AlphaFold-aware prediction of protein-DNA binding sites using graph transformer

Qianmu Yuan[1], Sheng Chen[1], Jiahua Rao[1], Shuangjia Zheng[1], Huiying Zhao[2], Yuedong Yang[1,3,*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China

[2]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China

[3]Key Laboratory of Machine Intelligence and Advanced Computing of MOE, Sun Yat-sen University, Guangzhou 510000, China

[*]To whom correspondence should be addressed.

Abstract

Motivation: Protein-DNA interactions play crucial roles in the biological systems, and identifying protein-DNA binding sites is the first step for mechanistic understanding of various biological activities (such as transcription and repair) and designing novel drugs. How to accurately identify DNA-binding residues from only protein sequence remains a challenging task. Currently, most existing sequence-based methods only consider contextual features of the sequential neighbors, which are limited to capture spatial information.

Results: Based on the recent breakthrough in protein structure prediction by AlphaFold2, we propose an accurate predictor, GraphSite, for identifying DNA-binding residues based on the structural models predicted by AlphaFold2. Here, we convert the binding site prediction problem into a graph node classification task and employ a transformer-based variant model to take the protein structural information into account. By leveraging predicted protein structures and graph transformer, GraphSite substantially improves over the latest sequence-based and structure-based methods. The algorithm was further confirmed on the independent test set of 196 proteins, where GraphSite surpasses the state-of-the-art structure-based method by 12.3% in AUPR and 9.3% in MCC, respectively.

Contact: yangyd25@mail.sysu.edu.cn

1. Introduction

Protein-DNA interactions play crucial roles in many biological processes such as transcription, repair and signal transduction (Charoensawan, et al., 2010; Zhao, et al., 2010). Identification of the residues involved in protein-DNA interactions helps analyze genetic materials, understand protein functions and design novel drugs (Schmidtke and Barril, 2010). However, conventional experimental methods for DNA-binding site detection such as X-ray crystallography (Orengo, et al., 1997) and fast ChIP (Mandel-Gutfreund and Margalit, 1998) are costly and time-consuming. Therefore, it is necessary to develop complementary computational methods capable of making reliable and accurate DNA-binding site prediction.

Computational methods for DNA-binding site prediction can be classified according to the information used for prediction: sequence-based and structure-based methods. Sequence-based methods such as TargetS (Yu, et al., 2013), TargetDNA (Hu, et al., 2016), DNAPred (Zhu, et al., 2019) and DNAgenie (Zhang, et al., 2021) learn local patterns of DNA-binding characteristics using sequence-derived features. For example, TargetDNA adopts evolutionary conservative information and predicted solvent accessibility extracted from protein sequences and employs SVMs to learn local patterns from sequence contexts using sliding-window strategy. Although these

sequence-based approaches require readily available protein sequences only, their predictive accuracies are usually quite limited.

By comparison, structure-based approaches inferring binding sites from known structures are often more accurate, which can be generally categorized into template-based methods, machine learning based methods and hybrid methods. Template-based methods identify DNA-binding sites using the sequence and structure information of templates, which are selected by alignment or comparison algorithms (Jones, et al., 2003; Tsuchiya, et al., 2004). Nevertheless, for the proteins that have no high-quality template, the performance of these methods will be seriously restricted. With features derived from protein structures, recent structure-based machine learning methods represent protein structures as voxels in three-dimensional Euclidean space or nodes in connected graph. For example, DeepSite (Jiménez, et al., 2017) maps protein atoms into 3D voxels and employs 3D convolutional neural networks (3DCNNs) to extract features from neighborhood of the target residue. Alternately, GraphBind (Xia, et al., 2021) encodes protein structure as graph and adopts graph neural networks (GNNs) to learn the local tertiary patterns for binding residue prediction. Hybrid methods integrate template-based methods and machine learning based methods simultaneously, such as DNABind (Liu and Hu, 2013), COACH-D (Wu, et al., 2018) and NucBind (Su, et al., 2019).

Recently, a novel machine learning approach named AlphaFold2 (Jumper, et al., 2021) has been developed, which can predict protein structures with atomic accuracy even when no similar structure template is known. Incorporating physical and biological knowledge about protein structure, information of multi-sequence alignment, and the sophisticated design of the deep learning algorithm, AlphaFold2 demonstrates accuracy competitive with experiment in a majority of cases in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14). Based on this breakthrough of protein structure prediction, accurate protein-DNA binding site prediction solely from protein sequence via predicted structures becomes possible. On the other hand, the transformer (Vaswani, et al., 2017) is well acknowledged as the most powerful neural network in modelling sequential data, such as natural language (Devlin, et al., 2018) and protein sequence (Chen, et al., 2020). In the last few years, transformer variants have also been shown great performance in computer vision (Dosovitskiy, et al., 2020) and even in graph representation learning (Chen, et al., 2021; Ying, et al., 2021).

In this study, we have developed a novel method GraphSite, which applies graph transformer network and predicted protein structures from AlphaFold2 for sequence-based prediction of DNA-binding residues. Specifically, we consider a protein as an undirected graph, where we integrate multi-sequence alignment and structural information to construct node features and calculate pairwise amino acid distances to mask out the spatially remote amino acids when calculating attention scores in the transformer. With the spatial information and structure-aware transformer, GraphSite was found to outperform other sequence-based and structure-based methods through various performance evaluations. To the best of our knowledge, this is the first work that utilizes graph transformer and AlphaFold2-predicted structures for DNA-binding site prediction, which can be easily extended to sequence-based prediction of other functional sites.

2. Methods

Here, the DNA-binding site prediction task is treated as a graph node classification problem, where our goal is to train a graph transformer network that takes inputs of node features and amino acid distance maps and outputs the probability of being a DNA-binding site for each node.

2.1 Datasets

We adopted two publicly available benchmark datasets from the previous study (Xia, et al., 2021) to train and test our method: Train_573 and Test_129, which are named by the numbers of proteins in the datasets. These two datasets were collected from the BioLiP database (Yang, et al., 2012), which is a collection of biologically relevant ligand-

protein interactions solved structurally in complexes. Concretely, Train_573 contains proteins released before 6 January 2016 while Test_129 from 6 January 2016 to 5 December 2018. In these datasets, a DNA-binding residue was defined if the smallest atomic distance between the target residue and the DNA molecule is less than 0.5 Å plus the sum of the Van der Waal's radius of the two nearest atoms. To deal with the data imbalance problem, the authors (Xia, et al., 2021) applied data augmentation on Train_573, which transferred binding annotations from similar protein chains to increase the number of binding residues. Finally, CD-HIT (Fu, et al., 2012) was used to ensure no redundant protein with more than 30% sequence identity within the training set and between the training and testing set. To further demonstrate the generalization of our model, we built another independent test set (Test_196) based on newly released DNA-binding proteins in BioLiP (12 December 2018 to 3 February 2021). We have removed redundant proteins sharing sequence identity > 30% over 30% overlap with any sequence in the above two datasets using CD-HIT. Details of the number of binding and non-binding residues of these datasets are given in **Table 1**.

Table 1. Statistics of the three benchmark datasets used in this study. The columns give, in order, the dataset name, the number of binding, and the number of non-binding residues in each dataset, and the percentage of the binding residues out of total.

Dataset	Binding residues	Non-binding residues	% of binding residues
Train_573	14479	145404	9.06
Test_129	2240	35275	5.97
Test_196	3434	75483	4.35

2.2 Protein representation

In our framework, a protein consisting of n amino acid residues is represented by an undirected graph $G = (V, E)$, with V denoting the amino acids (nodes) and E denoting the distances between pairwise amino acids (edges). Thus, the protein graph G can be represented by a node feature matrix $\boldsymbol{X}$ and a distance matrix $\boldsymbol{D}$.

2.2.1 Predicted protein structures and distance maps

Following the tutorial at https://github.com/deepmind/alphafold, we downloaded the model parameters and genetic databases including UniRef90 (Suzek, et al., 2007), MGnify (Mitchell, et al., 2020), BFD (Steinegger, et al., 2019), Uniclust30 (Mirdita, et al., 2017), PDB70 (Steinegger, et al., 2019) and PDB (Berman, et al., 2000) to implement AlphaFold2 in Tianhe-2 supercomputer. We set the "--max_template_date" parameter to 2020-05-14 just as the AlphaFold2 model in CASP14 and predicted the protein structures in Train_573 and Test_129. However, to further avoid the possibility that AlphaFold2 might use the native known structures in PDB as templates, we set "--max_template_date" to 2018-12-5 when predicting proteins in the independent Test_196. This made Test_196 more challenging and closer to the situation where only protein sequences instead of similar templates are available. According to the predicted protein structural models from AlphaFold2, we acquired the coordinate of the Cα atom of each amino acid residue, and then calculated the Euclidean distances between all residue pairs, which formed a distance map $\boldsymbol{D} \in \mathbb{R}^{n \times n}$.

2.2.2 Node features

We employed two groups of amino acid features to train our model: multi-sequence alignment (MSA) information and structural properties (DSSP), which were concatenated and formed the final node feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times 398}$ with n denoting the length of a protein sequence.

MSA information. Co-evolving amino acids may structurally contact, and evolutionarily conserved residues may contain motifs related to important protein properties such as DNA binding propensity. Here, we employed the "single representation" feature output by AlphaFold2 along with the predicted structures. This single representation

is derived by a linear projection of the first row of the MSA representation, which is a highly processed MSA feature matrix through 48 Evoformer blocks in AlphaFold2. The single representation feature is called MSA representation hereinafter. In addition, we also explored the position-specific scoring matrix (PSSM) and hidden Markov models (HMM) profile. Concretely, PSSM was generated by running PSI-BLAST (Altschul, et al., 1997) to search the query sequence against UniRef90 database with three iterations and an E-value of 0.001. The HMM profile was produced by running HHblits (Remmert, et al., 2012) to align the query sequence against UniClust30 database with default parameters. Each amino acid was encoded into a 384-dimensional vector in MSA representation and 20-dimensional vector in PSSM or HMM, and the values were normalized to scores between 0 to 1 using Equation (1), where v is the original feature value, and *Min* and *Max* are the smallest and biggest values of this feature type observed in the training set.

$$v_{norm} = \frac{v - Min}{Max - Min} \tag{1}$$

Structural properties. Three types of structural properties were extracted by the program DSSP (Kabsch and Sander, 1983) using predicted structures: 1) 9-dimensional one-hot secondary structure profile where the first 8 dimensions represent 8 categories of secondary structure, and the last dimension represents unknown secondary structure. 2) Peptide backbone torsion angles PHI and PSI, which were converted to a 4-dimensional feature vector using sine and cosine transformations. 3) Solvent accessible surface area (ASA), which was normalized to relative solvent accessibility (RSA) by the maximal possible ASA of the corresponding amino acid type. This 14-dimensional structural feature group is named DSSP in this study.

2.3 The architecture of GraphSite

Figure 1 shows the overall architecture of the proposed framework GraphSite**,** where the protein sequence is input to AlphaFold2 to produce the MSA representation and the predicted protein structure, from which the distance map and DSSP are extracted. Finally, the MSA representation and DSSP are concatenated to form the node feature vector, which is then input to the graph transformer masked by the distance map to learn the DNA-binding site patterns.

2.3.1 Graph transformer

The traditional transformer encoder layer consists of a multi-head self-attention module and a position-wise feed-forward network. Let $\boldsymbol{H} = [\boldsymbol{h}_1^T, \dots, \boldsymbol{h}_n^T]^T \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{d}}$ denote the input of the self-attention module where d is the hidden dimension and $\boldsymbol{h}_i \in \mathbb{R}^{1\times d}$ is the hidden representation of the ith amino acid. Note that the initial $\boldsymbol{H}^{(0)}$ does not have to be the feature matrix $\boldsymbol{X}$, which means that a fully connected layer can be applied on $\boldsymbol{X}$ to obtain the initial hidden representation $\boldsymbol{H}^{(0)}$. The input $\boldsymbol{H}^{(l)}$ is projected by three matrices $\boldsymbol{W}_Q \in \mathbb{R}^{d\times d_K}$, $\boldsymbol{W}_K \in \mathbb{R}^{d\times d_K}$ and $\boldsymbol{W}_V \in \mathbb{R}^{d\times d_V}$ to the corresponding queries, keys and values representation $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$:

$$\boldsymbol{Q} = \boldsymbol{H}^{(l)}\boldsymbol{W}_Q, \quad \boldsymbol{K} = \boldsymbol{H}^{(l)}\boldsymbol{W}_K, \quad \boldsymbol{V} = \boldsymbol{H}^{(l)}\boldsymbol{W}_V \tag{2}$$

The self-attention is then calculated as:

$$\boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_K}} \tag{3}$$

$$\boldsymbol{H}^{(l+1)} = Attn\left(\boldsymbol{H}^{(l)}\right) = \text{softmax}(\boldsymbol{A})\boldsymbol{V} \tag{4}$$

where $\boldsymbol{A}$ is a matrix capturing the similarity between queries and keys. In order to take the protein structure information into account, we adopt *k*-nearest mask according to the distance matrix $\boldsymbol{D}$ to mask out the spatially remote amino acids, which means that for node i, only spatially adjacent nodes $j \in \text{Neighbor}(i, k)$ are used to calculate the attention scores in the transformer. In order to jointly attend to information from different representation subspaces at different positions, we use multi-head attention to linearly project the queries, keys and values *h* times, perform the attention function in parallel and finally concatenate them. In this study, $d_K = d_V = d\ /\ h$.

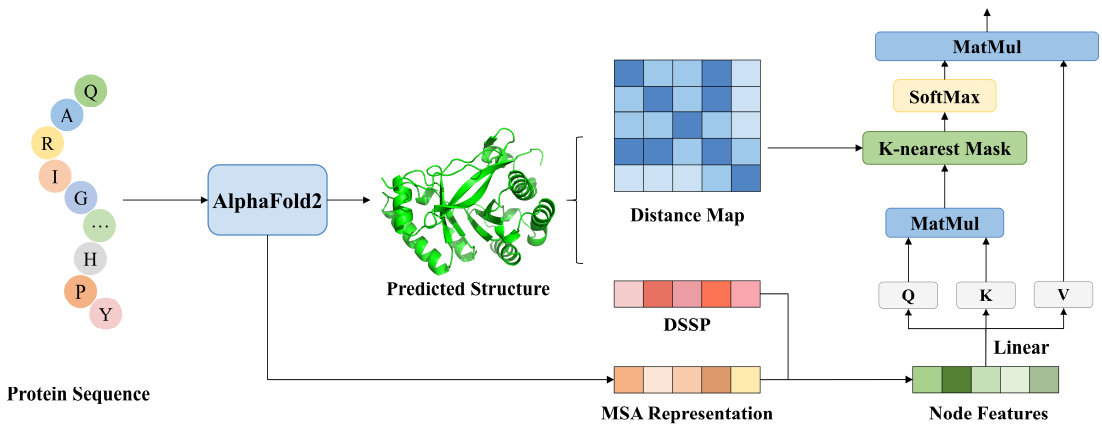

Figure 1. The network architecture of the proposed GraphSite framework. The protein sequence is input to AlphaFold2 to produce the MSA representation and the predicted protein structure, from which the distance map and DSSP are extracted. Then, the MSA representation and DSSP are concatenated to form the node feature vector, which is then input to the graph transformer model with *k*-nearest mask by the distance map to learn the DNA-binding site patterns.

2.3.2 Multilayer Perceptron

The output of the last graph transformer layer is input to the multilayer perceptron (MLP) to predict the DNA-binding probabilities of all *n* amino acid residues:

$$\boldsymbol{Y}' = Sigmoid\left(\boldsymbol{H}^{(L)}\boldsymbol{W} + b\right) \tag{5}$$

where $\boldsymbol{H}^{(L)} \in \mathbb{R}^{n\times d}$ is the output of the Lth graph transformer layer; $\boldsymbol{W} \in \mathbb{R}^{d\times 1}$ is the weight matrix; $b \in \mathbb{R}$ is the bias term, and $\boldsymbol{Y}' \in \mathbb{R}^{n\times 1}$ is the predictions of *n* amino acid residues. The sigmoid function normalizes the output of the network into binding probabilities ranging from 0 to 1.

2.4 Implementation details

We performed 5-fold cross-validation on the training data, where the data were split into five folds randomly. Each time, a model was trained on four folds and evaluated on the remaining one fold. This process was repeated five times and the performances on the five folds were averaged as the overall validation performance, which was used to choose the best feature combination and optimize all hyperparameters through grid search. In the testing phase, all five trained models in the cross-validation were assembled and used to make predictions, which were averaged as the final predictions of our method.

Specifically, we utilized a 2-layer graph transformer module with 64 hidden units and the following set of hyperparameters: $h = 2$, $k = 30$ and batch size of 16. We employed the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\varepsilon = 10^{-5}$, weight decay of 10^{-5} and learning rate of 3×10^{-4} for model optimization on the binary cross entropy loss. The dropout rate was set to 0.2 to avoid overfitting. We implemented the proposed model with Pytorch 1.7.1 (Paszke, et al., 2019). The training process lasted at most 100 epochs and took approximately 40 minutes on an Nvidia GeForce RTX 3090 GPU.

2.5 Evaluation metrics

Similar to the previous studies, we used precision, recall, F1-score (F1), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPR) to measure the predictive performance:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{9}$$

where true positives (TP) and true negatives (TN) denote the number of binding and non-binding sites identified correctly, and false positives (FP) and false negatives (FN) denote the number of incorrectly predicted binding and non-binding sites, respectively. AUC and AUPR are independent of thresholds, thus revealing the overall performance of a model. The other metrics were calculated using a threshold to convert predicted binding probabilities to binary predictions, which was determined by maximizing F1-score for each model. We used AUPR for the above hyperparameter selection as it is more sensitive and it emphasizes more on the minority class in imbalanced two-class classification tasks (Saito and Rehmsmeier, 2015).

3. Results

3.1 Performance on the 5-fold cross-validation and independent test

We evaluated the performance of GraphSite by AUC and AUPR using 5-fold cross-validation (CV) and the independent Test_129 and Test_196. The final GraphSite model obtains AUC of 0.910, 0.928 and 0.905; as well as AUPR of 0.571, 0.522 and 0.338 on the 5-fold CV and the two independent tests, respectively. The consistent performances on the CV and Test_129 indicate the robustness of our model. Note that the AUPR drops about 0.2 in Test_196, which might be ascribed to the lower positive sample ratio in Test_196, or the lower predicted qualities of the protein structures in this independent dataset (discussed in **Section 3.4**).

In order to demonstrate the advantages of protein geometric knowledge and the graph transformer model, we compared GraphSite with a baseline method BiLSTM, which contains a two-layer bidirectional long short-term memory network with 256 hidden units and an MLP module. This model uses the same residue features as GraphSite (MSA representation and DSSP) and serves as a geometric-agnostic baseline to evaluate the impact of the spatial information for binding residue prediction. As shown in **Figure 2**, GraphSite yields higher F1, MCC, AUC and AUPR, which are 0.072(0.041), 0.077(0.042), 0.027(0.036) and 0.082(0.058) higher than those of BiLSTM on Test_129(Test_196), respectively.

The performance improvement of GraphSite over BiLSTM is likely due to its better capability in capturing long-range contact information. To illustrate this, we compared the performance of GraphSite and BiLSTM on amino acids with different number of non-local contacts, defined as the contacts from the residues that are more than 20 residues away in sequence positions, but ≤ 12 Å in terms of their atomic distances between Cα atoms. **Figure 3** shows that GraphSite consistently surpasses BiLSTM on Test_129 and more importantly, the performance gap between them enlarges as the non-local contact number of the amino acids increases. Specifically, the performance of GraphSite surpasses BiLSTM by 16.6% in MCC on the amino acids with 0 to 9 non-local contacts, and the gap widens to as much as 237.8% on the amino acids with ≥ 40 non-local contacts. These comparisons highlight the importance of the spatial information, and the effectiveness of GraphSite in harnessing protein structural knowledge especially non-local contacts for DNA-binding residue recognition.

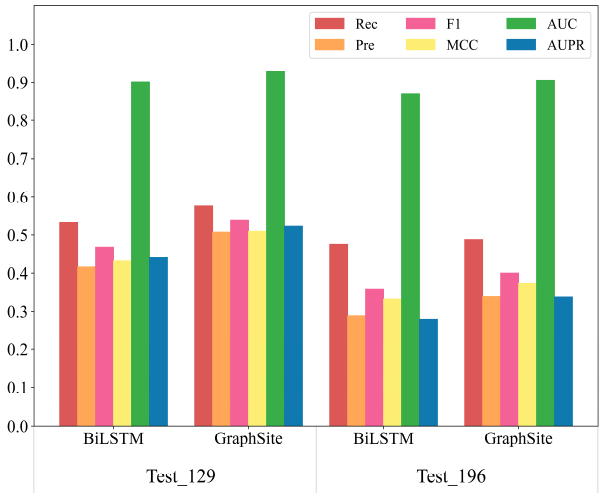

Figure 2. Performance comparison between BiLSTM and GraphSite on Test_129 and Test_196.

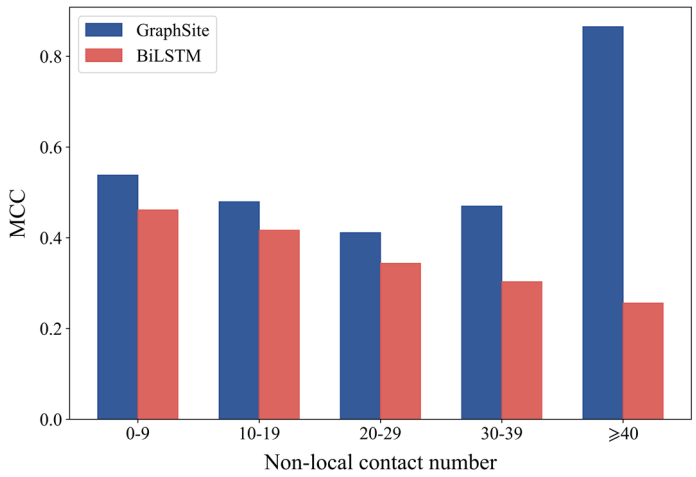

Figure 3. The MCC of GraphSite and BiLSTM on amino acids with different number of non-local contacts.

3.2 Feature importance

To demonstrate the relative importance of each feature group we adopt, we conducted feature ablation experiments by using only DSSP or MSA representation. As shown in **Table 2**, when using DSSP solely as node features, the model yields poor performance with AUPR of 0.126 and 0.078 in Test_129 and Test_196, indicating that structural properties of amino acids such as secondary structure and relative solvent accessibility are insufficient to capture the complicated patterns of DNA-binding sites. Surprisingly, using only the highly processed MSA representation by AlphaFold2 already gives satisfactory performance with AUPR of 0.520 and 0.327 in Test_129 and Test_196. Finally, according to the results on the CV, we combined MSA representation and DSSP as the final input features of GraphSite, which perform slightly better than using MSA representation alone with 0.005 and 0.011 boost of AUPR in the two test sets, respectively. These ablation experiments suggest that the two feature groups are nonredundant.

In addition, we also explored the effects of the widely used evolutionary features PSSM and HMM. When replacing the MSA representation with PSSM and HMM, the AUPR drops 0.060 and 0.075 in Test_129 and Test_196, suggesting that the processed MSA representation of AlphaFold2 is more powerful than PSSM and HMM in DNA-binding site detection. Although both the MSA representation, PSSM and HMM contain evolutionary information, combining GraphSite with PSSM and HMM still yields performance improvements of about 0.01 in AUPR on the two independent tests. For simplicity, we didn't include PSSM and HMM in our final GraphSite model in this study.

Table 2. The AUC and AUPR of the 5-fold cross-validation (CV), Test_129 and Test_196 using different feature combinations. MSA Repr. denotes the MSA representation produced by AlphaFold2, and Evo denotes the evolutionary feature groups PSSM and HMM.

Feature	CV AUC	CV AUPR	Test_129 AUC	Test_129 AUPR	Test_196 AUC	Test_196 AUPR
DSSP	0.691	0.173	0.721	0.126	0.678	0.078
MSA Repr.	0.906	0.557	0.925	0.520	0.900	0.327
GraphSite	0.910	0.571	0.928	0.522	**0.905**	0.338
DSSP+Evo	0.891	0.499	0.918	0.462	0.872	0.263
GraphSite+Evo	**0.914**	**0.578**	**0.932**	**0.532**	0.904	**0.349**

3.3 Comparison with state-of-the-art methods

We compared GraphSite with four sequence-based (TargetS, TargetDNA, SVMnuc and DNAPred) and four structure-based (COACH-D, NucBind, DNABind and GraphBind) predictors on Test_129. As shown in **Table 3**, the performances of the structure-based methods are generally better than those of the sequence-based methods, while GraphSite outperforms all sequence-based and structure-based methods. Note that GraphSite is a sequence-based predictor input with protein sequence only, and GraphSite surpasses the best sequence-based method DNAPred by 44.5% in F1 and 53.3% in MCC, respectively. Even when compared with the latest structure-based method GraphBind, our method outperforms GraphBind by 3.3% in F1 and 2.0% in MCC, respectively. This is reasonable because: 1) GraphBind uses PSSM and HMM as evolutionary features while our method employs the more informative MSA representation from AlphaFold2. 2) The graph transformer model is proven to be powerful. 3) The AlphaFold2-predicted protein structures used by GraphSite are high-quality (discussed in **Section 3.4**).

To further demonstrate the generalization and stability of our method, we compared GraphSite with the second-best method GraphBind on our newly built independent Test_196. Note that this is a more challenging dataset for our method since we set "--max_template_date" in AlphaFold2 before the release dates of all proteins in Test_196. As shown in **Table 3**, GraphSite consistently outperforms GraphBind with boosts of 0.042 on AUC and 0.027 on AUPR, even though the native protein structures were used by GraphBind. On the other hand, when using predicted

Table 3. Performance comparison of GraphSite with state-of-the-art methods on Test_129 and independent Test_196. The results of the methods marked with * were obtained from (Xia, et al., 2021).

Dataset	Method	Rec	Pre	F1	MCC	AUC	AUPR
Test_129	TargetS*	0.239	0.370	0.291	0.262	N/A	N/A
	TargetDNA*	0.417	0.280	0.335	0.291	0.825	N/A
	SVMnuc*	0.316	0.371	0.341	0.304	0.812	N/A
	DNAPred*	0.396	0.353	0.373	0.332	0.845	N/A
	COACH-D*	0.324	0.360	0.341	0.302	0.761	N/A
	NucBind*	0.323	0.373	0.346	0.309	0.797	N/A
	DNABind*	0.601	0.346	0.440	0.411	0.858	N/A
	GraphBind*	**0.676**	0.425	0.522	0.499	0.927	N/A
	GraphSite	0.576	**0.507**	**0.539**	**0.509**	**0.928**	**0.522**
Test_196	GraphBind[a]	**0.566**	0.294	0.387	0.371	0.863	0.311
	GraphBind[b]	0.491	0.293	0.367	0.343	0.878	0.301
	GraphSite	0.488	**0.339**	**0.400**	**0.375**	**0.905**	**0.338**

[a] Use native protein structures.

[b] Use predicted protein structures.

structures as input, the performance of GraphBind will further decrease (MCC drop of 0.028) and our method surpasses GraphBind by 9.3% in MCC and 12.3% in AUPR. This suggests that our method is more practical and powerful when only protein sequences are available.

3.4 Impact of the quality of predicted protein structure

Since GraphSite employs predicted protein structures for geometric deep learning, the predicted quality of AlphaFold2 should be crucial to the downstream DNA-binding site prediction. We used SPalign (Yang, et al., 2012) to calculate the average global distance test (GDT) (Zemla, 2003) and the template modeling score (TM-score) (Zhang and Skolnick, 2005) between the native structures and the predicted structures by AlphaFold2 in the three datasets. The average GDT for Train_573 and Test_129 are 86.3 and 85.0, and the average TM-score for Train_573 and Test_129 are 0.90 and 0.91, indicating that AlphaFold2 can make accurate structure predictions when similar structure templates are available. However, the predicted quality drops when we set the restriction of the max available date of templates in AlphaFold2. Concretely, the GDT and TM-score drop to 72.9 and 0.83 for the independent Test_196, which may be part of the reason that the performance of GraphSite decreases in this dataset. For further demonstration, **Figure 4** shows the positive correlation between the predicted quality (GDT) of AlphaFold2 and the per-protein AUPR on Test_196 for GraphSite (blue scatters). In addition, we sorted the proteins in Test_196 according to the structure predicted quality and divided them into 10 bins, from which the average GDT and AUPR are calculated for each bin (red line). A closer inspection shows that the top-30% proteins with the highest GDT (average GDT = 93.3) correspond to an average AUPR of 0.431 predicted by GraphSite. On the other hand, the last-30% proteins with the lowest GDT (average GDT = 48.0) correspond to an average AUPR of 0.300, which is significantly lower than that of the top-30% proteins (P-value = 7.3×10^{-3}) according to Mann–Whitney U test (Mann and Whitney, 1947). These results suggest the importance of the structure predicted quality for DNA-binding site prediction.

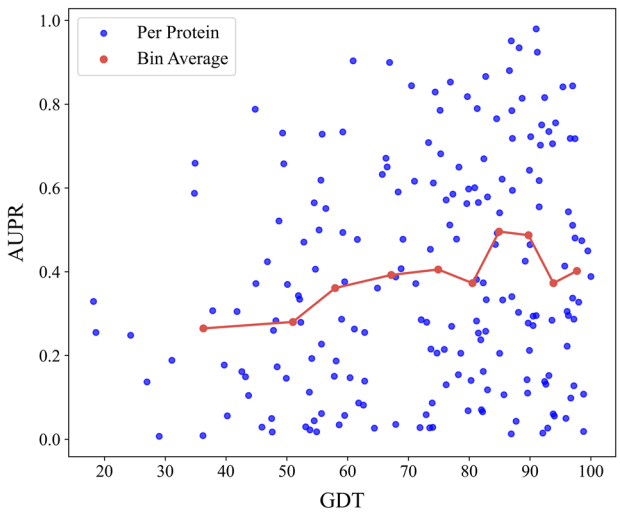

Figure 4. The correlation between the predicted quality of AlphaFold2 measured by GDT and per-protein AUPR for GraphSite on Test_196 (blue scatters). The red line denotes the average GDT and AUPR for each bin after sorting the GDT of all proteins and divided them into 10 bins

3.5 Case study

To visualize the superiority of our method, we selected human APOBEC3G (PDB ID: 6WMC, chain A) from Test_196 for illustration, since the AUPR of GraphSite for this protein is close to the overall AUPR (**Table 3**). **Figure**

5 shows the DNA-binding site prediction results of GraphSite (**A**) and the second-best method GraphBind (**B**) using the predicted structure of this protein. In this example, there are 7 DNA-binding residues over a total of 362 residues. GraphSite predicts 13 binding residues in which 4 are true positives, leading to an F1 of 0.400, MCC of 0.404 and AUPR of 0.334. By comparison, GraphBind predicts 5 binding residues in which all are false positives, leading to an F1 of 0.000, MCC of -0.017 and AUPR of 0.073. Furthermore, as shown in **Figure 5A**, false-positive binding residues (colored in red) predicted by GraphSite are mostly around the interface of protein-DNA interaction, while the false positives predicted by GraphBind are far away from the native DNA-binding sites (**Figure 5B**).

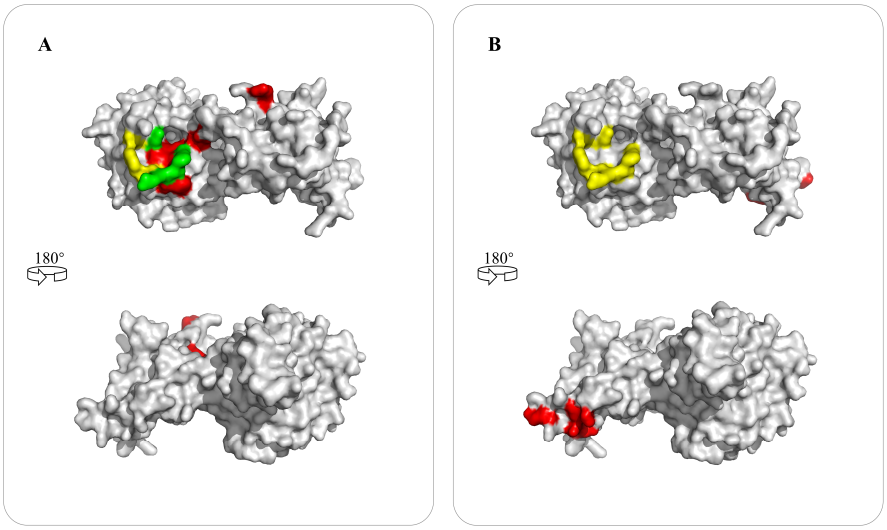

Figure 5. Visualization of one example (PDB ID: 6WMC, chain A) from Test_196 predicted by GraphSite (**A**) and the second-best method GraphBind (**B**). True positives, false positives and false negatives are colored in green, red and yellow, respectively. Top and bottom panels show the front and back views of the protein.

4. Conclusion and Discussion

In this study, we propose a geometric-aware framework called GraphSite for DNA-binding site prediction, where we predict protein structures from sequences using AlphaFold2 and employ graph transformer network to learn the amino acid representations. GraphSite shows preferable performance than existing sequence-based and structure-based methods in comprehensive evaluations.

However, there is still room for further improvement on this task. For instance, using advanced sampling or data augmentation techniques for imbalanced problems may help train a better model. Adding other informative sequence-derived features, such as physicochemical properties or relative amino acid binding propensity, may facilitate better representation of the residues and increase the model robustness to the structure predicted quality. On the other hand, our graph transformer model currently only uses the protein distance matrices as masks in the self-attention step, and modifications can be made to add the pairwise residue distances as biases to the attention scores. Our framework can also be explored to handle protein graphs with edge features constructed by residue distances, angles or the pair representations from AlphaFold2. In the future, we would further integrate multi-task learning (Sun, et al., 2020) to extend our method to various fields, including predicting protein binding sites with RNA and small ligands, or protein functional sites such as methylation and phosphorylation.

Conflict of Interest: none declared.

Reference

Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 1997;25(17):3389-3402.

Berman, H.M., *et al.* The protein data bank. *Nucleic acids research* 2000;28(1):235-242.

Charoensawan, V., Wilson, D. and Teichmann, S.A. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic acids research* 2010;38(21):7364-7377.

Chen, J., *et al.* Learning Attributed Graph Representations with Communicative Message Passing Transformer. *arXiv preprint arXiv:2107.08773* 2021.

Chen, L., *et al.* TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* 2020;36(16):4406-4414.

Devlin, J., *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* 2018.

Dosovitskiy, A., *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* 2020.

Fu, L., *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28(23):3150-3152.

Hu, J., *et al.* Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. *IEEE/ACM transactions on computational biology and bioinformatics* 2016;14(6):1389-1398.

Jiménez, J., *et al.* DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* 2017;33(19):3036-3042.

Jones, S., *et al.* Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic acids research* 2003;31(24):7189-7198.

Jumper, J., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021:1-11.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* 1983;22(12):2577-2637.

Kingma, D.P. and Ba, J. Adam: A Method for Stochastic Optimization. In, *3rd International Conference on Learning Representations (Poster)*. 2015.

Liu, R. and Hu, J. 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* 2013;81(11):1885-1899.

Mandel-Gutfreund, Y. and Margalit, H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic acids research* 1998;26(10):2306-2312.

Mann, H.B. and Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* 1947:50-60.

Mirdita, M., *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* 2017;45(D1):D170-D176.

Mitchell, A.L., *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic acids research* 2020;48(D1):D570-D578.

Orengo, C.A., *et al.* CATH–a hierarchic classification of protein domain structures. *Structure* 1997;5(8):1093-1109.

Paszke, A., *et al.* Pytorch: An imperative style, high-performance deep learning library. In, *Advances in neural information processing systems*. 2019. p. 8026-8037.

Remmert, M., *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 2012;9(2):173-175.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 2015;10(3):e0118432.

Schmidtke, P. and Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of medicinal chemistry* 2010;53(15):5858-5867.

Steinegger, M.*, et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics* 2019;20(1):1-15.

Steinegger, M., Mirdita, M. and Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods* 2019;16(7):603-606.

Su, H.*, et al.* Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics* 2019;35(6):930-936.

Sun, Z.*, et al.* To improve the predictions of binding residues with DNA, RNA, carbohydrate, and peptide via multiple-task deep neural networks. *bioRxiv* 2020.

Suzek, B.E.*, et al.* UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 2007;23(10):1282-1288.

Tsuchiya, Y., Kinoshita, K. and Nakamura, H. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *PROTEINS: structure, Function, and Bioinformatics* 2004;55(4):885-894.

Vaswani, A.*, et al.* Attention is all you need. In, *Advances in neural information processing systems*. 2017. p. 5998-6008.

Wu, Q.*, et al.* COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic acids research* 2018;46(W1):W438-W442.

Xia, Y.*, et al.* GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research* 2021;49(9):e51-e51.

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

Yang, Y.*, et al.* A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. *Proteins: Structure, Function, and Bioinformatics* 2012;80(8):2080-2088.

Ying, C.*, et al.* Do Transformers Really Perform Bad for Graph Representation? *arXiv preprint arXiv:2106.05234* 2021.

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

Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic acids research* 2003;31(13):3370-3374.

Zhang, J.*, et al.* DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences. *Briefings in Bioinformatics* 2021.

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

Zhao, H., Yang, Y. and Zhou, Y. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. *Bioinformatics* 2010;26(15):1857-1863.

Zhu, Y.-H.*, et al.* DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. *Journal of chemical information and modeling* 2019;59(6):3057-3071.