Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets

Andrea Volkamer,[†] Axel Griewel,[†] Thomas Grombacher,[‡] and Matthias Rarey*,[†]

Research Group for Computational Molecular Design, Bundesstr. 43, 20146 Hamburg, Germany, and Merck KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany

Received June 24, 2010

Automated prediction of protein active sites is essential for large-scale protein function prediction, classification, and druggability estimates. In this work, we present DoGSite, a new structure-based method to predict active sites in proteins based on a Difference of Gaussian (DoG) approach which originates from image processing. In contrast to existing methods, DoGSite splits predicted pockets into subpockets, revealing a refined description of the topology of active sites. DoGSite correctly predicts binding pockets for over 92% of the PDBBind and the scPDB data set, being in line with the best-performing methods available. In 63% of the PDBBind data set the detected pockets can be subdivided into smaller subpockets. The cocrystallized ligand is contained in exactly one subpocket in 87% of the predictions. Furthermore, we introduce a more precise prediction performance measure by taking the pairwise ligand and pocket coverage into account. In 90% of the cases DoGSite predicts a pocket that contains at least half of the ligand. In 70% of the cases additionally more than a quarter of the respective pocket itself is covered by the cocrystallized ligand. Consideration of subpockets produces an increase in coverage yielding a success rate of 83% for the latter measure.

INTRODUCTION

The molecular recognition of a low molecular weight ligand by a protein is the basis for the maintenance of biological systems. The 3D structure of an enzyme, precisely, the structure of the active site, is the key to its function. Moreover, proteins having structurally similar active sites tend to have similar functions.[1] These aspects motivate structure-based methods for protein function prediction and family classification, which are of high practical relevance for biotechnology and pharmaceutical science. Pharmaceutical science has already integrated structure-based methods into the standard pipeline of the drug development process; e.g., docking approaches[2−4] are commonly used to predict protein−ligand complexes and their binding affinities in order to identify candidates for novel drugs.

In the early phases of drug research, the prediction of the general ability of a target protein to be inhibited by low molecular weight compounds is of high interest. To describe this protein feature, the term druggability was coined in pharmaceutical sciences.[5−8] This can also be interpreted as ligandability for biotechnology science. Protein druggability is defined as the probability that a disease-modifying target can be modulated by a drug molecule.[7] In 2002, Hopkins et al.[5] stated that only 10% of the human genome is involved in disease onset and progression, providing a set of 3000 potential targets. In a following study, Brown et al.[9] claimed that 60% of small molecule drug discovery projects fail because the underlying targets are found to be not druggable. Over 60 000 three-dimensional protein structures are currently available in the PDB.[10] If a new target is identified, it is a challenging task to rapidly and reliably estimate its chances of interfering with small molecules.

Besides some experimental approaches,[7,11,12] most computational approaches for protein function or druggability prediction rely on descriptor-based comparison methods. Relevant structural, geometrical, and physicochemical parameters are identified from known protein−ligand complexes[13,14] to predict the function of structurally resolved proteins with yet unknown function[15] or to estimate their druggability. However, none of these approaches completely solved the problem of developing a general descriptor to distinguish between druggable and nondruggable proteins. The basis of a funded descriptor specification is a reliable method to detect the protein's active site.[16] Such a prediction algorithm has to be adaptable to a wide range of diverse protein active site shapes ranging from shallow over buried cavities to protein channels. Moreover, ligand binding frequently occurs at interfaces of protein subunits. These sites must not be missed by the detection algorithm. Flexibility of protein structures adds to the complexity of the prediction problem. A very tolerant prediction scheme is needed, since minor structural changes in amino acid side chains have a high impact on the later derived volume and shape parameters.

Several groups have investigated active site prediction methods. Table 1 shows an overview of published algorithms, which can be divided into three categories. Geometry-based methods[13,17−36] locate surface cavities by analyzing the geometry of the molecular surface. The methods to identify pockets either rely on a grid or are sphere- or tessellation-based. Energy-based methods represent the second group of prediction algorithms.[37−44] Interaction energies are calculated between the protein and a probe or a chemical group to identify cavities. Sometimes fragment-based docking methods are employed to find pockets. The third group comprises

* To whom correspondence should be addressed. E-mail: rarey@zbh.uni-hamburg.de.

[†] University of Hamburg.

[‡] Merck KgaA.

10.1021/ci100241y © 2010 American Chemical Society
Published on Web 10/14/2010

Table 1. Binding Pocket Detection Algorithm Overview

category	explanation and approaches
geometry-based	Locates surface cavities by analysis of the geometry of the molecular surface. Calculations rely on a grid and are partly sphere- or tessellation-based. CavitySearch,[17] POCKET,[18] method by Delaney,[19] method by Del Carpio et al.,[20] VOIDOO,[21] SURFNET,[22] APROPOS,[23] LIGSITE,[24] CAST[25] (Castp[26]), DOCK(sphgen),[27,28] Surface patches,[29] PASS,[30] LigandFit,[31] SCREEN,[13] TravelDepth,[32] PocketPicker,[33] VisGrid,[34] VICE,[35] Fpocket[36]
energy-based	Computes interaction energies between the protein and a probe/chemical group. Partially also blind and fragment-based docking is used. GRID,[37] Method by Ruppert,[38] vdW-FFT,[39] CS-Map,[40] DrugSite,[41] QSiteFinder,[42] PocketFinder,[43] BindingResponse[44]
evolutionary-based	Multiple sequence alignments find conserved residues. Active site profiles as well as homology modeling are also used. Method by Casari et al.,[45] method by de Rinaldis et al.,[46] method by Aloy et al.,[47] ConSurf,[48] Rate4Site,[49] Evolutionary trace method[50]
combined approaches	LigSiteCSC,[51] SURFNET-ConSurf,[52] SiteMap,[14] ConCavity,[53] MetaPocket,[54] SiteIdentify,[55] FINDSITE[56]

evolutionary-based methods.[45−50] Most of them rely on multiple sequence alignments to find conserved residues. Some evolutionary-based methods use active site profiles or homology modeling as instruments to find conserved residues. More recently published approaches combine several methods to enhance the prediction power.[14,51−56] All of these methods have major drawbacks.[13] Methods relying on a grid are sensitive to grid spacing, protein position, and orientation. Many geometry-based methods are afflicted with a wrong cavity ceiling definition. Sphere-based methods have problems with the detection of wide cavities. Energy-based methods highly depend on the underlying scoring function, force field parametrization, filter procedure, and cutoffs. Finally, since evolutionary-based methods employ multiple sequence alignments, they highly depend on the quality of the alignment tool as well as on the number of available sequences.

In this work, we introduce a new pocket detection algorithm called DoGSite which utilizes pattern recognition techniques for the identification of active sites. The algorithm is inspired by the fact that active sites frequently comprise invaginations which are large enough to accommodate at least one heavy atom. We identify these regions by filtering a grid representation of the protein with a 3D Difference of Gaussian (DoG) filter.[57] This filter locates spherically shaped structures in the grid, so-called DoG cores. The detected cores are then assembled to pockets which are well-suited to accommodate a typical ligand. The method is highly efficient, since the calculation can be carried out using two convolutions with a separable Gaussian filter of small size. For benchmarking purpose, a grid-based and an energy-based approach are also implemented, further referred to as LSite and DSite. These two methods resemble the published algorithms LigSite[24] and DrugSite[41] and were chosen due to their overall good performance.

Throughout the literature, a variety of different shapes and sizes of cavities have been considered as pockets, leading to a fuzzy definition of the term “pocket”. Numerous cavities and protrusions line the protein surface, connected through small narrow channels. Furthermore, ligand binding occurs at several neighboring cavities. Especially, auxiliary pockets near the binding pocket are of high interest in drug development. Hurdles in active site prediction studies are, for example, predictions of large and highly branched pockets. The narrow shape of these branches, also called “bottlenecks”, does not allow a ligand to penetrate them. Furthermore, the shape of large, solvent-exposed pockets indicates that they may consist of several subunits. To better approach the heterogeneous nature of active sites, we introduce the concept of subpockets. The subdivision allows a refined structural description of the topology of the entire active site. Additionally, the problem of predicting large pockets and overestimating the true ligand binding volume can be addressed by the subpocket concept. Subpockets correspond to the calculated cores that naturally emerge during the DoGSite pocket calculation. In order to compare DoGSite’s subpockets, we additionally introduce the subpocket concept into LSite. Herein, the detection of bottlenecks is based on a strategy to geometrically detect narrow regions in pockets with the scanning ray concept of LigSite.

A comparison of the predictive power of existing pocket detection algorithms is difficult. Many of them rely on an individual criterion for what is considered a correct prediction. Such success criteria are, for example, if the ligand can be found in one of the largest predicted pockets,[33] if predicted and experimental active site residues have a certain overlap,[41] or if any ligand atom lies within 4 Å of the center of the predicted pocket.[14] However, with the purpose of using pocket detection algorithms for function prediction and druggability studies, it is important to not only check whether the active site can be predicted at all, but which portion of the ligand is covered by the predicted binding pocket. Moreover, many pocket prediction methods do not evaluate the pocket volume. Volume is a major indicator for the quality of the method. Larger pocket volumes increase the chance of finding the ligand in the predicted pocket.[16] Similar to two recent studies, which considered ligand-occupied and unoccupied pocket fractions[35] or alpha sphere occupancy,[36] we introduced the concept of ligand and pocket coverage for a more objective performance assessment.

Furthermore, different data sets and structure preparation steps are used throughout the literature. To compare our method to the previously published methods, we follow a recent study[35] on a data set of 48 bound/unbound structures and show that DoGSite operates within the top-performing methods.

Many prediction algorithms were only evaluated on monomeric structures. Pocket prediction for multimeric structures is even more challenging, since large pockets may be predicted at interface regions between protein domains. Most methods rank the predicted pockets by size. This procedure weakens the success rate because interface regions suppress other smaller binding pockets. Nevertheless, examples show that ligands, e.g., sugars binding at lectins,[58] and also drugs, e.g., HIV-protease inhibitor,[59] bind at such interdomain or interprotein regions. Therefore, they must not be neglected.[60] We perform a comparison of DoGSite, DSite, and LSite on two benchmark data sets, PDBBind[61] and scPDB,[62] containing 828 and 6754 structures with druglike ligands, respectively. Predictions based on the complete structures are challenging due to potential interface regions. To analyze the difference in prediction power, we perform a second run where the data sets are truncated to monomeric protein structures.

In 93% of the PDBBind data set DoGSite ranks the cavity containing at least one ligand atom among the top three predicted pockets (Top3). When limiting the Top3 hits to pockets that cover at least half of the cocrystallized ligand (Top3_{LC}), the success rate is 90%. Further restricting Top3_{LC} hits to the fact that the cocrystallized ligand takes up at least a quarter of the pocket ($\text{Top3}_{\text{L+PC}}$) shifts the success rate toward more objective 70%. DoGSite outperforms DSite and LSite. Furthermore, DoGSite shows advantages, when considering ligand coverage, expressed in a more adequate pocket ceiling definition.

Since conformational changes in the binding site can affect the shape of the predicted pockets, the effect of protein flexibility on the pocket volume is studied. Pockets are predicted for 124 HIV-proteases and the predicted volumes are compared to each other. Only small variations in volume are observed for most of the structures, verifying the use of the currently rigid pocket model implemented in DoGSite.

As proof of the subpocket concept, we show that the ligand is completely contained in one subpocket in 87% of the predictions. Furthermore, splitting pockets into subpockets shrinks the average volume by one-third compared to the complete pocket volume. The advantage of the more granular subpocket description becomes apparent in an increase in success rate from 70% to 83%, when considering subpocket coverage instead of pocket coverage ($\text{Top3}_{\text{L+SPC}}$).

MATERIAL AND METHODS

Pocket and Subpocket Detection. Like the previously published geometry-based LigSite[24] and the energy-based DrugSite[41] approach, the new active site detection method DoGSite performs its calculations on a grid representation of the protein. For all three methods, a rectangular grid containing the protein is created by utilizing a dynamical adaption of grid spacing. The bounding box of the grid is determined by the atoms with minimal and maximal Cartesian coordinates plus a padding of 2 Å. Per default, the grid spacing is set to 0.4 Å. To avoid high computational costs, this value is dynamically increased for large proteins, forcing the number of grid points to drop below a specified limit. The predicted pockets calculated with a lower resolution resemble the pockets predicted with a higher resolution closely. Minor changes can occur when ranking the pockets; nevertheless, those do not affect the overall prediction performance of the algorithm.

Grid points are labeled as either free or occupied. A point is occupied if it lies within the van der Waals (vdw) radius of any protein atom. The respective methods assign inherent geometrically or energetically motivated property values to all free grid points. Pockets are identified by merging all grid points fulfilling the method-dependent criteria as described below. Pockets whose constituting grid points span a volume of less than 100 Å^3 are discarded. All remaining pockets are sorted according to contained volume, since former studies showed that the actual ligand binding site usually coincides with the largest protein pocket.[63,64]

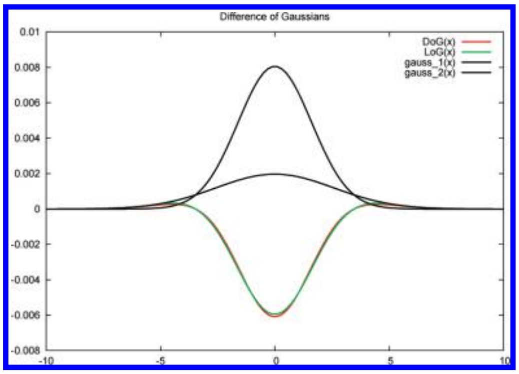


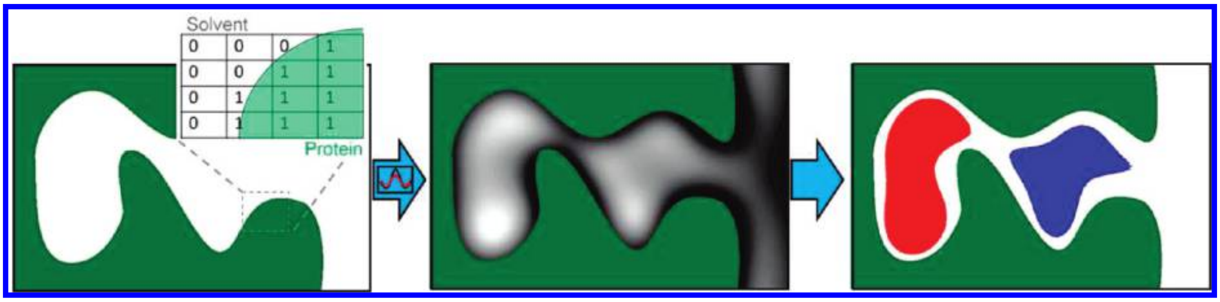
Figure 1. Two Gaussian functions with different σ are drawn in black lines. The difference of the two Gaussian functions and the Laplacian of a Gaussian with corresponding σ are plotted in red and green, respectively.

To analyze the topology of the pockets, we introduce the concept of subpockets individually to DoGSite and LSite. DoGSite first finds regions which resemble subpockets cores and pockets originate after dilating and merging them. LSite, in contrast, assigns pockets first and subsequently splits them into subpockets. For both methods, we use the terms core and bottleneck grid points. The subpocket core is described by core grid points; bottlenecks are the grid points joining the individual subpockets to pockets. In the end a volume check is performed. Subpockets having a volume smaller than 40 Å^3 are merged and the resulting subpockets are sorted by volume.

DoGSite Approach. DoGSite utilizes a 3D Difference of Gaussians[57] (DoG) to filter the grid representation of the protein. This filter approximates the second derivative of the Gaussian function (Laplacian of Gaussian). For a specific width of the Gaussian σ, the filter has a large response in regions which resemble spherical structures with an approximate radius of σ (Figure 1). Therefore, this filter can be used to identify invaginations of the protein surface, which are suitable to accommodate ligand atoms. Approximating frequent ligand atom radii, we use a σ of 1.75 Å in our implementation.

The DoG filtering procedure requires two passes of a Gaussian filter with different radii.[65] The subtraction of the filtered grids yields an approximation of the convolution with the Laplacian of Gaussian up to a constant factor. This is an efficient operation, since the Gaussian function is separable: A convolution with a Laplacian of Gaussian filter takes $O(m^3n)$, where m is the diameter of the filter expressed in the number of filter grid points and n is the total number of grid points. A convolution with the separable DoG filter as implemented in DoGSite requires asymptotic runtime of only $O(mn)$.

The filter response is assigned to each grid point (Figure 2). In the following, grid points with highly negative values outside the protein are identified as cores. This is done by thresholding the grid at $T = \text{Mean(grid)} - s \times \text{Std\_Dev(grid)}$ similarly to the DrugSite approach. To adjust the sensitivity of the algorithm, s is set to 3.25. After this process, each grid point below the threshold is set to active, considering it as part of a pocket (Figure 2).



0	0	0	1
0	0	1	1
0	1	1	1
0	1	1	1

Figure 2. A simplified 2D test scenario for the DoGSite pocket calculation is shown. From left to right: Protein and solvent are drawn in green and white, respectively; grid points are labeled accordingly and the filter procedure is started. The resulting grid is colored black to white, depending on the DoG value of each grid point. The white points represent good positions for placing a ligand atom. These points are clustered to two subpocket cores, shown in red and blue.

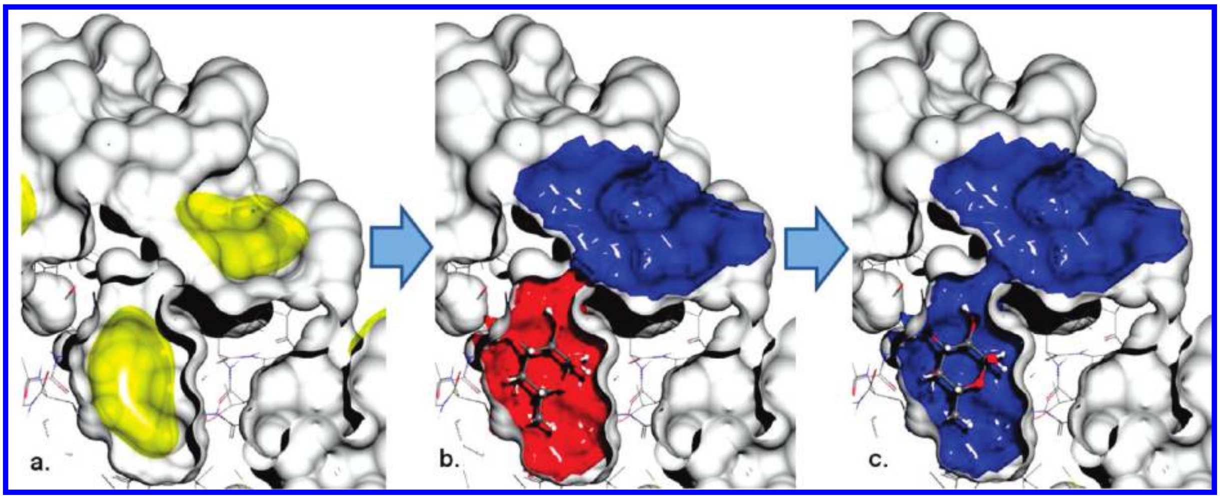


Figure 3. DoGSite pocket and subpocket detection for a sugar binding protein (1abf). A cross-section through the active site is shown; the protein surface is drawn in gray. (a) Pocket cores (yellow) originate from merging grid points fulfilling the DoGSite threshold criterion. (b) Dilating the identified cores automatically yields in this case two subpockets (blue and red). (c) The final pocket (blue) is formed by merging neighboring subpockets.

In Figure 3, pocket and subpocket detection upon the thresholded grid is shown exemplary on a sugar binding protein. Contiguous regions of active grid points are identified by merging neighboring active grid points. This results in so-called DoG cores (Figure 3a). These cores describe regions in which the placement of a ligand atom is favorable and represent subpocket cores. Regions where the intensity of the DoG filtered grid gets closer to zero specify bottlenecks. In a next step, subpocket cores are dilated by a radius of 2 Å (Figure 3b). This dilation is only performed for grid points lying between the core and the protein surface; expansion toward the solvent is not carried out. Neighboring subpockets are then merged and assembled to final pockets (Figure 3c). Using these steps, the DoGSite algorithm yields a list of pockets that are ranked by the volume of their DoG cores. Each pocket is composed of a list of subpockets. These are ranked by volume as well and represent the topology of the pocket.

LSite Approach. Our LigSite implementation LSite locates so-called "protein−solvent−protein" events (PSP)[24] by scanning along seven directions − the *x*, *y*, and *z* axes as well as the four cubic diagonals − on each grid point. Scanning in that manner detects solvent regions that are on both sides enclosed by protein atoms. These PSP events yield information on the buriedness of each grid point indicated by a value between zero and seven. Thus, a PSP value of seven corresponds to a high degree of buriedness, whereas a value of zero indicates a shallow region.

Pockets are calculated by merging all grid points having a buriedness value beyond a specified cutoff. In contrast to the original LigSite implementation, we do not use a fixed buriedness cutoff. The cutoff is calculated during runtime, guarantying to provide at least 15% of the free grid points as potential pocket points. In practice, this leads to a cutoff between four and seven, allowing more freedom in detecting shallow as well as deep pockets.

After LSite pockets have been calculated (Figure 4), the subpocket procedure starts. Grid points are labeled as either core or bottleneck, based on the distance to the protein atoms around each grid point. For pocket detection, the PSP value represents the number of scan lines that are restricted by the protein on both sides; i.e., the line touches a protein atom in both directions with a total line length below 10 Å. For the identification of subpockets, we use this criterion with a distance threshold of 4 Å, further referred to as short line. This value is chosen according to frequent ligand atom diameters. Grid points with at least four short lines out of seven are identified as bottleneck points (Figure 4a). All other grid points of the pocket are merged and represent subpocket cores (Figure 4b). Subsequently, the bottleneck points of the identified pocket are allocated to their closest core region, yielding the final set of subpockets (Figure 4c).

DSite Approach. The DrugSite algorithm[41] assigns an energy value to each grid point by placing a carbon probe on it and calculating the vdw energies between the probe and the surrounding protein atoms within a distance of 8 Å.

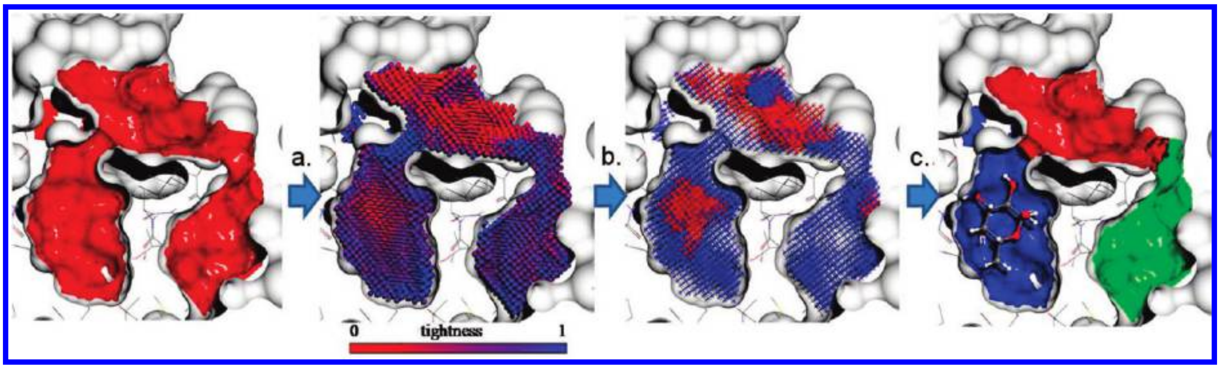


Figure 4. LSite subpocket detection for a sugar binding protein (1abf). A cross-section through the active site is shown; the protein surface is drawn in gray. Starting from predicted pocket grid points (red), a bottleneck value is assigned on the basis of the lengths of the PSP scan lines (a). The lower the value the more space is around a point (red). A cutoff separates core (red) from bottleneck (blue) grid points (b), and the three subpockets (blue, red, green) emerge by clustering the bottleneck points to their closest core (c).

In DSite, this energy is determined by the potential function $E = k_{pj}(1.0/d^{12} - 2.0/d^{6})$, where d is the distance between probe p and atom j divided by the sum of the vdw radii; k_{pj} is the geometric mean of the SYBYL constants k from the Tripos 5.2 force field associated with each atom.[66] All calculated energy values above an energy threshold of −0.4 kcal are truncated to zero. The genuine DrugSite algorithm applies a moving average filter 10 times to the grid to find energetic hot-spot volumes. This filter smoothens the potential with the effect of energy conservation on concave protein surface parts. In contrast to the moving average filter, we utilize a Gaussian filter with a σ of 6 Å. This choice is viable, since the iterative application of the moving average filter approximates a Gaussian filter. Furthermore, the direct application of the implemented Gaussian filter is more efficient and introduces fewer artifacts.

The contour level CL = Mean(grid) − s × Std_Dev(grid) defines which grid points are preserved as pocket grid points. Pockets are eventually identified by merging neighboring grid points with energy value below the CL. A threshold that lies $s = 3.5$ standard deviations below the grid's mean seems to be appropriate for the identification of pockets.

Data Preparation. To evaluate the methods, two benchmark data sets PDBBind[61,67] and scPDB[62] are used. Both data sets contain only structures with drug-like ligands and were investigated in several docking and druggability studies. The protein structures of both data sets are downloaded from the PDB,[68] while corresponding ligands are taken from the Ligand Expo database.[69] Structures are discarded, if the ligand has less than four or more than 80 heavy atoms or contains an ion or a metal. This leads to a total of 828 structures from PDBBind and 6754 structures from the scPDB data set. The biological unit of a protein is assembled of one monomere or multimer. Further, some protein structures in the PDB contain multiple copies of the biological unit. Therefore, a test subset is created for each original set containing only the monomeric structures. For these monomeric sets, only the first chain, with at least 50 heavy atoms, is considered for each structure representing the monomeric unit of the protein. Predictions of pockets located at interface regions cannot be performed for monomeric structures. A correct active site prediction based on one chain must fail here, since the counter residues from the second chain are missing. Therefore, structures are excluded from the monomeric test set, if their ligands bind at a dimer interface, defined as lying in the proximity (4 Å) of more than one chain. This leads to a final number of 581 monomeric structures from the PDBBind and 4915 structures from the scPDB data set.

To compare DoGSite to existing methods, the 48 structures from the unbound/bound data set from Tripathi and Kellogg[35] are used. All structures are downloaded from the PDB. Protein chains are removed for five structures (1cdo, 5cna, 1igj, 1swb, and 1a4j) as previously described.[33] To compare active site predictions for the unbound structures with the actual binding pocket, unbound structures are aligned with the corresponding complex using the "align" command of YASARA.[70]

Finally, a set of 124 HIV-proteases is used to analyze the effect of protein flexibility on the predicted pockets. The structures are collected from the HIV-protease database of the National Cancer Institute.[71]

RESULTS AND DISCUSSION

At first, the three implemented algorithms are compared to all algorithms evaluated on the 48 bound/unbound data set[35] with respect to pockets and subpockets. Subsequently, a detailed evaluation of DoGSite on the PDBBind data set is performed, including a comparison to the predictive power of LSite and DSite. We focus on the evaluation of ligand and pocket coverage. Additionally, we compare the results achieved for multimeric versus monomeric protein structures. Furthermore, success rates for the scPDB data set are given. Second, the effect of protein flexibility on the volume of the predicted pockets is analyzed. Finally, the subpockets methodology is investigated and its superiority compared to pockets with respect to coverage is shown.

Pocket Detection Performance. Evaluations of existing pocket detection algorithms vary substantially in the definition of a "correct prediction". Furthermore, different data sets with differently prepared structures are used, which makes a direct comparison difficult. LigSite[24] was evaluated on a data set of only 10 protein-ligand complexes which allows a manual checking of the correctness of predicted pockets. In the initial publication of DrugSite,[41] a set of 4711 PDB structures was used, containing only structures where the ligand does not bind at dimer interfaces. As a success criterion the relative overlap between experimental and predicted binding site residues was employed. A success rate of 98.8% is reported for nonzero overlap. In 2009, Halgren et al.[14] defined a prediction to be a hit if the geometric center

Table 2. Pocket Prediction Success Rates for 48 Bound and 48 Unbound Protein Structures in Percent

	Top1[b]		Top3[b]	
method[a]	unbound	bound	unbound	bound
VICE	83	85	90	94
DoGSite	71	83	92	92
Fpocket	69	83	94	92
LSite	75	75	85	88
PocketPicker	69	72	85	85
DSite	65	69	77	79
LIGSITE	58	69	75	87
CAST	58	67	75	83
PASS	60	63	71	81
SURFNET	52	54	75	78

[a] Success rates present the percentage of cases in which the active site was calculated by the respective algorithms.[35] Values for DoGSite, LSite, and DSite have been calculated in this study. [b] Top1/Top3: ligand was found in the largest pocket/in one of the three largest pockets, respectively.

of the presumed pocket lies within 4 Å of any atom of the ligand. The method was evaluated on a subset of 583 structures from the PDBBind data set containing only monomers. A success rate of 85.9% considering only the best scoring site was achieved. In a former publication, Weisel et al.[33] used a smaller data set of 48 bound/unbound protein structures to compare PocketPicker to the existing methods CAST,[25] LIGSITE,[24] LIGSITEcs,[51] PASS,[30] and SURFNET.[22] PocketPicker outperformed the mentioned methods with the criterion of predicting the true active site under the top one and top three ranked pockets. Recently, this study was followed up by LeGuilloux[36] and Tripathi,[35] showing that Fpocket and VICE, respectively, perform better than the previously published methods on the 48 bound/unbound data set.

We compare DoGSite to the latest published results.[35] According to Weisel,[33] a prediction is considered correct if for a specific protein structure the geometric center of the largest pocket lies within 4 Å of any ligand atom (Top1) or if this holds for one of the three largest predicted pockets (Top3). With 92% correct Top3 predictions on both sets, DoGSite is in line with the top-performing methods VICE and Fpocket (Table 2). Restricting the evaluation to Top1 predictions diminishes the success rates of all provided algorithms. DoGSite correctly predicts pockets in 83% of the bound data and 71% of the unbound data as Top1, which is due to ranking problems discussed in the following. Furthermore, LSite and DSite are among the well performing methods on both data sets. LSite outperforms the original LigSite implementation. This may be due to the dynamic buriedness cutoff, which is individually calculated for each protein. Making an objective statement for DSite is difficult, since the original implementation has neither been evaluated on this data set nor is it available to us. Our parametrized DSite method resembling the DrugSite method as close as possible yields slightly lower hit rates than DoGSite and LSite.

Even though the objective of our subpocket concept is to rather add a more granular level to the pockets than to replace them, we are interested in the subpocket specific performance of DoGSite$_{\mathrm{SPoc}}$. For this purpose, DoGSite subpockets are calculated as described before. Instead of merging them into pockets, they are treated as individual pockets and ranked by size. DoGSite$_{\mathrm{SPoc}}$'s hit rates increase to 94% correct Top3 predictions on bound and unbound data. Nevertheless, Top1 predictions decrease to 71% for unbound and 79% for bound data. This is due to the fact that not merging the pockets raises the number of predicted pockets and hinders ranking. Therefore, the performance is analyzed for the case that more than three pockets are considered. Allowing the consideration of up to 10 pockets, DoGSite$_{\mathrm{SPoc}}$ is able to successfully predict 98% of the true active sites among the Top4 for the bound structures and 96% among the Top7 for the unbound structures.

In the following, the findings are discussed in detail for the PDBBind data set, while the results for the scPDB data set are briefly summarized. Top1 and Top3 predictions specify that at least one ligand atom is covered by the largest and the largest three pockets, respectively. All three algorithms show similar results on the PDBBind data set (Table 3) with high success rates of 93% (DoGSite), 92.5% (LSite), and 89.5% (DSite) for the criterion of finding the ligand within one of the Top3 predicted pockets. While DSite performs slightly worse than DoGSite and LSite for Top3 predictions, it heads with 83% correctly detected Top1 pockets, compared to 76% Top1 hits predicted by DoGSite and LSite.

Considering the percentage of cases in which a true known active site cannot be predicted at all, referred to as "none" in Table 3, DoGSite performs best with a failure rate of 1.8%. LSite and DSite show higher failure rates of 3% and 10%, respectively. The active sites of 12 structures could not be predicted by any of the three algorithms. Most of these structures possess solvent-exposed ligands. Six structures for example describe subunits of lectins (1rdi, 1rdl, 1rdj, 1rdn, 1bcj, 1jlx), which need a second unit to form the active site. If the crystal structures include only a single fragment (e.g., 1fwv, 1fwu) or subunit (e.g., 1tyr, 1bm7) of the target protein, the three algorithms were not able to detect the respective binding pocket.

The cases for which DoGSite and LSite both fail form a subset of those of DSite. LSite and DSite do not find 12 active sites that are successfully predicted by DoGSite; five of them are among the Top3 predictions. Figure 5 shows one example out of those five, where DoGSite correctly predicts the true binding pocket, while the other two algorithms completely fail in pocket prediction. The cocrystallized ligand is solvent-exposed. Nevertheless, DoGSite predicts a Top1 pocket that covers most part of the cocrystallized ligand.

The true active site usually coincides with the largest protein pocket.[63,64] Thus, most algorithms rank their solutions by size, i.e., pocket volume. Accordingly, when evaluating the prediction performance, algorithms that detect fewer pockets and pockets with more distinct volumes benefit from this criterion. First, the three methods we discuss here differ in the number of pockets that they predict per structure. DoGSite detects on average 12 pockets per protein, while LSite and DSite predict five and two pockets, respectively. Second, compared to LSite and DSite, DoGSite finds more pockets of similar volume, which makes the ranking especially among the top ranking positions more difficult. In Figure 6, a volume box plot of the predicted pockets is shown. For each structure, the pocket that accommodates the cocrystallized ligand is taken as reference. If the ligand

Table 3. Pocket Prediction Success Rates for DoGSite, LSite, and DSite on the PDBBind Data Set in Percent

algorithm	Poc1[a]	Poc2[a]	Poc3[a]	Poc>3[a]	none[a]	Top3[b]	Top3_{LC}[b]	$\text{Top3}_{\text{L+PC}}$[b]
DoGSite	76.33	12.44	4.47	4.95	1.81	93.24	89.61	70.29
LSite	76.57	10.75	4.95	4.71	3.02	92.27	83.70	59.78
DSite	82.97	4.95	1.58	0.36	10.14	89.49	78.62	44.44

[a] Poc1/Poc2/Poc3: Ligand was found in largest/second/third largest pocket. Poc>3: Ligand was found in a pocket with rank higher than three. None: Ligand was not found in any pocket. [b] Top3: Ligand was found in one of the three largest pockets. Top3_{LC}: Ligand coverage above 50% required. $\text{Top3}_{\text{L+PC}}$: Additionally, pocket coverage above 25% is required.

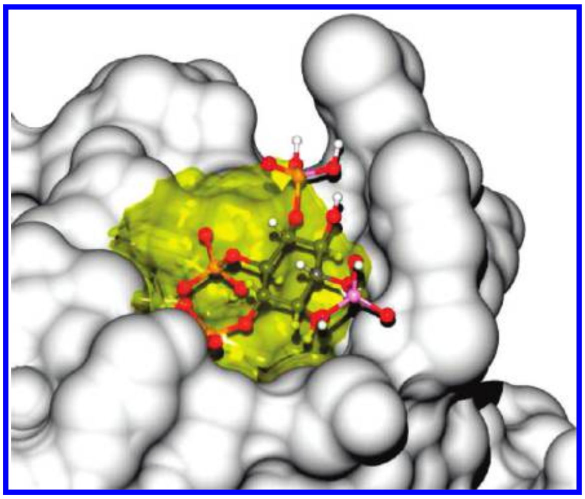


Figure 5. Structure of the active site of a signaling protein (1fao) with solvent-exposed ligand. The protein surface is drawn in gray; the pocket predicted by DoGSite is shown in yellow.

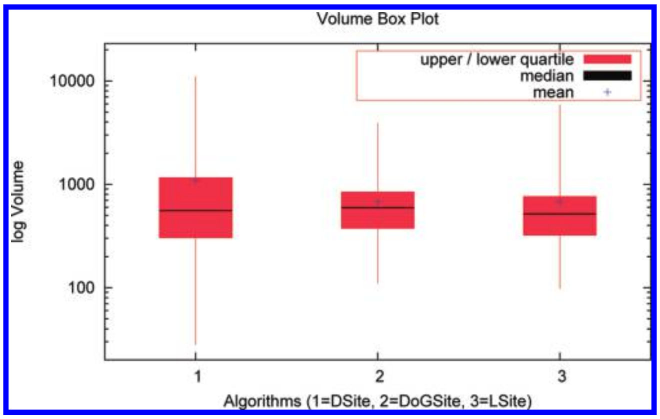


Figure 6. A volume box plot of the pockets predicted by the three algorithms DSite, DoGSite, and LSite is shown. Upper and lower quartile are drawn in red, and the median value is represented by the horizontal black line, while the arithmetic mean is indicated by a blue cross.

is not found in any predicted pocket, the largest pocket is used instead. While all three algorithms show similar median volume values around 550 Å^3, DoGSite exhibits the smallest standard deviation. LSite and DoGSite both predict pockets having an average volume of about 680 Å^3. Pockets computed by DSite have an average volume of 1090 Å^3. Tripathi et al.[35] reported pocket predictions of comparable size on their curated data set with an average volume of 1170 Å^3.

Predicting multiple, rather reasonably sized pockets is advantageous for druggability studies. Many proteins have more than one biochemical role and possess multiple active sites, e.g., allosteric ligand binding sites.[35] Thus, DoGSite's top ranking solutions provide starting points for further drug development. Descriptors can be calculated quickly for that number of predicted pockets, and analysis of these pockets will cope with the trend in drug development toward auxiliary pockets.

The restriction of the pocket volume toward the solvent by a proper definition of the pocket boundary is important for the specificity of the predicted pockets. Each of the three algorithms has an individual pocket ceiling definition, which is illustrated in Figure 7 for dihydrofolate reductase (4dfr). DoGSite predicts a pocket that completely covers the cocrystallized inhibitor, whereas LSite misses an important part of the inhibitor. Due to the buriedness criterion, LSite is inclined to predict a more concave pocket ceiling, cutting off parts of the ligand accessible volume. In contrast, DoGSite predicts rather convex pocket ceilings and better reproduces the true ligand binding volume. DSite also exhibits a convex pocket ceiling character, even though it only covers a part of the inhibitor. This weaker performance may be due to changes in the energy parameters and filter procedure that we introduced in our implementation and does not have to reflect the performance of the original DrugSite algorithm.

In the following, we discuss the correlation between pocket volume and pocket surface. DoGSite predicts the smallest sized protein covering pocket surfaces with an average of 305 Å^2, followed by LSite with 405 Å^2. DSite predicts on average a protein covering pocket surface of 534 Å^2, which resembles LSite closer than comparing the predicted pocket volumes. This effect can be better described by comparing the average volume to surface ratio of pockets predicted by the individual algorithms. DoGSite and DSite both have a ratio above 2, whereas LSite has a ratio of 1.7. This is a results of the different pocket ceiling definition. The more concave ceiling better resembles the true ligand binding volume.

As discussed in the last section, pocket volume is an important descriptor and indicator for druggability. Drugs are found to bind at large surface cavities.[13] As already shown by high prediction success rates (Table 3), the vague prediction of the active site region is not the problem for most structures, rather the prediction of pockets with large portions of ligand free volumes. For descriptor-based protein druggability or function estimates, the rough location of the active site is not sufficient. The derivation of descriptors requires a correct and restrictive representation of the active site's topology and its boundaries. Therefore, we use the pairwise coverage of the ligand and the predicted pocket as indicator for correctly defining the ligand binding pocket (Figure 8).

Similar ideas have been introduced in two lately published pocket detection methods: Fpocket[36] uses a mutual overlap criterion (MOc) to evaluate the pocket detection performance based on the underlying alpha spheres. VICE[35] uses a grid-

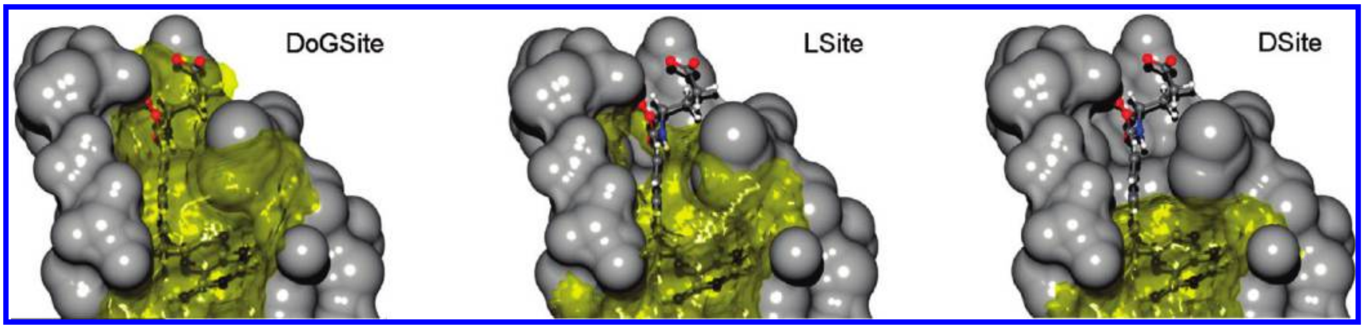


Figure 7. The difference in pocket ceiling in DoGSite (left), LSite (middle), and DSite (right) for the dihydrofolate reductase (4dfr) is shown. The surface of the active site is drawn in gray and the predicted pockets in yellow. DoGSite predicts a pocket that completely covers the cocrystallized inhibitor, whereas LSite and DSite miss a part of the inhibitor.

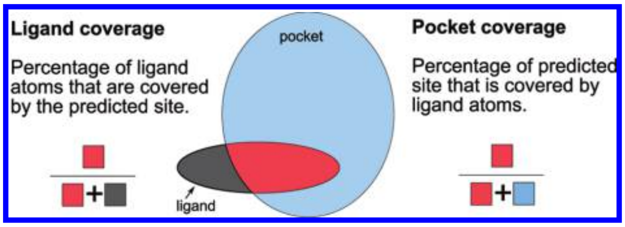

Ligand coverage
Percentage of ligand atoms that are covered by the predicted site.
pocket
ligand
Pocket coverage
Percentage of predicted site that is covered by ligand atoms.

Figure 8. Schematic depiction of the new evaluation concept: Pairwise ligand and pocket coverage.

based vector technique to detect pockets. Ligand occupied and free fractions of the predicted sites were discussed in the paper but not used for the evaluation of the pocket prediction performance.

Ligand coverage in this work is calculated as the percentage of ligand atoms covered by the predicted pocket. Covered, in this context, means that at least one pocket grid point lies within the vdw radius of the ligand atom. On the PDBBind data set, high average ligand coverages of 90% (DoGSite), 83% (LSite), and 74.5% (DSite) are achieved by the individual algorithms. A good prediction should contain most parts of the cocrystallized ligand. To increase the specificity of the pocket prediction, we utilize a stricter criterion for a correct prediction. Top3 pockets are only valuable hits, if they cover at least half of the ligand ($\mathrm{Top3}_{\mathrm{LC}}$). This shifts the success rates from 93% (DoGSite), 92.5% (LSite), and 89.5% (DSite) toward 90%, 84%, and 79%, respectively. The high stability of DoGSite's success rate is due to the more convex pocket ceiling, which better resembles the true ligand covering binding volume.

Pocket coverage is defined as the percentage of pocket grid points covered by ligand atoms. Compared to the ligand coverage, the average pocket coverage of all three algorithms is about a factor of 2 lower. DoGSite shows the best average pocket coverage of 43%, followed by LSite with 38.5% and DSite with 31%.

Other studies showed that ligands do not completely fill known active sites. For example, pockets predicted by VICE[35] have average occupancy values of 35−50%. With respect to the fact that only the cocrystallized ligand and not a union of all known ligands is considered for each protein, a pocket coverage above 25% seems reasonable as an indicator for a correct prediction. Similarly, Fpocket[36] requires in its MOc that at least 50% of the ligand atoms lie within 3 Å of at least one alpha sphere and 20% of the pocket alpha spheres lie within 3 Å of the ligand. Therefore, we restrict our criterion in two ways. First, the pocket has to cover at least half of the ligand. Second, the ligand itself must cover at least a quarter of the predicted pocket ($\mathrm{Top3}_{\mathrm{L+PC}}$). This restriction leads to success rates of 70% for DoGSite and 60% for LSite. Both algorithms outperform DSite showing a success rate of 44.5%. These values are substantially smaller than those achieved with other criteria. We believe that such a restrictive criterion is more objective and therefore more suitable for druggability studies.

As a result of employing pocket size as ranking criterion, several algorithms encounter problems considering multidomain proteins. In these cases, the largest pocket is often found at interface regions. This is a challenge for finding the true ligand binding pocket under the top ranking positions. Therefore, some methods were only evaluated on structures where the ligand does not bind at the interface.[14,41] In order to circumvent a misrepresentation of the prediction rates by these difficult cases, we additionally evaluate the algorithms on the monomeric subsets of the two benchmark data sets. Evaluating the PDBBind data subset of 581 structures, the success rates slightly increase. DoGSite and LSite both achieve approximately 95%; DSite's performance stays constant with 89.5%. The better performance is due to the absence of one large interface pocket which simplifies the ranking.

For the sake of completeness, all values for the larger scPDB data set are presented in Table 4. All statements discussed for the PDBBind data set equally hold true for the scPDB data set. With respect to the Top3 criterion, success rates above 91% were achieved on the whole set of 6754 structures for all three algorithms. The $\mathrm{Top3}_{\mathrm{LC}}$ criterion yields 89%, 84%, and 83% for DoGSite, LSite, and DSite, respectively. Lower rates are achieved for the $\mathrm{Top3}_{\mathrm{L+PC}}$ criterion with 66%, 57%, and 34%.

Pocket Flexibility Analysis. Protein flexibility adds to the complexity of pocket predictions and has been discussed in several studies.[72−74] Therefore, the impact of pocket flexibility on individual descriptors like the pocket volume is discussed here. A data set of 124 HIV-proteases is used to analyze the distribution of the volumes of the predicted pockets. The predicted pocket volumes accumulate around a median of 810 Å^3 with a standard deviation of 175 Å^3. In Figure 9, the size of the cocrystallized ligand is plotted against the volume of the predicted pocket.

A couple of outliers can be found with volumes below 600 Å^3 and small ligands. Many HIV-proteases have a covalently bound peptide inhibitor in the active site of length two to nine (indicated by the circular data points in Figure 9). Since the bound peptide belongs to the protein, the pocket

Table 4. Pocket Prediction Success Rates for DoGSite, LSite, and DSite on the scPDB Data Set in Percent

algorithm	Poc1[a]	Poc2[a]	Poc3[a]	Poc>3[a]	none[a]	Top3[b]	Top3$_{\text{LC}}$[b]	Top3$_{\text{L+PC}}$[b]
DoGSite	76.69	10.50	4.50	6.77	1.54	91.69	88.48	66.21
LSite	76.25	11.47	4.06	4.25	3.96	91.78	83.72	56.97
DSite	84.26	6.33	1.25	0.67	7.48	91.86	82.66	33.56

[a] Poc1/Poc2/Poc3: Ligand was found in largest/second/third largest pocket. Poc>3: Ligand was found in a pocket with rank higher than three. None: Ligand was not found in any pocket. [b] Top3: Ligand was found in one of the three largest pockets. Top3$_{\text{LC}}$: Ligand coverage above 50% required. Top3$_{\text{L+PC}}$: Additionally, pocket coverage above 25% required.

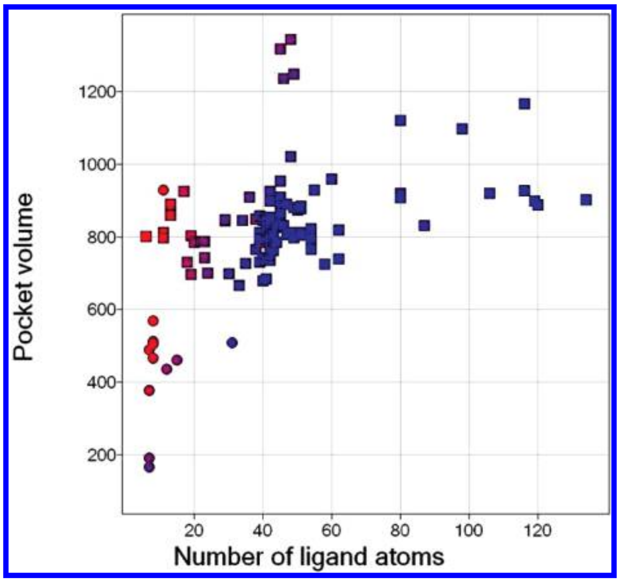


Figure 9. Scatter plot of ligand size (*x*-axis) versus predicted pocket volume (*y*-axis) on a set of 124 HIV-proteases. Shapes: circle, peptide bound to active site; square, no covalent binding partner in active site. Color coding by pocket coverage: red (0%) to blue (100%).

is already partly covered. Therefore, these structures are excluded from the analysis. Four multidrug-resistant HIV-proteases (1k6p, 1k6c, 1k6t, and 1k6v) form a group of outliers with pocket volumes above 1350 Å^3. These HIV-proteases underwent mutations in the active site to become multidrug-resistant. Most likely, these mutations cause the large difference in pocket volume.

For the remaining structures (rectangular data points in Figure 9) an interesting observation can be made by focusing on the size of the cocrystallized ligands. The number of atoms per ligand ranges from six to 134. Nevertheless, the change in pocket volume is not proportional to the change in ligand size. In contrast, the pocket coverage changes with the ligand size, as indicated by the color coding in Figure 9. While small ligands only occupy parts of the pocket (indicated by a red color), the pocket is almost completely filled by the larger ligands (indicated by a blue color).

This constant pocket volume behavior shows that the binding site is stable to a certain extent and does not adapt to the size of the bound ligand. This observation, together with the small volume range in which the pockets of the structures of interest fall, encourages a rigid pocket approach. Nevertheless, flexibility in the binding site matters. Some variation in the pocket volumes has been observed in this study. The impact of flexibility on local pocket descriptors will limit the success of rigid models even more. The explicit consideration of protein flexibility is therefore a reasonable approach for further studies.

Subpocket Analysis. The major difference of DoGSite to previously published methods is its ability to divide pockets into fine-granular subpockets. We compare the results achieved with DoGSite subpockets to those introduced to LSite. To show the relevance and the advantage of subpockets, we discuss the frequency of their occurrence and their influence on prediction sensitivity and specificity. For each protein structure the pocket which is covered by the ligand is considered. If the prediction fails, the largest pocket is taken into account. A total of 63% of the respective pockets predicted by DoGSite can be divided into subpockets; 50% of the pockets contain two to four subpockets. LSite identifies dividable pockets in 53% of the data set; two to four subpockets are predicted in 43% of the cases. DoGSite and LSite predict pockets of comparable size, but DoGSite splits a larger number of them into subpockets. LSite splits pockets at buried bottleneck parts, whereas DoGSite additionally partitions between surface exposed cores.

As proof that subpockets better describe real ligand binding regions, we show that the ligand is mostly contained in one subpocket and that we achieve higher pocket coverages with the subpocket approach. On the left side of Figure 10, ligand coverage by the predicted pocket is plotted against ligand coverage by the respective subpocket. The data is shown for a DoGSite run on the PDBBind data set. 87% of the data points lie on the main diagonal, showing that the ligand is completely contained in one subpocket. LSite performs similarly and predicts in 92% of the cases subpockets that do not split the ligand. However, in up to 13% of the predictions the coverage decreased indicating that the ligand is only partly contained in the predicted subpocket.

Figure 11 shows the structure of a human coagulation factor Xa (1mq9). In this case, DoGSite's ligand coverage drops from 93% to 56% when considering individual subpockets. The cocrystallized inhibitor binds to some extend at the surface. DoGSite predicts two subpockets that cut the inhibitor in two halves. The human coagulation factor Xa exhibits an S1 and S4 pocket, as described in literature.[75] The central phenyl ring forms a bridge between the substituents in the distinct subpockets through amide bonds. Therefore, we consider this example as a reasonable split. The predicted pocket of LSite (Figure 11) does not divide the ligand; it misses the more solvent exposed S4 pocket, underlining the previously mentioned advantage of DoGSite's pocket ceiling definition.

As mentioned in the pocket detection performance part, prediction of pockets with large ligand-free volumes is less useful for druggability studies. Although, cocrystallized ligands do not completely fill the active site, low predicted pocket coverage diminishes the descriptor accuracy. Using subpockets reduces the considered volume and simultaneously enhances the predicted pocket coverage. For volume

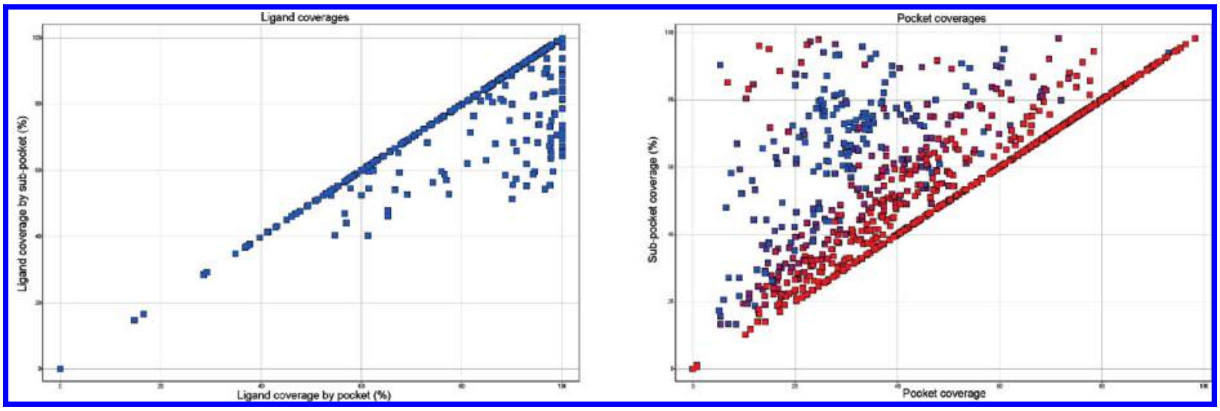


Figure 10. Ligand coverage (left) and pocket coverage (right) of the predicted pocket (*x*-axis) versus its respective subpocket (*y*-axis) on the PDBBind data set predicted by DoGSite are plotted. For pocket coverage, the ratio of the volumes of the two largest subpockets is indicated by the color coding. The color shading from red to blue represents a volume ratio from 0 to 1.

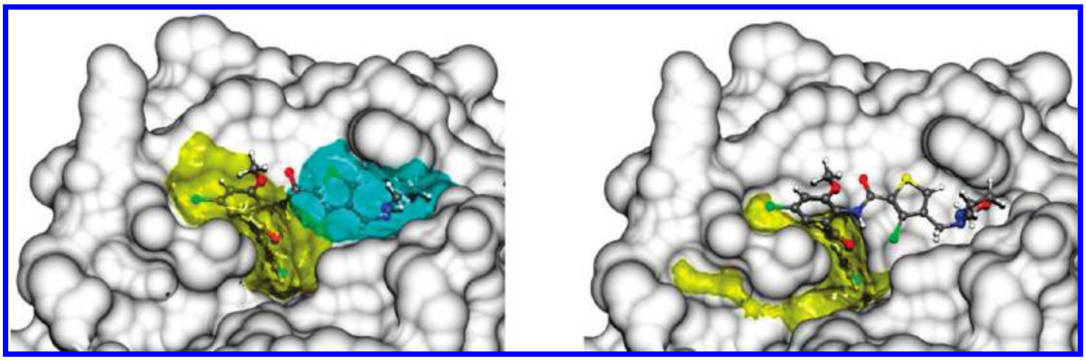


Figure 11. On the left side the subpockets predicted by DoGSite for a human coagulation factor Xa (1mq6) are shown. The protein surface is drawn in gray and the subpockets in yellow and blue. On the right side the respective pocket predicted by LSite is shown.

analysis, the volume of the ligand covering subpocket is considered as reference. If the pocket cannot be divided, the complete pocket volume is taken into account. The subpocket volume is on average two-thirds of the original pocket volume predicted by both algorithms. On the right side of Figure 10, the pocket coverage against the subpocket coverage calculated by DoGSite is plotted. The coverage remains unchanged for pockets that cannot be further divided, indicated by the dots on the main diagonal. For the remaining 60%, the prediction specificity increases, shown by dots lying above the diagonal. A large increase from 37% to 80% coverage has been observed for a catalytic antibody (1a0q) (Figure 12). The ligand is solvent exposed and the algorithm predicts a rather large pocket at the protein domain interface. A split of the pocket enables the ligand to fill exactly one subpocket.

To show that pockets are split into reasonable parts, we compare the ratio of the volumes of the largest to the second largest subpocket. Generally, the advantage of the subpocket concept is higher if the pockets can be divided into equally sized parts, as indicated by the color coding of Figure 10. A higher ratio value, indicated by a blue color, corresponds to a more significant enhancement in pocket coverage. A red color implies that the largest subpocket obeys most parts of the pocket. Nevertheless, some cases occur where a pocket is split into a large subpocket at the interface and a comparably small ligand-containing subpocket. This is indicated by the red dots in the upper left corner.

The objective of this work is to predict pockets that describe the active site as specifically as possible, valued by high pairwise ligand and pocket coverage. To evaluate the subpocket performance, the initial ranking by volume is performed on the pocket level. Subpockets further divide the initial pockets and the ligand covering subpocket is taken as reference. In total, the calculated success rates on the $\text{Top3}_{\text{L+PC}}$ criterion rise from 70% to 83% for DoGSite and from 60% to 74% for LSite using the subpockets approach. This increase in pocket coverage of up to 14% underlines the benefit of the subpocket concept.

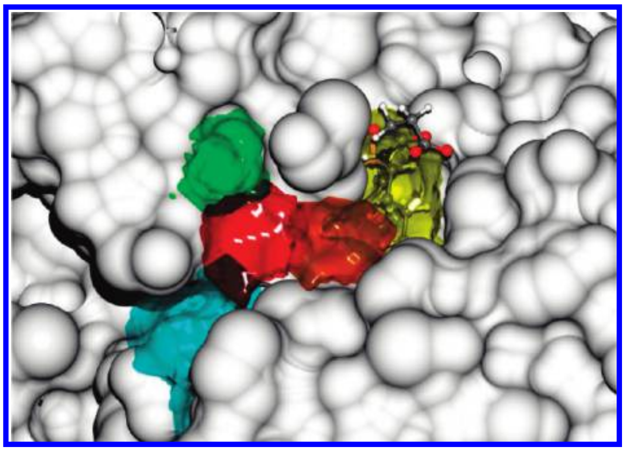


Figure 12. Active site of a catalytic antibody (1a0q) with inhibitor. The protein surface is shown in gray; subpockets predicted by DoGSite are drawn in yellow, orange, red, green, and blue. The inhibitor fits exactly one subpocket (yellow).

CONCLUSION

We presented DoGSite, a new pocket prediction algorithm featuring subpocket detection. On a set of 48 bound/

unbound structures,[35] DoGSite is in line with the best-performing methods published earlier (VICE, Fpocket). Furthermore, the pocket detection algorithm yields a Top3 success rates over 92% on the two large benchmark data sets PDBBind and scPDB. In comparison to LSite and DSite, DoGSite predicts pockets that better resemble the nature of true binding pockets. The predicted pockets show a more convex shape, with less branching and a more meaningful pocket ceiling. This is indicated by a high average ligand coverage of 90%.

The first step toward the detection of additional descriptors for druggability studies is to correctly describe the active site as good as possible Descriptor-based protein comparison and clustering rely on a realistic pocket volume definition. Indicators for correctly defining the ligand binding pocket are the ligand coverage and the pocket coverage criteria. A success rate of 90% is achieved by DoGSite, when restricting the success criterion to predicted pockets that cover at least half of the ligand ($\mathrm{Top3}_{\mathrm{LC}}$). Requiring that the ligand has to cover at least a quarter of the predicted pocket ($\mathrm{Top3}_{\mathrm{L+PC}}$) shifts the success rate from 90% to 70%. Predicted pockets fulfilling these criteria are more reliable starting points for protein function prediction and druggability studies.

With DoGSite, we introduced the subpocket terminology. This enables a novel representation of pockets and leads to a more realistic description of the active site. Since subpockets resemble the binding region of ligands more closely, they are better suited for protein function predictions and druggability studies. More than half of the pockets can be split into subpockets. Subpockets increase the specificity of the predictions while mostly conserving the selectivity. The average pocket coverage rises from 43% to 56% when considering subpockets. The success rate for the restrictive $\mathrm{Top3}_{\mathrm{L+PC}}$ criterion rises from 70% to 83% for DoGSite if subpockets instead of pockets are considered. Subpockets better describe the ligand accessible active site volume.

Especially fragment-based methods may benefit from this new subpocket representation. The revelation of additional, e.g., allosteric ligand binding sites is of high practical relevance for drug discovery. A subpocket-based approach enables a more restrictive definition of active sites making automated pocket detection more suitable for structure-based druggability studies.

Introducing flexibility into the pocket model will be the next step toward a realistic pocket representation and is vital for comparison of active sites and protein function prediction.

ACKNOWLEDGMENT

The project is part of the BIOCATALYSIS2021 cluster and is founded by the BMBF under grant 0315292A.

REFERENCES AND NOTES

(1) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

(2) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.

(3) Schellhammer, I.; Rarey, M. TrixX: Structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 223–238.

(4) Brooijmans, N.; Kuntz, I. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.

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

(6) Hopkins, A.; Groom, C. Target analysis: A priori assessment of druggability. *Ernst Schering Res. Found. Workshop* **2003**, 11–17.

(7) Hajduk, P.; Huth, J.; Tse, C. Predicting protein druggability. *Drug Discovery Today* **2005**, *10*, 1675–1682.

(8) Sakharkar, M.; Sakharkar, K.; Pervaiz, S. Druggability of human disease genes. *Int. J. Biochem. Cell Biol.* **2007**, *39*, 1156–1164.

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

(10) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. The worldwide Protein Data Bank (ww-PDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **2007**, *35*, D301–3.

(11) Hajduk, P.; Huth, J.; Fesik, S. Druggability indices for protein targets derived from NMR based screening data. *J. Med. Chem.* **2005**, *48*, 2518–2525.

(12) Cheng, A.; Coleman, R.; Smyth, K.; Cao, Q.; Soulard, P.; Caffrey, D.; Salzberg, A.; Huang, E. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.

(13) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892–906.

(14) Halgren, T. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.

(15) Hermann, J.; Marti-Arbona, R.; Fedorov, A.; Fedorov, E.; Almo, S.; Shoichet, B.; Raushel, F. Structure-based activity prediction for an enzyme of unknown function. *Nature* **2007**, *448*, 775–779.

(16) Laurie, A.; Jackson, R. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.* **2006**, *7*, 395–406.

(17) Ho, C.; Marshall, G. Cavity search: An algorithm for the isolation and display of cavity-like binding regions. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 337–354.

(18) Levitt, D.; Banaszak, L. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **1992**, *10*, 229–234.

(19) Delaney, J. Finding and filling protein cavities using cellular logic operations. *J. Mol. Graph.* **1992**, *10*, 174–7.

(20) Del Carpio, C.; Takahashi, Y.; Sasaki, S. A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I). Search for pocket regions. *J. Mol. Graph.* **1993**, *11*, 23–9.

(21) Kleywegt, G.; Jones, T. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.* **1994**, *50*, 178–185.

(22) Laskowski, R. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **1995**, *13*, 323–30.

(23) Peters, K.; 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.

(24) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359–63.

(25) 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.

(26) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **2006**, *34*, W116–8.

(27) Kuntz, I.; Blaney, J.; Oatley, S.; Langridge, R.; Ferrin, T. A geometric approach to macromolecule−ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(28) Hendrix, D.; Kuntz, I. Surface solid angle-based site points for molecular docking. *Pac. Symp. Biocomput.* **1998**, 317–326.

(29) Stahl, M.; Taroni, C.; Schneider, G. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng.* **2000**, *13*, 83–88.

(30) Brady, G.; Stouten, P. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383–401.

(31) Venkatachalam, C.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **2003**, *21*, 289–307.

(32) Coleman, R.; Sharp, K. Travel depth, a new shape descriptor for macromolecules: Application to ligand binding. *J. Mol. Biol.* **2006**, *362*, 441–458.

(33) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.

(34) Li, B.; Turuvekere, S.; Agrawal, M.; La, D.; Ramani, K.; Kihara, D. Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins* **2008**, *71*, 670–683.

(35) Tripathi, A.; Kellogg, G. A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins* **2010**, *78*, 825–842.

(36) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.

(37) Goodford, P. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(38) Ruppert, J.; Welch, W.; Jain, A. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997**, *6*, 524–533.

(39) Bliznyuk, A.; Gready, J. Identification and energetic ranking of possible docking sites for pterin on dihydrofolate reductase. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 325–333.

(40) Kortvelyesi, T.; Silberstein, M.; Dennis, S.; Vajda, S. Improved mapping of protein binding sites. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 173–186.

(41) An, J.; Totrov, M.; Abagyan, R. Comprehensive identification of “druggable” protein ligand binding sites. *Genome Inform.* **2004**, *15*, 31–41.

(42) Laurie, A.; Jackson, R. Q-SiteFinder: An energy-based method for the prediction of proteinligand binding sites. *Bioinformatics* **2005**, *21*, 1908–1916.

(43) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752–761.

(44) Zhong, S.; MacKerell, A. Binding response: A descriptor for selecting ligand binding site on protein surfaces. *J. Chem. Inf. Model.* **2007**, *47*, 2303–2315.

(45) Casari, G.; Sander, C.; Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **1995**, *2*, 171–178.

(46) de Rinaldis, M.; Ausiello, G.; Cesareni, G.; Helmer-Citterich, M. Three-dimensional profiles: A new tool to identify protein surface similarities. *J. Mol. Biol.* **1998**, *284*, 1211–1221.

(47) Aloy, P.; Querol, E.; Aviles, F.; Sternberg, M. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **2001**, *311*, 395–408.

(48) Armon, A.; Graur, D.; Ben-Tal, N. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **2001**, *307*, 447–463.

(49) Pupko, T.; Bell, R.; Mayrose, I.; Glaser, F.; Ben-Tal, N. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **2002**, *18* (Suppl 1), S71–7.

(50) Lichtarge, O.; Sowa, M. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **2002**, *12*, 21–27.

(51) Huang, B.; Schroeder, M. LIGSITEcsc: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19.

(52) Glaser, F.; Morris, R.; Najmanovich, R.; Laskowski, R.; Thornton, J. A method for localizing ligand binding pockets in protein structures. *Proteins* **2006**, *62*, 479–488.

(53) Capra, J.; Laskowski, R.; Thornton, J.; Singh, M.; Funkhouser, T. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **2009**, *5*, e1000585.

(54) Huang, B. MetaPocket: A meta approach to improve protein ligand binding site prediction. *Omics* **2009**, *13*, 325–330.

(55) Bray, T.; Chan, P.; Bougouffa, S.; Greaves, R.; Doig, A.; Warwicker, J. SitesIdentify: A protein functional site prediction tool. *BMC Bioinf.* **2009**, *10*, 379.

(56) Brylinski, M.; Skolnick, J. FINDSITE: A threading-based approach to ligand homology modeling. *PLoS Comput. Biol.* **2009**, *5*, e1000405.

(57) Marr, D.; Hildreth, E. Theory of edge detection. *Proc. R. Soc. London B Biol. Sci.* **1980**, *207*, 187–217.

(58) Ng, K.; Drickamer, K.; Weis, W. Structural analysis of monosaccharide recognition by rat liver mannose-binding protein. *J. Biol. Chem.* **1996**, *271*, 663–674.

(59) Umland, T.; Wolff, E.; Park, M.; Davies, D. A new crystal structure of deoxyhypusine synthase reveals the configuration of the active enzyme and of an enzyme.NAD.inhibitor ternary complex. *J. Biol. Chem.* **2004**, *279*, 28697–28705.

(60) Jones, S.; Marin, A.; Thornton, J. Protein domain interfaces: Characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **2000**, *13*, 77–82.

(61) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(62) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: An annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–727.

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

(64) Sotriffer, C.; Klebe, G. Identification and mapping of small-molecule binding sites in proteins: Computational tools for structure-based drug design. *Farmaco* **2002**, *57*, 243–251.

(65) Bomans, M.; Hohne, K.; Tiede, U.; Riemer, M. 3-D segmentation of MR images of the head for 3-D display. *IEEE Trans Med. Imaging* **1990**, *9*, 177–183.

(66) Clark, M.; Cramer, R.; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(67) Wang, R.; Fang, X.; Lu, Y.; Yang, C.; Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(68) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(69) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H.; Westbrook, J. Ligand Depot: A data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155.

(70) Krieger, E.; Koraimann, G.; Vriend, G. Increasing the precision of comparative models with YASARA NOVA−A self-parameterizing force field. *Proteins* **2002**, *47*, 393–402.

(71) Vondrasek, J.; van Buskirk, C.; Wlodawer, A. Database of three-dimensional structures of HIV proteinases. *Nat. Struct. Biol.* **1997**, *4*, 8.

(72) Li, X.; Keskin, O.; Ma, B.; Nussinov, R.; Liang, J. Protein−protein interactions: Hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: Implications for docking. *J. Mol. Biol.* **2004**, *344*, 781–795.

(73) Gunasekaran, K.; Nussinov, R. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J. Mol. Biol.* **2007**, *365*, 257–273.

(74) Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. Principles of flexible protein−protein docking. *Proteins* **2008**, *73*, 271–289.

(75) Adler, M.; Kochanny, M.; Ye, B.; Rumennik, G.; Light, D.; Biancalana, S.; Whitlow, M. Crystal structures of two potent nonamidine inhibitors bound to factor Xa. *Biochemistry* **2002**, *41*, 15514–15523.

CI100241Y