CrossBind: Collaborative Cross-Modal Identification of Protein Nucleic-Acid-Binding Residues

Linglin Jing[1,2*], Sheng Xu[3,1*], Yifan Wang[2], Yuzhe Zhou[4], Tao Shen[3], Zhigang Ji[5], Hui Fang[2], Zhen Li[4†], Siqi Sun[3,1 †]

[1]Shanghai Artifcial Intelligence Laboratory
[2]Department of Computer Science, Loughborough University
[3]Research Institute of Intelligent Complex Systems, Fudan University
[4] SSE & FNII, The Chinese University of Hong Kong (Shenzhen)
[5]Shanghai Jiao Tong University
{jinglinglin, xusheng1, sunsiqi1}@pjlab.org.cn, lizhen@cuhk.edu.cn

Abstract

Accurate identification of protein nucleic acid binding residues poses a significant challenge with important implications for various biological processes and drug design. Many typical computational methods for protein analysis rely on a single model that could ignore either the semantic context of the protein or the global 3D geometric information. Consequently, these approaches may result in incomplete or inaccurate protein analysis. To address the above issue, in this paper, we present CrossBind, a novel collaborative cross modal approach for identifying binding residues by exploiting both protein geometric structure and its sequence prior knowledge extracted from a large scale protein language model. Specifically, our multi modal approach leverages a contrastive learning technique and atom wise attention to capture the positional relationships between atoms and residues, thereby incorporating fine grained local geometric knowledge, for better binding residue prediction. Extensive experimental results demonstrate that our approach outperforms the next best state of the art methods, GraphSite and GraphBind, on DNA and RNA datasets by **10.8/17.3**% in terms of the harmonic mean of precision and recall (F1 Score) and **11.9/24.8**% in Matthews correlation coefficient (MCC), respectively. We release the code at https://github.com/BEAM-Labs/CrossBind.

Introduction

Proteins and nucleic acids (DNA or RNA) interact in numerous biological processes, including regulation of gene expression, signal transduction, and post-transcriptional modification and regulation. Identifying protein nucleic-acid-binding residues with accuracy is critical for comprehending the mechanisms behind various biological activities and developing new drugs. However, direct measurement of protein binding sites is challenging and often not feasible, especially when large-scale analyses are conducted. This is because it requires time-consuming and expensive experimental techniques, such as X-ray crystallography (Chayen and Saridakis 2008) or nuclear magnetic resonance spectroscopy (Tugarinov, Hwang, and Kay 2004). Therefore, computational prediction of binding residues in proteins with high efficiency and accuracy is essential.

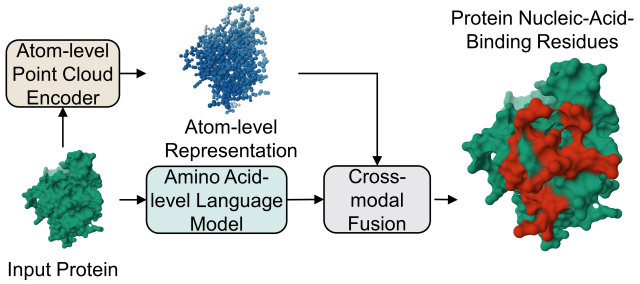

Figure 1: CrossBind model incorporates atom-level structure features in a point cloud representation and sequence features from a protein language model into a cross-model fusion module for protein Nucleic-Acid-Binding residues prediction (red).

Proteins can be represented as strings of letters using the 20 distinct types of amino acids (AAs) and a residue refers to a specific AA in a protein chain. Atoms are the basic units of matter that make up everything in the universe, including AA. Several approaches have been developed for representing proteins computationally, including one-hot encodings (Yan, Friedrich, and Kurgan 2016), position-specific scoring matrix (PSSM)(Su et al. 2019), pseudo-AA composition(Chou 2001), and hidden Markov models (HMM). Physico-chemical properties (Chen and Lim 2008), including hydrophobicity and electrostatics, have also proven effective for protein-related tasks. Currently, two primary types of protein-centric computational methods are available: sequence-based and structure-based methods.

Sequence-based methods analyze sequence-derived features to identify potential binding regions. Early machine

*These authors contributed equally. The work was done during their internships at the Shanghai Artificial Intelligence Laboratory.
†Corresponding author.

learning methods for predicting binding residues were primarily based on the primary protein sequence (Zhu et al. 2019; Su et al. 2019; Zhang, Chen, and Liu 2021). However, their performance is limited because the patterns of binding residues are implicit in the spatial structure and cannot be identified from sequence information alone (Wei et al. 2022).

Structure-based methods use protein structures to identify binding residues and generally outperform sequence-based methods. 3D convolutional neural networks, graph neural networks, and their variants have been widely adopted in structure-based methods (Lam et al. 2019; Liu and Hu 2013; Xia et al. 2021). However, structure-based methods typically require a large amount of biological information as training features, which consumes a lot of computing resources. Moreover, they may not accurately predict binding residues in cases where the protein or nucleic acid undergoes significant conformational changes upon binding (Chen and Ludtke 2021; Dai and Bailey-Kellogg 2021).

In recent years, the remarkable progress made in large-scale language modeling has extended to many fields (Floridi and Chiriatti 2020; Tenney, Das, and Pavlick 2019), including the study of amino acids in proteins. Amino acids, which are characteristic of certain letters in proteins, have been the subject of study in recent works such as AlphaFold2 (Jumper et al. 2021), which combines physical and biological knowledge about protein structure with deep learning algorithms implemented through transformer networks and 3D-equivariant structure transformation. Other works, such as RoseTTAFold (Baek et al. 2021), ESMFold (Verkuil et al. 2022), and ESM2 (Lin et al. 2022), have also been proposed, further improving the number of model parameters and computation efficiency. These developments are expected to have a significant impact on downstream protein function studies, such as the prediction of residues.

To address the limitations of the single-mode method, we present a new cross-modal training approach, named CrossBind, for identifying nucleic-acid-binding residues using both protein structure and sequence information. The proposed method leverages the power of deep learning to facilitate interactions between structure and sequence features at multiple scales, resulting in improved cross-modal fusion and utilization. The overview architecture of CrossBind is illustrated in Figure 2. The sequence encoder component employs ESM-2 (Lin et al. 2022), one of the largest protein language models to date, which was trained on millions of protein sequences with 15B model parameters. The structure encoder, on the other hand, uses a sparse convolution encoder (Schmohl and Sörgel 2019) to represent residues as a point cloud segmentation task at the atom-level. To capture the positional relationships between atoms and residues, an atom-wise attention (AWA) mechanism is introduced since the interactions between proteins and nucleic acids can occur on both backbone and side-chain atoms. Additionally, we introduce a self-supervised learning (SSL) strategy to account for conformational changes in 3D protein structures, increasing the diverse mobility of atoms and enhancing their ability to transmit signals when interacting with other molecules. Furthermore, since our dataset is imbalanced, we employ SSL to enhance the robustness of our model (Liu et al. 2021).

In summary, our main contributions are listed as follows:

- We propose a novel cross-modal strategy, CrossBind, that combines protein structure and sequence information to identify nucleic-acid-binding residues.
- Our method employs an atom-level point cloud segmentation on residues, along with an atom-wise attention component, to efficiently extract fine-grained local geometric knowledge of protein structure.
- We incorporate several biological task-related modules and demonstrate that our approach achieves state-of-the-art performance on multiple datasets consistently.

Related Work

Sequence-Based Method

Sequence-based methods offer a flexible approach to predicting protein-nucleic-acid-binding residues that can be applied to any protein sequence. There are two main types of sequence-based models: alignment-based methods and machine-learning-based models. Alignment-based methods rely on the assumption that proteins with similar sequences share similar binding partners and binding residues (Xue, Dobbs, and Honavar 2011). These methods predict binding residues by comparing annotations from proteins in the database that are sufficiently similar to the input protein. To do this, they typically require a database containing annotations of known binding residues. Sequence similarity between protein chain pairs can be calculated using E-value (McGinnis and Madden 2004) and TM-align (Zhang and Skolnick 2005). Machine-learning-based methods, on the other hand, predict the probability of each residue binding or not by leveraging sequence contextual information. Each protein residue is encoded with a feature vector as the model input, which typically contains physicochemical characteristics of the predicted residue and its neighboring residues. Examples of such methods (Pan and Shen 2018; Grønning et al. 2020) typically employ 1D convolution layers and bidirectional LSTM to capture local and global features from the protein sequence for binding prediction.

Recent progress in protein language models, such as ESM2, has enabled the utilization of pre-trained models for processing protein sequences. Leveraging the vast amount of data regarding the physical and chemical properties of protein structures, these models have displayed remarkable accuracy in predicting protein structures and executing a wide range of downstream tasks based exclusively on protein sequences (Lin et al. 2022).

Structure-Based Method

Recent structure-based methods (Lam et al. 2019; Xia et al. 2021) use low-resolution structural information, such as spatial neighbors, solvent accessibility, and secondary structure (Liu and Hu 2013), derived from protein structures to predict binding residues. These methods employ different approaches, such as constructing graphs or 3D-CNNs as spatial representation encoders. GraphBind (Xia et al. 2021), for

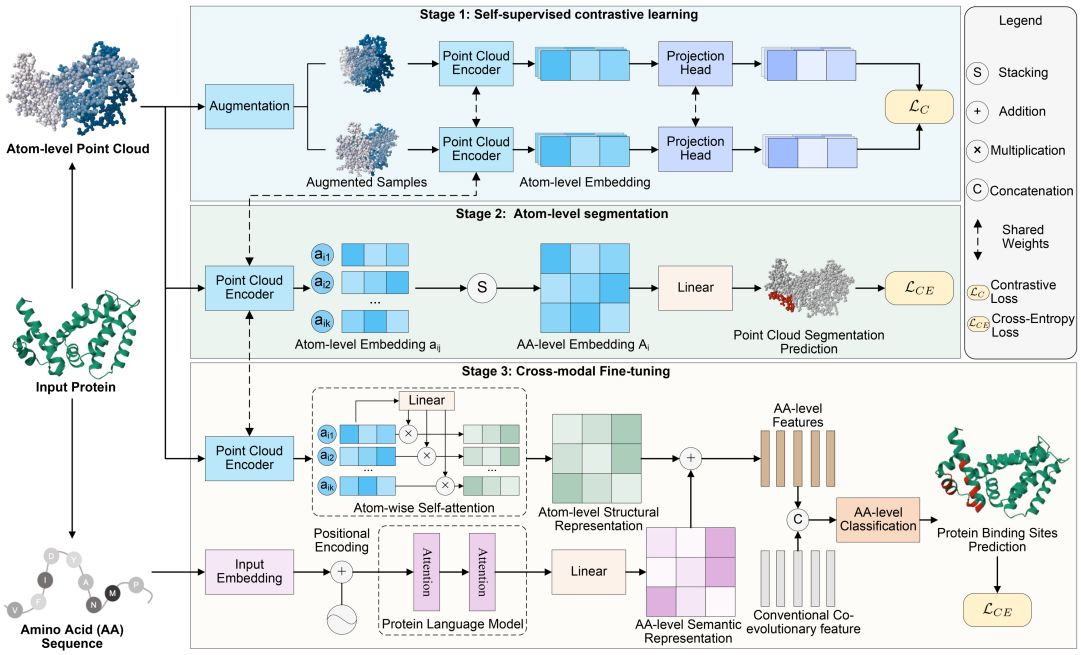

Figure 2: The overall architecture of CrossBind. Given a query protein, the input comprising atom-level structure information is fed to the Point Cloud Encoder. The encoder, which is pre-trained by using a self-supervised learning strategy, generates a structural point cloud representation. Further, an atom-wise attention module is introduced to capture the positional relationships between atoms and residues. Finally, the cross-modal module combines the structural and sequence representations to concatenate with co-evolutionary features for the prediction of protein binding sites.

example, proposes a hierarchical graph neural network that learns protein structural context embeddings for recognizing nucleic-acid-binding residues by using the residue and its physicochemical properties as nodes and the positional distance between residues as edges to construct a graph. However, due to the non-Euclidean nature of protein structures, learning latent knowledge from structures remains one of the most significant challenges (Wei et al. 2022; Bheemireddy et al. 2022).

On small-molecule-level tasks, such as ligand prediction, point cloud-based encoders are widely used (Yan et al. 2022; Wang et al. 2022). Due to the distribution state of small molecules (atoms) in protein, they can be expressed as a form of the point cloud. However, due to the complexity of residues, which contain multiple atoms, representing them in point cloud form presents a challenge. While deep learning methods based on point clouds as high-resolution structure encoders have been successful in computer vision and autonomous driving (Guo et al. 2020), applying these methods to protein residues remains an active area of research.

Cross-Modal Learning

Learning from multiple modalities can provide rich learning signals that enable the extraction of semantic information from a given context (Zhou, Ruan, and Canu 2019). Recent studies have shown that cross-modal learning, which combines information from different modalities, can achieve better results than using a single modality alone (Ding et al. 2021; Panda et al. 2021). The use of cross-modal learning has many potential applications, including in the fields of computer vision, natural language processing, and robotics, where it can help to improve the accuracy and efficiency of tasks such as object recognition, speech recognition, and machine translation.

Protein structure and sequence information can be seen as two distinct modalities that provide complementary information for predicting protein properties. Graphsite (Yuan et al. 2022) introduced a transformer that combines sequence and structure information to predict DNA-binding residues. Specifically, Graphsite uses AlphaFold2 to represent the sequence and maps the protein residues to a distance matrix that captures the pairwise distance relationship between each residue. However, using only the distance matrix to represent the 3D structure may result in a loss of spatial information, which is not considered true cross-modal learning.

Methods

This paper presents a novel cross-modal learning framework, called CrossBind, that aims to enhance the identification of protein nucleic-acid-binding residues. The approach leverages both atom point clouds and amino acid sequences to learn a unified representation of protein structure and sequence. The paper is organized as follows: first, we introduce the pre-training of the atom point cloud segmentation in Section . Next, we present the details of our cross-modal that integrates information from both protein structure and

sequence in Section . Finally, we describe a filter module in Section that is designed to leverage protein structure and biological properties and achieve further performance improvements. The overall approach is illustrated in Figure 2.

Definitions and Problem Formulation

Our goal is to identify binding residues in a given query protein, using a cross-modal training strategy that combines atom-level point cloud segmentation (ALS) with a protein large language model (LLM). The input data for each protein consists of both structural and sequence information. The structural data is composed of atom point clouds, each with three spatial coordinates (X, Y, Z), while the sequence data consists of amino acids with varying lengths ranging from tens to thousands. There are 21 AA types and 5 atom types in each protein.

To learn a segmentation encoder that can effectively represent protein 3D structure information, we employ a self-supervised learning strategy using the sparse convolution encoder F_Θ and multi-layer perceptron (MLP) projection heads G_ψ on unlabeled atom cloud points. The learned representations are then used for downstream tasks. After applying F_Θ, each atom is represented by a 32-dimensional vector $a_{ij} \in \mathbb{R}^{32}$. To obtain an amino acid level representation $A_i \in \mathbb{R}^{448}$, we stack the atom representations a_{ij} corresponding to the j-th atoms in the i-th amino acid.

In the cross-modal module, we employ an atom-wise attention (AWA) mechanism to process the atom-level representations $\{a_{i1}, a_{i2}, \ldots, a_{ik}\}$ obtained from the F_Θ, where k denotes the maximum index of atom in amino acid. This mechanism captures the positional relationships between the atoms and residues and produces a new representation $F_{struct} \in \mathbb{R}^{L\times 448}$ that encodes the structural information, where L denotes the number of AAs in a protein. Simultaneously, the sequence information is captured by a large language model (LLM), which produces a sequence representation $F_{seq} \in \mathbb{R}^{L\times 1280}$. In addition, we consider the conventional biological co-evolutionary feature $F_{evo} \in \mathbb{R}^{L\times 54}$ as an extra feature for identifying nucleic-acid-binding residues. The details of how these cross-modal features are fused are described in Section .

Atom-Level Segmentation (ALS) Module

Point cloud encoder. For protein structure analysis, we adopt a sparse convolution-based U-net (Schmohl and Sörgel 2019) as the segmentation encoder F_Θ. To accomplish this, we convert the original protein atoms into point clouds that include atom coordinates and features. These features are one-hot embeddings of amino acids and atoms. Consequently, the input for segmentation encoder includes 27-dimensional atom features and 3-dimensional spatial coordinates. The ALS output, $a_{ij} \in \mathbb{R}^{32}$, provides atom-level spatial information. However, binding sites identification occurs at the amino acid level; Hence, we employ a padding strategy to stack a_{ij} onto the amino acid level A_i:

$$A_i = [\hat{a}_{i1}, \hat{a}_{i2}, \ldots, \hat{a}_{iK}], \hat{a}_{ij} = \begin{cases} a_{ij} & j \le k \\ \mathbf{0} & k < j \le K \end{cases} \quad (1)$$

where i denotes the index of an amino acid, j denotes the index of an atom within an amino acid and k represents the maximum number of atoms within any given amino acid. $[.]$ denotes the concatenate function, and K is a constant that serves as an upper limit on the number of atoms any amino acid can contain in the entire data set.

Self-supervised learning (SSL). To address the issue of atom mobility in protein 3D structures, we use SSL to improve the identification of conformational changes that occur during binding. Furthermore, the success of self-supervised learning in handling imbalanced data has motivated us to use it to enforce invariance to a set of point cloud geometric transformations. Our approach involves using protein atoms point clouds as input and constructing augmented versions Q^{t_1} and Q^{t_2} using randomly combined transformations that include normal transformations such as rotation, scaling, and translation, as well as spatial transformations such as elastic distortion and jittering. The point cloud encoder F_Θ maps atom point clouds to the feature embedding space, and the feature embedding is then projected to an invariant space with projection heads G_ψ. We denote the projected vectors as z_i^{t1} and z_i^{t2}, where $z_i^t = F_\Theta(G_\psi({Q_i}^t))$. Similar to SimCLR (Chen et al. 2020), we use the NT-Xent loss as the contrastive loss (Section).

Cross-Modal Module

Atom-Wise Attention (AWA). The structure of our AWA module, which is illustrated in Figure 2. The AWA module is designed to dynamically highlight the centroid of each residue, encompassing both backbone and side-chain atoms, and generate the residue representation by stacking its constituent atoms. The atom representations a_{ij} from a residue are combined to form the global representation between all atoms. The atom-wise attention score $\Lambda = \{\sigma_j | \, j = 1, 2, \ldots K\}$ is calculated using a simple MLP mechanism (Chen et al. 2022), and a $Sigmoid$ function is used to map the attention value to a range of (0, 1).

$$\Lambda = \text{Sigmoid}\left(\text{MLP}\left([a_{i1}, a_{i2}, \ldots, a_{iK}]\right)\right), \quad (2)$$

According to Eq 3, each element in Λ indicates the importance of an atom in the same residue. The attention score is used to weight the atom representations a_{ij} by element-wise multiplication. Finally, we concatenate the weighted atom representations and use an MLP layer to generate the protein structure representation F_{struct}.

$$F_{struct} = \text{MLP}\left([a_{i1} \odot \sigma_1, a_{i2} \odot \sigma_2, ..., a_{iK} \odot \sigma_K]\right)), \quad (3)$$

where K has the same definition as Eq 1, σ_K is the K-th scale element from Λ, $\odot$ denotes the element-wise multiplication to scale weight the atom point cloud representations.

Cross-modal fusion. We used three groups of protein features to train our model: protein structural representation F_{struct}, sequence representation F_{seq}, and conventional biological co-evolutionary feature F_{evo}. $F_{seq} \in \mathbb{R}^{L\times 1280}$ first compress the feature dimension to the same as $F_{struct} \in \mathbb{R}^{L\times 448}$ by MLP, where L denotes the number of amino acids in a protein. Then F_{seq} and F_{struct} were fused together with the following fusion rules:

$$P_o = \lambda F_{struct} + (1-\lambda)F_{seq}, \tag{4}$$

where λ denotes a learnable parameter, '+' denotes the element-wise addition. At last, conventional biological co-evolutionary feature $F_{evo} \in \mathbb{R}^{L\times 54}$ was concatenated with P_o as the final group feature $P_{final} \in \mathbb{R}^{L\times 502}$ for the binding residue identification:

$$P_{final} = ([P_o, F_{evo}]). \tag{5}$$

Residue Propensity Filter (RPF)

As highlighted earlier, traditional approaches for identifying binding sites often require additional information, such as geometric or charge distribution, to accurately define binding regions. However, our proposed CrossBind method utilizes the inherent characteristics of proteins to identify binding residues. Specifically, binding residues are known to predominantly occur on the surface of proteins, which inspired us to design a biological filter that enhances the interpretability of the identification task. We adopted the amino acid propensity (Kim, Yura, and Go 2006), which measures the likelihood of an amino acid to interact with nucleic-acids, as a filtering condition. This biological property, such as the higher propensity of positively charged amino acids to interact with nucleic-acids, has been well established through biological experiments. Additionally, we leveraged the Geodesic-distance (Sverrisson et al. 2021), a measure of the distance between amino acids on the protein surface, to screen for outliers within a certain range. The outlier amino acids were then used to scale the logits of the CrossBind output according to their corresponding propensity, thereby improving the accuracy of the model. The amino acid propensity is calculated as follows:

$$\xi^i = (\frac{\bar{n}^i}{\sum_{i=1}^{20}\bar{n}^i})/(\frac{n^i}{\sum_{i=1}^{20}n^i}), \tag{6}$$

where n_i denotes the number of amino acid i with label 0 and $\bar{n}_i$ is label 1. ξ is calculated on the training set.

The algorithmic specifications of the RBF method can be found in the Supplementary Material.

Loss Functions

Classification loss. Given a training set V_{tr}, in ALS and Cross-Modal module, we use cross-entropy loss as:

$$\mathcal{L} = -\sum_{V_{tr}} \left(y_i \ln \widehat{y}_i + (1-y_i)\ln\left(1-\widehat{y}_i\right)\right), \tag{7}$$

where y_i is the label of a residue and $\widehat{y}_i$ is the probability corresponding to y_i.

Contrastive loss. In SSL strategy, we leverage NT-Xent loss, which maximize the similarity of $(\mathbf{z}_i^{t_1}, \mathbf{z}_i^{t_2})$ and minimizing the similarity with all the other samples in the mini-batch of point clouds. The loss function for a positive pair of examples (i, j) is defined as:

$$S_{i,i}^{t1,t2} = \cos\mathrm{sim}(z_i^{t1}, z_i^{t2}), \tag{8}$$

$$\mathcal{L}(i) = -\log \frac{\exp\left(S_{i,i}^{t_1,t_2}/\tau\right)}{\sum_{\substack{j=1\\ j\neq i}}^{N}\exp\left(S_{i,j}^{t_1,t_1})/\tau\right) + \sum_{j=1}^{N}\exp\left(S_{i,j}^{t_1,t_2}/\tau\right)}, \tag{9}$$

where $\cos\mathrm{sim}(\cdot)$ denotes the cosine similarity function. N is the mini-batch size, τ is a pre-set temperature constant. z_i^t denotes the projected vector.

Experiments

Experiments Setup

Datasets and evaluation metrics. We utilized two benchmark datasets, DNA_129 and RNA_117 dataset, from a previous study for training and testing our method. These datasets were obtained from the BioLip database (Yang, Roy, and Zhang 2012) and consist of experimentally determined complex structures. The DNA_573 comprises 573 training proteins and 129 testing proteins, while the RNA_117 dataset consists of 495 training proteins and 117 testing proteins. A binding residue was identified if the smallest atomic distance between the target residue and the DNA molecule was less than 0.5 Å plus the sum of the Van der Waal's radius of the two nearest atoms. To demonstrate the generalizability of our method, we used an additional independent testing set: DNA_181, which contained 181 proteins whose structures were predicted by Alphafold2. The datasets are highly imbalanced, with a large ratio between positive and negative samples.

To evaluate the performance of our method, we used several commonly metrics, including precision (Pre), recall (Rec), F1-score (F1), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPR). AUC and AUPR are threshold-independent measures, providing an overall assessment of the model's performance. The remaining metrics require the use of a threshold to convert predicted binding probabilities into binary predictions, which we determined by maximizing the F1-score. The performance metrics are summarized in Table 1.

Implementation details. In all experiments, we used the Adam optimizer with a weight decay of 1×10^{-4} and employed cosine annealing as the learning rate scheduler. We set the initial learning rate to 1×10^{-3} for the ALS module and 1×10^{-4} for the cross-modal module. In the RPF module, we ranked the nearest neighbors and selected the top five amino acids for the propensity filter using prediction logits with thresholds of [-0.8, 0.8], corresponding to a positive over a negative propensity. For data processing, we centered and re-scaled each point cloud to fit into a sphere, and then represented it as a sparse voxel representation with 0.1 voxel size. We trained and validated our method on the training dataset, with a validation ratio of 0.1.

Comparison With State-of-the-Art Methods

We evaluated the performance of our method against state-of-the-art methods on three nucleic-acid-binding test sets: DNA_129, DNA_181, and RNA_117, as reported in previous

Dataset	Method	Struct	Seq	LLM	Rec	Pre	F1	MCC	AUC	AUPR
DNA_129	COACH-D (Wu et al. 2018)	✓			0.328	0.318	0.323	0.279	0.712	0.248
	NucBind (Su et al. 2019)	✓			0.322	0.366	0.343	0.304	0.809	0.284
	SVMnuc (Su et al. 2019)		✓		0.316	0.371	0.341	0.304	0.812	0.302
	NCBRPred (Zhang, Chen, and Liu 2021)		✓		0.312	0.392	0.347	0.313	0.823	0.310
	DNABind (Liu and Hu 2013)	✓			0.487	0.389	0.433	0.395	0.832	0.391
	DNAPred (Zhu et al. 2019)		✓		0.396	0.353	0.373	0.332	0.845	0.367
	GraphBind (Xia et al. 2021)	✓			0.625	0.434	0.512	0.484	0.916	0.497
	GraphSite[a] (Yuan et al. 2022)		✓	✓	0.665	0.460	0.543	0.519	0.934	0.544
	CrossBind	✓	✓	✓	**0.684**	**0.538**	**0.602**	**0.581**	**0.953**	**0.628**
DNA_181	COACH-D (Wu et al. 2018)	✓			0.254	0.280	0.266	0.235	0.655	0.172
	NCBRPred (Zhang, Chen, and Liu 2021)	✓			0.259	0.241	0.250	0.215	0.771	0.183
	SVMnuc (Su et al. 2019)		✓		0.289	0.242	0.263	0.229	0.803	0.193
	NucBind (Su et al. 2019)		✓		0.293	0.248	0.269	0.234	0.796	0.191
	DNABind (Liu and Hu 2013)	✓			0.535	0.199	0.290	0.279	0.825	0.219
	DNAPred (Zhu et al. 2019)		✓		0.334	0.223	0.267	0.233	0.655	0.172
	GraphBind (Xia et al. 2021)	✓			0.505	0.304	0.380	0.357	0.893	0.317
	GraphSite[a] (Yuan et al. 2022)		✓	✓	0.517	0.354	0.420	0.397	0.917	0.369
	CrossBind	✓	✓	✓	**0.538**	**0.432**	**0.475**	**0.448**	**0.932**	**0.424**
RNA_117	RNABindR Plus (Yu et al. 2013)		✓		0.273	0.227	0.248	0.202	0.717	-
	SVMnuc (Su et al. 2019)		✓		0.231	0.240	0.235	0.192	0.729	-
	COACH-D (Wu et al. 2018)	✓			0.221	0.252	0.235	0.195	0.663	-
	NucBind (Su et al. 2019)	✓			0.231	0.235	0.233	0.189	0.715	-
	aaRNA (Li et al. 2014)	✓			0.484	0.166	0.247	0.214	0.771	-
	NucleicNet (Lam et al. 2019)	✓			0.371	0.201	0.261	0.216	0.788	-
	GraphBind (Xia et al. 2021)	✓			0.463	0.294	0.358	0.322	0.854	-
	CrossBind	✓	✓	✓	**0.490**	**0.366**	**0.420**	**0.402**	**0.903**	**0.352**

Table 1: Performance comparison of CrossBind with state-of-the-art methods on nucleic-acid-binding tasks. Structure_based and Sequence_based method are listed in the table as two main training methods, and protein large language model (LLM) is also used in some work to improve the performance. [a] Using the predicted structure by AlphaFold2.

works (Xia et al. 2021; Yuan et al. 2022). The previous methods contained both protein structure-based and sequence-based methods and are shown in Table 1. Similar to GraphSite, we used the large language model as the protein sequence encoder. As shown in Table 1, our method outperformed the second-best sequence-based method, GraphSite, CrossBind improved F1-score, MCC, and AUPR by 10.8%, 11.9%, and 15.4%, respectively. CrossBind demonstrated superior predictive accuracy, outperforming the second-best structure-based method, GraphBind by 17.5%, 20.0%, and 26.3% in F1-score, MCC, and AUPR, respectively. To demonstrate the generalization and stability of our method, we also compared CrossBind with other methods on two more challenging test sets: DNA_181 and RNA_117. The performance ranks of these methods are generally consistent with those in Test_129, and CrossBind still outperforms all other methods significantly. When using GraphSite as the baseline on the DNA_181 dataset, CrossBind improved F1-score, MCC, and AUPR by 13.0%, 12.8%, and 14.9%, respectively. When using GraphBind as the baseline on the RNA_117 dataset, CrossBind improved F1-score and MCC by 17.3% and 24.8%, respectively.

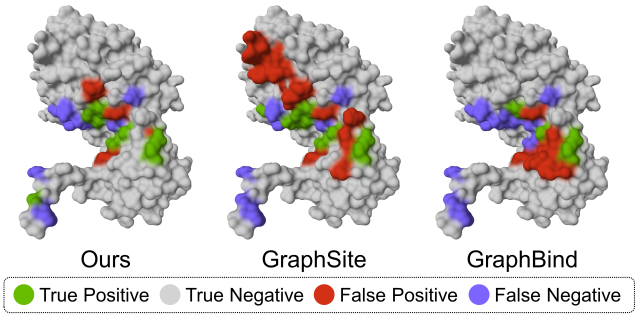

Figure 3: Classification results for protein chain 6YMW_B using CrossBind(Ours), GraphSite and GraphBind.

Case studies. To analyze our results in more detail, we conducted a visualization of three cases predicted by GraphSite, GraphBind, and our proposed model on DNA_129. We selected the example protein 6YMW_B, which was also discussed in (Yuan et al. 2022). This protein contains 668 residues, out of which 13 are binding residues. As demonstrated in Figure 3, our proposed model CrossBind predicted 10 true binding residues and 4 false positive residues, achieving a Rec of 0.64, a Pre of 0.53, and an F1 score of 0.58. In contrast, GraphSite predicted 8 true binding residues and 13 false positive residues, achieving a Rec of 0.60, a Pre of 0.47, and an F1 score of 0.52. On the other hand, GraphBind predicted only 6 true binding residues and 5 false positive residues, achieving a Rec of 0.40, a Pre of 0.26, and an F1 score of 0.32. Although GraphSite also predicted enough true binding residues, it had a higher false positive rate, whereas CrossBind exhibited higher accuracy.

These results are reasonable because: (i) CrossBind considers atom-level segmentation, capturing local geometric

Module	F1	MCC	AUC	AUPR
ALS[a]	0.429	0.501	0.902	0.477
LLM[b]	0.511	0.524	0.928	0.524
ALS[a] + LLM[b]	0.574	0.560	0.943	0.575
CrossBind	**0.602**	**0.581**	**0.953**	**0.628**
- AWA[c]	0.582	0.564	0.945	0.581
- SSL[d]	0.588	0.568	0.951	0.606
- RPF[e]	0.595	0.575	0.951	0.620
- COE[f]	0.599	0.577	0.952	0.624
GraphBind	0.512	0.484	0.912	0.497
GraphSite	0.543	0.519	0.934	0.544

Table 2: Ablation study on DNA_129 testset.

information within amino acids, while traditional methods only consider atom features and spatial information between amino acids. (ii) The LLM model characterizes the semantic information between amino acids better than models that use only local biological features. (iii) CrossBind integrates sequence and structure representations on residue prediction task and filters interacting surface residues based on RPF, which is more in line with biological experimental logic. Our results are also a significant improvement over other methods on the more challenging DNA_181 and RNA_117 tests, demonstrating the great advantages of CrossBind.

Ablation Study

Modules incremental ablation. We present several incremental ablation studies on the DNA_129 dataset to evaluate the effectiveness of our proposed modules. As shown in the table 2, using only the protein language model did not yield good results on the task. Similarly, using only the atom point cloud segmentation encoder led to even worse results. However, when combining the sequence and structure features, the results outperformed the state-of-the-art method by 5.7% in AUPRC and 7.8% in MCC. This demonstrated the effectiveness of our proposed cross-modal module, as the pre-trained language model contained no spatial information. The AWA module, which incorporates local geometric knowledge from atoms to amino acids, also contributed to the performance improvement. Removing this module led to significantly lower results. The SSL module improved the segmentation encoder's robustness on imbalanced data, and removing it resulted in a 2.2% reduction in AUPR. Finally, the RPF module optimized the results of cross-modal predictions based on a priori knowledge of biology. Introducing conventional co-evolutionary features led to an insignificant performance drop, as the pre-trained language model already contained most of the feature information. Overall, these ablation studies confirmed the effectiveness of each proposed module in our CrossBind method.

Large language model (LLM) ablation. Based on Table 3, it can be observed that using a smaller LLM model with fewer transformer layers leads to lower performance on the task. Specifically, using only 12-layers of transformer leads to a 2.4% decrease in AUC compared to 33-layers. Additionly, fine-tuning the LLM model on the task also greatly improves the performance. For example, fine-tuning the 33-layer LLM model leads to a 3.1% improvement in AUPRC compared to using the pre-trained LLM model only. These results suggest that a larger and fine-tuned LLM model can better capture the language information in protein sequences, which is beneficial for the residues prediction.

Layers	Fine-Tune	EMB_D	MCC	AUC	AUPR
6	N	320	0.432	0.904	0.474
6	Y	320	0.439	0.915	0.487
12	N	480	0.457	0.920	0.509
12	Y	480	0.492	0.929	0.524
30	N	640	0.525	0.931	0.547
30	Y	640	0.548	0.944	0.589
33	N	1280	0.559	0.947	0.597
33	**Y**	**1280**	**0.581**	**0.953**	**0.628**

Table 3: Ablations for pre-trained Large Language Model (ESM2) on DNA_129 test. EMB_D is the output dimension of the LLM.

Attention Method	AUC	AUPR
Atom feature (Mean)	0.931	0.568
Atom feature (Stack)	0.945	0.583
Self-attention	0.941	0.577
Atom-wise attention	**0.953**	**0.628**

Table 4: Ablations for Atom-wise attention module.

Atom-wise attention. We adopt a simple MLP-based attention to incorporate the fine-grained local geometric knowledge between atoms and amino acids. As shown in the table 4, only average or stacking the atoms feature leads to a significant reduction on performance, which lost the local geometric knowledge. We replace the atom-wise attention with a single self-attention layer on cross-modal module, the result is less than stacking all atoms feature, probably because self-attention layer over-smooth the local geometric knowledge between all atoms.

Conclusion

In this study, we propose CrossBind, a cross-modal framework for identifying protein nucleic-acid-binding residues by using both protein structure and sequence information. In addition, we introduce an atom-wise attention module that captures the positional relationship between atoms and residues for extracting fine-grained local geometric representations to encode the 3D protein structures. Our method achieves state-of-the-art results on three benchmark datasets and outperforms other single-mode methods based on a comprehensive evaluation.

In future work, we plan to further improve our structural encoder to extend various downstream tasks. This study provides evidence that cross-modal strategies are effective in protein-related tasks. Additionally, similar to the large language models based on protein sequence, a general 3D structure pre-training model also warrants further research. Besides, another tread is to solve the condition without native protein structure or reliable folding protein structures.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (NO.2022ZD0160101), by Shenzhen-Hong Kong Joint Funding No.SGDX20211123112401002, and by Shenzhen General Program No. JCYJ20220530143600001. This work is supported by funds from the Focus Project of AI for Science of Comprehensive Prosperity Plan for Disciplines of Fudan University, Netmind.AI, and Protagolabs Inc (to S.S.).

References

Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557): 871–876.

Bheemireddy, S.; Sandhya, S.; Srinivasan, N.; and Sowdhamini, R. 2022. Computational tools to study RNA-protein complexes. *Frontiers in Molecular Biosciences*, 9.

Chayen, N. E.; and Saridakis, E. 2008. Protein crystallization: from purified protein to diffraction-quality crystal. *Nature methods*, 5(2): 147–153.

Chen, M.; and Ludtke, S. J. 2021. Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. *Nature methods*, 18(8): 930–936.

Chen, T.; Zhou, D.; Wang, J.; Wang, S.; He, Q.; Hu, C.; Ding, E.; Guan, Y.; and He, X. 2022. Part-aware Prototypical Graph Network for One-shot Skeleton-based Action Recognition. *arXiv preprint arXiv:2208.09150*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, Y. C.; and Lim, C. 2008. Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic acids research*, 36(5): e29.

Chou, K.-C. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3): 246–255.

Dai, B.; and Bailey-Kellogg, C. 2021. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, 37(17): 2580–2588.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16321–16330.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Grønning, A. G. B.; Doktor, T. K.; Larsen, S. J.; Petersen, U. S. S.; Holm, L. L.; Bruun, G. H.; Hansen, M. B.; Hartung, A.-M.; Baumbach, J.; and Andresen, B. S. 2020. DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. *Nucleic acids research*, 48(13): 7099–7118.

Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; and Bennamoun, M. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12): 4338–4364.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kim, O. T.; Yura, K.; and Go, N. 2006. Amino acid residue doublet propensity in the protein–RNA interface and its application to RNA interface prediction. *Nucleic acids research*, 34(22): 6450–6460.

Lam, J. H.; Li, Y.; Zhu, L.; Umarov, R.; Jiang, H.; Héliou, A.; Sheong, F. K.; Liu, T.; Long, Y.; Li, Y.; et al. 2019. A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nature communications*, 10(1): 4941.

Li, S.; Yamashita, K.; Amada, K. M.; and Standley, D. M. 2014. Quantifying sequence and structural features of protein–RNA interactions. *Nucleic acids research*, 42(15): 10086–10098.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2022. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022–07.

Liu, H.; HaoChen, J. Z.; Gaidon, A.; and Ma, T. 2021. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*.

Liu, R.; and Hu, J. 2013. DNABind: A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning-and template-based approaches. *PROTEINS: structure, Function, and Bioinformatics*, 81(11): 1885–1899.

McGinnis, S.; and Madden, T. L. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, 32(suppl_2): W20–W25.

Pan, X.; and Shen, H.-B. 2018. Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, 34(20): 3427–3436.

Panda, R.; Chen, C.-F. R.; Fan, Q.; Sun, X.; Saenko, K.; Oliva, A.; and Feris, R. 2021. Adamml: Adaptive multimodal learning for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7576–7585.

Schmohl, S.; and Sörgel, U. 2019. Submanifold sparse convolutional networks for semantic segmentation of large-scale ALS point clouds. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4: 77–84.

Su, H.; Liu, M.; Sun, S.; Peng, Z.; and Yang, J. 2019. Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics*, 35(6): 930–936.

Sverrisson, F.; Feydy, J.; Correia, B. E.; and Bronstein, M. M. 2021. Fast end-to-end learning on protein surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15272–15281.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

Tugarinov, V.; Hwang, P. M.; and Kay, L. E. 2004. Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. *Annual review of biochemistry*, 73(1): 107–146.

Verkuil, R.; Kabeli, O.; Du, Y.; Wicky, B. I.; Milles, L. F.; Dauparas, J.; Baker, D.; Ovchinnikov, S.; Sercu, T.; and Rives, A. 2022. Language models generalize beyond natural proteins. *bioRxiv*, 2022–12.

Wang, Y.; Wu, S.; Duan, Y.; and Huang, Y. 2022. A point cloud-based deep learning strategy for protein–ligand binding affinity prediction. *Briefings in Bioinformatics*, 23(1): bbab474.

Wei, J.; Chen, S.; Zong, L.; Gao, X.; and Li, Y. 2022. Protein–RNA interaction prediction with deep learning: structure matters. *Briefings in bioinformatics*, 23(1): bbab540.

Wu, Q.; Peng, Z.; Zhang, Y.; and Yang, J. 2018. COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic acids research*, 46(W1): W438–W442.

Xia, Y.; Xia, C.-Q.; Pan, X.; and Shen, H.-B. 2021. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9): e51–e51.

Xue, L. C.; Dobbs, D.; and Honavar, V. 2011. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*, 12(1): 1–24.

Yan, J.; Friedrich, S.; and Kurgan, L. 2016. A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues. *Briefings in bioinformatics*, 17(1): 88–105.

Yan, X.; Lu, Y.; Li, Z.; Wei, Q.; Gao, X.; Wang, S.; Wu, S.; and Cui, S. 2022. PointSite: a point cloud segmentation tool for identification of protein ligand binding atoms. *Journal of Chemical Information and Modeling*, 62(11): 2835–2845.

Yang, J.; Roy, A.; and Zhang, Y. 2012. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1): D1096–D1103.

Yu, D.-J.; Hu, J.; Yang, J.; Shen, H.-B.; Tang, J.; and Yang, J.-Y. 2013. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(4): 994–1008.

Yuan, Q.; Chen, S.; Rao, J.; Zheng, S.; Zhao, H.; and Yang, Y. 2022. AlphaFold2-aware protein–DNA binding site prediction using graph transformer. *Briefings in Bioinformatics*, 23(2): bbab564.

Zhang, J.; Chen, Q.; and Liu, B. 2021. NCBRPred: predicting nucleic acid binding residues in proteins based on multi-label learning. *Briefings in bioinformatics*, 22(5): bbaa397.

Zhang, Y.; and Skolnick, J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, 33(7): 2302–2309.

Zhou, T.; Ruan, S.; and Canu, S. 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3: 100004.

Zhu, Y.-H.; Hu, J.; Song, X.-N.; and Yu, D.-J. 2019. DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. *Journal of chemical information and modeling*, 59(6): 3057–3071.