Supplementary Materials for

GeoBind: Segmentation of nucleic acid binding interface on protein surface with geometric deep learning

Pengpai Li and Zhi-Ping Liu*

*Corresponding author. Email: zpliu@sdu.edu.cn

The supplementary file includes:

Notes S1, to S4.
Figs. S1 to S5.
Tables S1 to S10.

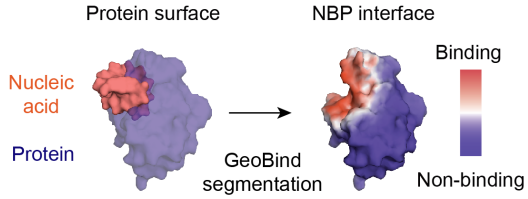
Note S1.

This note provides a comprehensive explanation of the proposed methods. In contrast to the METHODS section presented in the main text, this description elucidates each step using both graphical and textual explanations.

Problem formulation.

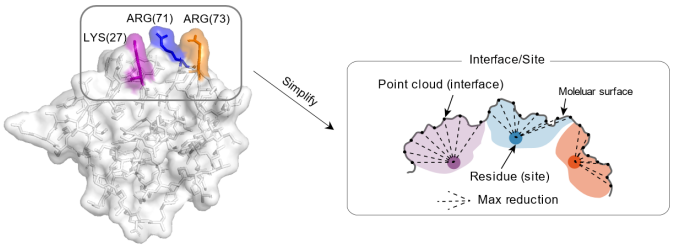
Step 1: GeoBind overview.

GeoBind is a tool used for classifying the nucleic acid binding interface on protein surfaces in a segmentation manner. It takes the entire surface of a protein as input, which is formatted as a point cloud, and then produces the binding scores for each point. In simple terms, a nucleic acid binding protein is represented as a point cloud, where each point is associated with an identity related to nucleic acid binding: $\mathcal{P}=\{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where N is the number of points on surface; $\mathbf{x}_i \in \mathbb{R}^3$ and $y_i \in \{0,1\}$. A point $\mathbf{x}_i$ is a binding interface, namely $y_i = 1$, if there exists a nucleic acid atom whose distance between them is less than 3 Å. The point cloud with features can be thought of as a map $f: \mathbb{R}^3 \to \mathbb{R}^n$, where the map assigns each point with a n-dimensional vector. In GeoBind, we design an SE(3)-equivariant operator $\mathcal{T}$ to produce a new function $\mathcal{T}(f) = o_f : \mathbb{R}^3 \to [0,1]$ that describes the point cloud with binding interface score $\hat{y}_i$.



Step 2: Transfer binding interface score to binding site score.

The above step provides a relatively complete description of GeoBind's segmentation task on protein surfaces. However, we further compute the binding preferences for protein sites (residues) participating in generating the point cloud. We collect the sites that participate in forming the surface, while those residues hidden inside of the surface are not considered. These sites are annotated by BioLiP (21) as the gold standards for binding or non-binding: $\mathcal{S}=\{s_i\}_{i=1}^{M}$, in which M denotes the number of sites on surface and $s_i \in \{0,1\}$. All evaluation metrics in the main text are computed according to the true label and predicted score of sites. This was done for two reasons: a) A traditional problem is the classification of binding sites, and evaluation based on them is a non-prejudiced comparison with existing methods. b) Binding sites are more authoritatively annotated by BioLip, while interface labels can only be calculated in terms of distances. One site binding score $\hat{s}_i$ is computed by max-pooling the binding scores of points generated by residue i: $\hat{s}_i = \max_{j \in R_i}\{\hat{y}_j\}$, where R_i denotes those surface points generated by residue i and is calculated along with the generating solvent excluded surface by program *msms* (25), as shown in the following figure.



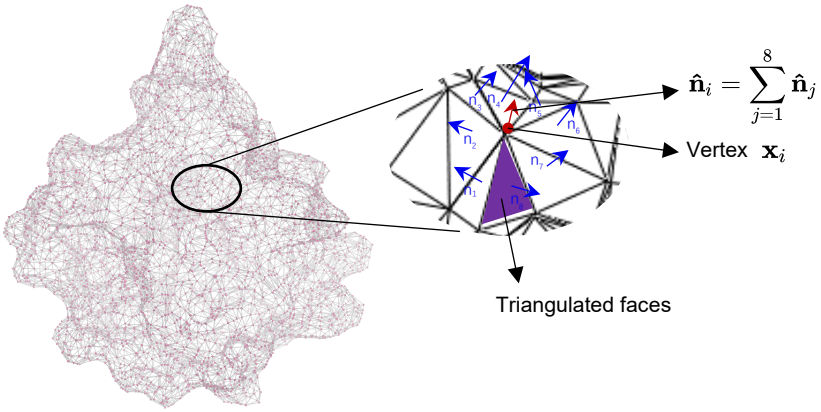
Oriented point cloud of protein surface.

Step 1: Adding missing hydrogen atoms.

X-ray crystallography cannot resolve hydrogen atoms in most protein crystals. As a result, most PDB files do not include hydrogen atoms. In some cases, hydrogens can be added to these files using modeling techniques. In PDB files resulting from NMR analysis, hydrogens are always present. To address the issue of missing hydrogen atoms, all proteins are protonated using a program called *reduce* (26), which adds the missing hydrogen atoms to the protein structure.

Step 2: Computing solvent excluded surface.

The classical solvent excluded surfaces (SES) (27) are triangulated using *msms* (25) program with parameters of density of 3 and water probe radius of 1.5 Å. The *msms* program takes input protein atoms with 3D coordinates and outputs a mesh which comprises vertices and triangulated faces. Then all protein meshes are resampled using PyMESH (28) at a resolution of 1.2 Å. As described in *Problem formulation*, the surface of a protein is represented by the vertices and their labels $\mathcal{P}=\{\mathbf{x}_i, y_i\}_{i=1}^{N}$. The normal $\hat{\mathbf{n}}_i$ of a reference point $\mathbf{x}_i$ on surface is computed by averaging the normal vectors of faces whose vertices contain the reference point $\mathbf{x}_i$. Then, the surface of a protein can be represented by $\mathcal{P}=\{\mathbf{x}_i, \hat{\mathbf{n}}_i, y_i\}_{i=1}^{N}$.



Descriptors.

Multiple sequence alignment (MSA) feature. The MSA information is of great significance in computational protein biotechnology. And it is a key intermediate step for predicting evolutionarily conserved properties such as tertiary structures, functional sites and interactions. We assign the MSA features to the point cloud according to the membership of points and residues. Specifically, the evolutionary score of a residue is assigned to the points of clouds generated by atoms in this residue. For a protein with the residue number of L, a profile hidden Markov model (HMM) matrix of shape $L\times 30$ is computed by using the tool *HHblits3* (29) searching against *Uniclust30* (30) database. The HMM matrix consists of three kinds of information, i.e., 20 columns of observed frequencies for twenty kinds of amino acids

in homologous sequences, 7 columns of transition frequencies and columns of local diversities.

Chemical feature. In GeoBind, we do not use the handcrafted protein physicochemical descriptors, such as electrostatics charge and hydropathy profile, etc. According to dMaSIF (*19*), the physicochemical environment of protein surfaces is easily regressed by a lightweight neural network using atomic point cloud. Therefore, a 1×6 vector of one hot encoding of six kinds of atoms (C, H, O, N, S, others), is considered as the chemical feature of GeoBind's input.

Geometric feature. For characterizing the geometric shape of point cloud, the shape index around each point on the surface is described by the local curvature. It is defined with respect to the principal curvature $\kappa_1, \kappa_2, \kappa_1 \geq \kappa_2$ as

$$\frac{2}{\pi}\tan^{-1}\frac{\kappa_1+\kappa_2}{\kappa_1-\kappa_2}. \quad (1)$$

After assigning the above features to the point cloud, we can represent the protein surface as: $\mathcal{P}=\{\mathbf{x}_i,\hat{\mathbf{n}}_i,\mathbf{f}_i,y_i\}_{i=1}^{N}$, where $\mathbf{f}_i \subset \mathbb{R}^{37}$.

Quasi-geodesic convolution.

GeoBind utilizes a local neighbor aggregation technique known as quasi-geodesic convolution to learn about the biological and geometric characteristics present on a protein surface. The concept of quasi-geodesic convolution was first introduced by dMaSIF (19) and involves the updating features of a reference point by merging the descriptors of nearby points, their distances, and their positions relative to the reference point. The three components will be explained in details, followed by the quasi-geodesic convolution formula.

The most accurate way to calculate distance on a protein surface is through geodesic distance. However, due to the high computational and memory requirements, an approximate method known as quasi-geodesic distance is used instead. **Step 1** will introduce and explain this method.

The relative position from a neighbor point to the reference point is a three-dimensional vector with the direction. This calculation requires a local reference frame (LRF) to be established for the reference point. The LRF must be independent of the initial coordinate system of the protein surface, making the model equivariant to SE(3) transformations (i.e. translation and rotation). Geometric gradients of a scalar field function on the protein surface can be used to determine the LRFs. **Steps 2** and **3** involve computing the relative position and scalar field function, respectively.

Finally, **Step 4** gives the formula of the Quasi-geodesic convolution that combines the points descriptors, distance and relative positions.

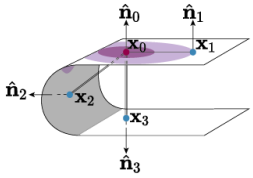
Step 1: Quasi-geodesic distance.

Computing the geodesic distance between every pair of points on a surface can be time-consuming. As shown in the following figure, an alternative approximation defines the geodesic distance between two points on a curved surface as

$$\mathrm{d}_{ij}=\left\|\mathbf{x}_i-\mathbf{x}_j\right\|\bullet\left(2-\left\langle\hat{\mathbf{n}}_i-\hat{\mathbf{n}}_j\right\rangle\right). \quad (2)$$

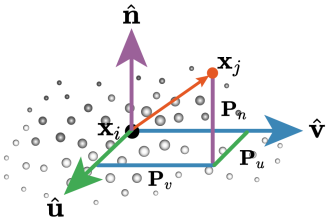
To localize the filters in convolutional layer, the geodesic distance is transformed by a smooth Gaussian window of $\sigma=12$ Å. The geodesic distance is defined as

$$w(\mathrm{d}_{ij}) = \exp(-\mathrm{d}_{ij} / 2\sigma^2)\,. \tag{3}$$



Step 2: **Local reference frame (LRF).**
For object recognition and surface registration task in 3D computer vision, a remarkable number of works introduced the LRF for designing 3D descriptors in order to reach model SE(3)-invariance (31-34). The LRFs indicate the local orientations of a 3D object. We build LRFs for all points on the protein surface. For any point $\mathbf{x}_i$, an LRF is represented as $\mathbf{C}_i = \{\hat{\mathbf{n}}_i, \hat{\mathbf{u}}_i, \hat{\mathbf{v}}_i\}$ to encode the relative positions between point $\mathbf{x}_i$ and its neighbors (see the following figure). The relative position $\mathbf{P}_{ij}$ between point $\mathbf{x}_i$ and $\mathbf{x}_j$ is a 3D vector and is defined as

$$\mathbf{P}_{ij} = [(\mathbf{x}_j - \mathbf{x}_i)^\top] \cdot [\hat{\mathbf{n}}_i \,|\, \hat{\mathbf{u}}_i \,|\, \hat{\mathbf{v}}_i]\,. \tag{4}$$



Here we give the details of generating the LRF of a point $\mathbf{x}_i$: $\mathbf{C}_i = \{\hat{\mathbf{n}}_i, \hat{\mathbf{u}}_i, \hat{\mathbf{v}}_i\}$. At first, $\hat{\mathbf{n}}_i$ is the normal vector of point $\mathbf{x}_i$ as described in Section *oriented point cloud of surface*. The normal vectors are equivariant to the SE(3) transformation of the protein. Then, we initialize the tangent vector $\hat{\mathbf{u}}'$, $\hat{\mathbf{v}}'$ using the orthonormal basis (36): $\hat{\mathbf{u}}' = [1 + sax^2, sb, -sx]$, $\hat{\mathbf{v}}' = [b, s + ay^2, -y]$, where $s = sign(z)$, $a = -1/(s+z)$ and $b = axy$. Next, we orient $(\hat{\mathbf{u}}', \hat{\mathbf{v}}')$ along the geometric gradient $\nabla^{\mathbf{u}',\mathbf{v}'} Q(\mathbf{x}_i)$ as following:

$$\nabla^{\hat{\mathbf{u}}',\hat{\mathbf{v}}'} Q(\mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^{N} w(\mathrm{d}_{ij}) [\mathbf{p}_{ij}^{\hat{\mathbf{u}}'}, \mathbf{p}_{ij}^{\hat{\mathbf{v}}'}] Q(\mathbf{x}_j) \tag{5}$$

$$\hat{\mathbf{u}}_i = (\nabla^{\hat{\mathbf{u}}'} Q(\mathbf{x}_i) \cdot \hat{\mathbf{u}}_i' + \nabla^{\hat{\mathbf{v}}'} Q(\mathbf{x}_i) \cdot \hat{\mathbf{v}}_i') / ((\nabla^{\hat{\mathbf{u}}'} Q(\mathbf{x}_i))^2 + (\nabla^{\hat{\mathbf{v}}'} Q(\mathbf{x}_i))^2) \tag{6}$$

$$\hat{\mathbf{v}}_i = (-\nabla^{\hat{\mathbf{v}}'} Q(\mathbf{x}_i) \cdot \hat{\mathbf{u}}_i' + \nabla^{\hat{\mathbf{u}}'} Q(\mathbf{x}_i) \cdot \hat{\mathbf{v}}_i') / ((\nabla^{\hat{\mathbf{u}}'} Q(\mathbf{x}_i))^2 + (\nabla^{\hat{\mathbf{v}}'} Q(\mathbf{x}_i))^2) \tag{7}$$

where Q is a scalar field function on protein surface $Q: \mathbf{x}_i \to \mathbb{R}$, $\mathbf{p}_{ij}^{\mathbf{u}'}$, $\mathbf{p}_{ij}^{\mathbf{v}'}$ are the relative positions of point $\mathbf{x}_j$ over the orientation $\hat{\mathbf{u}}'$ and $\hat{\mathbf{v}}'$ within the initial LRF of point $\mathbf{x}_i$. After building the LRF for each point, we can update the representation of protein as $\mathcal{P} = \{\mathbf{x}_i, \mathbf{C}_i, \mathbf{f}_i, y_i\}_{i=1}^N$.

Step 3: Choice of the scalar field function.
The generation of LRF requires a differentiable scalar field function Q. The choice of the function is diverse. An essential requirement for this function is that it is equivariant to SE(3) transformation. In this study, we choose BOARD (31) as the scalar function as it performs the best in both DNA- and RNA-binding site predictions. For a point on cloud, BOARD averages the signed distances to the tangent plane based on a subset of points within a cutoff radius distance. The tangent plane of a point is defined up to its normal vector. Here we

choose the cutoff radius the same as the size of Gaussian Window $\sigma = 12$ Å. The computing formulae of BOARD is given in Formula (8).

$$Q(\mathbf{x}_i) = \sum_{j \in \{j: \left\| \mathbf{x}_i - \mathbf{x}_j \right\| < \sigma\}} (\mathbf{x}_i - \mathbf{x}_j) \cdot \hat{\mathbf{n}}_i . \tag{8}$$

Step 4: Trainable convolution.

In the final stage, we utilize quasi-geodesic convolution as a method to combine points descriptors, distance and relative positions to obtain a high-level representation of the point cloud:

$$\mathbf{f}_i^t = \sum_{j=1}^{N} w(\mathrm{d}_{ij}) \mathrm{MLP}(\mathbf{P}_{ij}) \mathbf{f}_j^{t-1} . \tag{9}$$

In Equation (9), $w(\mathrm{d}_{ij})$ is the smoothed distance between point $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathbf{f}_i^t$ is the feature of point $\mathbf{x}_i$ at the t^{th} quasi-geodesic cnovolutional layer. The dimension of $\mathbf{f}_i$ is 64 for all quasi-geodesic convolutional layers. The MLP is a trainable multilayer perception for encoding the relative relations vector between point $\mathbf{x}_i$ and $\mathbf{x}_j$. The MLP layer consists of an input layer (3 units, which is dimension of relative position vector), a hidden layer (8 units), a ReLU non-linearity and an output layer (64 units). The MLP output layer dimension (64 units) is consistent with the dimension of $\mathbf{f}_i$. Accordingly, the quasi-geodesic convolution operation involves element-wise multiplication of $\mathrm{MLP}(\mathbf{P}_{ij})$ and $\mathbf{f}_j^{t-1}$ using the Hadamard product.

Note S2.

Description of the comparing four types of scalar field functions.

1) Local curvature. Described in Section **Methods** of the main text.

2) STED (sum of total Euclidean distances). From the definition, STED roughly describes the shape index of a protein from an overall perspective. The STED value varies from 0 (concave positions near to the mass center) to 1 (convex position far from the mass center). Specifically, STED is defined as:

$$Q_{(\mathbf{x}_i)} = \sum_{j=1}^{N} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2 .$$

3) FLARE. Similar to BOARD, for a point on cloud, BOARD averages the signed distances to the tangent plane, computed on a subset of points lying at the periphery of the support region. The two radiuses for the periphery are set as $\sigma 1 = 9\text{Å}$, $\sigma 2 = 12\text{Å}$, respectively. FLARE is defined as:

$$Q_{(\mathbf{x}_i)} = \sum_{j=1}^{N} (\mathbf{x}_i - \mathbf{x}_j) \bullet \hat{\mathbf{n}}_i ,$$

$$\text{where,} \quad M = \left\{ j : \sigma 1 < \| \mathrm{x}_i - \mathrm{x}_j \| < \sigma 2 \right\} .$$

4) MLP. Different from the handcrafted geodesic functions, MLP applies a trainable potential $Q_{(\mathbf{x}_i)} = Q_{(i)} = \mathrm{MLP}(\mathbf{f}_i)$, where $\mathbf{f}_i$ is the input feature of point $\mathbf{x}_i$.

Note S3.

Details of comparison experiments. All comparing predictors are trained and tested in the same datasets as GeoBind.

MaSIF-site. The standalone code of MaSIF is downloaded from its GitHub repository at https://github.com/LPDI-EPFL/masif. There are three applications in MaSIF, i.e., MaSIF-ligand, MaSIF-site and MaSIF-search. The framework of MaSIF-site can be transferred to the nucleic acid binding site predictions. All hyperparameters of the model and training strategies are the same as in the original paper. In its original code, limited by the computation cost, proteins with more than 8,000 surface points in the training set and more than 20,000 in the testing set are excluded.

dMaSIF-site. The source code of the dMaSIF-site was downloaded from https://github.com/FreyrS/dMaSIF and used as the default settings. The input of dMaSIF-site is the raw protein structure with only atom types and coordinates. The point cloud of protein surface used in dMaSIF is generated by its built-in smooth distance function.

3DZD. The source code of the 3DZD descriptors was downloaded from https://github.com/sebastiandaberdaku/AntibodyInterfacePrediction. The 3D Zernike descriptors is a classical protein surface representation method. It possesses several attractive features such as a compact representation, roto-translational invariance, and have been shown to adequately capture global and local protein surface shape. We used the program *single_structure_descriptors* with its default settings to generate 3D point cloud of proteins and their corresponding 3DZDs. The points (interfaces) were assigned with labels by measuring their distances to ligands atoms (cutoff of 3 Å). The evaluation metric of 3DZD methods are given by the true labels and predicted binding probabilities of interfaces.

As our nucleic acid binding site prediction task involves a large number of samples, we employed the Scikit-learn Bagging Classifier. This classifier builds 64 Support Vector Machine (SVM) classifiers by fitting them to random subsets of the training dataset. These SVMs are then used to predict the nucleic acid binding sites of proteins in the test dataset. We also experimented with the Random Forest classifier, but it did not perform as well as the Bagging SVMs.

GraphBind. The standalone code for GraphBind is downloaded from its webserver site http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/. The hyperparameters for GraphBind are set as recommended. The multiple alignment features of HMM and PSSM are considered in GraphBind. Referring to the description of feature contributions in GeoBind, the contribution of HMM is much greater than that of PSSM, and the combination of HMM and PSSM does not significantly improve the model performance using HMM. Due to the above evidence and the extensive computation of PSSM, we only use the combination of HMM, SS and AF features for nucleic acid residue encoding.

DRNAPred. DRNAPred is implemented by our best efforts according to the description by its published paper [doi.org/10.1093/nar/gkx059]. DRNAPred is a fast sequence-based method that accurately predicts DNA- and RNA-binding residues. The secondary structure (SS) and solvent accessibility (SA) features for residues encoding in DRNAPred are generated with relative programs. Here, we accurately compute the SS and SA features with *dssp* program with the input of a protein 3D structure.

Note S4.

We retrieved the Protein Databank (PDB)to identify the corresponding unbound structures of the bound NBPs in our compiled test dataset. For a bound NBP, we used the BLAST tool to search against the all protein sequences in PDB. The unbound proteins were selected if it satisfied a sequence identity cutoff $\geq$ 0.99 and with the bound proteins. To ensure sequence integrity, unbound candidates with missing, excessive, or mutated residues in internal positions are eliminated when compared to their corresponding bound proteins. When multiple candidates satisfy for one bound protein, the candidate with highest sequence identity is retained.

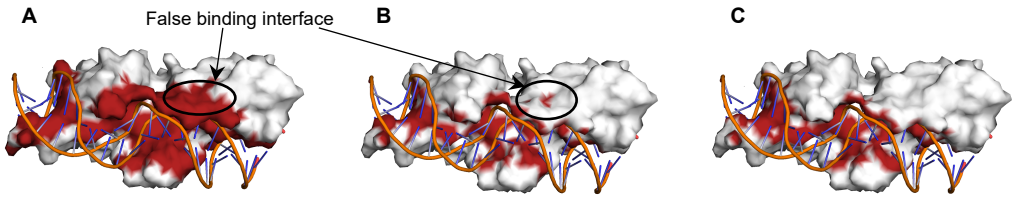


Fig. S1.

Illustration of three nucleic acid binding interface definitions on protein surface. (**A**) The points on the protein surface related to the binding residues (containing at least one heavy atom distance less than 3.5 Å to any atoms in nucleic acid) are defined as interface. (**B**) The points on the protein surface related to the binding atoms (distance less than 3.5 Å to any atoms in nucleic acid) are defined as interface. (**C**) A protein surface point is defined to be an interface point if its distance to any atom in the nucleic acid structure less than 3 Å. False positive interfaces are produced by the first two definitions.

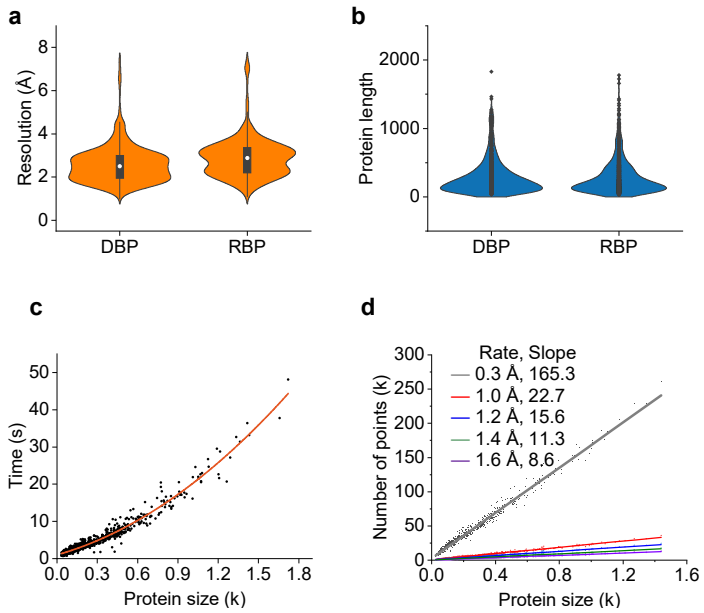


Fig. S2.

Basic description of proteins in our datasets and computation complexity of GeoBind for protein preprocessing. (**a**) The structure resolution distribution of DBPs and RBPs datasets. (**b**). The protein length distribution of DBPs and RBPs datasets. (**c**) Preprocessing time *vs* protein size. Protein size means the number of amino acid residues of a protein. (**d**) Number of points on the surface with five subsampling rates *vs* protein size. 0.3 Å is the initial resolution generated by the *msms* program.

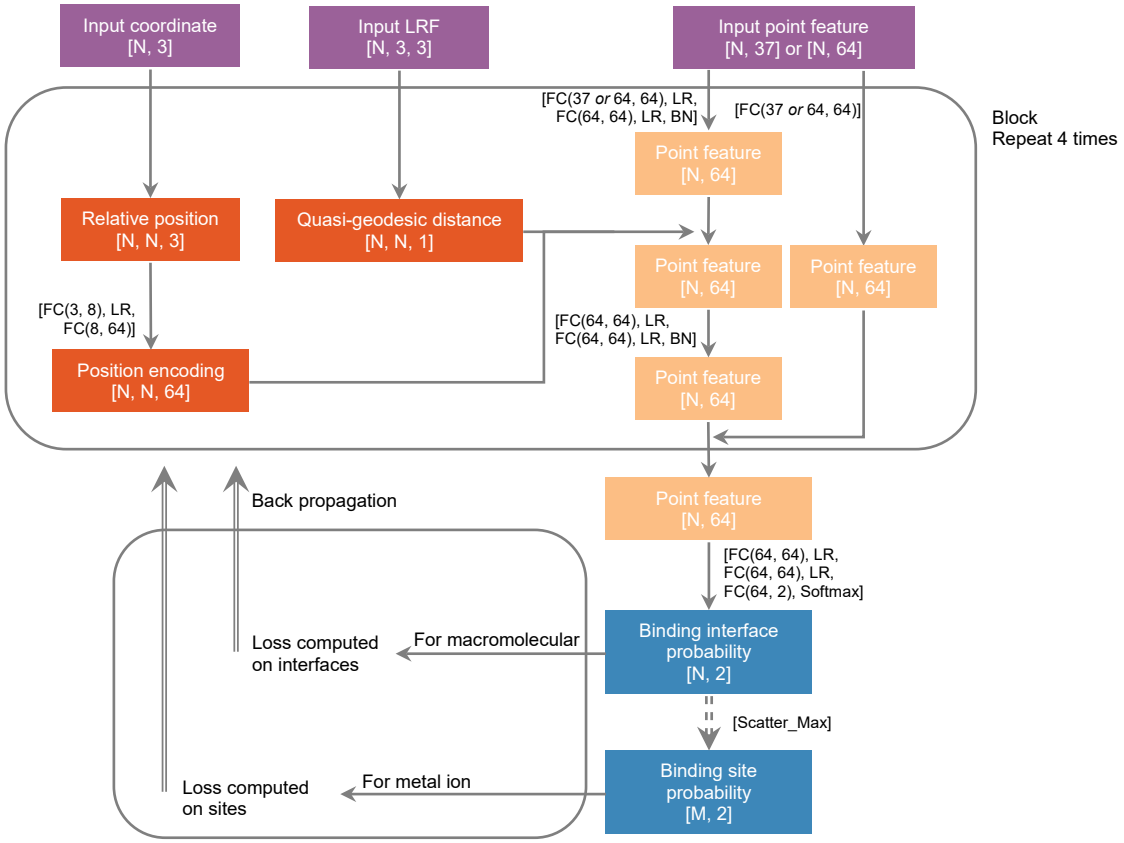


Fig. S3.

Details of model framework in GeoBind. GeoBind takes as input a point cloud of protein surface consisting of three components, i.e., coordinate, local reference frame (LRF) and feature of points. In the above diagrams, "N" denotes the number of points located on protein surface. "M" represents the number of residues associated with the generation of protein surface. "FC(I, O)" denotes a fully connected (linear) layer with "I" input channels and "O" output channels. "LR" denotes Leaky ReLU activation function with a negative slope of 0.2. "BN" denotes a batch normalization layer. The Scatter-Max operation is achieved by PyG[54] package. The probability of a residue being a binding site or not is given by maximizing the probabilities of points belonging to this residue.

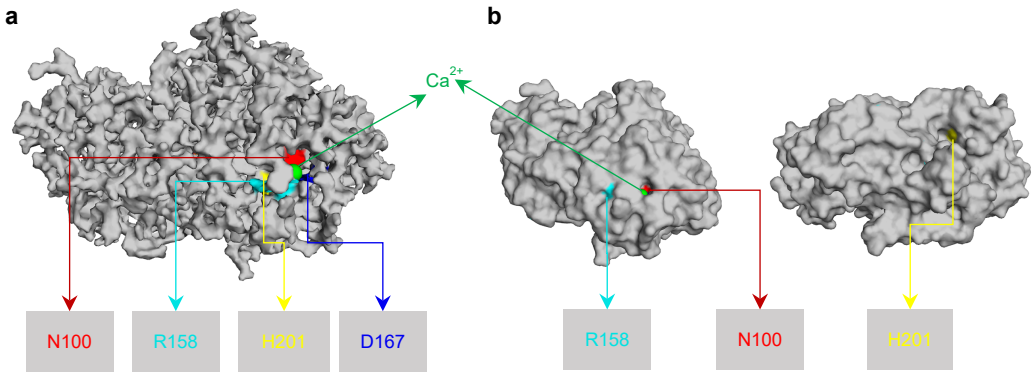


Fig. S4.

An illustration of why choosing a small probe radius for computing the surface of metal ion binding proteins. The Ca^{2+} binding protein (PDB ID: 34m5_A) is shown in the style of solvent excluded surface. Subfigure **a** is with the probe radius of 0.5 Å and **b** is with the probe radius of 1.5 Å. When the probe radius is set to 1.5 Å, the binding site D167 is buried inside the surface, while residue D167 contains atoms less than 3.5 Å away from Ca^{2+}. To account for this fact, we applied different probe radius to compute protein surfaces, e.g., 1.5 Å for macromolecular ligand and 0.5 Å for metal ion ligand.

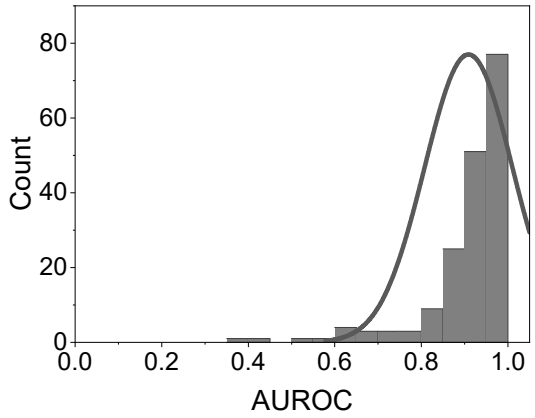


Fig. S5.

Distribution of AUROC values of classifying 179 DBPs in the DNA-179_Test dataset.

Table S1.

Summary of NBP datasets in GeoBind. PNratio represents the ratio of positive and negative samples.

Ligand	Dataset	Proteins	Positive sites	Negative sites	PNratio
DNA	DNA-719_Train	719	12,796	165,505	0.077
	DNA-179_Test	179	3,407	40,471	0.084
RNA	RNA-663_Train	663	15,154	144,511	0.105
	RNA-157_Test	157	3,791	33,916	0.112

Table S2.

Performance of GeoBind compared with the existing methods on our compiled datasets.

Dataset	Description	Method	Rec	Pre	F1	MCC	AUROC	AUPRC
DNA	Sequence	DRNApred	0.431	0.440	0.436	0.395	0.868	0.412
	Residual-graph	GraphBind	0.654	0.447	0.531	0.495	0.912	0.528
	Protein-surface	3DZD	**0.702**	0.096	0.169	0.166	0.751	0.124
		MaSIF-site	0.624	0.357	0.454	0.413	0.887	0.408
		dMaSIF-site	0.645	0.299	0.408	0.377	0.877	0.337
		GeoBind (Ours)	0.697	**0.492**	**0.576**	**0.544**	**0.941**	**0.572**
RNA	Sequence	DRNApred	0.450	0.457	0.453	0.402	0.831	0.434
	Residual-graph	GraphBind	0.639	0.425	0.510	0.456	0.884	0.502
	Protein-surface	3DZD	0.634	0.130	0.216	0.181	0.736	0.159
		MaSIF-site	0.601	0.348	0.440	0.377	0.829	0.360
		dMaSIF-site	0.638	0.277	0.386	0.332	0.829	0.329
		GeoBind (Ours)	**0.676**	**0.472**	**0.556**	**0.506**	**0.912**	**0.563**

Table S3.

Summary of NBP datasets collected in the GraphBind paper. Some proteins that failed to generate point clouds were not included in the list. Therefore, the number of proteins involved in training and testing is slightly less than the number of proteins in the original list of GraphBind.

Ligand	Dataset	Proteins	Positive sites	Negative sites	PNratio
DNA	DNA-573_Train	568	10,978	125,920	0.087
	DNA-129_Test	129	2,211	30,510	0.072
RNA	RNA-495_Train	488	11,306	103,025	0.110
	HEM-117_Test	109	1,930	26,563	0.072

Table S4.

Performance of GeoBind compared with the existing methods on datasets compiled by GraphBind.

Dataset	Method	Rec	Pre	F1	MCC	AUROC
DNA-129 Test	TargetDNA[a]	0.417	0.280	0.335	0.291	0.825
	TargetS[b]	0.239	0.370	0.291	0.262	N/A
	DNAPred[c]	0.396	0.353	0.373	0.332	0.845
	SVMnuce[d]	0.316	0.371	0.341	0.304	0.812
	COACH-D[e]	0.324	0.360	0.341	0.302	0.761
	NucBind[f]	0.323	0.373	0.346	0.309	0.797
	DNABind[g]	0.601	0.346	0.440	0.411	0.858
	GraphBind[h]	0.676	0.425	0.522	0.499	0.927
	Geobind	**0.618**	**0.509**	**0.558**	**0.526**	**0.940**
RNA-117 Test	RNABindRPlus[i]	0.273	0.227	0.248	0.202	0.717
	SVMnuc[g]	0.231	0.240	0.235	0.192	0.729
	COACH-D[e]	0.221	0.252	0.235	0.195	0.663
	NucBind[f]	0.231	0.235	0.233	0.189	0.715
	aaRNA[h]	0.484	0.166	0.247	0.214	0.771
	NucleicNet[i]	0.371	0.201	0.261	0.216	0.788
	GraphBind[h]	0.463	0.294	0.358	0.322	0.854
	Geobind	**0.522**	**0.345**	**0.416**	**0.373**	**0.874**

Notes: The experiments of the methods a~h are conducted by GraphBind. More details are available in GraphBind[12]. GeoBind was trained and tested on the training and testing datasets which are totally identical to those in GraphBind.

Table S5.

Ablation study for the contributions of feature subsets in GeoBind.

Dataset	Feature subset	Rec	Pre	F1	MCC	AUROC	AUPRC
DNA	All	**0.697**	**0.492**	**0.576**	**0.544**	**0.941**	**0.572**
	Chemical+Curvature	0.656	0.389	0.488	0.451	0.902	0.455
	HMM+Curvature	0.655	0.479	0.554	0.517	0.931	0.535
	HMM+Chemical	0.690	0.478	0.565	0.532	0.937	0.562
	HMM	0.671	0.446	0.536	0.501	0.925	0.526
	Chemical	0.603	0.393	0.476	0.434	0.899	0.440
	Curvature	0.628	0.268	0.376	0.335	0.844	0.325
RNA	All	**0.676**	**0.472**	**0.556**	**0.506**	**0.912**	**0.563**
	Chemical+Curvature	0.594	0.416	0.490	0.430	0.873	0.452
	HMM+Curvature	0.659	0.446	0.532	0.479	0.903	0.526
	HMM+Chemical	0.645	0.464	0.540	0.487	0.903	0.532
	HMM	0.664	0.426	0.519	0.466	0.894	0.518
	Chemical	0.626	0.341	0.441	0.380	0.845	0.410
	Curvature	0.633	0.292	0.400	0.336	0.826	0.326

Table S6.

Ablation study for the choice of scalar field functions affecting performances of GeoBind.

Dataset	Scalar function	Rec	Pre	F1	MCC	AUROC	AUPRC
DNA	Curvature	0.630	0.504	0.560	0.523	0.938	0.550
	STED	0.724	0.464	0.565	0.536	0.939	0.561
	BOARD	**0.697**	**0.492**	**0.576**	**0.544**	**0.941**	**0.572**
	FLARE	0.644	0.494	0.559	0.522	0.935	0.554
	MLP	0.698	0.464	0.558	0.525	0.937	0.554
RNA	Curvature	0.731	0.422	0.535	0.491	0.906	0.538
	STED	0.666	0.444	0.533	0.481	0.901	0.529
	BOARD	**0.676**	**0.472**	**0.556**	**0.506**	**0.912**	**0.563**
	FLARE	0.658	0.474	0.551	0.500	0.907	0.556
	MLP	0.693	0.447	0.554	0.495	0.909	0.543

Table S7.

Ablation studies for the depths of neural network and Gaussian window size.

Dataset		Scalar function	Rec	Pre	F1	MCC	AUROC	AUPRC
DNA	nBlocks	1	0.676	0.451	0.541	0.506	0.927	0.528
		2	0.721	0.440	0.546	0.517	0.933	0.544
		3	**0.725**	0.457	0.561	0.531	0.938	0.563
		4*	0.697	**0.492**	**0.576**	**0.544**	**0.941**	**0.572**
		5	0.686	0.474	0.561	0.527	0.936	0.560
	Radius	9.0	0.673	0.465	0.550	0.515	0.933	0.542
		12.0*	**0.697**	**0.492**	**0.576**	**0.544**	**0.941**	**0.572**
		15.0	0.677	0.484	0.565	0.530	0.936	0.548
RNA	nBlocks	1	0.651	0.445	0.528	0.475	0.896	0.528
		2	0.696	0.439	0.539	0.490	0.904	0.531
		3	0.656	0.483	0.556	0.505	0.910	0.550
		4*	0.676	0.472	**0.556**	**0.506**	**0.912**	**0.563**
		5	**0.738**	0.430	0.543	0.500	0.908	0.552
	Radius	9.0	0.668	0.467	0.550	0.499	0.907	0.553
		12.0*	0.676	**0.472**	**0.556**	**0.506**	**0.912**	**0.563**
		15.0	**0.699**	0.449	0.457	0.499	0.909	0.547

Table S8.

Summary of the five benchmark ligand datasets.

Ligand	Dataset	Proteins	Positive sites	Negative sites	PNratio
ATP	ATP-388_Train	364	4,909	113,119	0.043
	ATP-41_Test	40	593	11,765	0.050
HEM	HEM-175_Train	175	3,868	38,015	0.102
	HEM-96_Test	96	1,827	22,381	0.082
Ca^{2+}	CA-1022_Train	1,018	5,265	254,758	0.021
	CA-515_Test	514	3,512	185,372	0.019
Mn^{2+}	MN-440_Train	436	2,079	147,683	0.014
	MN-144_Test	144	706	50,824	0.014
Mg^{2+}	MG-1194_Train	1,190	4,307	319,989	0.013
	MG-651_Test	648	2,538	242,288	0.010

Notes: The names of the five datasets remain the same as their original ones. Few proteins are failed to be pre-processed due to the failure of *msms* program. Thus, for each ligand type, the number of proteins in the training and testing sets of GeoBind is listed in column “Proteins” accordingly.

Table S9.

Comparison results of GeoBind with other state-of-the-art methods for the five types of protein-ligand binding predictions (i.e., ATP, HEM, Ca^{2+}, Mn^{2+} and Mg^{2+}).

Dataset	Method	Rec	Pre	F1	MCC	AUROC
ATP	TargetS	0.516	0.689	0.590	0.580	N/A
(ATP-41_Test)	S-SITE	0.570	0.505	0.536	0.513	0.801
	COACH	0.632	0.703	0.666	0.652	N/A
	ATPbind	0.631	0.756	0.688	0.677	0.915
	DELIA	0.642	**0.758**	**0.695**	**0.685**	0.947
	GraphBind	0.603	0.666	0.631	0.616	0.939
	GeoBind	**0.732**	0.548	0.627	0.612	**0.963**
HEM	TargetS	0.493	**0.756**	0.597	0.588	N/A
(HEM-96_Test)	S-SITE	0.619	0.580	0.599	0.568	0.813
	COACH	0.677	0.403	0.505	0.476	0.835
	DELIA	0.648	0.660	0.654	0.628	0.951
	GraphBind	0.775	0.610	0.682	0.661	0.962
	GeoBind	**0.776**	0.669	**0.719**	**0.696**	**0.970**
Ca^{2+}	TargetS	0.174	0.506	0.259	0.291	N/A
(CA-515_Test)	S-SITE	0.303	0.124	0.176	0.174	0.661
	COACH	0.297	0.162	0.210	0.203	0.671
	IonCom	0.190	0.331	0.241	0.242	0.717
	DELIA	0.182	0.556	0.274	0.313	0.795
	GraphBind	0.325	0.563	0.410	0.420	0.863
	GeoBind	**0.394**	**0.624**	**0.483**	**0.488**	**0.900**
Mn^{2+}	TargetS	0.395	0.499	0.441	0.438	N/A
(MN-144_Test)	S-SITE	0.369	0.526	0.434	0.435	0.817
	COACH	0.562	0.272	0.367	0.381	0.821
	IonCom	0.531	0.495	0.512	0.506	0.872
	DELIA	0.513	0.632	0.566	0.565	0.903
	GraphBind	0.563	0.626	0.591	0.588	**0.951**
	GeoBind	0.518	**0.691**	**0.592**	**0.594**	0.946
Mg^{2+}	TargetS	0.154	0.449	0.229	0.259	N/A
(MG-651_Test)	S-SITE	0.243	0.132	0.171	0.169	0.682
	COACH	**0.273**	0.124	0.171	0.169	0.675
	IonCom	0.155	0.317	0.208	0.217	0.685
	DELIA	0.143	**0.562**	0.228	0.280	0.780
	GraphBind	0.259	0.410	0.317	0.320	0.827
	GeoBind	0.237	0.491	**0.320**	**0.337**	**0.869**

Table S10.

RNA binding information of proteins 4mdx and 6pif.

Chains of 4mdx	Binding sites for each chain of 4mdx	The affiliation or homologous information with the training set
A	N32 I34 G35 F38 S39 P40 T41	In RNA-157_Test.
B	F10 G18 S19 E20 Q21 G22 V24 R25 P26 I47 T48 A49 Q50 K53 L56 P57 T58 H59 F69 E70 S73 E78 Q79 D90	Homologous to Chain A

Chains of 6pif	Binding sites for each chain of 6pif	The affiliation or homologous information with the training set
A	A8 Y9 E10 R11 L39 L40 E44 H71 Y101 K102 W143 K144 R222 S224 Q225 V226 F227 K263 R283 R291 G344 G345 M346	Homologous to 6v9q:G:K in RNA-663_Train
B	A8 Y9 E10 R11 L39 L40 G41 H71 V73 Y101 K102 W143 K144 R222 S224 Q225 V226 F227 F262 K263 A266 R283 R291 K343 G344 G345 M346	
C	Y9 E10 R11 L39 L40 G41 H71 V73 Y101 K102 W143 R222 S224 Q225 V226 F227 S243 F262 K263 A266 R283 R291 K343 G344 G345 M346	
D	Y9 E10 R11 L39 L40 G41 Q42 V73 Y101 K102 W143 K144 R222 Q225 V226 F227 F262 K263 A266 R283 R291 G344 G345 M346	
E	Y9 E10 R11 L39 L40 G41 H71 V73 Y101 W143 K144 R222 Q225 V226 F227 T228 F262 K263 A266 R283 R291 K343 G344 G345 M346	
F	A8 Y9 E10 R11 H77 Y101 K102 W143 R147 K148 M220 R222 S224 Q225 V226 F227 R244 K263 R283 R291 K343 G344 G345 M346	
G	F89 L198 T199 Q200 I201 S202 L203 Y210 P215 V216 A217 S403 P404 S411 A414 G417 R424 L452 T453 E455 P501 R503 L504 A505 R506 Y570 L583 R584 Y586	Homologous to 6v9q:A:K in RNA-663_Train
H	H29 Y33 R103 Q105 D108 K109 R117 R120 R121 L122 K124 R125 F138 S151 R161 N162 F163 R164 N188 S189 E190	In RNA-157_Test
J	D45 N47	In RNA-157_Test