JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

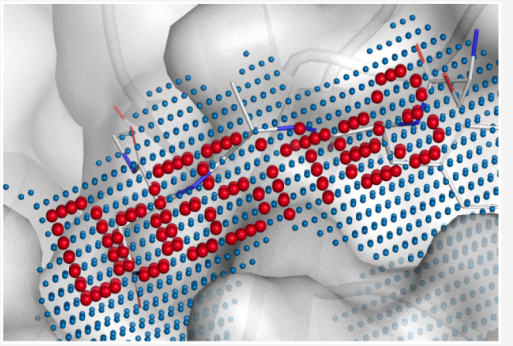
Binding Site Detection Remastered: Enabling Fast, Robust, and Reliable Binding Site Detection and Descriptor Calculation with DoGSite3

Joel Graef,[▽] Christiane Ehrt,[▽] and Matthias Rarey*

Cite This: *J. Chem. Inf. Model.* 2023, 63, 3128−3137 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Binding site prediction on protein structures is a crucial step in early phase drug discovery whenever experimental or predicted structure models are involved. DoGSite belongs to the widely used tools for this task. It is a grid-based method that uses a Difference-of-Gaussian filter to detect cavities on the protein surface. We recently reimplemented the first version of this method, released in 2010, focusing on improved binding site detection in the presence of ligands and optimized parameters for more robust, reliable, and fast predictions and binding site descriptor calculations. Here, we introduce the new version, DoGSite3, compare it to its predecessor, and re-evaluate DoGSite on published data sets for a large-scale comparative performance evaluation.



INTRODUCTION

One of the most critical steps in early drug discovery is the analysis of 3D protein structures, many of which are available from the Protein Data Bank (PDB).[1] In addition, structure models of high quality are available today for hundreds of thousands of proteins.[2,3] By identifying binding sites, functions can be predicted and proteins classified. Further, examining binding sites helps to identify novel active and regulatory target binding sites and supports drug design. For the latter, one prerequisite is finding druggable pockets, i.e., sites that can be bound by compounds with drug-like properties as defined by Lipinski's Rule of Five.[4−7] Assessing a binding site's druggability is one of the first steps before exploring it further, for example, by molecular docking or structure-based virtual screening.[8,9] Predicting where a ligand molecule can interact with a protein structure is undoubtedly an essential step in the drug design process. It necessitates the development of very precise in silico algorithms capable of detecting ligand binding sites based on a protein's 3D structure. Especially the calculation of reasonable pocket boundaries is often difficult but critical for downstream processing. Because most druggability prediction methods rely on binding site descriptors,[7] major attention has to be drawn to their calculation.

Automatic prediction methods rapidly identify binding sites with unknown functions. An overview of some established methods with their availability and an outline of their algorithmic approach is given in Table 1. These methods employ different strategies that can be categorized into structure-based, sequence-based, and combination methods. Structure-based methods rely on structural information such as atom coordinates. Geometry-based methods, as a subgroup of structure-based approaches, locate surface cavities by analyzing the molecular surface of the target protein, such as VOIDOO,[10] SURFNET,[11] PocketPicker,[12] and fpocket.[13] For example, fpocket analyzes the shape using Voronoi tessellation to calculate alpha shapes. A commercial representative is SiteMap.[14,15] It uses a grid-based approach, embedding the protein in a grid and searching for groups of nonprotein grid points. Energy-based methods search for energetically favorable regions on the protein surface. One such method is AutoSite[16] which computes affinity maps, filters out low-affinity points by applying a threshold, and identifies clusters according to local density. The ranking is done using a geometry score. SiteHound[17] calculates so-called Molecular Interaction Fields describing the interaction between protein residues and a probe. Template or knowledge-based methods infer the location of binding sites from known protein templates. COFACTOR,[18] available as a web server and standalone tool, combines different prediction pipelines. One of the pipelines compares the input structure against known pockets contained in the BioLiP library.[19] Another unique approach is based on machine learning.

Received: March 2, 2023
Published: May 2, 2023

Table 1. Exemplary Selection of Tools for Predicting Protein Binding Sites[a]

method	availability	algorithmic approach
AutoSite	standalone	energy-based
		computes maps and clusters high-affinity points
COACH	web server/standalone	combined methods
		generates predictions using structure and sequence ligand-binding templates and combining results
COFACTOR	web server/standalone	template-based
		genetic algorithm
ConCavity	standalone	sequence- and structure-based
		the protein is embedded in a grid
		grid points are annotated with sequence conservation values of nearby residues
fpocket	standalone	geometry-based
		shape analysis by Voronoi tesselation and alpha shape clustering
P2RANK	standalone	structure-based machine learning
		random forest with feature vectors of physicochemical and geometric properties of protein surface points
PocketPicker	standalone	geometry-based
		the protein is embedded in a grid
		nonprotein points are selected by using the buriedness and clustered into pockets
ROBBY	standalone	sequence-based
		SVM- and random forest-based prediction using evolutionary information
SiteHound	standalone	energy-based
		calculates Molecular Interaction Fields that describe the interaction between protein residues and a probe
SiteMap	standalone	geometry-based
		the protein is embedded in a grid
		nonprotein grid point clusters are searched
SURFNET	standalone	geometry-based
		the protein is embedded in a grid
		gap regions are found by using gap spheres which describe gaps between atom pairs of the protein surface
VOIDOO	standalone	geometry-based
		the protein is embedded in a grid
		the van der Waals radii of all surface atoms are repeatedly increased, which closes the entrance of cavities
		properties of detected cavities are then explored

[a]References are provided in the text.

P2RANK[20] is a random forest approach that describes the solvent-accessible protein surface as points. Each point is annotated by a feature vector of physicochemical and geometric properties from its surrounding atoms and residues.

Sequence-based methods are based on the assumption that residues belonging to a ligand binding site are conserved in the evolution of a protein family as they are indispensable for the protein's function. ROBBY[21] employs features derived from the protein sequence. Through a multiple sequence alignment performed by PSI-BLAST,[22] evolutionary information is retrieved from the LigASite[23] database. Then a support vector machine (SVM) and a random forest algorithm are used to predict binding site residues. ConCavity[24] calculates a grid for the protein and annotates its grid points with the sequence conservation values of nearby residues. Hence, sequence and 3D shape analysis based on the grid are combined to detect pockets. Finally, combined methods such as COACH[25] use structure- and sequence-based approaches to complement each other. The input of COACH can be either a structure or a sequence. COACH employs different methods, including COFACTOR[18] and TM-SITE,[25] to train an SVM.

In 2010, we presented the binding site detection algorithm DoGSite[26] and, in 2012, a respective web service.[6] The service includes a druggability estimator and became part of the Proteins*Plus* server,[27] which experiences popularity to this day. DoGSite predicts putative binding pockets and subpockets of a protein of interest. After the detection process, it reports geometric and physicochemical binding site properties. DoGSite correctly predicted binding pockets for 92% of the PDBbind and the scPDB[28−32] data set of 2010. A prediction was assumed to be correct if the geometric center of the largest three pockets lies within 4 Å of any ligand atom. However, this criterion is insufficient as it does not consider the successful detection of reliable binding site boundaries. Solvent-exposed pocket atoms in a 4 Å radius of the pocket center might lead to false positives for inaccurately defined sites. The criterion is also unsuitable for large pockets, which might still cover the ligand. Still, the ligand is situated in a subpocket too distant from the geometric center leading to false negatives. In the meantime, more elaborate measures of prediction success have emerged.[26,33−35]

This work introduces improvements to DoGSite's prediction quality, stability, functionality, and computational speed. While the overall DoGSite algorithm remains the same, we changed the code base and reparametrized the method based on new data. In the following, we introduce this comprehensively remastered version of the DoGSite pocket detection method and compare it to its previous version and a set of alternative binding site prediction techniques.

METHODS

DoGSite3 is based on the Difference-of-Gaussian (DoG) filter approach published earlier in 2010.[26] For simplicity, we describe the DoGSite3 improvements together with the original DoGSite algorithm.

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137

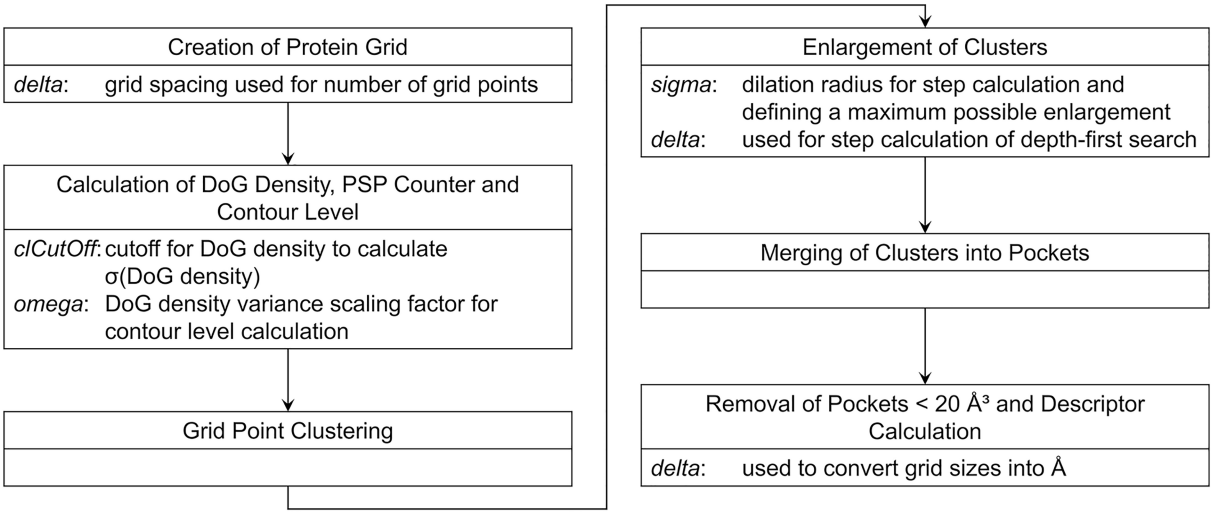


Figure 1. Overview of the DoGSite algorithm and parameters used in each processing step illustrated as boxes.

DoGSite3 calculates potential protein binding pockets of all macromolecular chains, including nucleic acids, except residues with a nonstandard backbone structure contained in a PDB entry. For a graphical overview of the algorithm, see Figure 1. First, the protein is mapped onto a 3D grid. Then, we apply principal component analysis to minimize the required grid size and enable an orientation-independent pocket prediction, which was not available in the previous version. The resulting first three principal axes of the macromolecule are aligned to the *x*-, *y*-, and *z*-axis, respectively. Then, the grid boundaries are calculated by finding the maximum and minimum macromolecule atom coordinates in all directions. In the following, the grid points are created using these coordinates and a predefined grid spacing named *delta*. For very large macromolecules resulting in more than nine million grid points, the grid size is readjusted such that the number of grid points is limited by nine million, leading to a maximum memory consumption of approximately 3.5 GB. The dynamically adjusted grid size (in *x*, *y*, *z* direction) is calculated by linear downscaling using the formula $((gridSize.x*gridSize.y*gridSize.z)/9{,}000{,}000)^{1/3}$. Next, we iterate over all atoms and annotate each grid point as occupied if the distance to the current atom is smaller than its van der Waals radius plus a tolerance radius that was set to the grid spacing parameter *delta* as opposed to the fixed radius of 0.4 Å in earlier versions of DoGSite to account for the uncertainties of the grid-based approach. Otherwise, the grid point is annotated as a solvent point.

Two annotations of the grid points are required to determine the pockets. The first is the Difference-of-Gaussian filter (DoG value) which favors sphere-like cavities of an atom-like radius of 1.75 Å. It is calculated on all grid points not covered by protein atoms. The mean and variance of all solvent points with a density higher than a DoG density cutoff (called *clCutOff*) are calculated. Using a density variance scaling parameter named *omega*, a contour level is derived as follows: $contourlevel = \mu - omega*\sigma$. The second annotation is PSP events,[36,37] indicating the buriedness of a grid point. It is calculated using a scanline procedure across all three principal axes and four cube corners. If a PSP event is observed in any direction, we increase the PSP count in all solvent grid points on the scanline between the two protein-occupied points. A maximum PSP count of seven corresponds to a highly buried grid point, while a minimum of zero indicates a shallow region.

We now iterate over all grid points to cluster the grid points into pockets. If the density value of the respective solvent grid point is below the calculated contour level, we iterate over its 26 neighbors. A depth-first search is started if at least five of the neighbors of this grid point are below the contour level. It runs over the neighboring points until no more solvent points below the contour level are found. This clustering process ends when all grid points have been checked. Finally, all grid points are annotated with their respective cluster number. A second depth-first search is started at each grid point to further enlarge these pockets with solvent points previously not found. This search checks whether in a maximum of $(sigma/delta) + 1$ (with *sigma* being a dilation radius) steps direct neighbors or distant neighbors, e.g., neighbors of neighbors, of the current grid point are (1) a solvent point and (2) not assigned to any cluster so far, (3) its distance to the grid point lies within a threshold of *sigma* squared, and (4) at least two PSP events are present.

After that, we iterate over all clusters and look at each grid point's neighbors again. If two clusters have at least a specific number of neighboring grid points to each other, we merge the two clusters and repeat this step until all points are searched. The number of grid points required is calculated by $((1.4\ \text{Å} - rTolerance)/delta)^2$, where a default solvent radius of 1.4 Å describes a two-dimensional section between the two clusters. Next, we fill areas between clusters. This filling step is particularly useful when clusters have been merged in the previous step, e.g., if we merged two pockets with only a few grid points next to each other, there might still be solvent points in the proximity not assigned to a pocket. To this end, we iterate over all neighbors of each grid point. If a neighbor is in the same cluster as the current grid point, we scan along that axis within a fixed distance of 2 Å. If this scan finds a solvent point belonging to the same cluster, all points in between will be added to this cluster.

After this pocket assignment process, we sort all pockets by the number of assigned grid points and remove all pockets smaller than 20 $Å^3$. Finally, we assign macromolecule atoms to each pocket by checking if the distance between a grid point and an atom is smaller or equal to the sum of the atom's van

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137

der Waals radius plus the tolerance radius plus the cubic grid diagonal: $dist(gp, atom) \leq vdW + cgd + rTolerance$, with *gp* being the grid point, *vdW* the van der Waals radius and *cgd* the cubic grid diagonal. Also, we calculate the descriptors such as the volume, surface, depth, ellipsoid main axis ratio, enclosure, and hydrophobicity of the pockets and count and categorize the atoms belonging to each pocket.[6] All calculated descriptors with their explanations can be found in the SI (Table S1).

Optimizations. Eleven years have passed since our web service DoGSiteScorer was released. Due to the extensive development in performant programming, we decided to reimplement DoGSiteScorer. As a basis, we use our in-house software library NAOMI.[38,39] NAOMI allows, e.g., easy import of structures in standard formats like PDB and mmCIF, preprocessing steps like the optimization of hydrogen placements, and the export of the processed structures.

In NAOMI, molecules are built based on the local geometry of each atom. It does not rely on definite assignments in structure files to create the molecular model because especially low-resolution structures are often error-prone. Molecules are built in four steps: First, the atoms are described with their element, valence state, and atom type. Their covalent bonds are identified based on the interatomic distances. In the second step, the potential valence states for each atom are defined and scored based on the environment. Third, valence states and associated bond order combinations are generated. In the final step, these are scored to determine the most appropriate molecule representation.[39]

DoGSite3 contains a new mode for identifying ligand-occupied pockets that are difficult to detect by pocket geometry alone. This mode is intended to enhance the detection of binding pocket boundaries that are not considered if only selecting protein atoms in the proximity of ligand atoms. To detect reliable pockets without including highly solvent-exposed regions, we introduced a workflow to consider buried ligand fragments only. Based on the approach of Mahmoud and co-workers,[40] a depth-first search divides all ligands or a specific molecule of interest into nonflexible units. Next, the molecule is cut at each rotatable or exocyclic single bond to generate fragments. Each fragment must contain at least two atoms. The solvent-accessible surface area is then calculated for each fragment in both the bound and unbound states. If the ratio of these values is below a cutoff (0.35 by default), the grid points covered by buried fragments are adjusted so that pocket detection is more likely at these locations. The cutoff is used to ignore fragments and molecules that are highly solvent-exposed, i.e., if some atoms of a molecule are bound to and other atoms are protruding from the protein, only the bound atoms are used for binding pocket detection.

We applied the following changes to the original algorithm: (1) The tolerance radius is set as a fixed parameter of the program and corresponds to the chosen grid spacing. Modifying this parameter might lead to unreliable results in defining binding site atoms. (2) We further improved the pocket atom assignment by additionally considering the grid spacing in the form of the cubic grid diagonal, which was neglected in the former DoGSite version. This adjustment is necessary to account for the discretization of the solvent space by the grid. Thereby, we identify all protein atoms surrounding each pocket grid point. (3) As a major change, we now also include nucleic acid chains as macromolecules in the pocket prediction. (4) The pocket prediction no longer depends on the macromolecule orientation.

For more intuitive and precise pocket descriptors, we adjusted some descriptor calculations. The surface is estimated by calculating the solvent-accessible surface area (SAS) of the pocket atoms[41] and is no longer approximated via the number of pocket grid points. In this context, we optimized the surface calculation for pocket surfaces with respect to the runtime. Also, we now calculate the enclosure by $1 - (lid/hull)$ instead of $lid/hull$ since the value thus becomes higher when the lid becomes smaller and the pocket is more enclosed. Additionally, we added the lipophilic surface and the ligand SAS ratio, i.e., the bound ligand solvent accessible surface divided by the unbound ligand solvent accessible surface.

Parameter Optimization. The robustness of binding site descriptors is vital to reliable binding site characterization. Therefore, we decided to address this issue through parameter optimization. For this purpose, we used data sets 1 and 2 of the binding site comparison benchmark ProSPECCTs,[42] including various structures of proteins with identical sequences. Data set 2 encompasses various protein structure models from NMR ensembles (17 groups, 329 structures, NMR ensembles), while data set 1 contains multiple structures for 12 different proteins (12 groups, 326 structures, identical proteins).

Based on these data sets, we evaluated the parameters grid spacing (*delta*, default: 0.4 Å), the density variance scaling (*omega*, default: 3.25), the DoG density cutoff (*clCutOff*, default: −0.001), and the dilation radius (*sigma*, default: 2.25 Å). Note that for all DoGSite parameters, the written form is used instead of the Greek letter to separate them from statistical values. The grid spacing was varied between 0.4 and 1.0 Å in 0.2 Å steps. The density variance scaling was increased from 2.0 to 4.0 in steps of 0.25, while the DoG density cutoff values were tested from −0.1 to 0.1 in 0.025 steps. Finally, the dilation radius was varied between 1.0 and 3.0 in 0.25 steps. Besides that, default parameters were used for the DoGSite runs. Consequently, $4 \times 9 \times 9 \times 9 = 2916$ parameter combinations were investigated.

The pockets were sorted according to their ligand coverage.[26] The pockets with the highest ligand coverage were compared to ensure that only corresponding pockets were considered. All parameter combinations ran successfully on the identical proteins, although using 270 parameter combinations did not predict pockets for all structures. For the NMR ensembles, all of the combinations could be evaluated. For 558 parameter combinations, however, no pockets were found for at least one protein structure.

For each parameter combination, we calculated the average of the pocket volume *V* standard deviations (σ(V)) within each group. Furthermore, we calculated the average of the mean ligand (μ(ligcov)) and pocket coverage (μ(poccov)) per group. Finally, we normalized the mean of the standard deviations of all pocket descriptors (volume, surface, lipophilic surface, depth, ellipsoid volume, ellipsoidal main axes c/a, ellipsoidal main axes b/a, surface grid points, lid grid points, hull grid points, number of protein heavy atoms, number of solvent-exposed hydrogen bond acceptor atoms, number of solvent-exposed hydrogen bond donor atoms, number of aromatic atoms, hydrophobicity) per group.

Benchmark of the Pocket Prediction Accuracy. We analyzed the method's pocket detection capabilities for the ligand-defined scPDB pockets (version as of December 2017)[31,32] and various data sets for benchmarking binding site identification tools[34] to evaluate the optimized parameter

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137

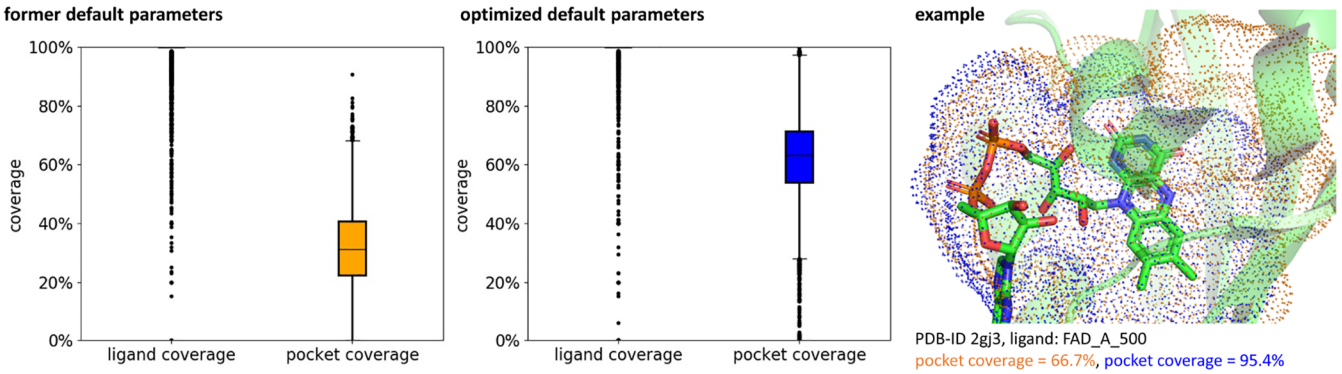


Figure 2. Analysis of the ligand and pocket coverage for ligand-defined binding sites in the scPDB with the former default parameters of DoGSite (orange) and the more robust and reliable default parameters of DoGSite3 (blue). Both ligand and pocket coverage are represented as box plots. The new feature of DoGSite3 for biasing the grid annotation by reference ligands was applied for the analysis. Therefore, the lower and upper quartiles and whiskers for the ligand coverage lie at 100%. The structure on the right represents an example of a pocket with low pocket coverage with the original DoGSite parameters and a much higher one with the optimized DoGSite3 default parameters (figure generated with PyMOL Molecular Graphics System, version 2.3.0).[51]

sets. The data sets include structures derived from the PDBbind core set,[43] cryptic binding sites as provided by the work of Cimermancic et al.,[44] and allosteric binding sites from the ASBench Core Diversity Set.[45] For the assessment of the pocket prediction accuracy, we used the residue-based prediction accuracy as defined by the relative residue overlap (RRO)[34] between the residues inside a 5 Å distance of the cavity-defining ligand (cavity-defining residues, CDR) and the predicted binding site residues by the pocket detection algorithm (PR) according to Equation 1.

$$\mathrm{RRO} = \frac{|\mathrm{CDR} \cap \mathrm{PR}|}{|\mathrm{CDR} \cup \mathrm{PR}|} \quad (1)$$

An RRO of 0 indicates that none of the predicted residues is part of the ligand-defined cavity. In contrast, an RRO of 1 denotes a complete correspondence between predicted and cavity-defining residues. A prediction was assumed to be successful for an RRO of at least 0.5.

Runtime Analysis. Runtime calculations were performed for the data set of identical proteins on a PC equipped with an Intel i5-8500 (3.0 GHz) processor, 16 GB of main memory, and a Toshiba KXG50ZNV256G solid-state drive (256 GB, model NVMe) with an xfs file system.

RESULTS AND DISCUSSION

Including Ligand Bias to Obtain Ligand-Based Binding Site Descriptors. As can be learned from various studies[34,35] and later in this work, some ligand-occupied pockets cannot be correctly predicted by any binding site detection algorithm. An unsuccessful pocket identification is especially obstructive if the druggability of specific binding sites is to be analyzed or the detection method is used to define the dimensions of, e.g., a fragment-bound binding site for molecular docking and fragment growing. The same holds for developing binding site databases for automated pocket comparisons, e.g., with GeoMine.[46,47] In these cases, DoGSite3 enables the user to annotate the DoGSite grid based on given reference ligands, ensuring the detection of pockets in the proximity of the reference ligands and allowing for analyses of binding sites that are difficult to detect. If this modification is applied for the druggable binding sites as stored in the scPDB, 99.0% of the binding sites are included with a ligand coverage of at least 80% as compared to 76.4% if this ligand bias option is not enabled.

The possibility to bias ligand binding sites by reference ligands provided in SDF files enables reliable binding site definitions even if the original algorithm did not detect some of the interacting residues as part of the binding site (SI, Figure S1). It has to be noted that this detection is not restricted to the ligand-biased grid points but that the pocket is extended based on the DoGSite algorithm. Therefore, this option represents a significant advantage compared to ligand radius-based pockets as the pocket is enlarged toward regions that might be additionally addressed, which is highly useful in the context of, for example, structure-based virtual screening. Furthermore, the user can derive binding site definitions biased toward ligand ensembles from similar binding sites, e.g., as retrieved by SIENA,[48] for molecular docking studies.

Ligand and Pocket Coverage Analysis. One of the major challenges in evaluating the quality of binding site prediction tools is the definition of binding site boundaries. Usually, ligand radius-based approaches are used to define binding site extents. However, this method is not always reliable, e.g., in the case of protein binding sites in complex with fragments binding to subpockets or small recesses within a larger binding site.

For a ligand-occupied pocket detected with the DoGSite algorithm, the ligand coverage describes the percentage of ligand atoms covered by a predicted pocket. In analogy, the pocket coverage represents the percentage of pocket grid points in the proximity of ligand atoms. We first explored the herein-analyzed data sets to investigate the distribution of ligand and pocket coverage across a large set of binding sites. We predicted the extent of ligand-occupied binding sites with DoGSite with default parameters using the novel bias to annotate grid points based on a given ligand as binding site points. Since we enforce that all ligand fragments that are not solvent-exposed will be included in the pocket volume, we can thereby obtain an assessment of the average pocket coverage for the druggable binding sites in the scPDB (Figure 2). According to these results, the average acceptable pocket coverage for well-defined binding sites is rarely higher than 70% but varies between 20% and 40% depending on the type of ligand. For 89 structures, the ligand coverage is below 70%. The lower ligand coverage, despite the ligand bias, can be

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137

attributed to solvent-exposed ligand atoms that are not considered for biasing but for ligand coverage calculations. The same analysis (biased and unbiased) was performed for the data sets of identical proteins and NMR ensembles[42] (SI, Figure S2).

As can be seen, DoGSite, with the original default settings, identifies pockets much larger than the volume of the occupying ligand, as the example in Figure 2 also indicates. Therefore, the present study's primary challenge was a more accurate binding site boundary prediction enabling a reasonable trade-off between maximum ligand and pocket coverage.

Descriptor Robustness Optimization. Attempting to improve the reliability of DoGSite, we evaluated the impact of the parameters grid spacing (*delta*, default: 0.4 Å), the density variance scaling (*omega*, default: 3.25), the DoG density cutoff (*clCutOff*, default: −0.001), and the dilation radius (*sigma*, default: 2.25 Å) on the method's performance. The former default parameters were analyzed with a version of DoGSite3 that does not rearrange the protein structures before the calculations, uses an older surface calculation implementation,[41] and performs the original merging procedure as it was optimized for this version of the tool. To enable the reliable detection of binding sites independent of minor structural fluctuations, we decided to focus first on the robustness of the binding site descriptors. These binding site characteristics were shown to fluctuate considerably in earlier studies.[7] These results could be reproduced in this study using the former default parameters and parameter combinations discussed in the following (SI, Figures S3−S8). Note that the ligand information was not included in the binding site prediction, but only the pockets with the highest ligand coverage were investigated for this part of the analysis.

Investigations of the separate parameters and their impact on the standard deviation of the pocket volumes per group of structures revealed a negligible influence of the grid spacing on the robust prediction of the binding sites in the data set of identical proteins, as only slightly smaller volume variations were observed for parameter combinations with a larger grid spacing. The standard deviation of the volume per group of identical proteins with increasing grid spacing for selected parameter combinations can be found in the SI (Figure S9).

In contrast, the contour level cutoff follows more distinct trends. An increase in the cutoff leads to more robust predictions with respect to the pocket volume according to its standard deviation in the groups of identical proteins. This trend persists for nearly all exemplary parameter combinations (SI, Figure S10). However, an opposite relation was observed for the DoG density cutoff. The higher this cutoff, the more the volumes of the predicted pockets in the groups of identical proteins vary (SI, Figure S11). Generally, a cutoff above 0 leads to a high variance in the pocket volume. Less prominent but apparent is a general trend in the dilation radius: the larger this radius, the higher the volume standard deviations within the groups of identical proteins (SI, Figure S12).

Similar trends were observed for the data set of NMR ensembles (SI, Figures S13−S16).

Next, we set out to identify the most reliable parameter combinations (see Table 2 for an overview). For both data sets, we chose the parameter combinations with the lowest sum of normalized mean descriptor standard deviations (see Methods section) that fulfill the criteria σ(V) of less than 110 Å^3, a μ(ligcov) of more than 80%, and a μ(poccov) of more than 40%. The cutoff for the volume standard deviations approximately describes the volume necessary to accommodate a molecule of cyclohexanol[49] and is chosen to reduce the number of considered combinations to about 200. To evaluate whether this selection of parameter combinations leads to an improvement in the robustness considering both data sets, we compared the σ(V), the μ(ligcov), and the μ(poccov) to those obtained with default parameters (Table 3). Parameter set 1 (see Table 2) performed best for the identical proteins (rank 328 in the parameter combinations sorted according to the sum of normalized mean descriptor standard deviations for the identical proteins). For the NMR ensembles, parameter set 2 led to the lowest standard deviations, nonetheless fulfilling all criteria (rank 6 in the parameter combinations sorted according to the sum of normalized mean descriptor standard deviations for the NMR ensembles). Besides the much higher robustness of the predicted pocket volume with the new parameter sets, we find that, despite the promising ligand coverage using the former default parameter set, the pocket coverage is low, suggesting artificially large predicted pockets. This situation considerably improves using the new parameter combinations, retaining a convincing mean ligand coverage. Finally, we could observe a significant decrease in the number of predicted pockets. In the following sections, we will show that this does not impact the prediction success of DoGSite3, hinting at a high number of irrelevant predicted sites with the former versions.

Table 2. DoGSite3 Parameter Sets Investigated in Detail

parameters	*delta* [Å]	*omega*	*clCutOff*	*sigma* [Å]
former default	0.4	3.25	−0.001	2.25
parameter set 1	0.8	2.5	−0.1	1.0
parameter set 2	0.8	3.0	−0.05	1.0
parameter set 3	0.4	3.75	−0.075	1.0
parameter set 4	0.4	3.25	−0.05	1.0
parameter set 5	0.6	4.0	−0.025	1.25

As one could argue that the consideration of the standard deviations of pocket volume, ligand coverage, and pocket coverage alone is not sufficient to guarantee a robust descriptor calculation, we also selected the parameter sets that led to the lowest sum of normalized mean descriptor standard deviations for both data sets leading to two other potential parameter sets (see parameter sets 3 and 4 in Table 2). Finally, the last parameter set was derived by ranking according to the sum of normalized mean descriptor standard deviations but adjusting the thresholds for μ(ligcov) ligand coverage and μ(poccov) pocket coverage to at least 60% and 30%, respectively. This selection led to parameter set 5 (see Table 2) for the data set of identical proteins. For the data set of NMR structures, the same parameter combination as already found in parameter set 4 was obtained. The box plots of the physicochemical and geometric descriptors for the default parameters and the new parameter combinations can be found in the SI (identical proteins: Figures S3−S5, NMR ensembles: Figures S6−S8).

For an external validation of these five parameter combinations, we use DoGSite3 to predict the druggable binding sites in the scPDB and monitored the ligand and pocket coverage (Table 4) for a final choice of parameter sets with the most promising detection success concerning ligand and pocket coverage. In addition, we analyzed the runtimes on the data set of identical proteins for the default and new parameter combinations.

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137

Table 3. Comparison of the Impact of Different Parameter Combinations on the Robustness of the Volume, the Ligand Coverage, the Pocket Coverage, and Number of Predicted Pockets for Structural Ensembles in the Data Sets of Identical Proteins and NMR Ensembles

parameters	data set	σ(V) [Å^3]	μ(ligcov) (%)	μ(poccov) (%)	no. pockets
former default	identical proteins	168.0	87.9	31.8	6221
	NMR ensembles	238.7	81.9	42.1	3474
parameter set 1	identical proteins	52.6	83.2	65.9	2173
	NMR ensembles	37.9	67.1	77.8	1305
parameter set 2	identical proteins	77.7	88.9	55.4	3993
	NMR ensembles	47.3	80.4	74.9	2576

Table 4. Performance of the Former Default Parameters and the Five Selected New DoGSite Parameter Sets on Predicting the Druggable Binding Sites from the scPDB (17589 Binding Sites) and Runtimes for the Data Set of Identical Proteins (326 PDB Files)[a]

parameters	mean ligand coverage (%)	mean pocket coverage (%)	percentage (ligcov ≥80%, poccov ≥40%)	runtime [s][b]
former default	88.2	41.1	35.5	5067.1 ± 68.4
parameter set 1	79.2	80.0	55.8	357.1 ± 10.1
parameter set 2	87.9	71.7	66.6	385.0 ± 1.7
parameter set 3	30.7	50.3	7.8	4193.9 ± 55.1
parameter set 4	65.8	64.6	34.3	4437.1 ± 58.3
parameter set 5	71.1	65.3	44.7	943.0 ± 7.5

[a]The runtimes were obtained from five independent DoGSite runs. [b]mean ± standard deviation.

From this analysis, it can be concluded that parameter sets 1, 2, and 5 led to the most convincing results (mean ligand coverage of >70% and >40% of the detected binding sites with a ligand coverage and a pocket coverage of at least 80% and 40%, respectively). Therefore, we concluded that these combinations are most suitable for robust and reliable binding site predictions with DoGSite3. For these parameters, there is a well-balanced trade-off between optimum ligand and pocket coverage, as illustrated by the success rate regarding the retrieval of pockets with a ligand coverage of at least 80% and a pocket coverage of at least 40%. These findings are in accordance with the previously discussed trends for exemplary parameter combinations. A higher grid spacing leads to more robust results that are not biased by local variations in the DoG values. A low DoG density cutoff leads to more reliable results as it emphasizes points with a more sphere-like character. For higher values of this parameter, we observe improved robustness with an increasing density variance scaling parameter *omega* (as observable for parameter combinations 2 and 5). The clear finding that a low dilation radius leads to lower descriptor standard deviations is explicable by a more restrictive extension of pockets. In contrast, we observe an insufficient retrieval of scPDB binding sites for parameter sets 3 and 4, even worse than with the former default parameter sets. Both sets are characterized by a lower grid spacing rendering the method sensitive toward local differences in the DoG density values. Parameter set 1 is the fastest of all promising parameter sets and shows the most convincing balance between the pocket and ligand coverage.

Reevaluation of Pocket Prediction Performance for the Optimized Parameter Sets. The remaining best-performing three parameter sets (parameter sets 1, 2, and 5 (Table 2)) were further evaluated for the reliable prediction of binding sites. For this evaluation, we used three data sets from a previous benchmark study on pocket detection algorithms.[34] These data sets include pockets extracted from the PDBbind,[43] cryptic binding sites,[45] and allosteric binding sites.[45] Analyzing the prediction success of the selected parameter combinations for these data sets, we find that all parameter combinations lead to an increase in prediction success compared to the former default parameters (SI, Table S2). However, parameter set 1 is slightly more successful than sets 2 and 5 for the cryptic and allosteric pockets. Furthermore, the percentage of successfully detected pockets in the PDBbind-derived sites is only marginally higher for parameter set 2 than for parameter set 1.

Based on this earlier benchmark study, we finally compared DoGSite3 with the new parameter combinations to 40 alternative geometry-based pocket detection algorithms and selected alternative tools. The results of the analysis are depicted in Figure 3. Regarding the percentage of successfully predicted pockets in the highest-scoring three pockets for geometry-based binding site prediction algorithms, DoGSite3 with parameter set 1 is on rank 1 (formerly rank 18), rank 1 (formerly rank 25), and rank 1 (formerly rank 25) for the PDBbind-derived data set, the data set of cryptic sites and the data set of allosteric sites, respectively. The DoGSite3 performance considerably improved compared to its previous versions (SI, Table S2) and shows that DoGSite3 is a fast and reliable geometry-based binding site detection method. Its prediction success is comparable to that of the analyzed knowledge-based prediction methods, which are substantially dependent on pre-existing similar protein−ligand complex structures.[50]

A visual inspection of the pocket and ligand coverage for the scPDB (Figure 2) and the data sets of identical proteins and NMR ensembles (SI, Figure S2 and Figure S3) provides evidence that the new version DoGSite3 with the optimized parameter set 1, besides its more robust descriptor calculation, predicts the most accurate and reliable binding site boundaries (improved pocket coverage). Furthermore, parameter set 1 is the best-performing set for binding site identification. Therefore, parameter set 1 ensures the overall most reliable predictions regarding descriptor robustness, ligand and pocket coverage, and prediction accuracy and can be recommended for DoGSite3 with a runtime of 0.5−4 s per structure.

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137

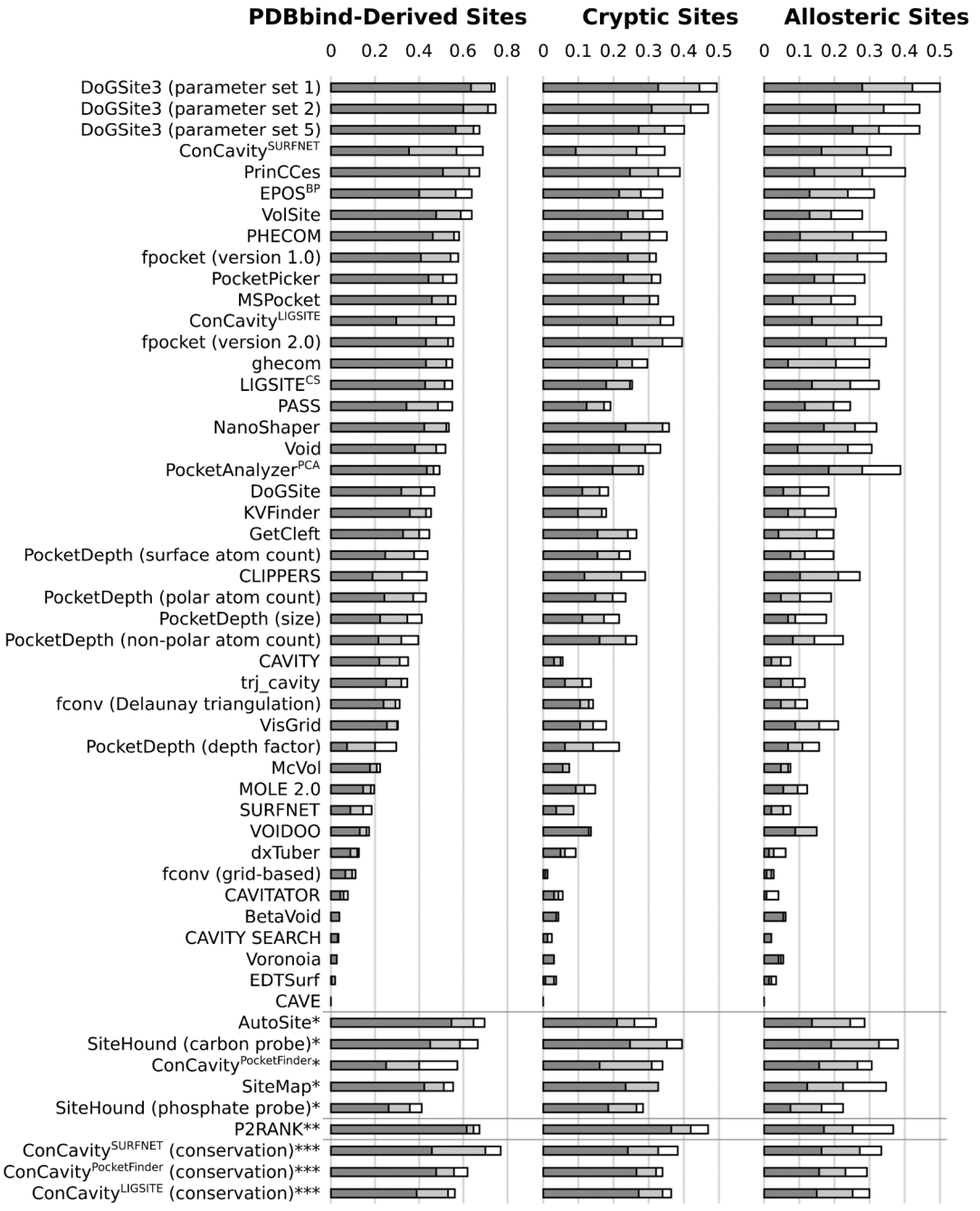


Figure 3. Binding site prediction accuracy of DoGSite3 compared to alternative methods (see SI, Table S2 for the corresponding references). Given is the percentage of successfully predicted pockets, i.e., predicted pockets with an RRO of at least 0.5, for the top1 (dark gray), top2 (light gray), and top3 (white) best-scored pockets. The original DoGSite algorithm was evaluated with an academically licensed version downloaded from the web site of BioSolveIT (DoGSiteScorer 2.0.0). *energy-based methods, **knowledge-based methods, ***combined methods.

■ CONCLUSION

In structure-based drug design, binding pockets play a crucial role. In the absence of cocrystallized ligands, the most accurate prediction models are needed. In the case of protein−ligand complex structures, the model building process is either completely focused on the ligand or ignores it at all, which is inappropriate. This results in a loss of knowledge for making precise predictions. With DoGSite3, we provide a binding pocket prediction that is significantly more accurate and robust than its predecessor and provides a valuable new feature enabling the optional consideration of existing ligands in the grid calculations. Furthermore, we developed a considerably improved and stable prediction tool by refining the pocket merging strategy and enabling a unique grid orientation. Finally, the newly derived parameter combination enables more robust and reliable binding site detection with a substantially lower runtime (approx. ten times faster than the previous version with former default parameters).

The web interface from the predecessor of DoGSite3 has been extended with the new functionality. Thereby, we retain the well-known and established easy-to-use web interface.

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137

We present DoGSite3 as a ready-to-use tool for reliable binding site prediction, robust binding site boundary definition, descriptor calculation, and the calculation of ligand-biased difficult-to-detect binding sites with considerably improved runtimes. Having said this, we hope that future screening efforts and binding site druggability prediction and classification efforts will benefit from the improved DoGSite3 capabilities.

■ ASSOCIATED CONTENT

Data Availability Statement

All data used is generated from structures of the Protein Data Bank, which is freely available here.[1] The training data sets are available on GitHub under this link: https://github.com/rareylab/DoGSite3-Datasets. The benchmark data sets are part of the ProSPECCTs sets and were used unaltered. They are available under this link: http://www.ewit.ccb.tu-dortmund.de/ag-koch/prospeccts/. DoGSite3 is available as a free web service that can be accessed using the link https://proteins.plus. In addition, a standalone tool of DoGSite3 is available as part of the NAOMI ChemBio Suite, which is free for academic use as well as licensable for commercial use.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00336.

Ligand-biased DoGSite pockets (Figure S1), analysis of the ligand and pocket coverage of default and new parameters for the data sets identical proteins and NMR ensembles (Figure S2), robustness of the geometric and physicochemical descriptors for the data sets of identical proteins (Figures S3−S5) and NMR ensembles (Figures S6−S8), impact of the DoGSite parameters on the volume standard deviations for the data sets of identical proteins (Figures S9−S12) and NMR ensembles (Figures S13−S16), abbreviations and explanations of calculated descriptors with DoGSite3 (Table S1), binding site detection hit rates of DoGSite3 and alternative binding site identification tools for three data sets (Table S2), and overview of the analyzed binding site prediction tools (Table S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Matthias Rarey − *Universität Hamburg, ZBH - Center for Bioinformatics, 20146 Hamburg, Germany;* orcid.org/0000-0002-9553-6531; Email: matthias.rarey@uni-hamburg.de

Authors

Joel Graef − *Universität Hamburg, ZBH - Center for Bioinformatics, 20146 Hamburg, Germany;* orcid.org/0000-0001-8327-4936

Christiane Ehrt − *Universität Hamburg, ZBH - Center for Bioinformatics, 20146 Hamburg, Germany;* orcid.org/0000-0003-1428-0042

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00336

Author Contributions

[▽]J.G. and C.E. contributed equally to this work. J.G. developed DoGSite3 and its new functionalities. C.E. performed the parameter optimization and prediction accuracy evaluations. M.R. supervised the project. J.G., C.E., and M.R. wrote the manuscript.

Notes

The authors declare the following competing financial interest(s): Proteins*Plus* and the NAOMI ChemBio Suite use some methods that are jointly owned and/or licensed by/to BioSolveIT GmbH, Germany. M.R. is a shareholder of BioSolveIT GmbH.

■ ACKNOWLEDGMENTS

The authors thank the whole development team of the NAOMI library, forming the basis of this work, with special thanks to our colleague Konrad Diedrich for his help in creating the web interface. This work was supported by the German Federal Ministry of Education and Research as part of CompLS and de.NBI [031L0172, 031L0105]. C.E. is funded by Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter (Grant-ID: HIDSS-0002).

■ ABBREVIATIONS USED

PDB, Protein Data Bank; SVM, support vector machine; DoG, Difference-of-Gaussians; PSP, protein−solvent−protein; SAS, solvent-accessible surface area; RRO, relative residue overlap; CDR, cavity-defining residues; PR, predicted residues

■ REFERENCES

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

(2) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(3) EMBL-EBI *Alphafold Protein Structure dDatabase*. https://alphafold.ebi.ac.uk/, (accessed 23.02.2023).

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(5) Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A.-C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J. Chem. Inf. Model.* **2015**, *55*, 882−895.

(6) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360−372.

(7) Ehrt, C.; Brinkjost, T.; Koch, O. Binding Site Characterization − Similarity, Promiscuity, and Druggability. *Med. Chem. Commun.* **2019**, *10*, 1145−1159.

(8) Volkamer, A.; von Behren, M. M.; Bietz, S.; Rarey, M. In *Applied Chemoinformatics*; Engle, T., Gasteiger, J., Eds.; John Wiley & Sons, Ltd, 2018; Chapter 6.7, pp 283−311.

(9) Hassan, N. M.; Alhossary, A. A.; Mu, Y.; Kwoh, C.-K. Protein-Ligand Blind Docking Using QuickVina-W with Inter-Process Spatio-Temporal Integration. *Sci. Rep.* **2017**, *7*, 15451.

(10) Kleywegt, G. J.; Jones, T. A. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Crystallogr., Sect. D: Struct. Biol.* **1994**, *50*, 178−185.

(11) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.

(12) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chem. Cent. J.* **2007**, *1*, 7.
(13) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf* **2009**, *10*, 168.
(14) Halgren, T. New Method for Fast and Accurate Binding-Site Identification and Analysis. *Chem. Biol. Drug Des.* **2007**, *69*, 146−148.
(15) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377−389.
(16) Ravindranath, P. A.; Sanner, M. F. AutoSite: An Automated Approach for Pseudo-Ligands Prediction—from Ligand-Binding Sites Identification to Predicting Key Ligand Atoms. *Bioinformatics* **2016**, *32*, 3142−3149.
(17) Ghersi, D.; Sanchez, R. EasyMIFs and SiteHound: A Toolkit for the Identification of Ligand-Binding Sites in Protein Structures. *Bioinformatics* **2009**, *25*, 3185−3186.
(18) Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: Improved Protein Function Prediction by Combining Structure, Sequence and Protein−Protein Interaction Information. *Nucleic Acids Res.* **2017**, *45*, W291−W299.
(19) Yang, J.; Roy, A.; Zhang, Y. BioLiP: A Semi-Manually Curated Database for Biologically Relevant Ligand−Protein Interactions. *Nucleic Acids Res.* **2012**, *41*, D1096−D1103.
(20) Jendele, L.; Krivak, R.; Skoda, P.; Novotny, M.; Hoksza, D. PrankWeb: A Web Server for Ligand Binding Site Prediction and Visualization. *Nucleic Acids Res.* **2019**, *47*, W345−W349.
(21) Pai, P. P.; Dattatreya, R. K.; Mondal, S. Ensemble Architecture for Prediction of Enzyme-Ligand Binding Residues Using Evolutionary Information. *Mol. Inf.* **2017**, *36*, 1700021.
(22) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(23) Dessailly, B. H.; Lensink, M. F.; Orengo, C. A.; Wodak, S. J. LigASite—A Database of Biologically Relevant Binding Sites in Proteins with Known Apo-Structures. *Nucleic Acids Res.* **2007**, *36*, D667−D673.
(24) Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Comput. Biol.* **2009**, *5*, e1000585.
(25) Yang, J.; Roy, A.; Zhang, Y. Protein−Ligand Binding Site Recognition Using Complementary Binding-Specific Substructure Comparison and Sequence Profile Alignment. *Bioinformatics* **2013**, *29*, 2588−2595.
(26) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041−2052.
(27) Fährrolfes, R.; Bietz, S.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Volkamer, A.; Rarey, M. ProteinsPlus: A Web Portal for Structure Analysis of Macromolecules. *Nucleic Acids Res.* **2017**, *45*, W337−W343.
(28) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(29) Paul, N.; Kellenberger, E.; Bret, G.; Müller, P.; Rognan, D. Recovering the True Targets of Specific Ligands by Virtual Screening of the Protein Data Bank. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 671−680.
(30) 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.
(31) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: A 3D-Database of Ligandable Binding Sites—10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399−D404.
(32) Rognan, D. sc-PDB. http://bioinfo-pharma.u-strasbg.fr/scPDB/, (accessed 23.02.2023).
(33) Krivák, R.; Hoksza, D. Improving Protein-Ligand Binding Site Prediction Accuracy by Classification of Inner Pocket Points Using Local Features. *J. Cheminf.* **2015**, *7*, 12.
(34) Ehrt, C. Protein Binding Site Comparison. Ph.D. thesis, Technische Universität Dortmund, 2019.
(35) Clark, J. J.; Orban, Z. J.; Carlson, H. A. Predicting Binding Sites from Unbound Versus Bound Protein Structures. *Sci. Rep.* **2020**, *10*, 15856.
(36) 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.
(37) 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.
(38) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.
(39) Urbaczek, S.; Kolodzik, A.; Groth, I.; Heuser, S.; Rarey, M. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* **2013**, *53*, 76−87.
(40) Mahmoud, A. H.; Masters, M. R.; Yang, Y.; Lill, M. A. Elucidating the Multiple Roles of Hydration for Accurate Protein-Ligand Binding Prediction via Deep Learning. *Commun. Chem.* **2020**, *3*, 19.
(41) Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an Integrated Description of Hydrogen Bonding and Dehydration: Decreasing False Positives in Virtual Screening with the HYDE Scoring Function. *ChemMedChem.* **2008**, *3*, 885−897.
(42) Ehrt, C.; Brinkjost, T.; Koch, O. A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (ProSPECCTs). *PLoS Comput. Biol.* **2018**, *14*, e1006483.
(43) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.
(44) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Biol.* **2016**, *428*, 709−719.
(45) Huang, W.; Wang, G.; Shen, Q.; Liu, X.; Lu, S.; Geng, L.; Huang, Z.; Zhang, J. ASBench: Benchmarking Sets for Allosteric Discovery. *Bioinformatics* **2015**, *31*, 2598−2600.
(46) Graef, J.; Ehrt, C.; Diedrich, K.; Poppinga, M.; Ritter, N.; Rarey, M. Searching Geometric Patterns in Protein Binding Sites and Their Application to Data Mining in Protein Kinase Structures. *J. Med. Chem.* **2022**, *65*, 1384−1395.
(47) Diedrich, K.; Graef, J.; Schöning-Stierand, K.; Rarey, M. GeoMine: Interactive Pattern Mining of Protein−Ligand Interfaces in the Protein Data Bank. *Bioinformatics* **2021**, *37*, 424−425.
(48) Bietz, S.; Rarey, M. SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles. *J. Chem. Inf. Model.* **2016**, *56*, 248−259.
(49) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. Fast Calculation of van der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds. *J. Org. Chem.* **2003**, *68*, 7368−7373.
(50) Tubiana, J.; Schneidman-Duhovny, D.; Wolfson, H. J. ScanNet: An Interpretable Geometric Deep Learning Model for Structure-Based Protein Binding Site Prediction. *Nat. Methods* **2022**, *19*, 730−739.
(51) Schrödinger, LLC. *The PyMOL Molecular Graphics System*, Version 2.3, 2015.

https://doi.org/10.1021/acs.jcim.3c00336
J. Chem. Inf. Model. 2023, 63, 3128−3137