure 4 right). Two conserved structural water molecules (WT1 and WT2 in Figure 4) are found in many crystal structures[14,43,44] of pcDHFR. WT1 is present in 1DAJ and mediates the interactions between the classical antifolate MTXO and Thr144.

Enrichment Using All Pharmacophore Elements. We performed a virtual screening study for each protein system using all pharmacophore elements identified in the protein's binding site (denoted as "FPP": full protein pharmacophore model). The pharmacophore-based scoring function from LigandScout was adapted to rank the ligands. To understand the influence of different starting protein conformations on the virtual screening performance, we compared the results from the three fXa structures, 1NFU, 1F0S, and 1MQ6. As shown in Table 1, the overall enrichment qualities indicated by the AUC of the ROC plots are quite similar among the three different

Table 1. Virtual Screening Results Using All Protein Pharmacophore Elements (FPP)[a]

	fXa:1F0S	fXa:1MQ6	fXa:1NFU	HIVPR:1AJV	pcDHFR:1DAJ
EF1	29.7	12.2	18.2	9.3	2.0
EF10	6.4	5.6	6.7	4.7	1.5
EF20	3.8	3.5	3.9	3.5	1.3
AUC	0.84	0.81	0.87	0.81	0.51
# of protein pharmacophores	148	139	149	131	160
runtime per ligand (s)	1225	1468	1206	934	896

[a] EF1: enrichment factors at 1% ranked decoys; EF10: enrichment factors at 10% ranked decoys; EF20: enrichment factors at 20% ranked decoys; AUC: area under the ROC curve. The number of protein pharmacophores and the runtime (seconds) per ligand needed in virtual screening are also shown for each system. The virtual screening was run on a single core of either 2.5 GHz quad-core AMD2380 or 2.1 GHz 12-core AMD6172 processors. On average, running the virtual screening on AMD2380 is faster than running on AMD6172 by about 18%.

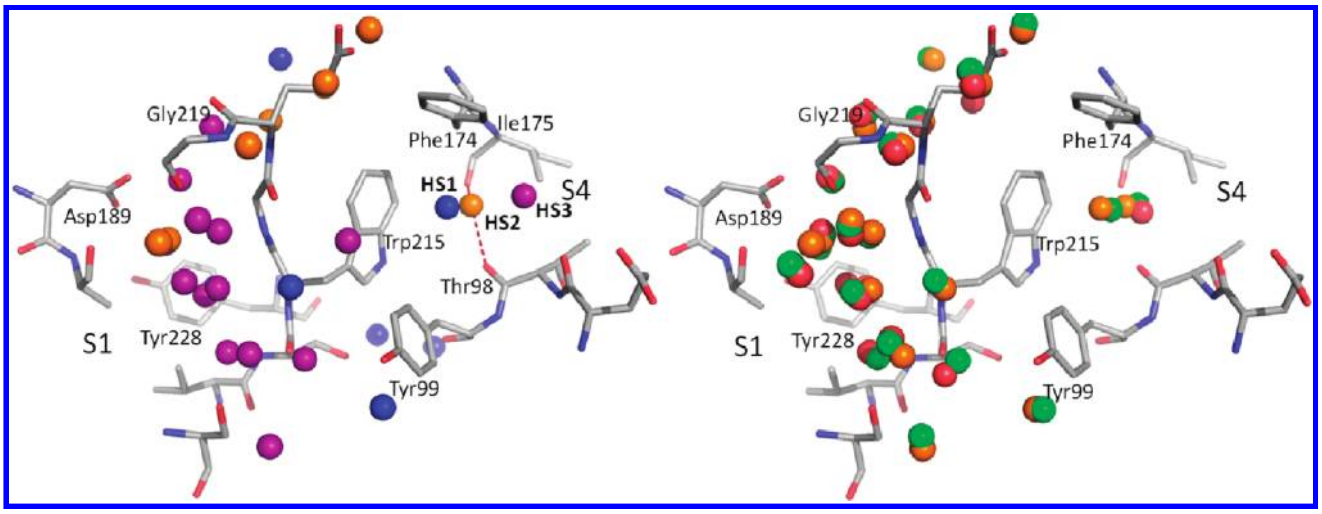


Figure 5. Hydration sites identified in the binding site of factor Xa. For clarity, hydration sites that are not accessible to the ligands or on the protein surface are removed. The protein structure shown is the minimized crystal structure of 1NFU. In the left panel, only hydration sites identified from 1NFU are shown. Hydration sites that contribute both entropically and enthalpically when replaced by ligands are shown in blue. The hydration sites whose entropic gain surpasses the enthalpic loss when replaced by ligands are shown in purple. The hydration sites whose enthalpic loss is larger than the entropic gain when replaced by ligands are shown in orange. In the right panel, the hydration sites identified from three different starting structures of factor Xa were overlaid (red: 1F0S, green: 1MQ6, orange: 1NFU).

fXa structures with values exceeding 0.8 for all three systems. The largest difference between the three protein structures was obtained for the enrichment factor at 1% ranked decoys showing a better early enrichment for using 1F0S than using the other two structures.

Considering the other two protein system, the overall virtual screening performance for HIVPR was also significantly better than random with an AUC of 0.81. However, the AUC for pcDHFR (0.51) was not significantly better than random (0.50). As there are two conserved water molecules (Figure 4) observed in many crystal structures of pcDHFR,[14,43,44] one possible explanation for the weak performance of the pharmacophore model is that the water-mediated interactions were not included in the model.

Although a high AUC value was achieved for both fXa and HIVPR, the FPP models contain more than 130 protein pharmacophore elements (Table 1). This extensive pharmacophore model allows a comprehensive search of binding poses and in combination with a sufficient scoring function might result in a low false negative rate. However, it should be recognized that the scoring function used for the pharmacophore-based virtual screening is typically a simple feature-matching function. Consequently, the inclusion of a large number of pharmacophore elements can potentially result in a high false positive rate. Furthermore, it is computationally prohibitive to search and score all the possible ligand binding poses against such an extensive pharmacophore model. The process took more than 1,200 s per ligand for fXa and approximately 900 s per ligand for HIVPR and pcDHFR (Table 1). As such, the FPP model would not be suitable for virtual screening using large compound libraries with hundreds of thousands of compounds.

Hydration Site Identification. The release of water molecules upon ligand binding can contribute both enthalpically and entropically to the free energy of binding. Water molecules in the binding site with a less negative free energy will contribute more to the free energy of ligand binding when replaced by the ligand. To compute the free energy of a water molecule in the binding site (see Materials and Methods: Free energies of desolvation of hydration sites), we performed a 10 ns MD simulation and used QT clustering to identify spherical regions (1 Å radius) in the protein binding sites that have higher density of water molecules compared to the density in bulk solvent. These spheres with high-water density were denoted as hydration sites. In the binding site of fXa, we identified 20, 33, and 30 hydration sites for 1F0S, 1MQ6, and 1NFU, respectively (Figure 5). This variation is due to the differences in the size of the binding sites which were determined by the size of the cocrystallized ligands (see Materials and Methods) as well as a slight variation among the

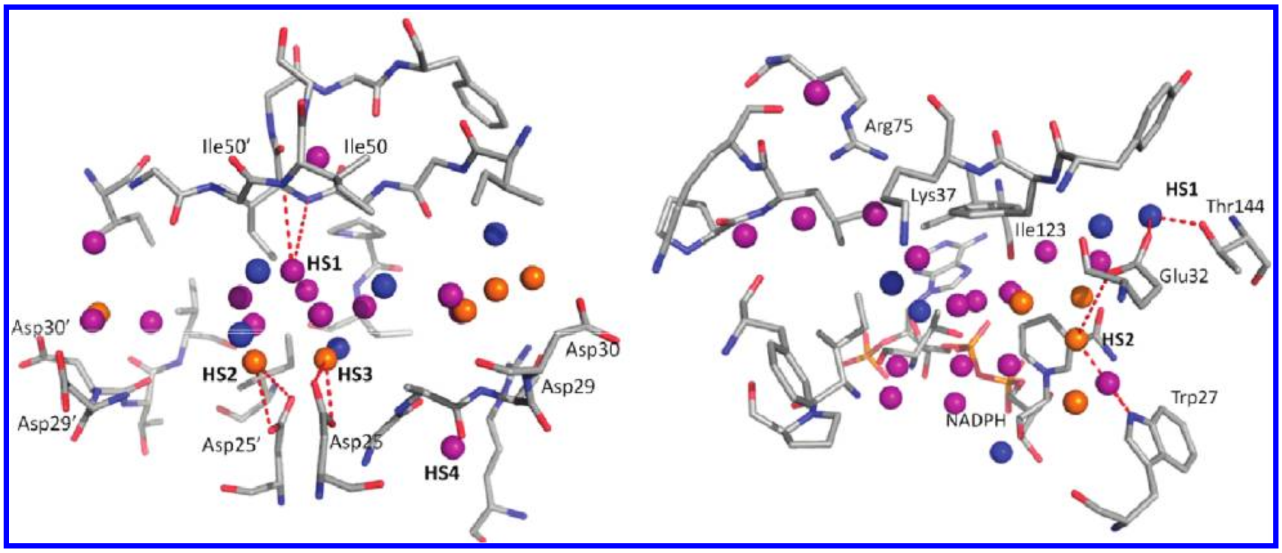


Figure 6. Hydration sites identified in the active site of HIV protease (left) and pcDHFR (right). Hydration sites that contribute both entropically and enthalpically when replaced by ligands are shown in blue. The hydration sites whose entropic gain surpasses the enthalpic loss when replaced by ligands are shown in purple. The hydration sites whose enthalpic loss is larger than the entropic gain when replaced by ligands are shown in orange. Some of the interactions between specific hydration sites and protein residues that are discussed in the text are shown as red dashed lines.

protein structures (Supporting Information S3). An overlay of the hydration sites of the three fXa structures showed that 26 hydration sites were shared by at least two structures and 16 hydration sites were shared by all three structures (Figure 5 right). Compared with the original complex structures (Figure 2), we found that the nonshared hydration sites were either on the surface of the protein or in a narrow cavity not accessible to the ligands. The hydration sites in common are mostly located within the S1 and S4 subpockets in which most ligands bind. Therefore, the starting protein conformation does not seem to significantly influence the identification of hydration sites associated with ligand binding.

Based on the probability distribution functions (PDF) for translational and rotational degrees of freedom of the water molecules in each hydration site, we estimated the entropic cost of transferring a water molecule from the bulk solvent into the hydration site. The enthalpic cost was estimated based on the average sum of van der Waals and electrostatic interactions between the water of interest with its surrounding environment throughout the MD trajectory. A list of the estimated free energy for each hydration site in the three protein systems can be found in Supporting Information S1. Because the water molecules are assumed to be freely rotatable in the bulk solvent (eq 4), each water molecule contributes entropically to the free energy of ligand binding upon being released into the bulk solvent. Therefore, all the identified hydration sites are entropically rewarding if the water molecules in those positions were replaced by the ligand. However, depending on the specific protein environment of the hydration sites, the enthalpic contribution can be either favorable or unfavorable. Therefore, the overall contribution of the replacement of a water molecule by a ligand to the free energy of binding depends on the sum of the enthalpic and entropic effects. For example, the S4 subpocket of fXa contains three aromatic residues, Tyr99, Phe174, and Trp215 (Figure 5 left). The water molecules adjacent to these residues cannot form the same number of hydrogen bonds with the residues as compared to the bulk solvent state. Therefore replacing these water molecules by ligand atoms will potentially contribute both entropically and enthalpically to the free energy of binding. As expected, the water in hydration site HS1 (Figure 5, left) identified in the S4 pocket of 1NFU was estimated to gain 1.14 kcal/mol enthalpically and 2.54 kcal/mol entropically if released into bulk solvent. On the other hand, water molecules in the hydration site HS2 can form hydrogen bonds with the surrounding residues Ile175 and Thr98. Releasing a water molecule in this position is estimated to result in an enthalpic loss of 3.44 kcal/mol, which is slightly larger than its entropic gain (2.85 kcal/mol). Therefore, releasing a water molecule in the HS2 position to bulk solvent would not contribute favorably to the free energy of binding. This agrees with the fact that a conserved water molecule was observed in the HS2 position of the *holo* structures[45] (1F0S, 1MQ6, and 1NFU) (Figure 2) as well as the *apo* structure of fXa (1HCG[46]). The hydration site HS3 colocalizes with the WT2 position in 1NFU (Figure 2b). Based on our estimation, the enthalpic loss (1.28 kcal/mol) of the water molecule in HS3 position is not as large as its entropic gain (2.12 kcal/mol). Therefore, releasing the water molecule in HS3 will potentially contribute favorably to the free energy of binding. However, the mediating role of WT2 in 1NFU suggests that the hydrogen bond donor/acceptor groups on the ligands are able to stabilize the water molecules in this position, therefore making the water molecule energetically favorable in the protein−ligand complex. To further confirm our postulate that the starting protein conformation does not significantly influence the identification of the hydration sites associated with ligand binding in fXa, we compared the estimated free energies of the overlapping hydration sites (Supporting Information S2). The standard deviations of the predicted free energies between the overlapping hydration sites of 1F0S and 1MQ6 is 0.69 kcal/mol, between that of 1F0S and 1NFU is 0.83 kcal/mol, and between that of 1MQ6 and 1NFU is 0.48 kcal/mol. This indicates that the estimated energy contributions of the hydration sites are consistent across all three protein structures.

For HIVPR, 26 hydration sites were identified (Figure 6, left). The water in hydration site HS1 surrounding residues Ile50/50′ is estimated to provide a favorable free energy contribution if replaced by the ligand. This is consistent with the results from nonpeptide cyclic urea inhibitors designed to mimic the hydrogen bonding features of water molecules in

Table 2. Number of Protein Pharmacophore Elements Selected for Virtual Screening Using Different Pharmacophore Models[a]

	fXa:1F0S (20)[b]		fXa:1MQ6 (33)[b]		fXa:1NFU (30)[b]		HIVPR:1AJV (26)[b]		pcDHFR:1DAJ (27)[b]	
distance cutoff (Å)	1.0	1.5	1.0	1.5	1.0	1.5	1.0	1.5	1.0	1.5
posHS	7	16	9	25	8	18	8	15	16	30
T0.25	5	12	3	10	3	8	4	6	4	10
T0.50	7	16	7	18	6	11	6	12	7	17
T0.75	10	20	9	25	8	18	8	15	12	26
allHS	10	23	11	32	13	27	13	22	18	35
FPP	148		139		149		131		160	

[a] "posHS" models were constructed using the hydration sites with positive estimated free energies. The hydration sites were also ranked in a descending order based on their estimated free energies. Top 25% (T0.25), 50% (T0.50), 75% (T0.75), and all (allHS) of the hydration sites were used to build hydration-site-restricted pharmacophore models. Two distance cutoffs, 1.0 Å and 1.5 Å, were used to select protein pharmacophore elements. The total number of protein pharmacophore elements in the binding site, i.e. the full protein pharmacophore (FPP) model, is also shown. [b] The values in the parentheses indicate the number of hydration sites identified for each protein system.

these positions.[40] One example is illustrated by 1AJV (Figure 3 right) where a sulfonyl group replaces the conserved water molecule to form hydrogen bonds with Ile50/50′. Six hydration sites are estimated to be energetically stable. Not surprisingly, these hydration sites are next to the charged residues Asp25/25′, Asp29/29′, and Asp30 respectively, where the ability of water molecules to form hydrogen bonds with these residues negates the entropic gain of releasing those water molecules. However, being energetically stable does not necessarily prohibit the replacement of a hydration site by a ligand. For example, in the case of 1AJV (Figure 3, right) the ligand replaced water molecules in hydration sites HS2 and HS3 to form hydrogen bonds with Asp25/25′.

Twenty-seven hydration sites were identified for pcDHFR (Figure 6 right). Two conserved structural waters (WT1 and WT2 in Figure 4 right) near Trp27, Glu32, and Thr144 are present in many pcDHFR X-ray structures.[14,36,37] We found that the hydration site (HS1) near WT1 is both enthalpically (2.09 kcal/mol) and entropically (2.75 kcal/mol) rewarding to the free energy of binding if replaced by the ligand. However the enthalpic loss (2.72 kcal/mol) of hydration site (HS2) near WT2 is larger than its entropic gain (2.33 kcal/mol) therefore making HS2 energetically stable in the binding site.

Enrichment Using Hydration-Site-Restricted Pharmacophore (HSRP) Models. The importance of binding site water molecules and their contribution to the free energy of ligand binding led us to hypothesize that the information derived from the identified hydration sites can be used to reduce the number of protein pharmacophores increasing the efficiency of virtual screening when compared to the FPP model without losing enrichment quality. To test our hypothesis, we selected protein pharmacophores based on their overlap with identified hydrations sites. We first constructed the hydration-site-restricted pharmacophore (HSRP) models based on the hydration sites with positive estimated free energies (posHS), i.e. the hydration sites that are energetically rewarding if the water molecules were replaced by the ligand. All pharmacophore elements within 1.0 Å radius of energetically favorable hydration sites sphere were selected to construct the posHS HSRP models for each of our test systems. Table 2 shows the number of protein pharmacophore elements selected for each model. It is worth noting that not all the hydration sites encompass pharmacophore elements. The most prominent example is in HIVPR where 20 hydration sites were estimated as energetically contributing; however, only eight pharmacophore elements were found to be encompassed by these hydration sites. One potential reason for this mismatch might be the process of generating the hydrophobic pharmacophore elements itself. For performing clique detection throughout the pharmacophore search, the number of protein pharmacophores need to be limited. Thus, the hydrophobic pharmacophores are defined by clustering over hydrophobic grid points using a radius of 2 Å. As such, the 1.0 Å radius criterion for defining overlap between pharmacophore elements and hydration sites may be too restrictive for the selection of protein pharmacophore elements. Therefore, another set of HSRP models was constructed using pharmacophore elements within 1.5 Å of the centers of the hydration sites. Using the 1.0 Å cutoff, AUC values of 0.68, 0.68, and 0.57 were achieved for 1F0S, 1MQ6, and 1NFU of fXa respectively (Figure 7, left). When the distance cutoff was increased to 1.5 Å, the AUC of all three fXa structures increased to 0.73 (Figure 7, right). The AUC for HIVPR and pcDHFR were 0.59 and 0.61 under the 1.0 Å cutoff and 0.58 and 0.57 under the 1.5 Å cutoff. The increase of the distance cutoff did not significantly alter the screening quality for these two protein systems.

In general, we found that the enrichment quality using the posHS HSRP models is comparable to that of using all pharmacophore elements in the binding site (FPP model) for three out of five protein structures, but lacks in enrichment quality for fXa:1NFU and HIVPR. However, the posHS HSRP models provide a large improvement in virtual screening efficiency. Depending on the protein system, the posHS HSRP models are 300−600 times more efficient than using the FPP models (Table 3). This makes our HSRP model more attractive than the FPP model for screening virtual libraries with hundreds of thousands to millions of compounds. However, to further improve the enrichment quality of our HSRP model, we addressed the question whether using only hydration sites with positive estimated free energies is too restrictive in selecting key pharmacophore elements for ligand binding. As discussed before, the ligands can potentially replace water molecules in energetically stable hydration sites to form hydrogen bonds with the protein (example of HS2 and HS3 in HIVPR). The ligands can also stabilize water molecules in an enthalpically unstable hydration site by forming hydrogen bonds with the water molecules (example of WT2 in fXa:1NFU). Furthermore, the underlying molecular mechanics approach which utilizes a classical force field may be limited in its accuracy in computing the enthalpy and entropy of desolvation. In a recent study, Friesner and co-workers developed the so-called "displaced-solvent functional" to estimate the contribution of each hydration site to the binding free energy.[9] The "*ab initio*" form of the functional uses the estimated excess entropy and enthalpy from the thermodynamics calculation directly in defining the rewarding hydration sites. They also trained their functional on a set of known fXa inhibitors to find cutoff values in selecting the energetically rewarding hydration sites.

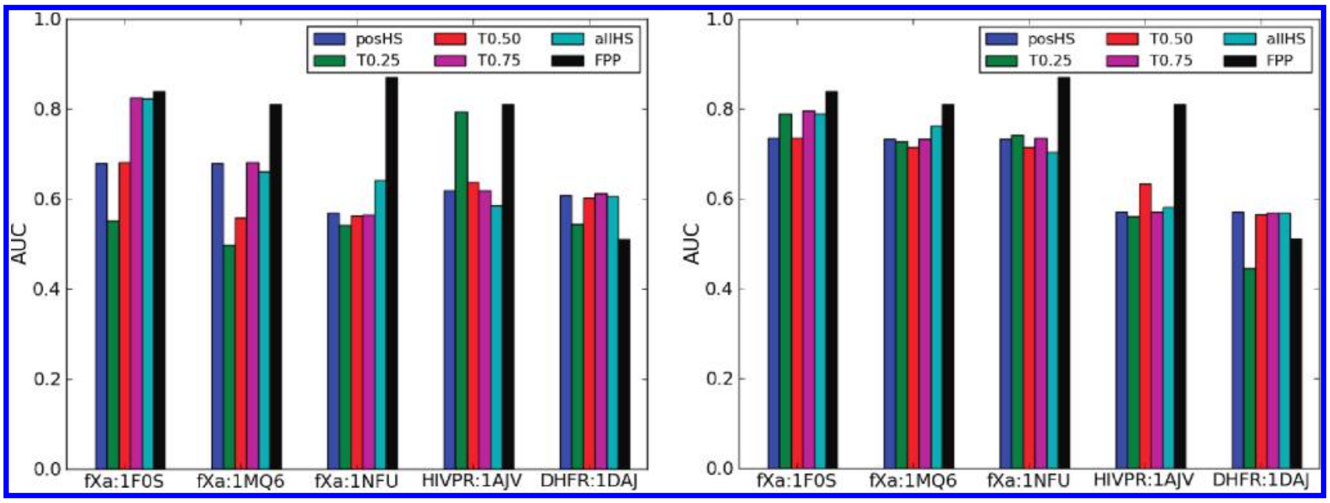


Figure 7. Comparison of AUC values from different virtual screening studies using HSRP models at 1.0 Å cutoff (left) and 1.5 Å cutoff (right).

Table 3. Runtime (s) per Ligand Needed in Virtual Screening Using Different Pharmacophore Models[a]

	fXa:1F0S		fXa:1MQ6		fXa:1NFU		HIVPR:1AJV		pcDHFR:1DAJ	
distance cutoff (Å)	1.0	1.5	1.0	1.5	1.0	1.5	1.0	1.5	1.0	1.5
posHS	1.7	2.0	1.7	5.3	1.6	2.3	1.1	2.1	0.9	1.8
T0.25	1.3	1.4	1.5	1.8	1.2	1.2	1.1	1.1	0.5	0.5
T0.50	1.2	1.5	1.4	2.7	1.2	1.5	1.1	1.7	0.5	1.0
T0.75	1.2	2.1	1.4	5.1	1.2	2.2	1.1	2.1	0.6	1.4
allHS	1.6	2.4	1.6	6.2	1.8	4.3	1.5	4.0	1.5	3.3
FPP	1225.2		1467.6		1205.7		933.9		895.9	

[a] All the virtual screening was run on a single core using either 2.5 GHz quad-core AMD2380 or 2.1 GHz 12-core AMD6172 processors. On average, running the virtual screening on AMD2380 is faster than running on AMD6172 by about 18%.

Their results showed that the ability of the "*ab initio*" form to predict binding affinities is not as good as the prediction using fitted parameters.[9] Therefore, simply using the estimated entropy and enthalpy as absolute cutoffs for defining energetically rewarding hydration sites can fail to identify the most relevant pharmacophore elements and potentially lead to a poor enrichment quality.

As a first attempt to overcome the above-discussed obstacles, instead of using a hard cutoff, we propose to select a number of the hydration sites based on the ranking of their estimated free energies. We first ranked all the hydration sites in a descending order based on their estimated free energies. The top 25% (T0.25), 50% (T0.50), 75% (T0.75) or all of the hydration sites (allHS) of each protein system were chosen and used to select colocalized pharmacophore elements using both 1.0 Å and 1.5 Å cutoffs. The number of selected pharmacophore elements is documented in Table 2. Note that the number of the hydration sites selected by different top percentages can be equal to the number of hydration sites with positive estimated free energy thus resulting in the same HSRP models.

The virtual screening results using the set of HSRP models with 1.0 Å cutoff were compared with the posHS model in Figure 7. For fXa, we observed that as the number of hydration sites used to restrict the pharmacophores was increased the enrichment quality also increased. The most prominent case is in 1F0S where the AUC value for the allHS model (0.82) was nearly identical to that of the FPP model (0.84). It is interesting to note that the hydration sites that are predicted to be energetically stable are also included in constructing the allHS models in 1F0S. This again confirms our postulate that being energetically stable does not necessarily prohibit the replacement of this water from a hydration site by a ligand. We also observed that the AUC values of the other two fXa structures were not as high as that of 1F0S. The top 25% results are comparable among the three fXa structures, but as the top percentage was increased the AUC values for 1MQ6 and 1NFU dropped below those for 1F0S. We attribute this to the larger number of hydration sites identified for these two fXa structures due to their larger defined active sites which might cause a large number of false positives.

Using the 1.5 Å cutoff, the AUC values for all HSRP models approached those of the FPP model for fXa. The differences in AUC values among the three structures are also less significant than they were for the 1.0 Å cutoff models. This supports the postulate that 1.0 Å cutoff is too restrictive in selecting the pharmacophore elements for fXa. We tested whether using 2.0 Å cutoff for the selection of protein pharmacophores would improve the enrichment quality. Those models, however, generated inferior enrichment results even when compared to the models using a 1.0 Å cutoff (data not shown). Therefore, 1.5 Å seems to be the optimal distance cutoff in selecting the pharmacophore models for fXa.

Among the HSRP models for HIVPR using a 1.0 Å cutoff, the one using the top 25% hydration sites showed a comparable AUC value to that of using the FPP model. As the number of the hydration sites being used was increased, the quality of the enrichment decreased. When the 1.5 Å cutoff was used, the enrichment quality in the T0.25 setting was significantly lower compared to the model using a 1.0 Å cutoff, whereas the AUC values for the other settings did not change significantly. A comparison between the selected pharmacophore elements of the T0.25 models using either the 1.0 Å or 1.5 Å cutoff revealed that the major

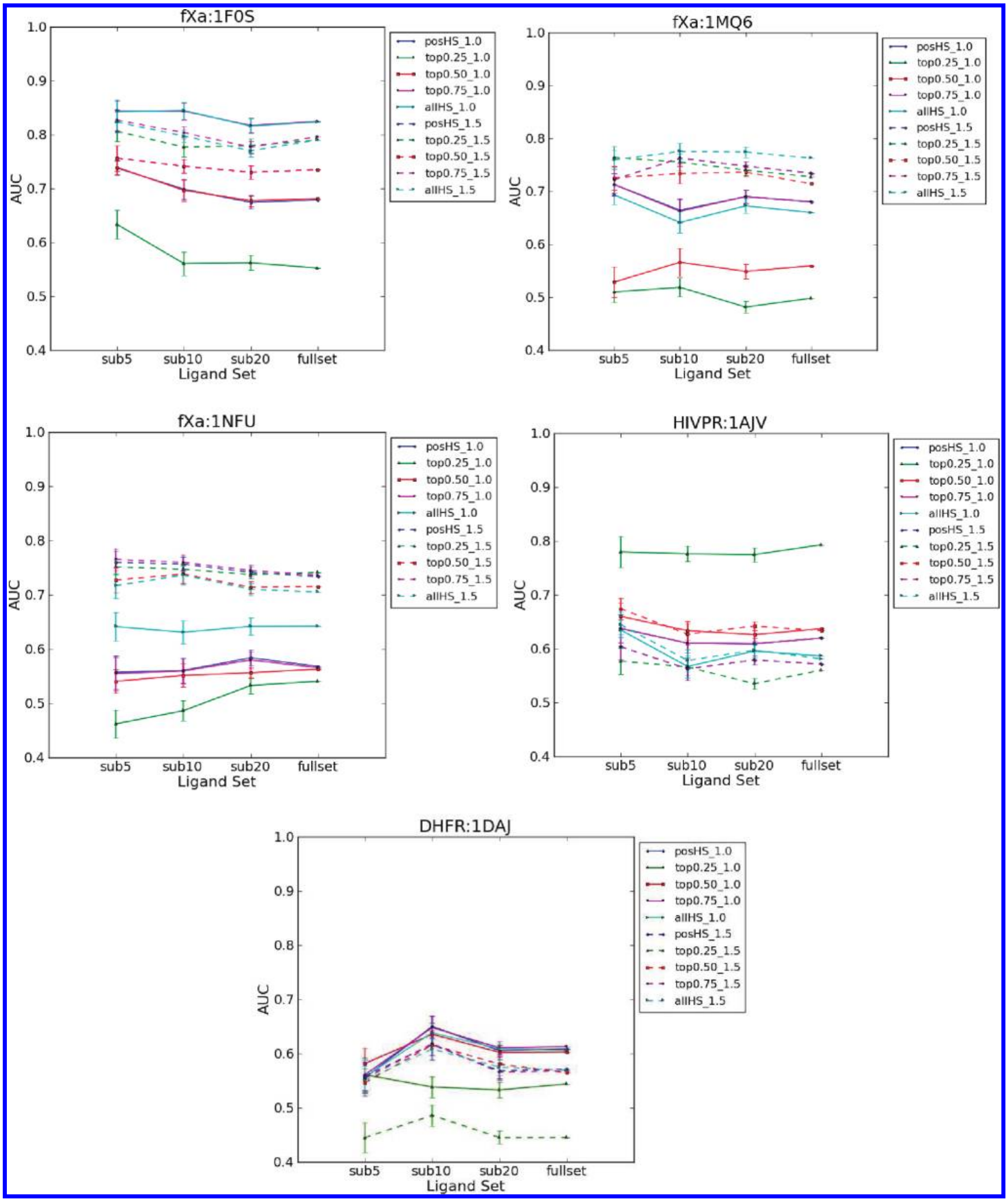


Figure 8. AUC results use different hydration-site-restricted pharmacophore models to screen against random selected subsets. The AUC was plotted against each subset with 5 (sub5), 10 (sub10), or 20 (sub20) active ligands and the full DUD data set. Results of the HSRP models using the 1.0 Å distance cutoff are shown using solid lines and that of the 1.5 Å cutoff are shown using dashed lines. For AUC of each subset, the standard error was computed on 20 randomly sampled ligand subsets.

difference is the inclusion of pharmacophores in HS4 surrounded by residues Ala28, Asp29, Arg87, and Arg8′ (Figure 6 left). In both the *apo* (1G6L[47]) and *holo* (1AJV) structure of the HIVPR, there is a conserved water molecule in this small cavity. However, this cavity does not form the actual binding pocket.[48] Most inhibitors bind in the protease subsites S2–S2′ near the middle of the dimer[49] (Figure 3 right). This explains the decreased enrichment quality when pharmacophore elements near HS4 are included in the model.

Finally, in the case of pcDHFR, the T0.50, T0.75, and allHS HSRP models did not display significant difference in enrichment quality compared to the posHS HSRP models using either a 1.0 Å or 1.5 Å cutoff; the AUC values are around 0.6 using a 1.0 Å cutoff and slightly lower when using a 1.5 Å cutoff. For both cutoffs the enrichment quality is significantly better than that of using all the binding site pharmacophore elements (0.51).

Overall, our studies using the HSRP models suggest that it is possible to find a subset of pharmacophore elements to achieve comparable (fXa and HIVPR) or better (DHFR) enrichment performance to that of using all binding site pharmacophore elements. To confirm that the hydration site information significantly contributes to the selection of relevant pharmacophore elements, we compared the best HSRP models to pharmacophore models with identical size but using random selection to select the pharmacophore elements (Supporting Information S5). We found that the enrichment qualities of the HSRP models are significantly better than those from the randomly selected pharmacophore models.

In conclusion, we constructed HSRP models using different distances and energy cutoffs for three protein systems. In the case of fXa, comparable enrichment qualities to the FPP models are achieved except for fXa:1NFU. In the case of pcDHFR, the HSRP models performed slightly better than the FPP model. It is interesting to speculate on the observed differences in enrichment quality for the different protein systems. One potential reason might be the ratio of number of protein to ligand pharmacophores (see Supporting Information S4 and Table 2). The smallest ratio of protein to ligand pharmacophores is found for HIVPR, where in the HSRP models the number of protein pharmacophores becomes approximately equal to the average number of ligand pharmacophores. Thus the reduction of protein pharmacophores in HSRP models might cause the removal of relevant protein−ligand contacts from the scoring process and consequently lead to a reduction in enrichment quality. In contrast, the largest ratio of protein to ligand pharmacophores is observed for DHFR, i.e. the ligands are much smaller than the binding site of the protein. In this situation, more potential ligand poses can be generated without steric clash of protein and ligand, increasing the potential for false positives. This is consistent with the overall lowest enrichment quality of the FPP models for DHFR. Reducing the number of protein pharmacophores elements in the HSRP models has the highest potential to improve the enrichment quality as observed in our studies.

In addition, a major advantage of using the HSRP models is the improvement of the computational efficiency throughout the virtual screening process. As shown in Table 3, even for the HSRP models with the largest number of protein pharmacophore elements, the efficiency was increased by at least 500 times for 1F0S, 200 times for 1MQ6, 280 times for 1NFU, 200 times for 1AJV, and 270 times for 1DAJ compared to using the FPP models. Using only eight 12-core AMD6172 processors would require less than 20 h to screen a virtual library with 1 million compounds for fXa using the largest HSRP model (Tables 2 and 3).

Training Process. Using the constructed HSRP models, we achieved similar enrichment quality for fXa and HIVPR and better enrichment quality for pcDHFR compared to using the FPP models. However, the settings of the best HSRP models vary dependent on the protein systems. The performance of the constructed HSRP models is also sensitive to the selected pharmacophore elements (as in the example of HIVPR T0.25 models). We attribute this to 1) the uncertainty in the estimated desolvation energies of the hydration sites, which influences the selection of protein pharmacophores for HSRP models; 2) the neglect of other factors which contribute to the importance of the different protein pharmacophores for ligand binding such as the strength of ligand interaction with the protein residues, potential water-mediated effects and the contributions from ligand desolvation to the free energy of binding; and 3) the dynamics of the protein structures which contributes to the uncertainty in selecting the pharmacophore elements near the hydration sites. Thus, a uniform optimal setting for different protein systems should not be expected. Therefore, we addressed the question whether we can find a separate optimal setting for each system by training with a small set of known active and decoy ligands. The underlying idea is that if a small set of known active and decoy ligands is available virtual screening can be first performed on this small training set. The HSRP model that performs best on the small set can then be used for screening large compound libraries.

To test this idea, we studied whether the optimal HSRP model identified by virtual screening on a small subset of the DUD data set is consistent with the best HSRP model for the full DUD data set. We randomly selected a series of ligand subsets with the same active-to-decoy ratio as in the full set. Virtual screening using each HSRP model was then performed on these subsets. The overall performance indicated by the AUC values was plotted for each subset size in Figure 8. In general, the HSRP model with the highest AUC value for the subsets also achieved the best enrichment quality on the full DUD data set. For example, for HIVPR the T0.25 HSRP model with a 1.0 Å cutoff performs significantly better than any other model on all subset sizes and the full DUD data set. We also observe that multiple HSRP models have similar enrichment qualities on the small subsets. For example, in the case of 1NFU, virtual screening results on subsets with 5, 10, or 20 active ligands suggest a similar performance of the posHS, T0.25, and T0.75 HSRP models using a 1.5 Å cutoff. Indeed, the enrichment quality using these models also displayed no significant difference on the full DUD data set. Overall, our results suggest that we can reliably identify optimal HSRP models by using a small training sets of known active and decoy ligands.

■ CONCLUSIONS

We presented a new concept to select protein pharmacophore elements important for ligand binding using hydration site analysis. The underlying hypothesis is that water molecules that are replaced by the ligand can significantly contribute to the free energy of ligand binding if those water molecules gain free energy upon release to bulk solvent. Thus, protein pharmacophore elements spatially colocalized with those hydration sites can be important for ligand binding. Tests of this concept on five different protein structures (three protein systems) revealed that HSRP models can be identified that display similar or only slightly worse enrichment quality for most protein systems but that a training process on a small set of known actives and decoys might be necessary to select the optimal settings. While no significant improvement in enrichment quality was observed using HSRP models, the major advantage is that the virtual screening efficiency is drastically improved by a factor of 200−500 compared to using all protein pharmacophores. Thus, the reduction of the size of the pharmacophore model allows HSRP to be utilized for screening of large ligand libraries.

Some limitations of utilizing HSRP models, however, should be noted. First, this concept requires initial MD simulations of the ligand-unbound form of the protein. However, those simulations need to be performed only once. Second, the training process requires knowledge of a small set of actives for the target protein of interest. If such data are not available, robust but not optimal enrichment can nevertheless be obtained using HSRP models without the training process. Third, the focus on hydration sites as pharmacophore selection criteria neglects important contributions of protein−ligand binding. For example, the strength of hydrogen bonds between protein and ligand can vary due to the protein and ligand environment. Thus, hydrogen bonding pharmacophores will have different importance independent of the hydration site data. Strong direct interactions between protein and ligand or conformational restraints on the binding pose of the ligand can cause the replacement of energetically favorable water molecules, as shown for HIVPR for HS2 and HS3 (Figure 6). Also neglected are potential water mediated protein−ligand interactions. This problem could be addressed by including pharmacophores representing water-mediated interactions. In future work, we plan to extend our current concept of HSRP to incorporate these effects so as to generate even more robust protein pharmacophores models.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Estimated free energies for the hydration sites of the studied protein systems; overlay of the hydration sites identified in three structures of fXa; alignment of the binding site residues in the three fXa structures; distribution of molecular weight and number of ligand pharmacophore elements for all actives of fXa, HIVPR, and DHFR; comparison of the optimal Hydration Site Restricted Pharmacophore model of each protein structures to Randomly-Selected Protein Pharmacophore models with identical number of pharmacophore elements. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: (765) 496-9375. Fax: 765 494-1414. E-mail: mlill@purdue.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Laura Kingsley and Gregory Wilson for critical reading of the manuscript. M.A.L. thanks the National Institutes of Health (GM092855) for funding the present research.

■ REFERENCES

(1) Martin, Y., Distance comparisons (DISCO): a new strategy for examining 3D structure-activity relationships. In *Classical and Three-Dimensional QSAR in Agrochemistry*; Hansch, C., Fujita, T., American Chemical Society: Washington, DC, 1995; pp 318−329.

(2) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (3), 563−571.

(3) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20* (10), 647−671.

(4) Richmond, N. J.; Abrams, C. A.; Wolohan, P. R. N.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput.-Aided Mol. Des.* **2006**, *20* (9), 567−587.

(5) Chen, X.; Rusinko, A. III; Tropsha, A.; Young, S. S. Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 887−896.

(6) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model* **2005**, *45* (1), 160−169.

(7) Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. Application of structure-based focusing to the estrogen receptor. *J. Comput. Chem.* **2001**, *22* (10), 993−1003.

(8) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein−ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (3), 808−813.

(9) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008**, *130* (9), 2817−2831.

(10) Maignan, S.; Guilloteau, J. P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Crystal structures of human factor Xa complexed with potent inhibitors. *J. Med. Chem.* **2000**, *43* (17), 3226−3232.

(11) Adler, M.; Kochanny, M. J.; Ye, B.; Rumennik, G.; David, R.; Biancalana, S. Whitlow, M., Crystal structures of two potent nonamidine inhibitors bound to factor Xa. *Biochemistry* **2002**, *41* (52), 15514−15523.

(12) Maignan, S.; Guilloteau, J. P.; Choi-Sledeski, Y. M.; Becker, M. R.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Molecular structures of human factor Xa complexed with ketopiperazine inhibitors: preference for a neutral group in the S1 pocket. *J. Med. Chem.* **2003**, *46* (5), 685−690.

(13) Bäckbro, K.; Löwgren, S.; Österlund, K.; Atepo, J.; Unge, T.; Hultén, J.; Bonham, N. M.; Schaal, W.; Karlén, A.; Hallberg, A. Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *J. Med. Chem.* **1997**, *40* (6), 898−902.

(14) Cody, V.; Galitsky, N.; Luft, J.; Pangborn, W.; Gangjee, A.; Devraj, R.; Queener, S.; Blakley, R. Comparison of ternary complexes of Pneumocystis carinii and wild-type human dihydrofolate reductase with coenzyme NADPH and a novel classical antitumor furo [2, 3-d] pyrimidine antifolate. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1997**, *53* (6), 638−649.

(15) Word, J.; Lovell, S.; Richardson, J.; Richardson, D. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. *J. Mol. Biol.* **1999**, *285* (4), 1735−1747.

(16) Case, D. A.; Cheatham, T. E. III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M. Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26* (16), 1668−1688.

(17) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789−6801.

(18) *OMEGA*, version 2.2.1; OpenEye Scientific Software, Inc.: Santa Fe, NM, USA, 2010. www.eyesopen.com (accessed month day, year).

(19) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47* (1), 45−55.

(20) Zavodszky, M. I.; Kuhn, L. A. Side chain flexibility in protein−ligand binding: The minimal rotation hypothesis. *Protein Sci.* **2005**, *14* (4), 1104−1114.

(21) Zhao, Y.; Sanner, M. F. FLIPDock: docking flexible ligands into flexible receptors. *Proteins: Struct., Funct., Bioinf.* **2007**, *68* (3), 726−737.

(22) Lill, M. A. Efficient incorporation of protein flexibility and dynamics into molecular docking simulations. *Biochemistry* **2011**, *50* (28), 6157−6169.

(23) Xu, M.; Lill, M. A. Significant Enhancement of Docking Sensitivity Using Implicit Ligand Sampling. *J. Chem. Inf. Model.* **2011**, *51* (3), 693−706.

(24) Berendsen, H. J. C.; Postma, J. P. M; Van Gunsteren, W. F.; DiNola, A.; Haak, J. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(25) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425−445.

(26) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins: Struct., Funct., Bioinf.* **1998**, *33* (3), 367−382.

(27) Berendsen, H.; Postma, J.; Van Gunsteren, W.; Hermans, J. Interaction models for water in relation to protein hydration. *Intermol. Forces* **1981**, *331*, 331−342.

(28) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7* (8), 306−317.

(29) Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52* (2), 255−268.

(30) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31* (3), 1695−1697.

(31) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52* (12), 7182−7190.

(32) 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.

(33) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103* (19), 8577−8593.

(34) Ben-Naim, A. *Statistical thermodynamics for chemists and biochemists*; Plenum Press: New York, 1992.

(35) Minh, D.; Bui, J.; Chang, C.; Jain, T.; Swanson, J.; McCammon, J. The entropic cost of protein-protein association: a case study on acetylcholinesterase binding to fasciculin-2. *Biophys. J.* **2005**, *89* (4), L25−L27.

(36) Lazaridis, T. Solvent reorganization energy and entropy in hydrophobic hydration. *J. Phys. Chem. B* **2000**, *104* (20), 4964−4979.

(37) Bron, C.; Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* **1973**, *16* (9), 575−577.

(38) Harley, E. R. Graph algorithms for assembling integrated genome maps. University of Toronto: 2003.

(39) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32* (5), 922−923.

(40) Young, R. J.; Campbell, M.; Borthwick, A. D.; Brown, D.; Burns-Kurtis, C. L.; Chan, C.; Convery, M. A.; Crowe, M. C.; Dayal, S.; Diallo, H. Structure-and property-based design of factor Xa inhibitors: pyrrolidin-2-ones with acyclic alanyl amides as P4 motifs. *Bioorg. Med. Chem. Lett.* **2006**, *16* (23), 5953−5957.

(41) Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P. M.; Schneider, D.; Müller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V. Design and quantitative structure-activity relationship of 3-amidinobenzyl-1 H-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J. Med. Chem.* **2002**, *45* (13), 2749−2769.

(42) Wlodawer, A.; Erickson, J. W. Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* **1993**, *62* (1), 543−585.

(43) Cody, V.; Galitsky, N.; Rak, D.; Luft, J. R.; Pangborn, W.; Queener, S. F. Ligand-induced conformational changes in the crystal structures of Pneumocystis carinii dihydrofolate reductase complexes with folate and NADP+. *Biochemistry* **1999**, *38* (14), 4303−4312.

(44) Champness, J.; Achari, A.; Ballantine, S.; Bryant, P.; Delves, C.; Stammers, D. The structure of *Pneumocystis carinii* dihydrofolate reductase to 1.9 Å resolution. *Structure* **1994**, *2* (10), 915−924.

(45) Kamata, K.; Kawamoto, H.; Honma, T.; Iwama, T.; Kim, S. H. Structural basis for chemical inhibition of human blood coagulation factor Xa. *Proc. Natl. Acad. Sci.* **1998**, *95* (12), 6630−6635.

(46) Tulinsky, A.; Padmanbhan, K.; Padmanbhan, K.; Park, C.; Bode, W.; Huber, R.; Blankenship, D.; Cardin, A.; Kiesel, W. Structure of Human des (1− 45) Factor Xa at 2.2 Å Resolution. *J. Mol. Biol.* **1993**, *232*, 947−966.

(47) Pillai, B.; Kannan, K.; Hosur, M. 1.9 Å x-ray study shows closed flap conformation in crystals of tethered HIV-1 PR. *Proteins: Struct., Funct., Bioinf.* **2001**, *43* (1), 57−64.

(48) Louis, J. M.; Ishima, R.; Torchia, D. A.; Weber, I. T. HIV-1 protease: structure, dynamics, and inhibition. *Adv. Pharmacol.* **2007**, *55*, 261−298.

(49) Krohn, A.; Redshaw, S.; Ritchie, J. C.; Graves, B. J.; Hatada, M. H. Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere. *J. Med. Chem.* **1991**, *34* (11), 3340−3342.