JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

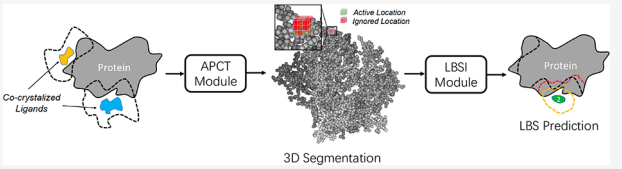
PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms

Xu Yan, Yingfeng Lu, Zhen Li,* Qing Wei, Xin Gao, Sheng Wang, Song Wu, and Shuguang Cui

Cite This: J. Chem. Inf. Model. 2022, 62, 2835−2845 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Accurate identification of ligand binding sites (LBS) on a protein structure is critical for understanding protein function and designing structure-based drugs. As the previous pocket-centric methods are usually based on the investigation of pseudo-surface-points outside the protein structure, they cannot fully take advantage of the local connectivity of atoms within the protein, as well as the global 3D geometrical information from all the protein atoms. In this paper, we propose a novel point clouds segmentation method, PointSite, for accurate identification of protein ligand binding atoms, which performs protein LBS identification at the atom-level in a protein-centric manner. Specifically, we first transfer the original 3D protein structure to point clouds and then conduct segmentation through Submanifold Sparse Convolution based U-Net. With the fine-grained atom-level binding atoms representation and enhanced feature learning, PointSite can outperform previous methods in atom Intersection over Union (atom-IoU) by a large margin. Furthermore, our segmented binding atoms, that is, atoms with high probability predicted by our model can work as a filter on predictions achieved by previous pocket-centric approaches, which significantly decreases the false-positive of LBS candidates. Besides, we further directly extend PointSite trained on bound proteins for LBS identification on unbound proteins, which demonstrates the superior generalization capacity of PointSite. Through cascaded filter and reranking aided by the segmented atoms, state-of-the-art performance can be achieved over various canonical benchmarks, CAMEO hard targets, and unbound proteins in terms of the commonly used DCA criteria.



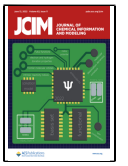
INTRODUCTION

In computational biology, a fundamental question remains to be answered: given a protein structure, can we accurately identify the atoms that form the ligand-binding site(s) (LBS)?[1] In a cellular environment, most proteins perform biological functions by interacting with other ligands, which are small molecules ranging from metal ions, organic or inorganic molecules, to polymers such as polysaccharides and short peptides.[2] The accurate identification of LBS on protein surfaces is critical for understanding the functions of the protein,[3] and in turn an indispensable step for rational structure-based drug design (SBDD).[4] However, it is highly expensive and time-consuming to detect the protein LBS using experimental techniques, which often requires solving the 3D structure of the protein−ligand complexes. Furthermore, although the number of 3D structures in the protein data bank (PDB) grows rapidly, a protein−ligand binding complex structure remains limited even if the apo protein has solved structures.[5] Therefore, computational methods are valuable and needed to predict LBS on a protein surface.

Formally, the protein LBS are defined as the heavy atoms on the protein that are within a certain distance (such as 6.5 Å, denoted as definition radius) to any heavy atom of the ligand. Till now, various approaches have been developed for the identification of protein LBS, which have been comprehensively reviewed and summarized.[2,6−22] In general, the existing methods can be mainly categorized into two classes: template-based methods and template-free methods. In this work, we focus on template-free methods.

For template-free methods, we may divide them into pocket-centric and protein-centric ones (Figure 1). The key principle of those pocket-centric approaches is based on the following two steps: (i) searching for pseudo-surface-points (PSPs) outside the protein structure, which could be regarded as the candidate LBS or pocket; (ii) identifying the binding atoms within a certain radius to these PSPs (denoted as identification radius). The existing pocket-centric methods can be categorized into geometric, energetic, and machine learning based on their main algorithmic strategy to search for PSPs.[23] The geometric strategy includes methods that identify the PSP by using the 3D geometric structure of the protein to search for cavities and pockets of the protein, which could be further classified into grid scanning (e.g., PocketPicker,[24] LIGSITE,[25] LIGSITE_csc[3]), probe sphere (e.g., PASS[26]), alpha shape (e.g., CAST[27]), and alpha sphere (e.g., FPocket[28]); the energetic strategy includes methods that explore the energy

Received: December 16, 2021
Published: May 27, 2022



https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

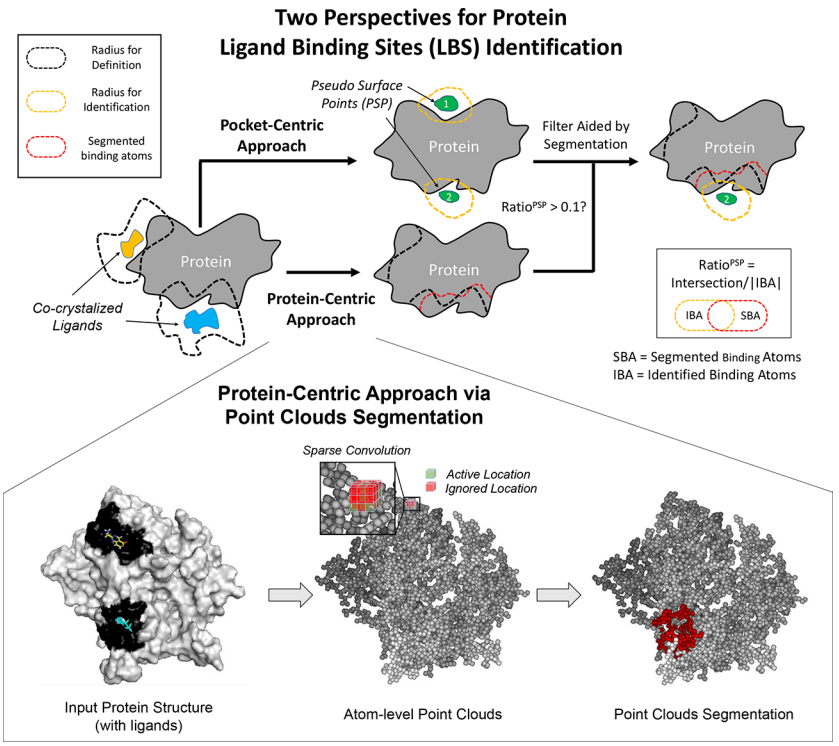


Figure 1. Given a protein structure, our purpose is to identify the binding atoms within a certain radius (i.e., definition radius) of the cocrystallized ligands. Two template-free perspectives for identifying protein ligand binding site(s) (LBS) exist: pocket-centric and protein-centric. (i) The pocket-centric perspective first searches for pseudo-surface-points (PSPs) outside the protein structure, and then uses these PSPs to identify the protein binding atoms within a certain radius (i.e., identification radius); (ii) the protein-centric perspective directly identifies the binding atoms on the protein structure. The key idea of our approach is to formulate the binding atom identification problem as a point clouds segmentation problem, where each protein atom is represented as a point and the corresponding labels are nonbinding or binding defined by the definition radius. Consequently, we show that our point clouds segmentation approach can serve as a powerful tool for prioritizing candidate PSPs identified by other pocket-centric approaches through filtering those PSPs that do not share enough common identified atoms with our approach. Best view in color.

of each position of the protein to the PSP (e.g., SiteHound,[29] Q-sitefinder,[29] PocketFinder[30]). Later, several machine learning methods have been proposed to exploit the protein features at residual and/or the atomic level that influence the PSPs, and then use machine learning algorithms to identify (e.g., DeepSite[31] based on 3D-CNN) or rank (e.g., P2Rank[23] based on random forest) those PSPs.

It is clear that the pocket-centric methods answer the original question (i.e., identifying the ligand-binding atoms) in an indirect way. Furthermore, pocket-centric methods cannot explicitly take advantage of any connectivity between the atoms as well as the physicochemical characteristics of the atoms and/or residues on the protein structure. Therefore, a more direct way to identify LBS is the protein-centric perspective (Figure 1). In this work, we present a point clouds segmentation approach for accurate identification of protein LBS at the atom-level. First, we represent all the atoms in the query protein as point clouds, and mark those atoms in LBS as binding atom with label 1 and others with label 0. Then, we formulate the LBS identification problem as a point clouds segmentation problem, which could be elegantly solved by Submanifold Sparse Convolutional networks (SSCN).[32] Considering a conventional 3D convolutional neural network (3D-CNN) requires a large amount of memory consumption for dealing with 3D point clouds; sliding windows and down-sampling operations are exploited in DeepSite.[31] Such operations cannot take all point clouds as inputs and inherently harm the context learning. More seriously, traditional 3D-CNN suffers the submanifold dilation issue.[32] On the contrary, SSCN will keep the sparsity of the input points with each convolutional layer, and mark them to be either active or ignored (the lower panel in Figure 1). Thus, such a deep neural network framework can input the entire point clouds at once and explicitly take into account the connectivity between the atoms as well as the physicochemical features of the protein atoms and residues simultaneously. Furthermore, SSC based U-Net learns the global and geometrical information on the query protein via the incrementally increased receptive field and skip connections, which in turn can capture the very complex relationship between the 3D protein structure and the binding atoms. Finally, our deep learning approach is a data-driven method, which can efficiently learn the characteristics of the entire database. Once the PointSite model has been trained, the inference time of our method is extremely fast (less than 1 s per query protein) and the entire package is light-weighted.

Comprehensive experiments show that the proposed method, PointSite, achieved the state-of-the-art performance under the atom-level IoU (Intersection over Union) measurement: the intersection over union of the predicted atom set and ground truth atom set. Moreover, our approach could serve as a tool for prioritizing candidate PSPs identified by other pocket-centric approaches through filtering those PSPs that do not share common identified atoms with our segmentation (described in Figure 1). In terms of DCA (distances between the center of the Top-N identified LBS and

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

any heavy atom of the cocrystallized ligand), the *de facto* gold standard pocket-centric measurement,[23,33] PointSite can significantly outperform all pocket-centric approaches by a large margin. For hard targets, such as the proteins with novel fold from CAMEO,[34] the steadily better performance of PointSite not only confirms the generalization ability of our point clouds segmentation approach, but also proves the fact that the combination of the two complemental perspectives, protein-centric and pocket-centric approach, can lead to even better LBS identification. Besides, the real challenge for LBS is to find binding sites in structures without a bound ligand. Thus, in this paper, we creatively extend PointSite, a novel model trained on proteins with bound ligands, to conduct LBS identification on unbound proteins, that is, proteins without a bound ligand. The state-of-the-art LBS identification performance on unbound proteins further verifies the superiority and predominant generalization capacity of the proposed PointSite for real applications.

METHODS

In this paper, we propose a novel point clouds segmentation method, PointSite, for accurate identification of protein ligand binding atoms, which conducts protein LBS identification at the atom-level in a protein-centric manner. Specifically, our proposed model mainly consists of three modules: (1) the atom point clouds transformation (APCT) module; (2) the ligand-binding atoms prediction (LAP) module; and (3) the LBS identification (LBSI) module. In the APCT module, we first transfer the original protein structure from the PDB format to atom-level point clouds. Taking into the transferred atom point clouds with simple features, the LAP module conducts ligand atoms prediction by utilizing a Submanifold Sparse Convolution (SSC) based U-net.[35,36] Thanks to the sparse convolution operations, LAP can not only take all point clouds into consideration, but also contributes to a better global context learning. With the assistance of segmented ligand atoms, the accurate LBS identification is accomplished by filtering and reranking the prediction results generated by pocket-centric approaches.

Definitions and Problem Formulation. Given the PDB file of a query protein, our goal is to identify all the possible ligand-binding atoms, as well as output all possible LBS in a ranked list with the geometric center of each LBS. Note that our method is a template-free approach, which only takes the original PDB file as the input. To identify LBS, our method requires an additional pocket-centric approach to generate candidate LBS in a ranked order.

Atom Point Clouds Transformation (APCT) Module. Considering that previous pocket-centric approaches will predefine ad-hoc pseudo-surface-points (PSPs) for candidate LBS, here we conduct protein LBS identification at atom-level in a protein-centric perspective. Thus, we first transfer the original protein to atom point clouds. Given proteins with a known ligand from a data set, for example, scPDB,[37] all information has been presented in the original PDB file, including protein atoms and coordinates, ligand atoms and coordinates, protein residues, and so on. In this paper, we transfer the original PDB file to the atom point clouds through a self-developed software LIG_Tool (see the Supporting Information). By taking a PDB file as input, the output is the features for each atom from the protein itself, including 21 residue type (A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V, and Unknown), 5 atom type (C,N,O,S, and Unknown) and the 3 coordinates for each atom. On the other hand, we label a protein atom as the ligand-binding atom (label 1) if it is within a sphere with the definition radius r_{def} (i.e., 6.5 Å) to any heavy atom of the ligands, otherwise the label is 0. Under this setting, the features for all atoms is a 29-dimension vector, and the ground-truth label for each atom is 0/1.

In literature, the definition radius r_{def} ranges from 4.0 to 6.5 Å, and here we use 6.5 Å for the following reasons: (a) This definition is used in scPDB that is an annotated database of druggable binding sites from PDB;[37] (b) the 6.5 Å radius will consider those atoms that are in contact with binding atoms in implicit interaction with the ligands;[38] and (c) using this radius to define the ground-truth binding atoms will alleviate the class-imbalance issue.

Ligand Atoms Prediction (LAP) Module. In the APCT module, the input features and ground-truth labels of each protein atom have been defined. Considering that the atom number for each protein is usually over 10k, the submanifold dilation would occur if exploiting conventional 3D convolutions on voxelization representation. This will substantially reduce the sparsity of the input points from the previous convolutional layer and geometrical information in the next layer. More seriously, since protein atoms only occupy a relatively small ratio in the whole space of voxelization representation, conventional 3D-CNN (e.g., DeepSite[31]) would waste lots of calculations and cannot take entire point clouds as the input due to memory limitation. Such phenomenon is extremely disadvantageous for the context learning and identification of LBS.

In this paper, we innovatively exploit 3D sparse convolutions[39] to tackle the above problem. We transform the original atoms' coordinates (i.e., a point cloud) into a sparse volumetric representation. Specifically, a 3D point cloud can be represented in the format of $x = \{x_k\} = \{(p_k, f_k)\}$, where p_k is the 3D coordinate of the kth point, and f_k is its corresponding feature. First, we shift all points into the local coordinate system with the geometric center as origin. Then, we normalize the points into the unit sphere by dividing all coordinates by $\max\|p_k\|_2$, and then scale and translate the points to $[0, 1]$. The point features $\{f_k\}$ remain unchanged, and we denote the normalized coordinates as $\{\hat{p}_k\}$. After that, we transform the above normalized point cloud to a sparse voxel representation with resolution r:

$$\hat{p}_k^* = (\hat{x}_k^*, \hat{y}_k^*, \hat{z}_k^*) = (\lfloor \hat{x}_k \times r \rfloor, \lfloor \hat{y}_k \times r \rfloor, \lfloor \hat{z}_k \times r \rfloor),$$
$$f_m^* = \frac{1}{N_m} \sum_{k=1}^{n} \mathbb{I}[\hat{x}_k^* = \hat{x}_m, \hat{y}_k^* = \hat{y}_m, \hat{z}_k^* = \hat{z}_m] \cdot f_k \quad (1)$$

where $\lfloor \cdot \rfloor$ is the floor operation, and $I(\cdot)$ is a binary indicator of whether $\hat{p}_k^*$ belongs to the mth voxel grid. N_m is the number of points in the mth voxel, and we average features of these points. In our method, we set the resolution $1/r = 0.125$ Å, and the average sparsity of the training set is 0.2%. After eq 1, only nonempty voxels are preserved ($N_m > 0$), which are saved in an input Hash table. After that, locations of output sites are obtained by iterating through the input hash table and stored into output hash table. One location is considered as an output site if the convolution kernel at this location covers an active input site. A rule book is then created to record corresponding output sites for each element in a convolution kernel.

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

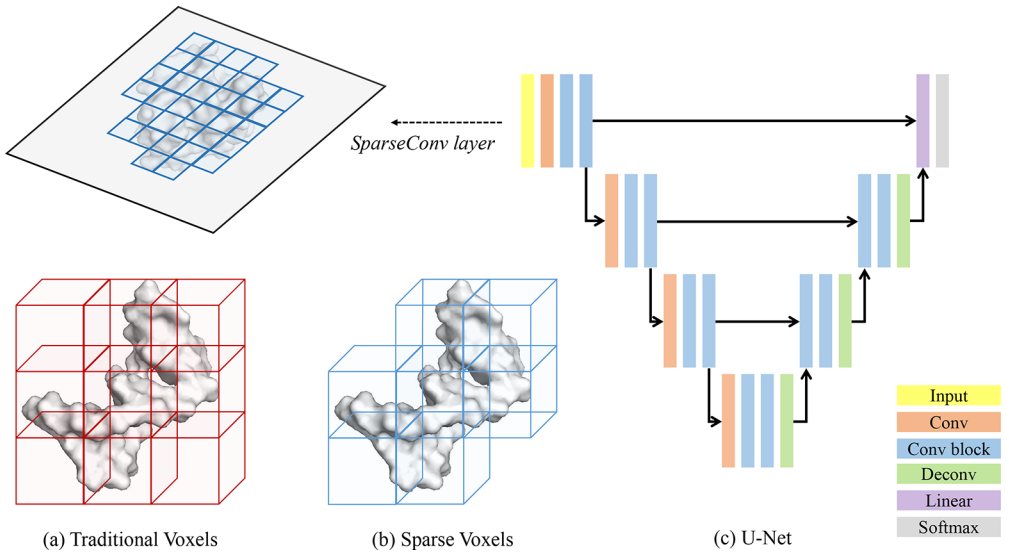


Figure 2. Submanifold Sparse Convolution based U-Net in 2D perspective. The difference between it and traditional 3D-CNN is illustrated in parts a and b. In part c we demonstrate the architecture of U-Net.

Sparse convolution (SC) is operated by visiting elements in the rule book instead of utilizing sliding window to traverse the input volume. A sparse convolution kernel is defined as SC(c_{in}, c_{out}, k, s) where c_{in} and c_{out} are input and output feature channel number, k is the kernel size, and s represents stride size. The element set of the kernel is $F = \{0, 1, ..., k^3 - 1\}$, one element $i \in F$ has a parameter matrix $W^i \in R^{c_{in} \times c_{out}}$. The feature of the kth output site is calculated by multiplying its corresponding input site feature and parameter matrix W^i. Submanifold sparse convolution (SSC) is similar to sparse convolution (SC), except that its output site is active only when the central site of the input is active. In such a manner, we can represent the point cloud with larger volumetric resolution while keeping computational efficiency.

The Figure 2 (a) and (b) illustrate the difference between traditional voxels with 3D-CNN and sparse voxels with sparse convolution. Exploiting sparse representation can effectively reduce the memory usage and computation, especially with larger advantage when dealing with large resolutions. Finally, the sparse voxels will be fed into sparse convolution-based U-Net[32] as shown in Figure 2 (c), and the results are transformed back to the atom-level by nearest-neighbor interpolation. Specifically, we use multiscale (i.e., seven scales) U-Net structure to capture the information from different resolution. Each encoder layer contains a SSC with stride 2 to down-sample the feature maps, and skip connections are applied in the further convolution block. The inversed operations are conducted in the decoder side, where it concatenates the features from encoder phase on the decoder features at each scale for better context learning. Through such a manner, we can not only learn a more effective geometric representation with a larger resolution, but also speed up the network computation.

LBS Identification (LBSI) Module. As mentioned above, pocket-centric approaches usually require additional information (e.g., geometric and/or energetic patterns of the ligand-binding pocket outside the protein) to define the PSP for the downstream LBS identification, while our point clouds segmentation method leverages native information on the atoms in the protein structure. Thus, it inspires us to design a merged model for LBSI through combining pocket-centric and protein-centric perspectives.

Filter Aided by Segmented Binding Atoms. The output of a certain pocket-centric approach (e.g., FPocket and SiteHound) is a ranked list of candidate PSP, each of which consists of a cluster of PSPs that satisfy the corresponding criteria (such as high-scoring alpha spheres[28] and the surface points with high interaction energy[40]). Then starting from the Top-K candidate PSPs, we identify the ligand-binding atoms for a certain PSP with a certain identification radius r_{iden} in the same procedure as defining the ground-truth ligand-binding atoms. Therefore, for the ith candidate PSP, the following filter strategy is used with the aid of our segmented ligand-binding atoms:

$$\text{ratio}_i^{\text{PSP}} = |\text{SBA} \cap \text{IBA}_i|_1 / |\text{IBA}_i|_1, \quad i = 1, \cdots, K \tag{2}$$

where SBA stands for segmented binding atoms, IBA_i stands for identified binding atoms for the ith candidate PSP and $|\cdot|_1$ stands for the L1 norm (i.e., atom numbers). If the $\text{ratio}_i^{\text{PSP}}$ is below a threshold, then the ith candidate PSP will be filtered out. Consequently, the Top-K identifications would reduce to Top-K_1 identifications for the pocket-centric approach.

Reranking by Common Binding Atoms. From the above filtered Top-K_1 identifications, we may further conduct a very simple reranking strategy based on the common binding atoms. Specifically, based on our SBA, we rerank the remaining K_1 identifications according to the common atoms as follows:

$$\text{common atoms}_j = |SBA \cap IBA_j|_1, \quad j = 1, \cdots, K_1 \tag{3}$$

where SBA stands for segmented binding atoms set, IBA_j stands for identified binding atoms from the jth filtered candidate PSP. Such simple strategy can not only benefit the protein-centric criteria (e.g, DCA), but also improve the performance of atom-level IoU.

EXPERIMENTS

Data Sets. There are three types of data sets used in this work: training, testing, and blind test. We train and validate our method on the training data set. The test data set is applied for testing our method and comparing it with other approaches.

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

We use CAMEO as the blind test data set to show the performance on real-world cases. Each data set shall belong to either chain-level or protein-level, where the former indicates that the entries in the data set are protein chains while the latter are the protein complexes consisting of multiple chains. According to scPDB,[41] we consider all cocrystallized ligands, such as agonists, antagonists, selective modulators, simply binders, etc. It should be noted that for all data sets, we remove those entries with more than five binding ligands as well as those invalid ligands defined according to MOAD.[42] We define the binding ligand using our in-house method LIG_Tool in Supporting Information.

Training Data Set. The data set used to train our proposed method contains 6151 entries at the chain-level, which is a subset of scPDB[41] created in 2017. Note that these entries share no more than 25% sequence identity with any test data set.

Testing Data Set. To evaluate our approach and compare with other approaches, we collect a set of well-recognized LBS data sets from the available literature.[23] These data sets include B277,[3] DT198,[43] ASTEX85,[44] CHEN251,[33] COACH420,[45,46] and a large data set HOLO4k.[47]

- B277: a data set containing 277 proteins in a bound state from LIGSITE_csc benchmark, released at the year 2006. After LIG_Tool selection, we have 263 valid entries at the protein-level in this data set.
- DT198: a data set containing 198 drug-target complexes from MetaPocket2 benchmark, released at the year 2011. After LIG_Tool selection, we have 163 valid entries at chain-level in this data set.
- ASTEX85: a data set containing 85 entries introduced by [astex] at the year 2007, which is used for evaluating molecular docking methods, and the ligands all meet with drug-like criteria. After LIG_Tool selection, we have 82 valid entries at protein-level in this data set.
- CHEN251: a data set containing 251 entries introduced by [chen11] at the year 2011, which is a nonredundant data set containing typical representatives at the SCOP family level. Note that this data set has been used to evaluate a variety of LBS identification methods released before the year 2011. After LIG_Tool selection, we have 238 valid entries at the chain-level in this data set.
- COACH420: a data set containing 420 proteins that contain a mix of drug targets and natural ligands. This is a subset of the COACH benchmark, released at the year 2013, and has no redundancy with B277, DT198, ASTEX85, and CHEN251. After LIG_Tool selection, we have 402 valid entries at the chain-level in this data set.
- HOLO4k: this is a large data set containing 4543 protein−ligand complexes in the holo form, which was used to evaluate four popular pocket-centric approaches before the year 2010. Again, it has no redundancy with B277, DT198, ASTEX85, and CHEN251. After LIG_Tool selection, we have 4063 valid entries at the protein-level in this data set.

Blind Test. To show the performance of our approach on "real-world" hard cases, we extracted a subset of the CAMEO[34] from 9/22/2018 to 9/22/2019 that contains 103 entries at the chain-level, which has limited structural similarity (defined as *TMscore* < 0.6[48]) and no sequence identity (<25%) to our training data. After LIG_Tool selection, we have 81 valid entries at the chain-level in this data set.

Unbound Proteins. To apply our approach for real-world LBS predictions, we extend our PointSite to proteins without bound ligand, that is, unbound proteins. Specifically, we first extract the 721 bound−unbound protein pairs with exact 100% residual sequence identity from the PDB database[49] (note that there are missing residues or regions in the original PDB files), which means that one protein is with bound ligand while the other one is without ligand for the two same proteins. In this work, we only concentrate on those proteins with similar structures instead of focusing on allosteric proteins. Thus, we further conduct two more filters of the 721 bound−unbound pairs to guarantee the similarity in terms of ATOM sequence similarity (i.e., the amino sequence extracted from the protein structure without considering missing residues over 98% similarity) and structural similarity (defined as *TMscore* > 0.95[48]). At last, we achieve 521 bound−unbound pairs as the new test data set.

Blind and Unbound Proteins. To conduct a blind test on unbound proteins, we extract proteins of hard cases in CAMEO from 03/19/2019 to 03/19/2022. After screening proteins with low sequence identity (<25%), we achieve 55 proteins. To obtain apo structures of proteins, we feed proteins' sequences in AlphaFold2 (without template)[50] to predict their structures and regard them as unbound structures. Then structures with low accuracy (*global lDDT* < 0.8) are removed, rendering a blind and unbound data set containing 46 proteins.

Comparison Methods and Evaluation Criteria. *Comparison Methods.* We compare out method PointSite with the following template-free pocket-centric methods: FPocket, SiteHound, MetaPocket2, DeepSite, and P2Rank on the test data set. FPocket (version 3) is a typical geometric method based on filtering and clustering of alpha spheres found by Voronoi tessellation.[28] It is very popular and *de facto* the best geometric approach.[23] SiteHound is a typical energetic method based on the interaction energies between the protein and a PSP,[40] and could also be regarded as the best energetic approach.[23] MetaPocket2[43,51] is a meta server to identify LBS based on the identification results from the eight popular pocket-centric methods: LIGSITEcsc,[3] PASS,[26] Q-SiteFinder,[52] SURFNET,[53] Fpocket (version 1),[28] GHECOM,[54] ConCavity,[55] and POCASA.[56] DeepSite[31] and P2Rank[23] are two recently developed approaches based on machine learning, where the former is rooted in 3D-CNN and the latter employs random forest. It should be noted that FPocket and P2Rank are open-source software, SiteHound has the stand-alone version, while MetaPocket2 and DeepSite only have the server version.

Evaluation Criteria. We use atom-level IoU (atom-IoU) to measure the performance of the accuracy of binding atoms identification. In particular, atom-IoU is defined as the intersection over union of two point sets, where the former point set is the identified ligand binding atoms and the latter one is the ground-truth ligand binding atoms. It is obvious that such a metric is a typical protein-centric criterion, and all comparing methods are pocket-centric criteria, which require an identification radius (see Figure 1) to identify the binding atoms from the PSPs. As the size, shape, and number of PSPs identified by different pocket-centric approaches are quite different, here we only select Top-N PSPs where N is the number of cocrystallized ligands, and output the maximal value

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

Table 1. Comparison of Identification Performance on B277, DT198, ASTEX85, CHEN251, COACH420, and HOLO4K Datasets in Terms of atom-IoU (%)[a]

Method	B277	DT198	ASTEX85	CHEN251	COACH420	HOLO4K
FPocket	31.5	23.2	34.1	25.4	30.0	30.5
SiteHound	36.4	23.1	38.9	29.4	34.9	34.5
MetaPocket2	37.3	25.8	37.5	32.8	37.7	38.8
DeepSite	34.0	29.1	37.4	27.4	33.9	33.2
P2Rank	49.8	38.6	47.4	**56.5**	45.3	48.8
PointSite	**60.9**	**45.4**	**61.3**	54.2	**59.6**	**63.4**

[a]As the size of the pseudo-surface-points (PSP) for different pocket-centric approaches varies, we show the maximal value of atom-IoU generated by the three different identification ratios (4.5 Å, 5.5 Å, and 6.5 Å). See Supplemental Table S1 for more details of the value at each identification radius. The same applies to Table 3 and Table 4.

Table 2. Comparison of Identification Performance on B277, DT198, ASTEX85, CHEN251, COACH420, and HOLO4K Datasets in Terms of DCA Using the Original Results of Those Pocket-Centric Approaches as Well as the Merged Results in Consideration of the Segmentation of PointSite[a]

Method	B277	DT198	ASTEX85	CHEN251	COACH420	HOLO4K
Results of PSP Based Methods Only						
FPocket	44.0	32.8	43.4	34.2	43.6	42.4
SiteHound	53.9	39.0	54.3	45.6	53.2	53.1
MetaPocket2	60.9	49.1	59.7	46.8	59.2	60.1
DeepSite	59.9	52.6	60.2	43.4	55.0	58.8
P2Rank	**74.9**	**68.6**	**65.1**	**79.2**	**68.9**	**75.1**
Results of PSP Based Methods Combining PointSite						
FPocket	75.0	63.8	73.6	59.1	73.0	80.2
SiteHound	**78.4**	70.6	**76.7**	67.4	**76.8**	**86.6**
MetaPocket2	74.0	62.6	67.2	57.7	69.1	71.6
DeepSite	66.9	60.6	61.7	47.9	59.7	62.5
P2Rank	77.8	**74.9**	70.5	**71.3**	71.2	81.0

[a]The numbers represent identification success rate [%] measured by DCA criterion (i.e., the distance from the center of Top-K identified LBS to the closest ligand heavy atom, where K is the number of ligands in the query protein structure) with 4 Å threshold. We show the original result of a certain approach in the first row, while putting the merged results in consideration of the segmentation of PointSite in the second row. As there are two parameters (identification radius, ratio PSP) during the merging process, we fix the value to (6.5, 0.1) since it would lead to the best performance (shown in Tables S7 to S11). The same applies to Table 3.

of atom-IoU generated by 4.5 Å, 5.5 Å, and 6.5 Å identification radius.

We use DCA to show the original performance of all the pocket-centric approaches, as well as the performance in consideration of our segmentation results. Specifically, DCA is defined as the minimal distance between the geometric center of the Top-N identified LBS and any heavy atom of the cocrystallized ligands.[33] Again, N is the number of ligands. Usually, an identified LBS is considered as correct if the DCA is no further than 4 Å.

Performance. *Protein-centric Performance.* We have extensively evaluated the protein-centric performance in terms of atom-IoU of PointSite and compared it against several widely used pocket-centric approaches on a variety of test data sets. As shown in Table 1, on almost all data sets, our point clouds segmentation method significantly outperforms all pocket-centric approaches including geometric methods, energetic ones, or machine learning or deep learning-based ones. P2Rank is the second best method. The only exception is the data set CHEN251 which is the training set for P2Rank. However, even on this data set, our result of 0.54 is comparable to that of P2Rank, which is 0.56. For all other data sets, our approach is 7%, 11%, 14%, 14%, and 15% better than P2Rank on DT198, B277, ASTEX85, COACH420, and HOLO4k, respectively. Note that P2Rank is the current state-of-the-art pocket-centric approach for LBS indentation, which is about 10% better than all the previous methods. These results indicate that our approach can significantly improve the accuracy of the binding atoms identification.

Pocket-centric Performance. As our method PoinstSite is a protein-centric approach, it is not easy to directly compare with the other pocket-centric approach in terms of DCA due to the difficulty in clustering the atom-level segmentation into several ranked classes in order to calculate DCA. However, the significantly improved identification results in terms of atom-IoU indicate that our approach could serve as a tool for prioritizing candidate PSPs (or, LBS) identified by other pocket-centric approaches through filtering those PSPs that do not share enough common identified atoms with our segmentation. Table 2 confirms this claim that our approach could significantly enhance the identification success rate measured by DCA for all pocket-centric approaches to a great deal on almost all the data sets (except CHEN251 which is the training set for P2Rank). One notable example is the merged result of SiteHound. The original result of SiteHound is about 10−20% worse than that of P2Rank. However, the merged results of SiteHound is 4%, 2%, 11%, 8%, and 11% better than the original result of P2Rank on B277, DT198, ASTEX85, COACH420, and HOLO4k, respectively. In some data sets, such as B277 (1%), ASTEX85 (6%), COACH420 (5%), and HOLO4k (6%), the merged result of SiteHound is even better than the merged result of P2Rank (value is shown in parentheses), respectively. Such trends also appear in the measurement of atom-IoU (see Tables S2 to S6), and in some

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

data sets the results are even better than that of PointSite itself. Therefore, the atom-level segmentation of PointSite can help the ranking and filtering of the identified LBS of the pocket-centric approaches, which significantly improve the performance not only in pocket-centric metric DCA but also protein-centric metric atom-IoU.

Table 3. Comparison of Identification Performance on CAMEO Blind Datasets in Terms of DCA Criterion at 4 Å Threshold as Well as Atom-IoU (%)

method	DCA	Atom-IoU	PointSite
FPocket	32.5	25.2	w/o
SiteHound	24.3	20.4	w/o
MetaPocket2	28.3	24.4	w/o
DeepSite	31.1	22.3	w/o
P2Rank	46.3	36.5	w/o
FPocket	48.7	39.9	w
SiteHound	**52.8**	41.0	w
MetaPocket2	45.8	28.2	w
DeepSite	37.7	26.6	w
P2Rank	51.2	39.7	w
PointSite		**43.1**	

Blind Test. A natural question to ask is whether our point clouds segmentation approach learns the complex relationship between the 3D protein structure and the binding atoms or just simply "remembers" the training data. Although we removed all redundant and homologous proteins in the training data scPDB in the six testing data sets, it is still not sufficient to completely address this concern. To this end, we challenged our method on "real-world" hard targets from CAMEO that not only have no sequence identity (<25%) but also have limited structural similarity (*TMscore* < 0.6[48]) to our training data. Furthermore, these targets were released after 9/22/2018, which is even later than the latest version of scPDB released in 2017.

As shown in Table 3, among these 81 hard targets, our point clouds segmentation tool reaches 0.43 atom-IoU, which is 18%, 23%, 19%, 21%, and 6% better than FPocket, SiteHound, MetaPocket2, DeepSite, and P2Rank, respectively. When merged with our segmentation, the identification accuracy of all pocket-centric approaches is significantly improved in terms of both atom-IoU and DCA. An interesting discovery is that such improvement occurred more often in FPocket and SiteHound, which is a typical geometric and energetic approach, but not those consensus approaches (say, MetaPocket2) or machine learning approaches (say, DeepStie and P2Rank). These results further prove the generalization power of the combination of our point clouds segmentation in the protein-centric perspective with those individual unsupervised learning single approaches in the pocket-centric perspective, because the two perspectives contain complement information to each other.

Bound−Unbound Proteins. For drug-design applications, the actual challenge is to find binding sites in structures without a bound ligand, which requires methods to have an advantageous generalization capacity for unseen structure in unbound proteins. To this end, we extend our method to bound−unbound protein pairs (one with bound ligand while the other without) from PDB database that each pair has 100% residual sequence identity and compact structure similarity (*TMScore* > 0.95[48]). Considering missing widely exists in original PDB files,[57] we further filter the same bound−unbound protein pairs through the ATOM sequence identity. At last, we achieve 521 bound−unbound pairs, where each two proteins should have 100% residual sequence similarity, over 98% ATOM sequence similarity (i.e., the amino sequence extracted from the protein structure without considering missing residues) and over *TMScore* > 0.95 structure similarity. As there exists strong structural similarity between the bound−unbound protein pairs, we may transfer the labels of ligand-binding atoms, as well as superimpose the bound ligand, from bound protein to unbound protein.

As shown in Table 4, our point clouds segmentation tool PointSite can reach 47.3% atom-IoU for bound proteins and 44.3% atom-IoU for unbound proteins, which only causes around a 3% drop for unseen unbound structures, respectively. Such experiment not only demonstrates the superior generalization capacity of the proposed PointSite, but also verifies the effectiveness of PointSite for real-world applications. On the other hand, when the tool is merged with our segmentation, the identification accuracy of all pocket-centric approaches is also significantly improved in terms of both atom-IoU and DCA as the previous Blind test on CAMEO hard targets.

Table 4. Comparison of Identification Performance on Bound−Unbound Pair Datasets in Terms of DCA Criterion at 4 Å Threshold as well as Atom-IoU (%). (PS Means combining PointSite)

	bound proteins			unbound proteins		
method	DCA	atom-IoU	PS	DCA	atom-IoU	PS
FPocket	34.2	22.1	w/o	27.9	16.2	w/o
SiteHound	41.2	23.9	w/o	36.7	21.1	w/o
P2Rank	65.1	39.6	w/o	63.0	36.6	w/o
FPocket	63.8	45.6	w	58.0	39.9	w
SiteHound	**70.0**	47.4	w	**66.5**	43.0	w
P2Rank	67.9	**48.3**	w	64.1	**44.5**	w
PointSite		47.3			44.3	

Blind and Unbound (apo) Proteins. Within the real-world drug design process, it is nontrivial to handle unbound (apo) proteins with low sequence identity to the training data. To test the capability of our model to handle blind and unbound proteins, we collect the corresponding proteins from hard targets in a CAMEO data set (03/19/2019−03/19/2022). In practice, we achieve 55 proteins after screening proteins with low sequence identity to the training set (<25%) and with holo structures. In order to obtain apo structures of proteins with holo structures, we fed proteins' sequences in AlphaFold2 (without template)[50] to predict their structures, which are regarded as unbound structures. Note that predicted structures with low accuracy (*global lDDT* < 0.8) are removed, leading to a blind and unbound data set containing 46 proteins with an average lDDT score of 87.6%. Table 5 lists the results of different methods on the blind and unbound data set. Our method achieves 29.4% atom-IoU, which is 14.6%, 12.2%, and 5.0% better than FPocket, DeepSite, and P2Rank, respectively, which denotes the superiority of our model to deal with blind and unbound proteins. Furthermore, when combined with other methods, PointSite is capble to boost other methods significantly in term of both atom-IoU and DCA.

Case Study. To further verify the effectiveness of PointSite in utilizing the intrinsic protein geometry information, we illustrate two bound−unbound pairs visualization results for

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

Table 5. Comparison of Identification Performance on CAMEO Blind and Unbound Datasets in Terms of DCA Criterion at 4 Å Threshold as Well as Atom-IoU (%)

method	DCA	atom-IoU	PointSite
FPocket	6.5	14.8	w/o
DeepSite	10.9	17.2	w/o
P2Rank	13.0	24.4	w/o
FPocket	56.5	26.5	w
DeepSite	45.7	20.0	w
P2Rank	45.7	24.7	w
PointSite		**29.4**	

better understanding in Figure 3. In the left three columns is the unbound−bound pair 1q1c−4lay (the protein structures of the first 1−260 residues of human FKBP52, unbound and bound with I63), in which the ligand bound region causes an obvious deformation (see the red point cloud labels in the middle ground truth column). Our PointSite can also predict the approximate LBS even for the unbound protein 1q1c. Compared with the left pair, the right unbound−bound pair 1p9r-1p9w (the protein structures of *Vibrio cholerae* putative NTPase EpsE, unbound and bound with ANP) contains fewer deformations. Thus, the prediction difference for 1p9r−1p9w is smaller than that of 1q1c−4lay. Even though more false-positive atoms exist in the unbound predictions, PointSite can always achieve the approximate ligand binding sites for unseen structures. Besides, once combined with pocket-centric approaches, such protein-centric approaches can also boost the ligand binding site predictions.

We also provide two case studies of RNA-dependent RNA polymerase (RdRp) from SARS-CoV-2 (COVID-2019) (PDB ID: 7BTF) and main protease(Mpro) of SARS-Cov2 (PDB ID: 7BQY). The ground truth ligand binding sites are obtained from ref 58 and are illustrated in the middle column of Figure 4. As the superimposed prediction and ground-truth in the third column of Figure 4 shows, PointSite can predict the accurate ligand binding site position compared with ground truths, and includes the core residues. These results shows that PointSite can be an inherent tool for future structure-based drug design.

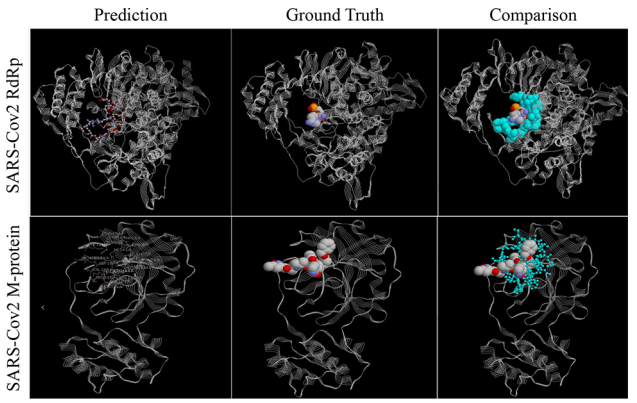


Figure 4. Case study of RNA-dependent RNA polymerase (RdRp) from SARS-CoV-2 (COVID-2019) (PDB ID: 7BTF) and main protease(Mpro) of SARS-Cov2 (PDB ID: 7BQY). The first column is the result predicted by PointSite, and the second and third columns illustrate the ground truth ligand binding and comparison of our prediction.

CONCLUSION

We introduced PointSite, a novel protein-centric approach for accurate identification of ligand binding atoms, which significantly improves the atom-level IoU over all previous approaches by a great margin. Our method distinguishes itself from previous protein-centric approaches in that we formulate the binding atom identification problem as a typical point clouds segmentation problem. We have shown that the segmented binding atoms from PointSite could serve as a postprocessing tool to guide any pocket-centric approaches through a filtering and reranking strategy to prioritize the identified candidate PSPs. Since pocket-centric approaches might output many false positive results, a subsequent prioritization step can greatly boost the performance of such tools. This is proven by the evidence that the merged results can greatly enhance the identification accuracy in terms of commonly used DCA criterion as well as the atom-level IoU for all pocket-centric approaches.

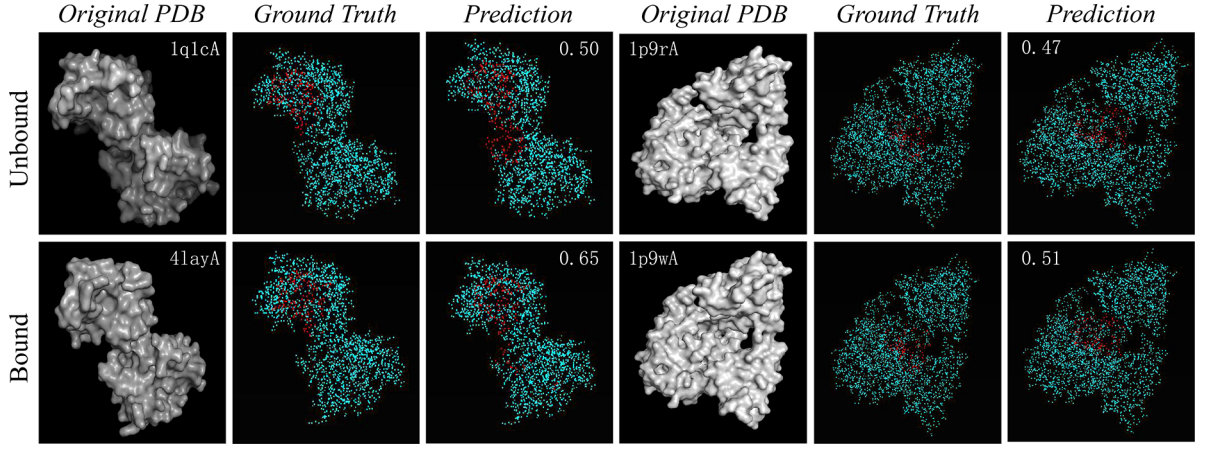


Figure 3. Visualization of two unbound−bound protein pairs. Even though a few false-positive atoms are segmented as ligand atoms on unbound proteins, PointSite can always figure out the accurate ligand binding sites for unseen structures.

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01512.

Description and Usage of the LIG_Tool (section S1); description of training and testing set (Table S1); comparison of identification performance in terms of atom-IoU with different identification radius (Table S2); identification performance in terms of atom-IoU using pocket-centric methods filtered by PointSite with different parameters (Tables S3−S7); identification performance in terms of DCA using pocket-centric methods filtered by PointSite with different parameters (Tables S8−S12); comparison of identification performance on bound-unbound pair data set and CAMEO blind test data set in terms of atom-IoU with different identification radius (Tables S13, S14); list of abbreviations (Table S15) (PDF)

PointSite code (ZIP)

LIG tool code (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

⊥**Zhen Li** − *The Chinese University of Hongkong (Shenzhen) & Future Network of Intelligence Institute, Shenzhen 518172, China;* orcid.org/0000-0002-7669-2686; Email: lizhen@cuhk.edu.cn

Authors

⊥**Xu Yan** − *The Chinese University of Hongkong (Shenzhen) & Future Network of Intelligence Institute, Shenzhen 518172, China*

⊥**Yingfeng Lu** − *The Chinese University of Hongkong (Shenzhen) & Future Network of Intelligence Institute, Shenzhen 518172, China*

Qing Wei − *The Chinese University of Hongkong (Shenzhen) & Future Network of Intelligence Institute, Shenzhen 518172, China*

Xin Gao − *King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia;* orcid.org/0000-0002-7108-3574

Sheng Wang − *Shanghai Zelixir Biotech Company Ltd., Shanghai 200030, China; CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Song Wu − *Shenzhen University, Shenzhen 518060, China;* orcid.org/0000-0003-3504-1630

Shuguang Cui − *The Chinese University of Hongkong (Shenzhen) & Future Network of Intelligence Institute, Shenzhen 518172, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01512

Notes

The authors declare no competing financial interest.

⊥X.Y., Y.L., and Z.L. are co-first authors.

Data and Software Availability. Our code and links of data set are publicly available on github: https://github.com/PointSite. Links of training and testing data set are available in "Training and Testing Data" section in README.md file. The setup and usage of the package are also described in README.md file. We have released a trained model in *model/scale_80.pth* in this repository for inference.

■ ACKNOWLEDGMENTS

This work was supported in part by NSFC-Youth 61902335, by Key Area R&D Program of Guangdong Province with Grant No. 2018B030338001, by the National Key R&D Program of China with Grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No.2017ZT07 × 152, by Guangdong Regional Joint Fund-Key Projects 2019B1515120039, by the NSFC 61931024&81922046, by helixon biotechnology company Fund and CCF-Tencent Open Fund.

■ REFERENCES

(1) Laurie, R.; Alasdair, T.; Jackson, R. M. 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.

(2) Naderi, M.; Lemoine, J. M.; Govindaraj, R. G.; Kana, O. Z.; Feinstein, W. P.; Brylinski, M. Binding Site Matching in Rational Drug Design: Algorithms and Applications. *Briefings Bioinf.* **2018**, bby078.

(3) Huang, B.; Schroeder, M. LIGSITE csc: Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation. *BMC Struct. Biol.* **2006**, *6*, 19.

(4) Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **2003**, *10*, 787−797.

(5) Jiang, M.; Li, Z.; Bian, Y.; Wei, Z. A Novel Protein Descriptor for the Prediction of Drug Binding Sites. *BMC Bioinf.* **2019**, *20*, 1−13.

(6) Zhou, W.; Yan, H. Alpha Shape and Delaunay Triangulation in Studies of Protein-Related Interactions. *Briefings Bioinf.* **2014**, *15*, 54−64.

(7) Skolnick, J.; Brylinski, M. FINDSITE: A Combined Evolution/Structure-Based Approach to Protein Function Prediction. *Briefings Bioinf.* **2009**, *10*, 378−391.

(8) Macalino, S. J. Y.; Gosu, V.; Hong, S.; Choi, S. Role of Computer-Aided Drug Design in Modern Drug Discovery. *Arch. Pharmacal Res.* **2015**, *38*, 1686−1701.

(9) Lipinski, C. A. Rule of Five in 2015 and Beyond: Target and Ligand Structural Limitations, Ligand Chemistry Structure and Drug Discovery Project Decisions. *Adv. Drug Delivery Rev.* **2016**, *101*, 34−41.

(10) Henrich, S.; Salo-Ahen, O. M.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C. Computational Approaches to Identifying and Characterizing Protein Binding Sites for Ligand Design. *J. Mol. Recognit.* **2010**, *23*, 209−219.

(11) Leis, S.; Schneider, S.; Zacharias, M. In Silico Prediction of Binding Sites on Proteins. *Curr. Med. Chem.* **2010**, *17*, 1550−1562.

(12) Fauman, E. B.; Rai, B. K.; Huang, E. S. Structure-Based Druggability Assessment—Identifying Suitable Targets for Small Molecule Therapeutics. *Curr. Opin. Chem. Biol.* **2011**, *15*, 463−468.

(13) Simões, T.; Lopes, D.; Dias, S.; Fernandes, F.; Pereira, J.; Jorge, J.; Bajaj, C.; Gomes, A. Geometric Detection Algorithms for Cavities on Protein Surfaces in Molecular Graphics: A Survey. *Computer graphics forum* **2017**, *36*, 643−683.

(14) Broomhead, N. K.; Soliman, M. E. Can We Rely on Computational Predictions to Correctly Identify Ligand Binding Sites on Novel Protein Drug Targets? Assessment of Binding Site Prediction Methods and a Protocol for Validation of Predicted Binding Sites. *Cell Biochem. Biophys.* **2017**, *75*, 15−23.

(15) Roche, D. B.; Brackenridge, D. A.; McGuffin, L. J. Proteins and Their Interacting Partners: An Introduction to Protein-Ligand Binding Site Prediction Methods. *Int. J. Mol. Sci.* **2015**, *16*, 29829−29842.

(16) Yuan, Y.; Pei, J.; Lai, L. Binding Site Detection and Druggability Prediction of Protein Targets for Structure-Based Drug Design. *Curr. Pharm. Des.* **2013**, *19*, 2326−2333.
(17) Xu, Y.; Wang, S.; Hu, Q.; Gao, S.; Ma, X.; Zhang, W.; Shen, Y.; Chen, F.; Lai, L.; Pei, J. CavityPlus: AWeb Server for Protein Cavity Detection With Pharmacophore Modelling, Allosteric Site Identification and Covalent Ligand Binding Ability Prediction. *Nucleic Acids Res.* **2018**, *46*, W374−W379.
(18) Zhao, Z.; Xu, Y.; Zhao, Y. SXGBsite: Prediction of Protein-Ligand Binding Sites Using Sequence Information and Extreme Gradient Boosting. *Genes* **2019**, *10*, 965.
(19) Cui, Y.; Dong, Q.; Hong, D.; Wang, X. Predicting Protein-Ligand Binding Residues With Deep Convolutional Neural Networks. *BMC Bioinf.* **2019**, *20*, 1−12.
(20) Xia, C.-Q.; Pan, X.; Shen, H.-B. Protein-Ligand Binding Residue Prediction Enhancement Through Hybrid Deep Heterogeneous Learning of Sequence and Structure Data. *Bioinformatics* **2020**, *36*, 3018−3027.
(21) Mylonas, S. K.; Axenopoulos, A.; Daras, P. DeepSurf: ASurface-Based Deep Learning Approach for the Prediction of Ligand Binding Sites on Proteins. *Bioinformatics* **2021**, *37*, 1681−1690.
(22) Kandel, J.; Tayara, H.; Chong, K. T. PUResNet: Prediction of Protein-Ligand Binding Sites Using Deep Residual Neural Network. *J. Cheminf.* **2021**, *13*, 1−14.
(23) Krivák, R.; Hoksza, D. P2rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites From Protein Structure. *J. Cheminf.* **2018**, *10*, 39.
(24) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites With Shape Descriptors. *Chem. Cent. J.* **2007**, *1*, 7.
(25) 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.
(26) Brady, G. P.; Stouten, P. F. Fast Prediction and Visualization of Protein Binding Pockets With Pass. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383−401.
(27) Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884−1897.
(28) le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, 168.
(29) Ghersi, D.; Sanchez, R. EasyMIFS and SiteHound: A Toolkit for the Identification of Ligand-Binding Sites in Protein Structures. *Bioinformatics* **2009**, *25*, 3185−3186.
(30) An, J.; Totrov, M.; Abagyan, R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752−761.
(31) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: Protein-Binding Site Predictor Using 3d-Convolutional Neural Networks. *Bioinformatics* **2017**, *33*, 3036−3042.
(32) Graham, B.; Engelcke, M.; van der Maaten, L. 3D Semantic Segmentation With Submanifold Sparse Convolutional Networks. *Proc. IEEE Conf. CVPR* **2018**, 9224−9232.
(33) Chen, K.; Mizianty, M. J.; Gao, J.; Kurgan, L. A Critical Comparative Assessment of Predictions of Protein-Binding Sites for Biologically Relevant Organic Compounds. *Structure* **2011**, *19*, 613−621.
(34) Haas, J.; Barbato, A.; Behringer, D.; Studer, G.; Roth, S.; Bertoni, M.; Mostaguir, K.; Gumienny, R.; Schwede, T. Continuous Automated Model EvaluatiOn (CAMEO) Complementing the Critical Assessment of Structure Prediction in CASP12. *Proteins: Struct., Funct., Bioinf.* **2018**, *86*, 387−398.
(35) Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical image computing and computer-assisted intervention* **2015**, 234−241.
(36) Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation From Sparse Annotation. *International conference on medical image computing and computer-assisted intervention* **2016**, 424−432.
(37) 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.
(38) Schmidt, T.; Haas, J.; Cassarino, T. G.; Schwede, T. Assessment of Ligand-Binding Residue Predictions in CASP9. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 126−136.
(39) Liu, B.; Wang, M.; Foroosh, H.; Tappen, M.; Pensky, M. *Sparse Convolutional Neural Networks* **2015**, 806−814.
(40) Hernandez, M.; Ghersi, D.; Sanchez, R. SITEHOUND-web: A Server for Ligand Binding Site Identification in Protein Structures. *Nucleic Acids Res.* **2009**, *37*, W413−W416.
(41) 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.
(42) Smith, R. D.; Clark, J. J.; Ahmed, A.; Orban, Z. J.; Dunbar, J. B., Jr; Carlson, H. A. Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing. *J. Mol. Biol.* **2019**, *431*, 2423−2433.
(43) Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M.; Huang, B. Identification of Cavities on Protein Surface Using Multiple Computational Approaches for Drug Binding Site Prediction. *Bioinformatics* **2011**, *27*, 2083−2088.
(44) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein- Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(45) 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.
(46) Roy, A.; Yang, J.; Zhang, Y. COFACTOR: An Accurate Comparative Algorithm for Structure-Based Protein Function Annotation. *Nucleic Acids Res.* **2012**, *40*, W471−W477.
(47) Schmidtke, P.; Souaille, C.; Estienne, F.; Baurin, N.; Kroemer, R. T. Large-Scale Comparison of Four Binding Site Detection Algorithms. *J. Chem. Inf. Model.* **2010**, *50*, 2191−2200.
(48) Xu, J.; Zhang, Y. How Significant Is a Protein Structure Similarity With TM-Score= 0.5? *Bioinformatics* **2010**, *26*, 889−895.
(49) Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; et al. The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information. *Nucleic Acids Res.* **2016**, gkw1000.
(50) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction With AlphaFold. *Nature* **2021**, *596*, 583−589.
(51) Huang, B. MetaPocket: AMeta Approach to Improve Protein Ligand Binding Site Prediction. *OMICS A Journal of Integrative Biology* **2009**, *13*, 325−330.
(52) Laurie, A. T.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics* **2005**, *21*, 1908−1916.
(53) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.
(54) Kawabata, T. Detection of Multiscale Pockets on Protein Surfaces Using Mathematical Morphology. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1195−1211.
(55) 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.
(56) Yu, J.; Zhou, Y.; Tanaka, I.; Yao, M. Roll: ANew Algorithm for the Detection of Protein Pockets and Cavities With a Rolling Probe Sphere. *Bioinformatics* **2010**, *26*, 46−52.

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845

(57) Wang, S.; Ma, J.; Xu, J. AUCpreD: Proteome-Level Protein Disorder Prediction by AUC-Maximized Deep Convolutional Neural Fields. *Bioinformatics* **2016**, *32*, i672−i679.
(58) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of Mpro from Sars-Cov-2 and Discovery of Its Inhibitors. *Nature* **2020**, *582*, 289−293.

https://doi.org/10.1021/acs.jcim.1c01512
J. Chem. Inf. Model. 2022, 62, 2835−2845