JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Identification of Protein–Ligand Binding Sites by the Level-Set Variational Implicit-Solvent Approach

Zuojun Guo,[†] Bo Li,[‡] Li-Tien Cheng,[§] Shenggao Zhou,[‡] J. Andrew McCammon,[∥] and Jianwei Che*[,†]

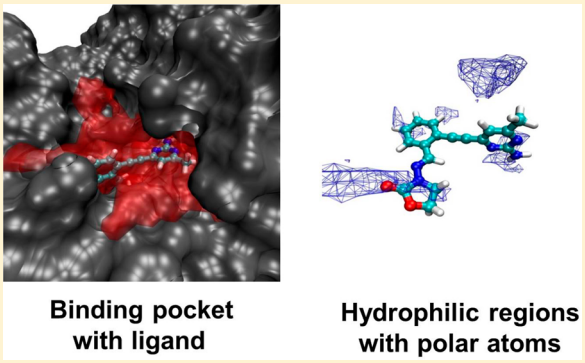
[†]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, United States

[‡]Department of Mathematics and Center for Theoretical Biological Physics, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0112, United States

[§]Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0112, United States

[∥]Department of Chemistry and Biochemistry, Department of Pharmacology, Howard Hughes Medical Institute, and Center for Theoretical Biological Physics, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0365, United States

ABSTRACT: Protein–ligand binding is a key biological process at the molecular level. The identification and characterization of small-molecule binding sites on therapeutically relevant proteins have tremendous implications for target evaluation and rational drug design. In this work, we used the recently developed level-set variational implicit-solvent model (VISM) with the Coulomb field approximation (CFA) to locate and characterize potential protein–small-molecule binding sites. We applied our method to a data set of 515 protein–ligand complexes and found that 96.9% of the cocrystallized ligands bind to the VISM-CFA-identified pockets and that 71.8% of the identified pockets are occupied by cocrystallized ligands. For 228 tight-binding protein–ligand complexes (i.e, complexes with experimental pK_d values larger than 6), 99.1% of the cocrystallized ligands are in the VISM-CFA-identified pockets. In addition, it was found that the ligand binding orientations are consistent with the hydrophilic and hydrophobic descriptions provided by VISM. Quantitative characterization of binding pockets with topological and physicochemical parameters was used to assess the "ligandability" of the pockets. The results illustrate the key interactions between ligands and receptors and can be very informative for rational drug design.



1. INTRODUCTION

Many biological processes, such as signal transduction, cell regulation, and the immune response, involve protein–ligand binding. The identification and characterization of protein binding sites for small molecules are crucial to the understanding of the functions of both endogenous ligands and drug molecules. Despite our increased knowledge about proteins, there are many proteins with unknown binding sites. Even for the ones with known sites, it is still possible to find additional binding sites (e.g., allosteric binding sites) that provide new means to modulate the protein function.

There are about 30 000 genes in the human genome. It is speculated that only 10% are amendable for small-molecule modulation. Among them, less than half have therapeutic potential.[1] Analysis of data from major pharmaceutical companies indicates that less than 2% of projects succeed to get drugs to market throughout the discovery and development phases. This poor success rate is mainly associated with two central problems: the identification and validation of disease-specific targets and the development of specific molecules that can modulate these targets with good therapeutic windows.[2] Because some biological targets are not "druggable", ∼60% of small-molecule drug discovery projects fail at the stage of "hit-to-lead".[2] Knowing the locations and physicochemical properties of the target protein binding sites prior to screening or optimization offers tremendous benefits with respect to time and cost.

Most marketed small-molecule drugs target protein–ligand interactions (PLIs). Recently, the demand for bioavailable protein–protein interaction (PPI) modulators is growing.[3,4] For PLI, the analysis of available crystallographic structures has indicated that most small molecules prefer hydrophobic protein pockets with more complex topological features than typical protein surfaces.[5−7] Small-molecule binding sites usually consist of deep concave pockets that can maximize favorable protein–ligand contacts. Similar to the cores of proteins, the binding sites of small molecules are often made of hydrophobic residues that could positively contribute to the binding of organic molecules in aqueous environments. On the other hand, general PPI sites are relatively large and flat with both polar and apolar residues (700–2000 Å^2 contact surface area).[4,8−11] Historically, these interfaces are considered to be "undruggable" by small molecules. With detailed characterization of unique aspects of the interaction surfaces, however, some encouraging breakthroughs have been achieved in the past decades for

Received: September 26, 2014
Published: January 7, 2015

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

certain cases.[4,12] Recently, 12 small molecules targeting PPIs, such as the inhibitors for p53−MDM2, BCL-2 family−BH3 domain, and tubulin-α−tubulin-β interactions,[13] have been clinically developed.

In principle, binding sites at protein surfaces can be detected experimentally. Hajduk and co-workers used heteronuclear-NMR-based screening to identify and characterize hot spots on protein surfaces.[14] By screening of a large number of diverse "fragmentlike" or "leadlike" compounds (approximately 10 000) against 23 target proteins, this method predicted that 90% of the ligands bind to specific locations on the protein surface. The NMR hit rates of particular sites showed high correlation with the probability of finding high-affinity ligands. The hit rates were also correlated with the pocket apolar surface area (with low-hit-rate pockets having ∼35% lower apolar surface area) for the known pockets. On the basis of the results for a large number of diverse compounds and targets, it is believed that certain properties of small-molecule binding sites should be common to general molecular recognition. Seco et al.[15] used molecular dynamics (MD) simulations with explicit binary water/organic solvent to study this phenomenon. Isopropyl alcohol (iPrOH) molecules were used to represent generic druglike small molecules with relatively high diffusion coefficients. After sufficient sampling, iPrOH molecules tended to replace water at the protein binding hot spots. The local density of iPrOH was proportional to the interaction strength with the protein.[15]

Most of the computational studies to identify protein active sites can be categorized into three major classes: geometry-based cavity detection algorithms, energetics-based methods, and pocket physicochemical property-based analysis. Geometric algorithms have been developed to detect putative binding pockets since the early days. Examples are POCKET,[16] SURFNET,[17] APROPOS,[18] LIGSITE,[19] CAST,[6] PASS,[20] CASTp,[21] etc. These methods are all based on steric complementarity and involve moving probes with different sizes around the protein surface to detect the accessible and inaccessible regions.[22] Various algorithms have been reviewed elsewhere.[23−27] When the binding sites have well-defined pockets, these methods are fast and accurate. PocketFinder by Abagyan and co-workers expands the geometric method by contouring a smoothed van der Waals (vdW) potential for the target protein to identify candidate ligand binding sites.[28] This method partially accounts for nonpolar energetic properties but neglects electrostatic and desolvation effects. SiteMap, developed by Schrödinger, Inc., identifies potential binding sites by linking together "site points" that are likely to contribute to tight protein−ligand or protein−protein binding. This method showed >96% sensitivity in a validation test of 538 proteins with cocrystallized ligand.[29,30]

More sophisticated methods combining physical properties and knowledge-based properties have also been developed.[31,32] One of the caveats for most of these methods is their relatively low specificity. They are able to find most small-molecule binding sites, but they also identify many other functionally irrelevant pockets (i.e., false positives) as potential binding sites. For any novel protein, it is a challenge to separate the correct pockets from multiple false positives.

Our current work is based on the recently developed level-set variational implicit-solvent model (VISM) with the Coulomb field approximation (CFA). In this VISM-CFA approach, the molecular solvation process is described by minimization of a solvation free energy functional of all possible solute−solvent interfaces. The final stable equilibrium solute−solvent surface balances the interactions among various solvation contributions, including surface tension, vdW interactions, and electrostatics. Comparison between the local geometries of the equilibrium VISM surface and the protein molecular surface illustrates the strength of local hydrophobicity or hydrophilicity near the protein surface. In principle, the balanced description by VISM surfaces can characterize potential small-molecule binding pockets on target proteins. With such a physics-based method, we aim to describe the properties of generic protein surfaces where a druglike small molecule could potentially bind and to provide a geometrical and physicochemical characterization of these regions. All of the protein−ligand complexes used in this study were cocrystallized structures. A set of topological and energetic parameters from quantitative analysis of the protein−ligand binding sites were used to predict the target protein "ligandability". We believe that this method can be very helpful for rational drug design and target evaluation.

2. MATERIALS AND METHODS

2.1. Materials and Preparation. Our data set consisted of 515 biologically relevant protein−ligand complexes from the PDBbind database. The proteins belong to 40 families, among which the top 10 are shown in Table 1. Within the data set, subsets of 228 proteins with experimental pK_d values larger than 6 were used in a tight-binding case study.

Table 1. Top 10 Functional Types of Proteins Studied in the Data Set Containing 515 Complexes

functional type	count (PDB)
hydrolase	298
transferase	26
chaperone	21
transporter	20
isomerase	17
transcription factor	15
oxidoreductase	12
aspartyl protease	11
hormone receptor	7
lyase	4

We used the "Protein Preparation" workflow in Maestro[33] to prepare the input files from original Protein Data Bank (PDB) files. Hydrogen atoms were added, N- and C- termini were incorporated, missing side chains were added, and appropriate bond orders were assigned. Then, impref was used to relax the complexes and remove unphysical contacts with the heavy atoms constrained. Cocrystallized ligands and the corresponding protein targets were separated with the standard Schrödinger protocol (i.e., pv_convert.py). The numbers of heavy atoms for the cocrystallized ligands ranged from 24 to 144. In this data set, we visually checked all 515 complexes and found no ligand that formed a covalent bond with its receptor. Most of the ligands (72%) were small organic molecules with molecular weight (MW) less than 600 Da, and we also included 92 known drugs in their respective targets in our study; 72 ligands were peptides or peptide-like small molecules, and 10 ligands were nucleotides. The total number of atoms for each target protein ranged from 1100 to 9200 with cocrystallized ligands. Waters play a crucial role in the mediation of protein−ligand and protein−protein interactions.[34−37] Proteins are covered with water molecules with varying degrees of tightness

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

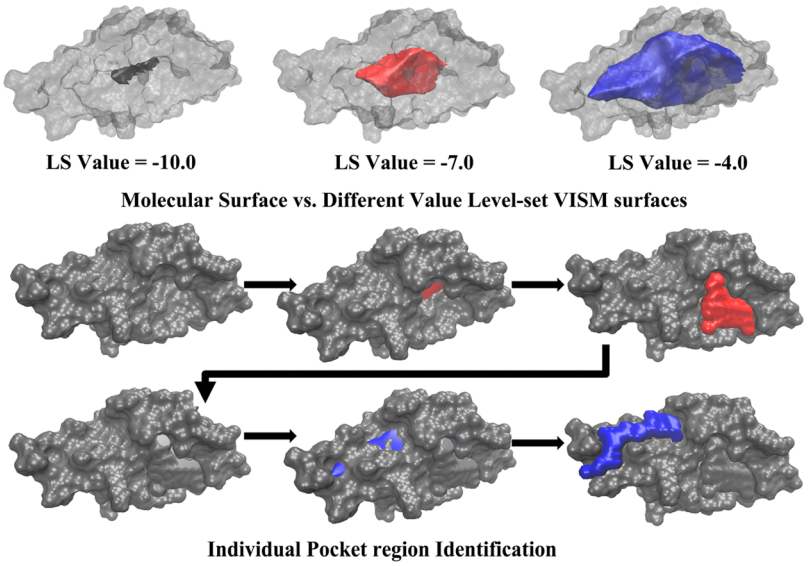


Figure 1. Definition of the binding pocket region by the differences between the molecular surface (gray) and VISM surfaces obtained using a loose initial surface.

according to the physicochemical properties of the protein surface and the ionic conditions. When a ligand binds to a protein surface, it must displace bound waters in that region to interact with the protein directly. Understanding of the behavior of water molecules near a protein surface greatly improves the characterization of protein−ligand binding.[38] In our study, we analyzed the VISM equilibrium surface to obtain hydrophobic pockets near the protein surface. The treatment of water molecules as an implicit solvent in this theory averages water−protein and water−water interactions (i.e., the potential of mean force is used). Explicit consideration of the solvent in protein binding sites can potentially provide refined insights into protein−ligand binding.[15,39]

For the target "ligandability" analysis, 27 protein targets described by Cheng et al.[31] were used. All of these targets were also contained in the whole data set. For each target, one representative crystallographic structure was chosen. The targets were angiotensin-converting enzyme 1 (ACE-1) (PDB entry 1o86), acetylcholinesterase (1gpk), aldose reductase (1pwl), DNA gyrase B (1kij), cyclooxygenase 2 (4cox), cAbl kinase (1iep), EGFR kinase (1m17), P38 kinase (1kv1), cyclin-dependent kinase 2 (CDK2) (1ke9), enoyl reductase (1c14), HIV reverse transcriptase (RT) NNRTI site (1ep4), HIV RT nucleotide site (1t03), HIV-1 protease (1hvr), fungal Cyp51 (1ea1), HMG CoA reductase (1hwi), IMPDH (1nf7), phosphodiesterase 4D2 (PDE-4D) (1oyn), phosphodiesterase 5A (PDE-5A) (1udt), penicillin binding protein (PBP) (1qmf), thrombin (1ktt), neuraminidase (1a4q), factor Xa (1ezq), MDM2 (1rv1), protein-tyrosine phosphatase 1B (PTP-1B) (1g1f), cathepsin K (1nlj), HIV integrase (1qs4), and caspase 1 (ICE-1) (1bmq).

2.2. Level-Set Implementation of VISM-CFA. The level-set implementation of VISM-CFA has been described extensively in previous publications.[40−44] In short, we optimize the free energy of the solvation system, G, as a functional of all possible solute−solvent interfaces Γ:

$$G[\Gamma] = P\mathrm{Vol}(\Omega_m) + \int_\Gamma \gamma_0(1 - 2\tau H)\, \mathrm{d}S + \rho_w \sum_{i=1}^{N} \int_{\Omega_w} U_i(|\mathbf{x} - \mathbf{x}_i|)\, \mathrm{d}V + \frac{1}{32\pi^2\varepsilon_0}\left(\frac{1}{\varepsilon_w} - \frac{1}{\varepsilon_m}\right)\int_{\Omega_w}\left|\sum_{i=1}^{N}\frac{Q(\mathbf{x} - \mathbf{x}_i)}{|\mathbf{x} - \mathbf{x}_i|^3}\right|^2 \mathrm{d}V \quad (1)$$

where there are assumed to be N solute atoms located at $\mathbf{x}_1$, ..., $\mathbf{x}_N$ inside the solute region Ω_m with point charges Q_1, ..., Q_N, respectively. The first term, $P\mathrm{Vol}(\Omega_m)$, is the volumetric part of the energy for creating the solute cavity Ω_m, in which P is the pressure difference between the solvent liquid and solute vapor. The second term is the surface energy, where $\gamma(\mathbf{x}) = \gamma_0(1 - 2\tau H(\mathbf{x}))$, in which γ_0 is the constant macroscopic surface tension for a planar solvent liquid−vapor interface, τ is the first-order correction coefficient, often termed the Tolman coefficient,[45] $H(\mathbf{x})$ is the mean curvature, defined as the average of the two principal curvatures, and S is the interface area. The third term is the energy of the vdW interaction between the solute atoms and the continuum solvent. The bulk solvent density ρ_w was set to 0.0333 Å^{-3}. The last term represents the electrostatic contribution to the solvation free energy. It is defined by the Born cycle[46] as the difference between the energies of the vacuum and solvated states, where ε_0 is the vacuum permittivity, ε_m is the relative permittivity of the solute molecule, and ε_w is the relative permittivity of the solvent.

To minimize the free-energy functional (eq 1), an initial surface that encloses all of the solute atoms located at $\mathbf{x}_1$, ..., $\mathbf{x}_N$ is chosen. In this pocket-finding study, we chose a loose initial surface in which the closest solute atom (from the edge of the vdW sphere) was at least 1.5 water diameters away from the surface. The initial interface can have a very large value of the free energy. The system is subsequently moved in the direction of steepest descent of the free energy by the level-set method until a minimum is reached. We performed the level-set VISM-CFA calculations for the target proteins after removing the

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

cocrystallized ligands. The partial charges and Lennard-Jones (LJ) 12−6 potential parameters of solute atoms were obtained from the Amber force field; the TIP3P water LJ parameter ε_{ww} = 0.152 kcal/mol and solvent molecular diameter σ_{ww} = 3.15 Å were used. We set the macroscopic planar surface tension as γ_0 = 0.076 kcal mol^{-1} Å^{-2} at 300 K, which was obtained from the TIP3P water simulation.[47] We chose the Tolman coefficient to be τ = 1 Å for the convex and concave atomic-level surface tension correction. It should be noted that for consistency we used the same VISM parameter as in previous studies.[37,43]

2.3. Identification of Putative Binding Pockets from Equilibrium VISM-CFA Surfaces. In previous studies,[37] we found that the stable equilibrium VISM surface resembles the predefined solvent-accessible surface near protein polar and convex molecular surfaces. However, the VISM surface differs from the molecular surface in the concave and hydrophobic regions (i.e., the binding pocket) because of the relatively strong surface tension and weak attractive polar interactions. The unique features captured by VISM surfaces are consistent with those from the analysis of known small druglike molecular binding sites.[5−7] In this part, we describe a method to identify the putative binding pockets of target proteins and extract the regions for further characterization using VISM-CFA.

The basic concept is illustrated in Figure 1. The gray transparent surface represents the protein molecular surface. The black, red, and blue surfaces represent VISM isosurfaces with different level-set values. The equilibrium solute−solvent interface is represented by the zero-level-set surface. In the first row of Figure 1, the opaque black, red, and blue surfaces are "contracted" VISM isosurfaces with lower level-set values. The level-set value equals the distance (in units of Å) of the "contracted" VISM surface from the equilibrium zero-level-set VISM surface. Negative values represent distances from the VISM surface toward the inside of the solute, and positive values represent distances in the other direction. We grow the level-set value from the center of the molecule. By comparing the appropriate VISM surface with the molecular surface, one can readily identify potential binding sites (the seed of the first one is shown as a tiny red tip in the middle structure in the second row of Figure 1). In practice, each pocket is identified and refilled from this deepest region until a "water level" defined by the equilibrium (zero-level-set) VISM surface is reached. In addition, through the different VISM level-set values from the beginning level to the final refilled level, one can characterize the individual pockets, such as the depth. In VISM theory, the surface defines the dielectric boundary between the solvent and solute. Our previous study demonstrated that the VISM surface tracks the first solvation shell nicely and encloses the ligand binding site. After the first pocket region is completely identified, we pave that region and make it part of the molecular surface (the first structure in the third row). As the level-set value continues to grow, a second pocket seed (the blue region in the middle structure in the third row) is obtained. Iterations are implemented as in the process to obtain the first pocket region until we find all of the putative binding pockets for individual targets.

2.4. Characterization of Identified Binding Pockets by Topological and Physicochemical Parameters. *A. Pocket Topological Parameters.* We define the pocket depth as the largest distance along the normal direction from the stable equilibrium VISM surface to the protein molecular surface. The pocket solvent-accessible surface area (SASA) is obtained by rolling a sphere with a radius of 1.4 Å on the protein molecular surface of atoms in the pocket region. The pocket volume is calculated by counting the grid points inside the identified pocket, with each accounting for a volume of 0.8 × 0.8 × 0.8 Å^3. The three principal axes of the pocket, represented by the principal moments of the inertia matrix of the pocket grids, are used to characterize the overall shape of the pocket. For example, three similar principal moments of inertia indicate a spherical-like pocket.

B. Pocket Physicochemical Parameters. The parameters used to characterize binding pockets include local hydrophilicity, hydrophobic SASA fraction, dehydration penalties (vdW, surface, electrostatic, and total), and predicted optimal binding affinities for small molecules. The total solvation free energy in the VISM theory (eq 1) is considered to be the sum of the corresponding solvation free energy density φ over the entire space. The solvation free energy density is heterogeneously distributed around the protein surface. It can be used to characterize the relative hydrophobicity or hydrophilicity. Here, we define the solvation energy density as

$$\varphi_{\text{solute-solvent}}(\mathbf{x}) = \varphi_{\text{vdW}}(\mathbf{x}) + \varphi_{\text{elec}}(\mathbf{x}) \tag{2}$$

where the vdW solute−solvent solvation energy density is given by

$$\varphi_{\text{vdW}}(\mathbf{x}) = \rho_{\text{w}} \sum_{i=1}^{N} U_i(|\mathbf{x} - \mathbf{x}_i|) \tag{3}$$

and the electrostatic solute−solvent solvation energy density is given by

$$\varphi_{\text{elec}}(\mathbf{x}) = \frac{1}{32\pi^2\varepsilon_0}\left(\frac{1}{\varepsilon_{\text{w}}} - \frac{1}{\varepsilon_{\text{m}}}\right)\left|\sum_{i=1}^{N} \frac{Q(\mathbf{x} - \mathbf{x}_i)}{|\mathbf{x} - \mathbf{x}_i|^3}\right|^2 \tag{4}$$

The estimated total penalty for desolvation (i.e., the opposite process of solvation) for each identified pocket is obtained as the sum of the vdW dehydration energy ΔG_{vdW}, the electrostatic dehydration energy ΔG_{elec}, and the dehydration energy penalty caused by the surface, ΔG_{surf}. The vdW and electrostatic contributions to the total dehydration energy penalty are obtained using the individual components of the solvation energy density (i.e., eqs 3 and 4) with the opposite sign by volume integration over the binding pocket region. The dehydration energy penalty caused by the surface area, ΔG_{surf}, is modeled using the curvature-corrected surface tension term:

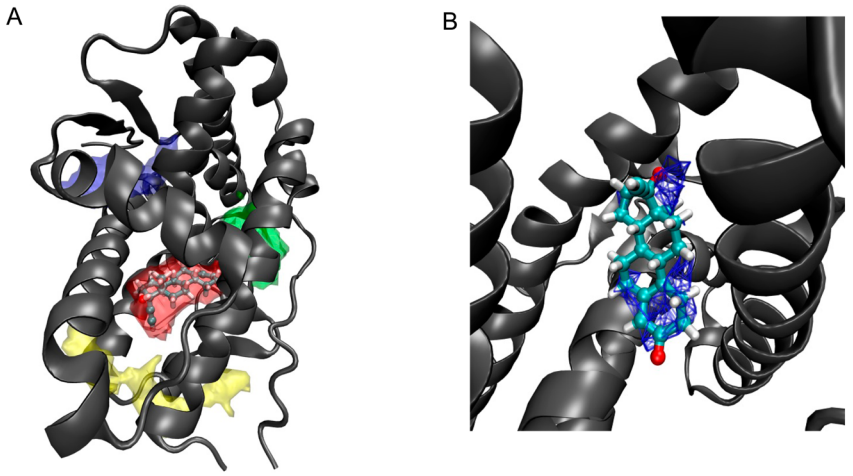
$$\Delta G_{\text{surf}} = -\int_{\Gamma_{\text{binding}}} \gamma_0(1 - 2\tau H(\mathbf{x}))\, \mathrm{d}S \tag{5}$$

where Γ_{binding} is the surface of the binding pocket region.

Local regions inside the pockets can be relatively hydrophobic or hydrophilic, as characterized by the local hydration energy distribution $\varphi_{\text{solute-solvent}}(\mathbf{x})$. On the basis of complementary ligand−protein binding interactions, the relatively strong hydrophobic regions of a binding pocket are usually occupied by hydrophobic funtional groups of a ligand. However, occupation of hydrophilic regions is unlikely to contribute to the binding affinity significantly for many druglike ligands (apolar molecules with rarely more than one formal charge).[48] It has been argued that the ligand−protein electrostatic interactions and desolvation penalties counteract each other and that the combination provides insubstantial contributions to the binding affinity.[49]

Theoretical and experimental studies have indicated that occupation of hydrophobic pockets is the main contributor to

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765



C

Index	Depth (Å)	SAS (Å^2)	Volume (Å^3)	Shape (Principal moments of inertia)	Hydrophobic Fraction (%)	Optimal Binding Affinity (kcal/mol)	Dehydration Penalties (vdW, Surf, Elec, Tot) (kcal/mol)
Pocket_1	11.0	88.7	306.7	(0.42, 0.91, 1)	74.8	-9.0	(4.09, -14.36, 3.42, -6.86)
Pocket_2	6.8	41.3	197.6	(0.29, 0.90, 1)	60.0	-3.9	(1.58, -8.63, 4.48, -2.57)
Pocket_3	6	67.0	239.1	(0.55, 0.76, 1)	12.6	-0.9	(2.69, -14.25, 13.67, 2.11)
Pocket_4	2.4	42.4	204.8	(0.12, 0.92, 1)	50.6	-3.3	(1.97, -7.83, 8.36, 2.50)

Figure 2. (a) Identified pockets (PDB entry 1sqn), in which the target protein is shown as a gray-colored cartoon, the ligand as sticks colored by atom name, and the identified pockets as transparent surfaces. The primary pocket is shown in red, the secondary pocket in blue, the tertiary pocket in green, and the last pocket in yellow. The cocrystallized ligand binds to the primary pocket. (b) Hydrophilic regions for the identified primary binding pocket and overlapping with the cocrystallized ligand are shown in blue wireframe. The top-right region is occupied by the ligand hydroxyl group (−OH), and the bottom one is occupied by a carbonyl group (C═O). (c) Topological and energetic parameters for the identified protein pockets.

the binding affinity, while hydrophilic regions mainly contribute to the drug specificity.[50] Here we estimate the optimal binding affinity as the contribution from the apolar surface area:

$$\Delta G_{\text{optimal}} = -\int_{\Omega_{\text{nonpolar\_surface}}} \gamma_0(1 - 2\tau H(\mathbf{x}))\, dS \quad (6)$$

In the previous study of p53/MDM2 hydration,[37] we compared the water density around the MDM2 with the hydration energy distribution $\varphi_{\text{solute-solvent}}(\mathbf{x})$ and found that for $\varphi_{\text{solute-solvent}}(\mathbf{x}) < -0.06$ kcal mol^{-1} Å^{-3}, the water density was larger that its bulk value (hydrophilic regions). Therefore, $\Omega_{\text{nonpolar\_surface}}$ is defined as the region where $\varphi_{\text{solute-solvent}}(\mathbf{x}) \geq -0.06$ kcal mol^{-1} Å^{-3}. In the current level-set VISM model, spatial and orientational distribution of water around protein are ignored. The entropic contributions from these factors might lead to systematic overestimation.

2.5. Statistical Validation of Identified Pockets. The predicted binding pockets are validated by volume overlaps between the identified pockets and the cocrystallized ligands. A significant overlap indicates that the identified pocket is a ligand binding pocket. The volume-overlap method is argued to be a more precise method than methods based on the pocket−ligand center distance, where the actual spatial configuration of the identified pocket and the corresponding cocrystallized ligand is ignored. Box plots are used for comparison of the topological and physicochemical characters of ligand-occupied pockets and unoccupied ones.

3. RESULTS AND DISCUSSION

3.1. Performance of Protein Binding Pocket Identification through VISM-CFA. VMD was used for visualization inspection. Figure 2a displays the identified pockets aligned with the original complex (PDB entry 1sqn), with the target protein shown in gray cartoon, the ligand shown as a stick model, and the identified pockets highlighted by transparent surfaces. The primary (deepest) pocket is shown in red, the secondary pocket in blue, the tertiary pocket in green, and the fourth pocket in yellow. In Figure 2b, the relatively strong hydrophilic regions are shown in blue wireframe for the identified primary pocket. There are two blue regions. The top-right region is occupied by the ligand hydroxyl group (−OH), and the bottom one is occupied by a ketone carbonyl group (C═O). This is consistent with ligand polar groups occupying the relatively strong hydrophilic regions for specificity and hydrophobic groups occupying the hydrophobic regions for affinity. These specific hydrophilic regions can be illustrated by the pocket hydration energy density distribution. The table below the illustrations of the pockets lists the various topological and physicochemical parameters for the corresponding putative pockets for further characterization of each pocket's properties. We will interpret these implications in the following sections of this study.

To systematically investigate the robustness of the VISM-CFA binding pocket identification method, we examined 515 diverse cocrystallized ligand−protein complexes (Table 2). In the data set of 515 complexes, a total of 703 potential binding

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

Table 2. Binding Site Identification Performance of VISM-CFA in 515 Target Proteins

result	set of 515 proteins[a] (%)	set of 228 tight binders[b] (%)
deepest pocket (primary)	92.4	96.9
largest pocket	90.7	94.2
best optimal binding affinity (OBA) pocket	91.1	97.0
both deepest and largest pocket	88.3	93.0
identified pocket site	96.9	99.1

[a]All 515 proteins contained a cocrystallized ligand. [b]Subset of 228 proteins that contain a ligand with experimental pK_d larger than 6.

pockets were identified, and 505 identified pockets were occupied by cocrystallized ligands. Among them, 398 proteins had only the primary pocket identified, 107 proteins had two putative pockets, 54 proteins had more than two putative pockets, and only three protein targets (PDB entries 1qmf, 1t03, and 2qft) showed as many as five pockets. In the crystallographic complexes, 96.9% of the ligands are bound to the VISM-CFA-identified pockets, 71.8% of the identified pockets are occupied by the cocrystallized ligands, 92.4% of the ligands bind to the deepest site, 4.7% of the ligands bind to the second-deepest pocket, 0.6% of the ligands bind to the third-deepest pocket, 90.7% of the ligands bind to identified pocket with the largest volume, 88.3% of the ligands bind to the pocket that is both deepest and largest, and 91.1% of the ligands bind to the pocket with the maximum estimated binding affinity.

In this data set, we chose 228 complexes with experimental $pK_d > 6$ as tight binders. For this subset, 99.1% of the ligands bind to the pocket identified by VISM-CFA, 96.9% of the ligands bind to the deepest site, 94.2% bind to the largest-volume site, 93.0% bind to the pocket that is both deepest and largest, and 97% bind to the pocket with the maximum estimated binding affinity. For comparison, in an exhaustive study of 5616 protein−ligand complexes using PocketFinder,[28] which incorporates a modified LJ potential that was believed to be more sensitive and specific in terms of pocket identification, 96.8% of the ligands were found to bind in the identified pockets. However, only 5% of the proteins had only one pocket identified, 17% were identified with two pockets, 20% of the proteins showed three identified pockets, and more than half of the proteins were predicted to have more than three pockets. The largest protein showed 17 identified pockets. Our method is based solely on our physical implicit-solvent model with unmodified vdW and electrostatic potentials. The sensitivity reached 96.9%, which is as high as those for the best literature-reported methods. At the same time, we significantly improved the specificity, as 77.3% of the proteins had only one protein binding pocket identified, 12.4% showed two pockets, 7.0% showed three identified pockets, and no protein had more than five putative binding pockets. In VISM theory, the equilibrium VISM surfaces are the consequence of balanced interactions between the solute and solvent. Strong attractive interactions between the solute and solvent (i.e., solute−solvent electrostatic interactions) draw the VISM surface closer to the solute until it is counterbalanced by the repulsive part of the vdW interactions and minimal solvent surface tension. The pockets identified by VISM-CFA tend to be deep and contain significant hydrophobic characteristics. This self-consistent physical chemical description is largely missing from other methods.

Among the 515 ligand−protein complexes, there are 16 cases where cocrystallized ligands do not bind to the VISM-CFA-identified pockets. Table 3 lists these PDB codes and the binding information on the cocrystallized ligands. They can be classified into five different categories. For the first two cases (PDB entries 1l83 and 220l), we found that the ligands are very small ring molecules with fewer than 10 heavy atoms and are completely enclosed by the molecular surface and deeply buried inside the protein. The target protein is T4 lysozyme with the L99A mutation to create a cavity that can accommodate one benzene molecule inside of the protein.[51] The cavity is tailored for specific ligands to stabilize the protein structure and add chemical functionality.[52] The size and shape of the buried cavity strongly depend on the ligands packed in it. There are other mutations to alter the size and polarity of the T4 lysozyme pocket suitable for specific bound ligands. On the other hand, in our data set there are five complexes of the same target with larger ligands (PDB entries 1li2, 1li3, 1li6, 1lgw, and 3dmz) that all have the cocrystallized ligand bound to the VISM-CFA-identified primary pocket. Case 3 shows a similar situation with a small ligand tightly enclosed by its target protein.

Table 3. Data for the 16 Ligand−Protein Complexes in Which the Cocrystallized Ligand Does Not Bind to the VISM-CFA-Identified Pocket

case	PDB ID	target	ligand binding mode
1	1l83	T4 lysozyme	small ligand bound to deep buried cavity
2	220l	T4 lysozyme	small ligand bound to deep buried cavity
3	1oss	*Streptomyces griseus* trypsin (SGT)	small ligand bound to deep buried cavity
4	1gtb	glutathione S-transferase (GST)	charged ligand bound to shallow pocket
5	2r5a	MBT repeats of sex comb on midleg (Scm)	charged ligand bound to shallow pocket
6	2vyt	MBT repeats of sex comb on midleg (Scm)	charged ligand bound to shallow pocket
7	2bt9	lectin from *Rastonia solanacearum*	charged ligand bound to shallow pocket
8	2jkj	adhesin subunit (DraE/AfaE)	charged ligand bound to shallow pocket
9	1hwi	HMG-COA reductase	charged ligand bound to shallow pocket
10	1fao	pleckstrin homology domains	Tetrakis(phosphate) as the ligand (strong charge)
11	1o9d	14-3-3 protein	non-natural peptide as the ligand
12	2itk	human pin1	non-natural peptide as the ligand
13	1m48	interleukin-2	non-natural peptide as the ligand
14	1nlj	human cathepsin K	non-natural peptide as the ligand
15	1fzq	ADP-ribosylation factor-like protein 3	nucleic acid base with phosphate as the ligand
16	1t03	HIV-1 nucleotide RT	nucleic acid base with phosphate as the ligand

In cases 4−9, the ligands carry a net charge higher than 1 and bind to shallow pockets. In case 10, the ligand is a highly charged molecule (−8) with four phosphate groups. In cases 11−16, the ligands are either non-natural peptides with phosphate groups or DNA elements. The binding sites are actually substrate binding sites that lack typical features of small-molecule binding pockets. In all of these cases, the ligands

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

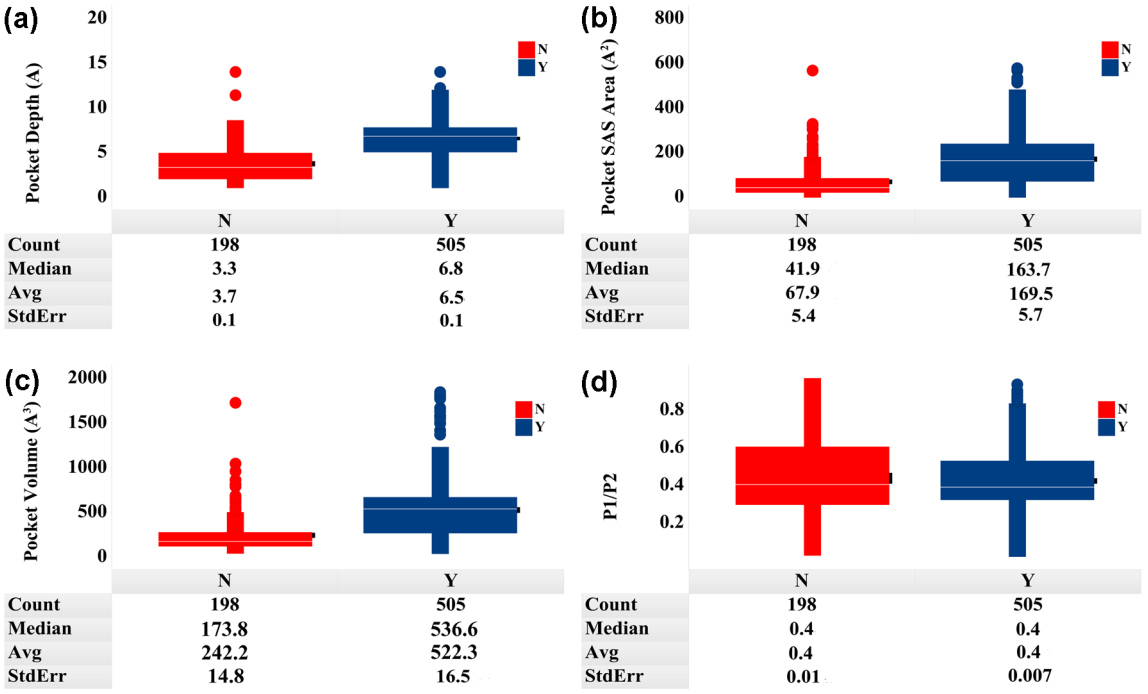


Figure 3. Comparison between the topological characters of ligand-occupied and unoccupied pockets.

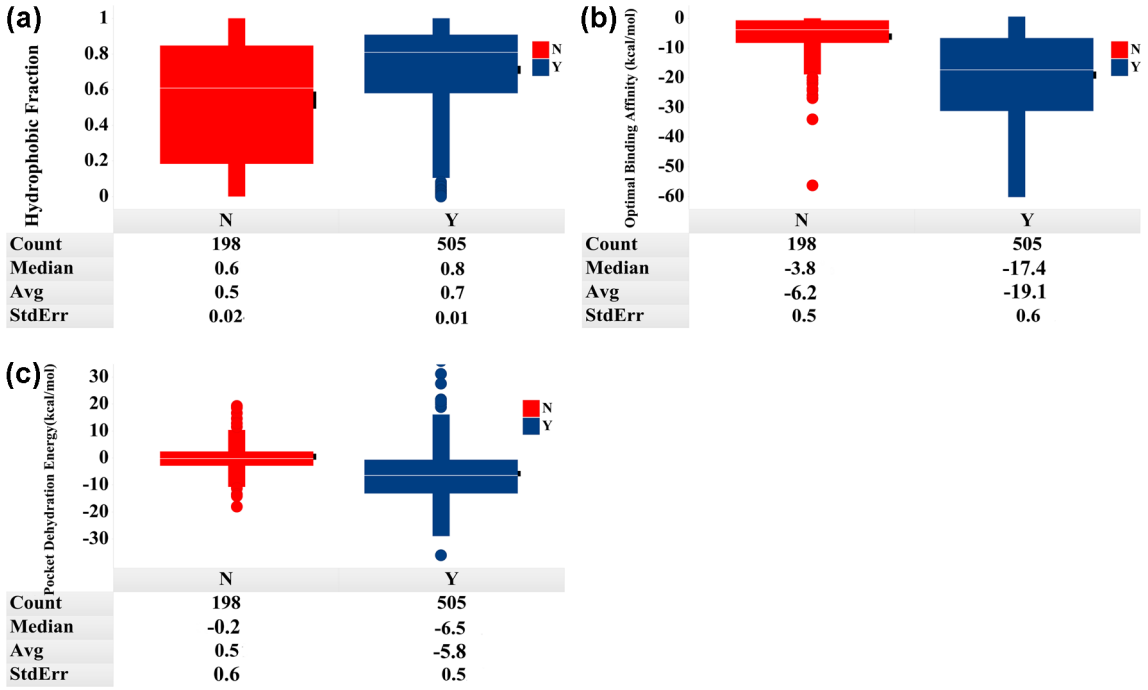


Figure 4. Comparison between physicochemical characters of ligand-occupied and unoccupied pockets.

bind to the protein surfaces with relatively strong polarity and shallow topological features that are unlikely to be captured by the VISM calculation.

3.2. Comparison between the Topological and Physicochemical Characters of Ligand-Occupied Pockets and Unoccupied Ones. In the 703 VISM-CFA-identified pockets for the 515 protein targets, 505 identified pockets are occupied by cocrystallized ligand and 198 pockets are unoccupied. We compared the topological and physicochemical characters of the ligand-occupied pockets and the unoccupied ones. In Figure 3, pocket depth, SAS area, pocket volume, and pocket shape comparisons are shown as box plots. In these figures, the red-colored boxes represent the distributions of unoccupied pockets while the blue boxes are for ligand-occupied pockets. The black bar at the right side of each box indicates the 95% confidence interval of the mean. In Figure 3a–c, the plots show statistically significant differences (p value < 0.05) in pocket depth, SAS area, and volume. The ligand-occupied pockets tend to be deeper with relatively larger volumes and SAS areas. The median depth of occupied binding pockets is 6.8 Å, compared with 3.3 Å for unoccupied ones. The median SAS area of occupied binding pockets is 163.7 Å^2,

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753–765

compared with 42.0 Å^2 for unoccupied ones. The median volume of occupied binding pockets is 536.6 Å^3, compared with 173.8 Å^3 for unoccupied pockets. Figure 3d shows the comparison of the normalized ratios of first and second principal moments. If the value is close to 1, the pocket is disklike, whereas the pocket shape is tubelike if the value is close to 0. While the shape differences between ligand-occupied and unoccupied pockets are not statistically significant (two-tailed p value = 0.41), the shape distribution of ligand-occupied pockets (i.e., first quartile = 0.32, median = 0.39, third quartile = 0.53) is generally narrower than that of the unoccupied ones (i.e., first quartile = 0.29, median = 0.40, third quartile = 0.61), which indicates that tubelike pockets are slightly favored by the small-molecule ligands.

Figure 4 compares the physicochemical properties of ligand-occupied and unoccupied pockets. In Figure 4a–c, the box plots indicate statistically significant differences between the pocket hydrophobic fractions, solvent-accessible surface tension-based optimal binding affinities, and pocket dehydration energies for the ligand-occupied pockets (blue boxes) and the unoccupied pockets (red boxes). In Figure 4a, 75% of the ligand-occupied pockets show pocket hydrophobic fractions greater than 60%, compared with just 50% for unoccupied pockets. This distribution is much narrower and concentrated around a large hydrophobic fraction for occupied pockets. The difference in distributions indicates that a high percentage of hydrophobicity is an important factor for ligand binding, but it obviously is not the only factor. Figure 4b,c shows the estimated optimal binding affinities when a hypothetical ligand occupies only the pocket hydrophobic fraction and the total pocket dehydration energies. Both of them confirm that the occupied pockets are thermodynamically favored by the small-molecule ligands.

3.3. Characterization of the Influence of Protein Conformational Changes on the Binding Pocket. Target proteins can undergo conformational changes when bound with different ligands. It is important to investigate multiple crystallographic structures for a given protein target to understand the pocket characteristics. In this study, we chose the "closed" and "open" conformations of heat shock protein 90 (HSP90) as an example to illustrate the quality of binding site identification and characterize the protein pockets of different conformations through VISM-CFA.

The molecular chaperone HSP90 has been reported to help cancer cell survival through stabilization of key proteins responsible for a malignant phenotype. In a comprehensive yeast protein interaction study, hundreds of proteins were revealed to interact with HSP90. It is known that HSP90 also interacts with numerous oncoproteins, including Cdk4, Akt, BCR-ABL, p53, and v-src. It is of high interest to study the binding sites and find potential small-molecule inhibitors. Both NMR and X-ray crystallographic studies have confirmed that there are "closed" and "open" conformations of HSP90. Fluorescence resonance energy transfer (FRET) assay studies indicated that a small compound binds HSP90 with a K_i of 18 ± 1 μM, while a larger one is 5-fold more potent with a K_i of 4 ± 1 μM.[53]

Figure 5a,b shows the binding sites of the "closed" and "open" conformations of HSP90 identified by VISM-CFA. Only one primary binding site (shown as a red transparent surface) was found for each conformation. In Figure 5c,d, the blue-colored wireframe indicates the pocket hydrophilic regions. In both ligand–protein complexes, the hydrophilic

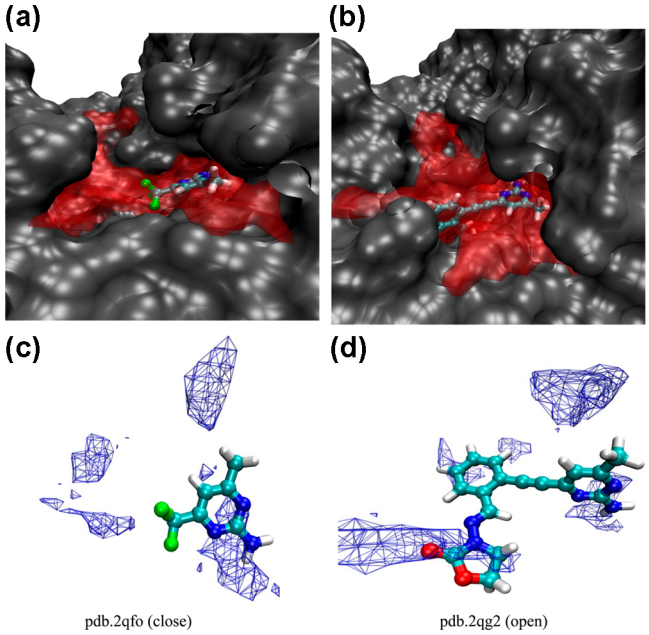


Figure 5. (a, b) VISM-CFA-identified pockets for "closed" and "open" conformations of HSP90. The target protein is shown as a gray-colored molecular surface, and the ligand is shown in the ball-and-stick representation. The identified pockets are enclosed by transparent red surfaces. (c, d) Hydrophilic regions are depicted with blue wireframe for the identified primary binding pocket together with the cocrystallized ligand.

aminopyrimidine ring of the ligand overlaps well with the pocket hydrophilic regions (blue wireframe). In Figure 5c, the strongly hydrophobic trifluoromethyl group is mainly located in the hydrophobic regions (no blue wireframe). In Figure 5d, the larger ligand also arranges itself in such a manner as to complement the binding pocket physiochemical features to maximize the binding affinity. Both NMR spectroscopy and X-ray crystallographic results indicate a high level of protein flexibility for the HSP90 binding pocket, and it undergoes a conformational change between the "closed" and "open" states when different ligands are bound. With a large ligand (Figure 5b), the HSP90 binding site is wider to open the binding area and accommodate the tighter-binding inhibitor (i.e., 5-fold increase in binding affinity). With different target protein conformations, the pockets identified by VISM-CFA have different characters although they are around the same protein surface region. As shown in Table 4, the "open" pocket is relatively large and deep. The pocket depth and size are nearly doubled in going from the "closed" to the "open" conformation. Both the optimal binding affinity and total pocket dehydration energy indicate the increasing ligand binding potency of the "open" conformation. The occupancy of a large-sized ligand is facilitated by the large solvent-accessible surface of the "open" pocket.

It is argued that conformational differences of target proteins affect the prediction performance of potential ligand binding sites. On the basis of the comparison of different conformational structures in a limited set of protein targets, VISM-CFA seems to tolerate a certain degree of conformational flexibility and to be able to predict nearly identical binding locations for each target. Abagyan et al.[28] demonstrated that the conformational differences between the occupied and empty pockets do not significantly affect the pocket prediction results for a large

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753–765

Table 4. Parameters for the "Closed" and "Open" Binding Pockets of HSP90

			pocket characteristic parameters obtained from VISM-CFA analysis						
			geometrical characters					physiochemical characters	
PDB ID	conformation	exptl K_i (μM)	depth (Å)	SASA (Å^2)	volume (Å^3)	shape (principal moments of inertia)	hydrophobic fraction (%)	optimal binding affinity (kcal/mol)	dehydration penalties (vdW, surf, elec, tot) (kcal/mol)
2qfo	closed	18 ± 1	5	100.0	348.2	(0.43, 0.72, 1)	94.7	−13.5	(3.3, −16.0, 10.7, −2.1)
2qg2	open	4 ± 1	9	173.5	640	(0.51, 0.95, 1)	81.6	−18.5	(6.1, −29.1, 13.6, −9.4)

Table 5. Topological and Energetic Information for VISM-CFA-Identified Pockets for the 27 Targets Classified As "Druggable", "Difficult", and "Undruggable" by Cheng et al.[31]

knowledge-based			pocket characteristic parameters obtained from VISM-CFA analysis						
			geometrical characters					physiochemical characters	
PDB ID	target	druggability	depth (Å)	SASA (Å^2)	volume (Å^3)	shape (principal moments of inertia)	hydrophobic fraction (%)	optimal binding affinity (kcal/mol)	dehydration penalties (vdW, surf, elec, tot) (kcal/mol)
1rth	HIV RT (NNRTI)	druggable	6.8	575.3	1799.2	(0.21, 0.92, 1)	81.9	−58.7	(20.8, −85.7, 40.9, −23.9)
1hvr	HIV-1 protease	druggable	6.4	211.5	571.4	(0.33, 0.85, 1)	92.7	−30.7	(6.5, −34.5, 12.3, −15.7)
1m17	EGFR kinase	druggable	8.8	427.8	1416.7	(0.54, 0.75, 1)	68.7	−29.9	(17.0, −52.9, 76.3, 40.4)
1hwi	HMG CoA reductase	druggable	3.8	270.7	787.5	(0.32, 0.85, 1)	88.6	−25.9	(11.9, −36.5, 10.5, −14.1)
4cox	cyclooxygenase 2	druggable	12.0	337.1	1097.2	(0.53, 0.82, 1)	74.1	−25.0	(14.6, −47.1, 22.8, −9.7)
1udt	PDE 5A	druggable	10.2	260.6	1009.2	(0.56, 0.63, 1)	50.8	−24.4	(11.1, −46.2, 56.3, 21.2)
1c14	enoyl reductase	druggable	8.6	243.0	817.2	(0.39, 0.86, 1)	70.9	−22.6	(11.1, −35.5, 22.8, −1.6)
1ke9	CDK2	druggable	7.6	193.6	801.3	(0.47, 0.85, 1)	74.9	−21.7	(7.3, −34.6, 36.5, 9.2)
1oyn	PDE 4D	druggable	9.2	341.8	1023.0	(0.55, 0.76, 1)	51.8	−18.9	(14.071, −46.0, 63.1, 31.2)
1kv1	P38 kinase	druggable	8.8	335.2	1230.3	(0.37, 0.98, 1)	32.9	−17.9	(13.9, −58.6, 89.0, 44.4)
1kij	DNA gyrase B	druggable	8.0	200.8	728.6	(0.42, 0.81, 1)	52.4	−13.0	(8.1, −37.9, 36.9, 7.1)
1iep	cAbl kinase	druggable	9.6	302.4	1207.3	(0.63, 0.74, 1)	32.3	−10.1	(11.7, −51.6, 81.6, 41.6)
1pwl	aldose reductase	druggable	8.8	139.3	587.8	(0.18, 0.96, 1)	42.2	−8.8	(6.7, −29.8, 15.5, −7.6)
1rv1	MDM2	druggable	3.8	66.8	218.6	(0.31, 0.94, 1)	90.2	−8.5	(2.4, −10.6, 1.8, −6.4)
1ea1	fungal Cyp51	druggable	10.2	223.5	973.8	(0.60, 0.80, 1)	33.9	−7.7	(9.9, −40.1, 37.8, 7.6)
1gpk	acetylcholinesterase	druggable	11.0	162.4	583.7	(0.32, 0.89, 1)	42.4	−7.7	(7.9, −27.9, 16.0, −3.9)
1ezq	factor Xa	druggable	5.5	93.9	353.8	(0.34, 0.88, 1)	49.7	−5.4	(3.9, −19.5, 13.5, −2.1)
1o86	ACE-1	difficult	14.0	1345.4	4328.5	(0.19, 0.88, 1)	41.1	−52.2	(62.5, −188.9, 261.4, 135.0)
1nf7	IMPDH	difficult	9.2	425.0	1276.4	(0.31, 0.92, 1)	68.9	−26.4	(20.9, −49.0, 44.4, 16.4)
1t03	HIV RT (nucleotide)	difficult	2.2	156.4	499.7	(0.21, 0.98, 1)	93.0	−18.3	(6.5, −20.9, 8.9, −5.5)
1ktt	thrombin	difficult	3.8	106.4	417.3	(0.21, 0.87, 1)	65.8	−12.0	(4.4, −21.4, 12.4, −4.6)
1qmf	penicillin binding protein	difficult	9.4	206.1	789.5	(0.25, 0.85, 1)	43.2	−11.6	(9.0, −16.8, 43.8, 35.9)
1a4q	neuraminidase	difficult	2.4	66.3	225.3	(0.41, 0.98, 1)	19.4	−1.6	(3.4, −9.9, 14.3, 7.9)
1g1f	PTP-1B	undruggable	4.0	31.0	138.8	(0.41, 0.78, 1)	5.5	−0.4	(1.2, −8.5, 9.5, 2.2)
1nlj	cathepsin K	undruggable	1.0	23.0	87.0	(0.43, 0.88, 1)	28.7	−0.6	(0.1, −0.8, 10.0, 9.1)
1bmq	caspase 1 (ICE-1)	undruggable	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1qs4	HIV integrase	undruggable	N/A	N/A	N/A	N/A	N/A	N/A	N/A

number of protein−ligand complexes. For PPI interfaces, our VISM-CFA-based algorithm can be applied to individual protein subunits to detect the potential "pockets" at the PPI interface and help evaluate the feasibility of modulating PPIs.[37] Keskin et al.[54] suggested that "hot spots" at a PPI interface and their "neighboring" residues in the protein binding "hot regions" are often preorganized by evolution in the unbound protein state.[38,55,56] When a protein binds with different partners, different combinations of "hot regions" may be involved. Similarly, there are characteristic regions in small-molecule binding pockets that are evolutionarily conserved to interact with natural ligands. They are routinely explored by synthetic ligands in drug design. Although we have also used a large set of static crystallographic structures to study the volumetric, topological, and energetic characters of the potential protein binding pockets, we must emphasize that it is still not a comprehensive and exhaustive study. Protein binding sites are flexible and dynamic. It would be beneficial to combine advanced protein conformational sampling techniques (e.g., accelerated molecular dynamics) with our VISM-CFA pocket identification method, which would provide new insights into the binding pocket dynamics and enable the exploration of transient pockets for new drug discovery opportunities.

The results of the above studies indicate that the equilibrium VISM surfaces can provide statistically confident predictions of the locations and characters of potential protein small-molecule binding pockets. The above results are also consistent with the observation that hydrophobic driving forces dominate small-molecule ligand−protein binding processes.[57]

3.4. Assessing Target Protein Ligandability with the Potential Binding Pocket Topological and Physicochemical Properties. The ability to assess the "ligandability" of target proteins is highly valuable.[1,2] The ligandability information on a gene family is often used to identify specific target proteins. Nevertheless, estimations of gene family

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

ligandability can vary dramatically. Computational approaches have been developed mainly according to the topological and physiochemical properties of the protein binding pocket in order to distinguish "druggable" and "difficult" from "undruggable".[30,31] In the structural-based maximal binding affinity model developed by Cheng et al.,[31] structural information on the target binding site is used to estimate the ligandability of the target. They derived the maximal achievable binding affinity for binding of a druglike ligand with a protein pocket (ΔG_{MAP}) from physical-based desolvation penalties:

$$\Delta G_{MAP} \approx (-300\text{Å}^2) f_{nonpolar} \left(\frac{\gamma_0}{1 - \frac{1.4}{r}} \right) + C \quad (7)$$

where $f_{nonpolar}$ is the ratio of pocket nonpolar SASA to the total pocket SASA, which discriminates the relatively hydrophobic pockets from the rest, γ_0 is the planar surface tension, r is the overall pocket curature to characterize the enclosure of the potential binding pocket, and C is a constant to account for system error. Despite many inherent exceptions and crude approximations and the subjective "druggability" classification for 27 targets, the performance of the maximal binding affinity method is very reasonable for druggability prediction. We used this method to test whether the information obtained by VISM-CFA surface analysis can also provide target protein druggability assessments.

Table 5 shows the topological and energetic information for the ligand occupying the primary putative pockets. The four "undruggable" targets (caspase 1, HIV integrase, cathepsin K, and PTP-1B) show the lowest optimal binding affinities (greater than −2 kcal/mol) with the smallest pocket volumes (<200 Å^3) or no identifiable binding pocket. The protein surfaces for this category are relatively flat compared with those of the other 23 targets. One prerequisite for a binding pocket for a small organic ligand is the presence of a suitable surface cavity with hydrophobicity. From the comparison of the pocket topological information, the pocket size is an important factor to discriminate "druggable" and "undruggable" protein targets. For the other 23 targets, there are marketed drugs or advanced drug candidates. All of them show reasonably good optimal binding affinities (less than −5 kcal/mol) compared with the "undruggable" targets.

In the "druggable" targets, the binding pocket of MDM2 is a PPI interface that interacts with a regulator of tumor suppressor p53. It shows an optimal binding affinity as −8.50 kcal/mol for the pocket regions with relatively large pocket volume of 218.63 Å^3. With about 20 years of effort, MDM2 has been recognized as a druggable PPI interface suitable for small-molecule drugs.[12] This benefits from the strong hydrophobicity inside of this pocket. The pocket hydrophobic fraction is as high as 90.21% with a small electrostatic dehydration penalty of ∼1.84 kcal/mol. The P1/P2 ratio of 0.33 indicates a tube-shaped binding pocket, consistent with the small-molecule pocket analysis. The compound nutlin-2 (PDB entry 1rv1) displaced p53 protein from the binding pocket of MDM2 with a median inhibitory concentration (IC_{50}) of 140 nM.[58] In Figure 6, we align the p53−MDM2 complex, the nutlin-2−MDM2 complex, and the VISM-CFA-predicted binding pocket. Nutlin-2 closely mimics the interactions of the p53 peptide to bind with MDM2. In the nutlin-2−MDM2 complex, two bromophenyl moieties insert deeply into the p53 Leu26 and Trp23 binding pockets. The ethyl ether side chain from nutlin-2 directly inserts into the p53 Phe19 pocket. The shape of nutlin-2 shows good complementary to the VISM-CFA-identified pocket, and the polar groups (e.g., bromine atoms) of the compound also align well with the hydrophilic regions of the pockets (blue wireframe). We must emphasize that PPIs and "druggability" are profoundly complex topics that are far from being resolved.[59] In addition, the difficulty of finding small-molecule PPI modulators is confounded by that fact that legacy compound collections are biased toward known drug targets such as enzymes and G-protein-coupled receptors. They do not necessarily reflect the best properties for PPI modulators. Therefore, the "druggability" assessment of the PPI should be treated with caution.

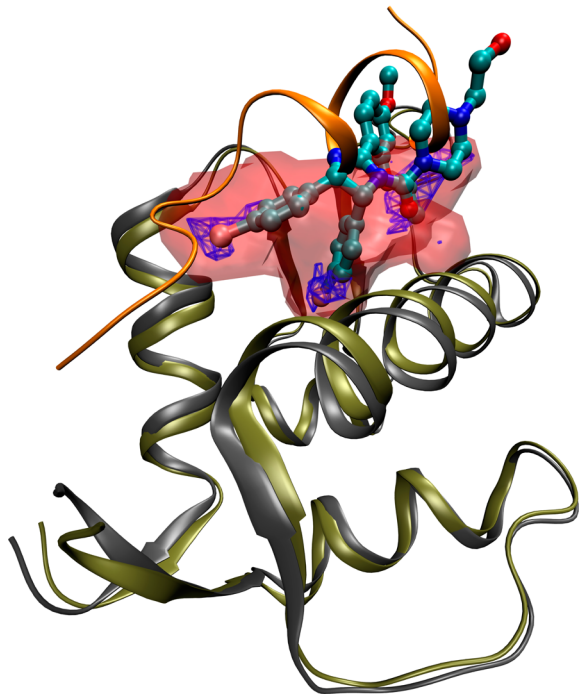


Figure 6. VISM-CFA pocket prediction (red transparent isosurface) vs small-molecule inhibition of the p53−MDM2 PPI. The p53−MDM2 protein−protein complex (gray and orange, PDB entry 1ycr) is superimposed on the protein−inhibitor complex (tan for protein, red for oxygen, blue for nitrogen, cyan for carbon, and pink for bromine; PDB entry 1rv1).

Thrombin is categorized as an outlier in several widely used "druggability" prediction methods.[30,31,39] Thrombin (1ktt) and factor Xa (1ezq) are two serine proteases that are thought to have similar druggability since they are in the same protein family. Most inhibitors for thrombin are prodrugs that are categorized as "difficult" by Cheng et al.[31] Meanwhile, the factor Xa inhibitors (e.g., rivaroxaban[60] and apixaban[61]) are not prodrugs and are classified as "druggable". Interestingly, several druggability assessment methods show a better druggability potential for thrombin, such as the Dscore in SiteMap,[30] the hotspot index in Watermap,[39] and even the structure-based maximal binding affinity model of Cheng et al.[31] In our VISM pocket "ligandability"prediction study, thrombin also shows a better optimal binding affinity by 6.6 kcal/mol with a slightly larger pocket compared with factor Xa. Strictly speaking, "druggability" is different from "ligandability", as many factors other than the interaction with its target play important roles for a ligand to be a drug, such as its pharmacokinetic and toxicology profiles. In order to be consistent and facilitate the comparison with previous studies,[30,31,39] we investigated the 27

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

pharmaceutical targets by Cheng et al.[31] and kept their original "druggability" classifications.

Many kinase inhibitors have been approved for cancer and other disease treatments. With about 518 kinases in the human genome, the inhibition of kinase activities can lead to marked physiological responses through a phosphotransfer cascade.[62] Three different kinases were included in this study. While all of them are in the "druggable" group with optimal binding affinities less than −10 kcal/mol, all of them show relatively high electrostatic dehydration penalties of around 80 kcal/mol. This indicates that ligand selectivity could be a key for ligand design.

In the structure-based maximal affinity model of Cheng et al.,[31] the binding sites are defined as the regions within 5.0 Å of the cocrystallized ligand or α spheres, and a common size factor for all druglike ligands is used rather than the actual pocket size itself. In the case of ACE-1 for instance, the pocket is a continuous long tube inside of the protein, and the ligand occupies less than 10% of the volume. The putative binding site defined by the cocrystallized ligand substantially underestimates the binding potency of the target. In the present study, all of the binding pockets were identified according to the target protein without ligand present, which is completely different from the model developed by Cheng et al. Both their method and ours can estimate binding affinities on the basis of the hydrophobic fractions of pocket surface areas. However, binding of druglike ligands and target proteins is hydrophobically driven in most but certainly not all cases.[57] In the case of HMG-CoA reductase, drug ligands have a conserved glutaminyl group that makes substantial hydrogen-bonding and ion-pair interactions with the protein.[63] The clinically most advanced irreversible kinase inhibitors of EGFR kinase are engineered to form a covalent bond with a cysteine residue located at the entrance of the ATP binding site. The formation of a covalent bond between the ligand and protein could yield an effectively "infinite" affinity potency for the ATP binding site.[62]

Statistically, druglike molecules prefer to bind target proteins with higher SASA hydrophobic fractions and surface complexities to make more contacts and gain higher binding affinity. However, many other factors should also be considered in the target druggability assessment, as it is an extremely complex issue involving not just high binding affinity but also pharmacokinetics and toxicology.

4. CONCLUSIONS

We have developed a method to identify and characterize potential small-molecule binding sites using the VISM-CFA equilibrium surface and the protein molecular surface. Compared with previous geometrical- or physical-based methods, our method properly incorporates electrostatic contributions and does not impose subjective size restrictions. The pocket prediction performance is enhanced through the exclusion of shallow and polar pockets where druglike ligands are unlikely to bind. In this study, individual pockets were listed and characterized. For 515 protein−ligand complexes, 96.9% of the ligands were found to be bound to the pockets identified by VISM-CFA. For 228 tight-binding protein−ligand complexes (i.e, ones with experimental pK_d larger than 6), 99.1% of the cocrystallized ligands were found to be in the VISM-CFA-identified pockets. The hydration energy density maps were consistent with cocrystallized ligand binding modes, and therefore, they also provide guidance for structural-based drug design. Quantitative characterization with volumetric, topological, and energetic parameters was also conducted. By the statistical comparison of these parameters for ligand-bound and unbound pockets, we found that the ligand -bound pockets are likely to be deeper with relatively larger fractions of hydrophobicity than unoccupied pockets. This is consistent with previous theoretical and experimental studies suggesting that druglike molecules prefer hydrophobic protein surfaces with complementary geometry. In addition, we used these binding pocket characteristics to assess the protein target "ligandability". Using the same 27 targets as studied by Cheng et al.[31] the Dscore in SiteMap,[30] and the hotspot index from Watermap,[39] we found that all of the undruggable targets lack relatively large pockets. Pockets with greater hydrophilic character (i.e., higher electrostatic dehydration penalty) are required for higher specificity for the ligand design, such as kinase targets.

The VISM-CFA-based algorithm provides a very sensitive and specific method to identify small-molecule binding sites on proteins. In addition, it offers a quantitative means to estimate the level of ligandability. A software package based on level-set VISM is to be made publicly available for the analysis of biomolecular solvation.[64] We believe that it can be very useful for rational drug design.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jianwei.che@gmail.com.

Funding

This work was supported by the U.S. National Science Foundation (NSF) through Grant DMS-1319731 (B.L. and L.-T.C.), the NSF Center for Theoretical Biological Physics (CTBP) through Grant PHY-0822283 (B.L. and J.A.M.), the National Institutes of Health through Grant R01GM096188 (J.C., L.-T.C., Z.G., B.L., and J.A.M.), and the Genomics Institute of the Novartis Research Foundation (J.C. and Z.G.). Work in the J.A.M. group was supported by NSF, NIH, HHMI, NBCR, and CTBP.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge Dr. Joachim Dzubiella for critical reading of the manuscript.

REFERENCES

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

(2) Brown, D.; Superti-Furga, G. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today* **2003**, *8*, 1067−1077.

(3) Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M. Predicting druggable binding sites at the protein−protein interface. *Drug Discovery Today* **2009**, *14*, 155−161.

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

(5) Miller, D. W.; Dill, K. A. Ligand binding to proteins: The binding landscape model. *Protein Sci.* **1997**, *6*, 2166−2179.

(6) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884−1897.

(7) Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand binding: Functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* **2003**, *13*, 389−395.

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

(8) Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. A dissection of specific and non-specific protein−protein interfaces. *J. Mol. Biol.* **2004**, *336*, 943−955.
(9) Nooren, I. M.; Thornton, J. M. Structural characterisation and functional significance of transient protein−protein interactions. *J. Mol. Biol.* **2003**, *325*, 991−1018.
(10) Janin, J.; Bahadur, R. P.; Chakrabarti, P. Protein−protein interaction and quaternary structure. *Q. Rev. Biophys.* **2008**, *41*, 133−180.
(11) Chakrabarti, P.; Janin, J. Dissecting protein−protein recognition sites. *Proteins* **2002**, *47*, 334−343.
(12) Nero, T. L.; Morton, C. J.; Holien, J. K.; Wielens, J.; Parker, M. W. Oncogenic protein interfaces: Small molecules, big challenges. *Nat. Rev. Cancer* **2014**, *14*, 248−262.
(13) Meier, C.; Cairns-Smith, S.; Schulze, U. Can emerging drug classes improve R&D productivity? *Drug Discovery Today* **2013**, *18*, 607−609.
(14) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48*, 2518−2525.
(15) Seco, J.; Luque, F. J.; Barril, X. Binding site detection and druggability index from first principles. *J. Med. Chem.* **2009**, *52*, 2363−2371.
(16) Levitt, D. G.; Banaszak, L. J. Pocket—A Computer-Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino-Acids. *J. Mol. Graphics* **1992**, *10*, 229−234.
(17) Laskowski, R. A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.
(18) Peters, K. P.; Fauck, J.; Frommel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **1996**, *256*, 201−213.
(19) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 359−363.
(20) Brady, G. P., Jr.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383−401.
(21) Binkowski, T. A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of Proteins. *Nucleic Acids Res.* **2003**, *31*, 3352−3355.
(22) Schneider, S.; Zacharias, M. Combining geometric pocket detection and desolvation properties to detect putative ligand binding sites on proteins. *J. Struct. Biol.* **2012**, *180*, 546−550.
(23) Leis, S.; Schneider, S.; Zacharias, M. In silico prediction of binding sites on proteins. *Curr. Med. Chem.* **2010**, *17*, 1550−1562.
(24) Zheng, X.; Gan, L.; Wang, E.; Wang, J. Pocket-based drug design: Exploring pocket space. *AAPS J.* **2013**, *15*, 228−241.
(25) Perot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O. Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. *Drug Discovery Today* **2010**, *15*, 656−667.
(26) Jalencas, X.; Mestres, J. Identification of Similar Binding Sites To Detect Distant Polypharmacology. *Mol. Inf.* **2013**, *32*, 976−990.
(27) Trosset, J. Y.; Vodovar, N. Structure-based target druggability assessment. *Methods Mol. Biol.* **2013**, *986*, 141−164.
(28) An, J.; Totrov, M.; Abagyan, R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752−761.
(29) Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* **2007**, *69*, 146−148.
(30) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377−389.
(31) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71−75.
(32) Coleman, R. G.; Salzberg, A. C.; Cheng, A. C. Structure-based identification of small molecule binding sites using a free energy model. *J. Chem. Inf. Model.* **2006**, *46*, 2631−2637.
(33) *Maestro*; Schrödinger, LLC: New York, 2012.
(34) Poornima, C. S.; Dean, P. M. Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 521−531.
(35) Poornima, C. S.; Dean, P. M. Hydration in drug design. 2. Influence of local site surface shape on water binding. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 513−520.
(36) Poornima, C. S.; Dean, P. M. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein−ligand interactions. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 500−512.
(37) Guo, Z.; Li, B.; Dzubiella, J.; Cheng, L.-T.; McCammon, J. A.; Che, J. Heterogeneous Hydration of p53/MDM2 Complex. *J. Chem. Theory Comput.* **2014**, *10*, 1302−1313.
(38) Ringe, D. What makes a binding site a binding site? *Curr. Opin. Struct. Biol.* **1995**, *5*, 825−829.
(39) Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins* **2012**, *80*, 871−883.
(40) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. Coupling hydrophobicity, dispersion, and electrostatics in continuum solvent models. *Phys. Rev. Lett.* **2006**, *96*, No. 087802.
(41) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. Coupling nonpolar and polar solvation free energies in implicit solvent models. *J. Chem. Phys.* **2006**, *124*, No. 084905.
(42) Wang, Z. M.; Che, J. W.; Cheng, L. T.; Dzubiella, J.; Li, B.; McCammon, J. A. Level-set variational implicit-solvent modeling of biomolecules with the Coulomb-field approximation. *J. Chem. Theory Comput.* **2012**, *8*, 386−397.
(43) Guo, Z.; Li, B.; Dzubiella, J.; Cheng, L. T.; McCammon, J. A.; Che, J. Evaluation of hydration free energy by level-set variational implicit-solvent model with Coulomb-field approximation. *J. Chem. Theory Comput.* **2013**, *9*, 1778−1787.
(44) Zhou, S.; Cheng, L.-T.; Dzubiella, J.; Li, B.; McCammon, J. A. Variational Implicit Solvation with Poisson−Boltzmann Theory. *J. Chem. Theory Comput.* **2014**, *10*, 1454−1467.
(45) Tolman, R. C. The effect of droplet size on surface tension. *J. Chem. Phys.* **1949**, *17*, 333−337.
(46) Born, M. Volumen und Hydratationswrame der Ionen. *Z. Phys.* **1920**, *1*, 45−48.
(47) Vega, C.; de Miguel, E. Surface tension of the most popular models of water by using the test-area simulation method. *J. Chem. Phys.* **2007**, *126*, No. 154707.
(48) Karplus, P. A. Hydrophobicity regained. *Protein Sci.* **1997**, *6*, 1302−1307.
(49) Hendsch, Z. S.; Tidor, B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **1994**, *3*, 211−226.
(50) Kangas, E.; Tidor, B. Electrostatic specificity in molecular ligand design. *J. Chem. Phys.* **2000**, *112*, 9120−9131.
(51) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **2002**, *322*, 339−355.
(52) Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. *Nature* **1992**, *355*, 371−373.
(53) Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J. Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. *Chem. Biol. Drug Des.* **2007**, *70*, 1−12.
(54) Keskin, O.; Ma, B.; Rogale, K.; Gunasekaran, K.; Nussinov, R. Protein−protein interactions: Organization, cooperativity and map-

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765

ping in a bottom-up Systems Biology approach. *Phys. Biol.* **2005**, *2*, S24−S35.
(55) Hu, Z.; Ma, B.; Wolfson, H.; Nussinov, R. Conservation of polar residues as hot spots at protein interfaces. *Proteins* **2000**, *39*, 331−342.
(56) Ma, B.; Elkayam, T.; Wolfson, H.; Nussinov, R. Protein−protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 5772−5777.
(57) Davis, A. M.; Teague, S. J. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew. Chem., Int. Ed.* **1999**, *38*, 737−749.
(58) Vassilev, L. T.; Vu, B. T.; Graves, B.; Carvajal, D.; Podlaski, F.; Filipovic, Z.; Kong, N.; Kammlott, U.; Lukacs, C.; Klein, C.; Fotouhi, N.; Liu, E. A. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **2004**, *303*, 844−848.
(59) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* **2003**, *52*, 2−9.
(60) Roehrig, S.; Straub, A.; Pohlmann, J.; Lampe, T.; Pernerstorfer, J.; Schlemmer, K. H.; Reinemer, P.; Perzborn, E. Discovery of the novel antithrombotic agent 5-chloro-*N*-({(5*S*)-2-oxo-3-[4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene-2-carboxamide (BAY 59-7939): An oral, direct factor Xa inhibitor. *J. Med. Chem.* **2005**, *48*, 5900−5908.
(61) Pinto, D. J.; Orwat, M. J.; Koch, S.; Rossi, K. A.; Alexander, R. S.; Smallwood, A.; Wong, P. C.; Rendina, A. R.; Luettgen, J. M.; Knabb, R. M.; He, K.; Xin, B.; Wexler, R. R.; Lam, P. Y. Discovery of 1-(4-methoxyphenyl)-7-oxo-6-(4-(2-oxopiperidin-1-yl)phenyl)-4,5,6,7-tetrahydro-1*H*-pyrazolo[3,4-*c*]pyridine-3-carboxamide (apixaban, BMS-562247), a highly potent, selective, efficacious, and orally bioavailable inhibitor of blood coagulation factor Xa. *J. Med. Chem.* **2007**, *50*, 5339−5356.
(62) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28−39.
(63) Istvan, E. S.; Deisenhofer, J. Structural mechanism for statin inhibition of HMG-CoA reductase. *Science* **2001**, *292*, 1160−1164.
(64) Zhou, S.; Cheng, L.-T.; Sun, H.; Che, J.; Dzubiella, J.; Li, B.; McCammon, J. A. LS-VISM: A software package for analysis of biomolecular solvation. *J. Comput. Chem.* **2015**, submitted.

DOI: 10.1021/ct500867u
J. Chem. Theory Comput. 2015, 11, 753−765