Nucleic Acids Research, 2024, **52**, e27
https://doi.org/10.1093/nar/gkae039
Advance access publication date: 28 January 2024
Methods

EquiPNAS: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks

Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Sumit Tarafder and Debswapna Bhattacharya *

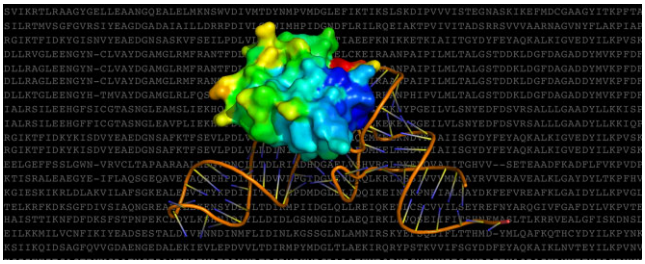
Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

*To whom correspondence should be addressed. Tel: +1 540 231 2865; Fax: +1 540 231 6075; Email: dbhattacharya@vt.edu

Abstract

Protein language models (pLMs) trained on a large corpus of protein sequences have shown unprecedented scalability and broad generalizability in a wide range of predictive modeling tasks, but their power has not yet been harnessed for predicting protein–nucleic acid binding sites, critical for characterizing the interactions between proteins and nucleic acids. Here, we present EquiPNAS, a new pLM-informed E(3) equivariant deep graph neural network framework for improved protein–nucleic acid binding site prediction. By combining the strengths of pLM and symmetry-aware deep graph learning, EquiPNAS consistently outperforms the state-of-the-art methods for both protein–DNA and protein–RNA binding site prediction on multiple datasets across a diverse set of predictive modeling scenarios ranging from using experimental input to AlphaFold2 predictions. Our ablation study reveals that the pLM embeddings used in EquiPNAS are sufficiently powerful to dramatically reduce the dependence on the availability of evolutionary information without compromising on accuracy, and that the symmetry-aware nature of the E(3) equivariant graph-based neural architecture offers remarkable robustness and performance resilience. EquiPNAS is freely available at https://github.com/Bhattacharya-Lab/EquiPNAS.

Graphical abstract



Introduction

Interaction of protein with deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) underpins a wide range of cellular and evolutionary processes such as gene expression, regulation, and signal transduction (1–4). The identification of the interaction sites between proteins and nucleic acids (i.e. binding sites) is important for determining protein functions (5) and novel drug design (6). A number of computational methods for predicting protein–DNA and protein–RNA binding sites have been developed to overcome the challenges of lengthy and expensive nature of experimental characterization of protein–nucleic acid binding sites. Such computational methods can be broadly categorized into two categories: sequence-based and structure-aware methods. Sequence-based methods such as SVMnuc (7), NCBRPred (8), DNAPred (9), DNAgenie (10), RNABindRPlus (11), ConSurf (12), TargetDNA (13), SCRIBER (3) and TargetS (14) exploit readily available and abundant protein sequence information to predict binding sites. However, these methods lack structural information, which can limit their prediction accuracy. To overcome the challenge, structure-aware methods such as COACH-D (15), NucBind (7), DNABind (16), DeepSite (17), aaRNA (18), NucleicNet (19), GraphBind (20), and GraphSite (21) integrate available structural information for binding site prediction. While structure-aware methods usually achieve higher prediction accuracy than sequence-based methods, a vast majority of structure-aware methods rely on known structural information from the Protein Data Bank (PDB) (22) that are not as abundant as sequence information, limiting their large-scale applicability.

Promisingly, the recent breakthrough of AlphaFold2 (23,24) has enabled highly accurate prediction of single-chain protein structures from sequence information, providing new opportunities for replacing the experimentally solved

Received: September 13, 2023. **Revised:** December 22, 2023. **Editorial Decision:** January 6, 2024. **Accepted:** January 11, 2024

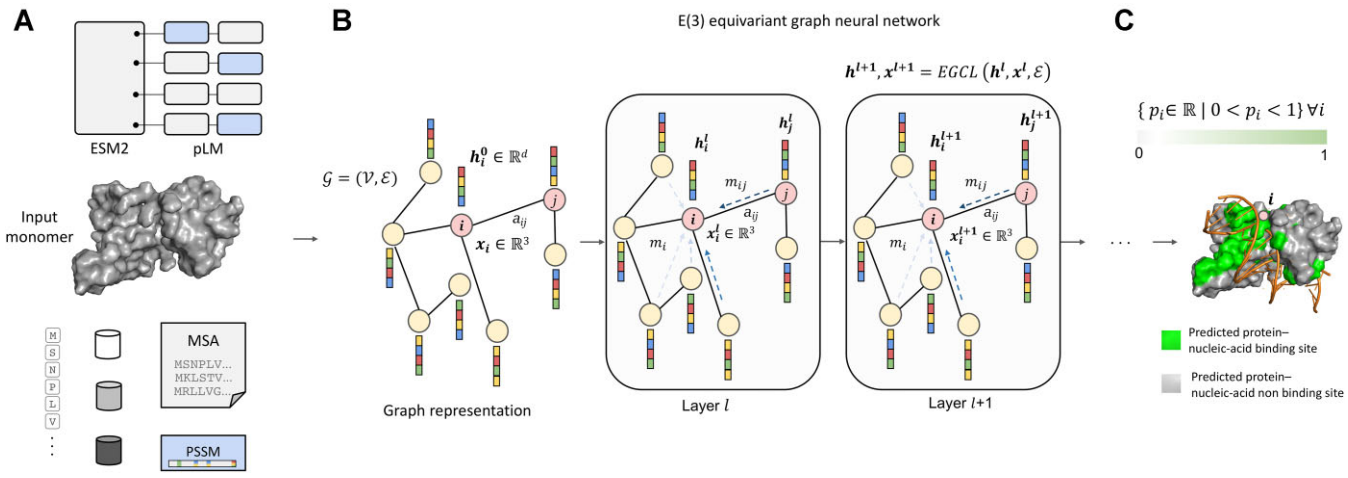


Figure 1. Illustration of the EquiPNAS method for protein–nucleic acid binding site prediction. (**A**) A set of node and edge features are generated from the input protein monomer. (**B**) E(3)-equivariant graph convolutions are employed on the featurized graph representation of the input. (**C**) Graph node classification is performed for residue-level binding site prediction.

structures with AlphaFold2-predicted structural models as input for binding site prediction at scale, without compromising on accuracy. While a recent protein–DNA binding site prediction method, GraphSite (21), has successfully used AlphaFold2-predicted protein structural models, effective utilization of predicted structures from AlphaFold2 for protein–RNA binding site prediction is yet to be explored. Alongside the AlphaFold2 breakthrough, a significant advancement has been made in pre-trained protein language models (pLM) (25–30) powered by attention-based transformers networks (31). pLMs have proven highly successful in various predictive modeling tasks including protein structure prediction (28,30), protein function prediction (26,29), and protein engineering (27,32,33). Despite their usefulness, the potential of pLMs in protein–DNA and protein–RNA binding site prediction tasks remains to be unlocked. Given the recent progress, a natural question arises: can we develop a generalizable computational framework that can harness the power of pLMs while leveraging the predicted structural information by AlphaFold2 for accurate prediction of protein–DNA and protein–RNA binding sites at scale?

Here, we present EquiPNAS, a new pLM-informed equivariant deep graph neural network framework for accurate protein–nucleic acid binding site prediction. EquiPNAS effectively leverages the pLM embeddings derived from the ESM-2 model (30) for improved protein–DNA and protein–RNA binding site prediction. The core of EquiPNAS consists of an E(3) equivariant graph neural network architecture (34), employing symmetry-aware graph convolutions that transform equivariantly with translation, rotation, and reflection in 3D space. Such an architecture has recently been shown to offer substantial accuracy gain while exhibiting remarkable robustness and performance resilience in our work on protein–protein interaction site prediction (35). Inspired by the notable successes of pLMs (32,36–38), here we integrate pLM embeddings from the encoder-only transformer architecture of ESM-2 to refine our sequence-based node features using the E(3) equivariant graph-based framework. By doing this, we are able to significantly reduce the dependence on the availability of evolutionary information which is not always abundant such as with orphan proteins or rapidly evolving proteins, thus enabling us to build generalizable and scalable models. In addition, our translation-, rotation-, and reflection-equivariant deep graph learning architecture provides richer representations for molecular data compared to invariant convolutions, offering robustness for graph structured data and particularly suitable when predicted protein structures are used as input (35).

Our method, EquiPNAS, consistently outperforms the state-of-the-art methods in several widely used benchmarking datasets for both protein–DNA and protein–RNA binding site prediction tasks. EquiPNAS exhibits remarkable robustness with only a minor performance decline when switching from experimental structures to AlphaFold2 predicted structural models as input, enabling accurate prediction of protein–DNA and protein–RNA binding sites at scale. The pLM embeddings used in EquiPNAS are sufficiently powerful that can dramatically reduce the dependence on the availability of evolutionary information, leading to a generalizable framework. In addition, the symmetry-aware nature of the E(3) equivariant graph-based neural architecture of EquiPNAS offers remarkable robustness and performance resilience, as verified directly through our ablation study. An open-source software implementation of EquiPNAS, licensed under the GNU General Public License v3, is freely available at https://github.com/Bhattacharya-Lab/EquiPNAS.

Materials and methods

Overview of EquiPNAS framework

Figure 1 illustrates our EquiPNAS method for protein–nucleic acid binding site prediction consisting of graph representation and featurization, E(3) equivariant graph neural network leveraging the coordinate information extracted from the input monomer together with sequence- and structure-based node and edge features as well as pLM embeddings from the ESM-2 model, and performing graph node classification to predict the probability of every residue in the input monomer to be a protein–nucleic acid binding site.

Graph representation and feature generation

Input protein graph representation

We represent the input protein monomer as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ represents a residue, and each edge $e \in \mathcal{E}$ represents an interacting residue pair. We consider

Table 1. Sequence-based node features

Features *[shape]*	Description
aa *[L, 20]*	One-hot encodings of 20 amino acid residue types.
PSSM *[L, 20]*	Normalized position specific scoring matrix (PSSM).
MSA *[L, 256]*	Multiple sequence alignment (MSA) representation distilled through ColabFold's EvoFormer blocks.
pLM *[L, 5120]*	pLM embeddings from ESM-2 with 15B parameters.

The shape of the corresponding type for a protein with L residues is shown next to each feature.

Table 2. Structures-based node features

Features *[shape]*	Description
SS *[L, 11]*	One-hot encodings of 3- and 8-state secondary structure.
RSA *[L, 10]*	One-hot encodings of 2- and 8-state relevant solvent accessibility.
Local geometry *[L, 11]*	Cosine angle between the C=O of consecutive residues, normalized values of virtual bond and torsion angles, and normalized peptide backbone torsion angles.
Residue orientation *[L, 9]*	Unit vectors pointing towards the directions of $C_\alpha{}^{(i+1)}-C_\alpha{}^{i}$, $C_\alpha{}^{(i-1)}-C_\alpha{}^{i}$ and $C_\beta{}^{i}-C_\alpha{}^{i}$.
Relative residue positioning *[L, 2]*	Two types of relative positional features for the *i*th residue: (i) inverse of *i* representing the relative sequence position, and (ii) inverse of the Euclidean distance of C_α atom from the centroid representing the relative spatial positioning.
Residue virtual surface area *[L, 1]*	Virtual surface area of the conceptual convex hull constructed by the atoms in a residue.
Contact count *[L, 1]*	The number of spatial neighbors of each residue.

The shape of the corresponding type for a protein with L residues is shown next to each feature.

a residue pair to be interacting if their C_α–C_α Euclidean distance is within 14Å for protein–DNA binding site prediction and 15Å for protein–RNA binding site prediction. The specific distance cut-offs are chosen through independent cross-validations for the protein–DNA and protein–RNA binding site tasks (Supplementary Tables 1 and 2). We additionally use a minimum sequence separation of 6 for the interacting residue pairs to focus on longer-range interactions.

Feature generation

We use a number of standard sequence-derived node features including amino acid residue type, position specific scoring matrix (PSSM), multiple sequence alignment (MSA) and combine them with protein language model-based features from ESM-2 pLM. Additionally, we extract structure-derived node features from the input protein monomer, using either the experimentally solved structure or AlphaFold2-predicted structural model, including secondary structure (SS), relative solvent accessibility (RSA), local geometry, residue orientations, relative residue positioning, residue virtual area and contact count.

Sequence-based node features

An overview of sequence-based node features and the corresponding shape can be found in Table 1. We use one-hot encoding to represent each of the 20 amino acid residue types (aa) as a binary vector with 20 entries. We run PSI-BLAST (39) on UniRef90 database (40) to obtain position specific scoring matrix (PSSM). We then extract the first 20 columns of the PSSM and normalize the values using the sigmoidal function. We additionally generate multiple sequence alignment (MSA) from the input amino acid sequence by running ColabFold (41) pipeline, which uses MMseq2 (42) for MSA generation. The generated MSA is then fed to the EvoFormer blocks of AlphaFold2 as implemented in the ColabFold pipeline, resulting in a distilled MSA representation encoded as a dictionary. We extract the first row of the distilled MSA representation ('msa_first_row' from the dictionary) to be used as our MSA feature. We also use protein language model-based features from the pretrained ESM-2 model, having 15B parameters (30). Specifically, we use the 'representations' embeddings as pLM features by supplying the amino acid sequence to the ESM-2 model.

Structure-based node features

Our structure-based node features and the corresponding shape can be found in Table 2. We use one-hot encoding to represent both 3-state and 8-state secondary structures (SS). Additionally, we use one-hot-encodings to represent both 2-state relative solvent accessibility (RSA) features using an RSA cut-off of 50 and finer-grained 8-state RSA features by discretizing the RSA value into 8 bins with the following ranges: 0–30, 30–60, 60–90, 90–120, 120–150, 150–180, 180–210 and > 210. We also extract local geometric features directly from the input protein monomer. These include the cosine angle between the C = O of consecutive residues, normalized virtual bond and torsion angles formed between consecutive C_α atoms, and normalized backbone torsion angles of the polypeptide chain. Inspired by the recent GVP-GNN study (43), we adopt two types of residue orientation features in our study: (i) unit vectors pointing towards $C_\alpha{}^{(i+1)}-C_\alpha{}^{i}$ and $C_\alpha{}^{(i-1)}-C_\alpha{}^{i}$, and (ii) unit vectors indicating the imputed direction of $C\beta^{i}-C_\alpha{}^{i}$, which is computed assuming tetrahedral geometries and normalization. We use two types of relative residue positioning features for the *i*th residue of the input protein monomer: (i) the relative sequence position captured by the inverse of *i*, and (ii) the relative spatial positioning captured by the inverse of the Euclidean distance between the centroid of the input protein monomer and the C_α atom of the *i*th residue. We additionally conceptualize an amino acid residue as a virtual convex hull that is constructed by its constituent atoms and quantify the virtual surface area of the convex hull and calculate its inverse to use as a feature. Finally, we include the normalized contact count as a structure-driven feature, defined as the number of spatial neighbors of each residue (i.e. residues that are in contact) where two residues are considered to be in contact if the Euclidean distance between their C_β atoms is <8 Å.

Edge features

As the edge feature for the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we use the ratio of the logarithm of the absolute difference between the indices of the two residues ($log|i-j|$) in the primary sequence and their Euclidean distance. The numerator of the ratio measures

how far apart the two residues are in the primary sequence, while the denominator measures their spatial distance in 3D space.

Coordinate features

We obtain coordinate features from the Euclidean coordinates (x, y and z) of the Cα atoms in input protein monomers.

Network architecture

Our network architecture consists of deep E(3)-equivariant graph neural networks (EGNNs) (34), independently trained for protein–DNA and protein–RNA binding site prediction tasks. The input to the EGNNs includes the node and edge features described above as well as coordinate features based on the Cartesian coordinates of the C_α atoms in the input protein monomer. The EGNN architecture consists of a stack of equivariant graph convolution layers (EGCL), performing a series of transformations of its input by updating the coordinate and node embeddings using the edge information and the coordinate and node embeddings from the previous layer. A linear transformation is first applied to the input node features (h_i^0), which results in a transformed set of node embeddings (h_i^l). These embeddings, along with input coordinates (x_i^0) and edge information (a_{ij}) are passed to the subsequent EGCL layers. Formally, each EGCL performs a coordinate and node embeddings update, such that $x_i^{l+1}, h_i^{l+1} = EGCL[x_i^l, h_i^l]$, which is defined below:

$$m_{ij} = \phi_e\left(h_i^l, h_j^l, \|x_i^l - x_j^l\|^2, a_{ij}\right)$$

$$x_i^{l+1} = x_i^l + C\sum_{j\neq i}\left(x_i^l - x_j^l\right)\phi_x\left(m_{ij}\right)$$

$$m_i = \sum_{j\neq i} m_{ij}$$

$$h_i^{l+1} = \phi_h\left(h_i^l, m_i\right)$$

where, h_i^l and h_j^l are the node embeddings of node i and j at layer l, respectively; a_{ij} denotes the edge attributes; x_i^l and x_j^l are the coordinates of node i and j at layer l, respectively; $\|x_i^l - x_j^l\|^2$ is the squared distance between node i and j at layer l; ϕ_e, ϕ_x, and ϕ_h are non-linear operations, implemented using multilayer perceptrons (MLP); and C is a constant factor chosen as $1/(M-1)$, where M is the number of nodes. The EGCL operation attains equivariance by incorporating the coordinate update during message passing, wherein for each node i, $\sum_{j\neq i}(x_i^l - x_j^l)$ is the sum of its relative coordinate difference with all the other nodes, are taken into account for updating the coordinate x_i^{l+1} of node i at layer $l+1$. We also use an attention mechanism $\tilde{e}_{ij} = \phi_{inf}(m_i)$ to infer a soft estimation of edges. Finally, a linear transformation is applied to squeeze the hidden dimension of the last EGCL for condensing the learned information into a single scalar value, followed by a sigmoidal function to obtain the node-level classification to predict the likelihood of every residue in the input monomer to be a protein–nucleic acid binding site. The architecture of our EGNN consists of 12 EGCL layers with hidden dimensions of 768. The size of hidden dimensions and the number of layers are selected through 5-fold cross-validation (see Supplementary Table 1, 2). To mitigate the risk of overfitting, we apply dropout regularization to the node embeddings of each EGCL layer with a dropout rate of 0.1, determined through 5-fold cross validation (see Supplementary Table 1, 2). Our EquiPNAS models are implemented using PyTorch 1.12.0 (44) and the Deep Graph Library (DGL) 0.9.0 (45). During training, we use the binary cross-entropy loss function and a cosine annealing scheduler from the Stochastic Gradient Descent with Warm Restarts (SGDR) algorithm (46). We also utilize the ADAM optimizer (47), with a learning rate of 1e-4 and a weight decay of 1e-16. The training process consists of at most 40 epochs on an NVIDIA A40 GPU. In addition to the full-fledged version of EquiPNAS, we train baseline models for both protein–DNA and protein–RNA binding site prediction using the same hyperparameters and features as EquiPNAS, but without equivariant updates, that is, invariant baseline networks with the coordinate updates of the EGCL turned off, enabling us to verify the importance of equivariance used in our model.

Datasets and performance evaluation

For a fair performance comparison of our method against the state-of-the-art methods for protein–DNA and protein–RNA binding site prediction, we use widely recognized public datasets as follows.

Protein–DNA benchmarking dataset

To evaluate the performance of protein–DNA binding site prediction method, we use train (Train_573) and test (Test_129) datasets from the published work of GraphBind (20), which contain a total of 573 and 129 protein chains, respectively. Additionally, we use another test set consisting of 181 protein chains (Test_181) from the published work of GraphSite (21). These datasets are originally curated from the public BioLiP database (48) that contains precomputed protein–DNA and protein–RNA binding sites from known protein–DNA and protein–RNA complexes after filtering out protein chains with >30% sequence similarity among the datasets, by applying CD-Hit (49) to ensure non-redundancy. The training dataset (Train_573) was released before 6 January 2016 whereas the Test_129 set was released between 6 January 2016 to 5 December 2018, and Test_181 was more recently released between 6 December 2018 to August 2021. The binding (and non-binding) residue count for Train_573, Test_129 and Test_181 are 14 479 (and 145 404), 2240 (and 35 275) and 3208 (and 72 050), respectively.

Protein–RNA benchmarking dataset

To evaluate the performance of protein–RNA binding site prediction method, we use the Train_495 set for training and the Test_117 set for testing, also from the published work of GraphBind (20), which contain a total of 495 and 117 protein chains, respectively. These datasets are also extracted from the BioLiP database (48) and pre-processed to ensure non-redundancy between the train and test sets, using CD-Hit (49) to filter out protein chains with > 30% sequence similarity. The Train_495 set contains 14 609, and 122 290 binding, and non-binding residues, respectively, while in the Test_117 set, 2031 and 35 314 residues are binding, and non-binding residues, respectively.

Table 3. Protein-DNA and protein–RNA binding site prediction performance of EquiPNAS against the top-performing methods on the test datasets using AlphaFold2 predicted structural models as input. Values in bold represent the best performance

	Datasets	Methods	ROC-AUC	PR-AUC
Protein–DNA	Test_129	GraphBind*	0.916	0.497
		GraphSite*	0.934	0.544
		EquiPNAS	**0.940**	**0.569**
	Test_181	GraphBind*	0.893	0.317
		GraphSite*	0.917	0.369
		EquiPNAS	**0.918**	**0.384**
Protein–RNA	Test_117	GraphBind	0.793	0.204
		EquiPNAS	**0.886**	**0.320**

Note: * Results are obtained directly from the published work of GraphSite.

Evaluation metrics and competing methods

We assess the performance of our method using two widely recognized metrics: the area under the Receiver Operating Characteristic curve (ROC-AUC) and the area under the Precision-Recall curve (PR-AUC) scores. Both ROC-AUC and PR-AUC are threshold-independent metrics, thereby providing a comprehensive and robust view of the performance of a model across the full range of possible classification thresholds.

We compare our protein–DNA interaction site prediction method against eight existing methods. Three of the methods, SVMnuc (7), NCBRPred (8), and DNAPred (9), are sequence-based methods, while the other five methods, COACH-D (15), NucBind (7), DNABind (16), GraphBind (20) and GraphSite (21) are structure-aware methods. SVMnuc is a support vector machine (SVM)-based method that utilizes features from PSI-BLAST (39), PSIPRED (50) and HHblits (51). NCBRPred employs bidirectional Gated Recurrent Units (BiGRU) (52) with multi-label sequence labeling. DNAPred is a two-stage ensembled hyperplane-distance-based support vector machine (E-HDSVM) (9) for predicting protein–DNA binding sites. COACH-D is a consensus-based approach incorporating four different template-based and one template-free prediction methods. NucBind integrates the *ab initio* SVMnuc and template-based COACH-D for higher accuracy prediction. DNABind is a hybrid method combining machine learning with template-based predictions. GraphBind proposes hierarchical graph neural networks, while GraphSite employs graph transformer neural networks. Among these competing methods, GraphBind and GraphSite are the most recent and represent the state-of-the-art for protein–DNA binding site prediction.

We compare our protein–RNA binding site prediction method with seven existing methods. Two of the methods RNABindRPlus (11) and SVMnuc (7) are sequence-based methods, while the other five methods, COACH-D (15), NucBind (7), aaRNA (18), NucleicNet (19) and GraphBind (20) are structure-aware methods. SVMnuc, COACH-D, NucBind and GraphBind are the methods we also compared against on protein–DNA binding tasks, as discussed earlier. RNABindRPlus is a hybrid method that combines sequence-homologs and support vector machine (SVM)-based predictions. aaRNA is a both sequence- and structure-based method that utilizes homology modeling to extract structural features along with various sequence-based features. NucleicNet is a deep learning framework that extracts physiochemical characteristics of the protein surface by quantifying it with grid points. Among these methods, GraphBind is currently the top-performing method for protein–RNA binding site prediction.

Results

Test set performance

Table 3 shows the performance of EquiPNAS for protein–DNA (on Test_129 and Test_181 sets) and protein–RNA (on Test_117) binding site prediction tasks using AlphaFold2 predicted structural models as input compared to two closest competing methods: hierarchical graph neural network-based method GraphBind for protein–DNA and protein–RNA binding site prediction (20) and graph transformer-based method GraphSite for protein–DNA binding site prediction (21) (see Supplementary Table 3 and Supplementary Table 4 for comprehensive performance comparison against all competing methods). The results demonstrate that EquiPNAS attains the highest scores in all three test datasets. The performance gain of EquiPNAS over the state-of-the-art methods is particularly noteworthy considering PR-AUC, a stringent and rigorous evaluation metric. For example, EquiPNAS yields 56.9% relative PR-AUC gain over GraphBind for protein–RNA binding site prediction; and 14.5%-21.1% relative PR-AUC gains over GraphBind and 4.1–4.6% relative PR-AUC gains over GraphSite for protein–DNA binding site prediction. In summary, EquiPNAS improves upon the state-of-the-art accuracy of both protein–DNA and protein–RNA binding site prediction using AlphaFold2 predicted structural models by consistently attaining better performance than the existing approaches.

To investigate whether the performance attained by EquiPNAS is significantly better than the closest competing methods GraphSite and GraphBind, we conduct statistical significance tests by randomly sampling 70% of the targets for each of the test sets (Test_129, Test_181, and Test_117) and calculating the ROC-AUC and PR-AUC for the EquiPNAS as well as the other competing methods. This sampling process is repeated 10 times, yielding a set of 10 scores for EquiPNAS, GraphSite, and GraphBind for protein–DNA binding site prediction for both Test_129 and Test_181 sets, and a set of 10 scores for EquiPNAS and GraphBind for protein–RNA binding site prediction for Test_117 set. If the measurement is normal, determined by the Anderson-Darling test (53), then paired t-test is used to calculate significance of the measurement. If the measurement is not normal, then we use the Wilcoxon rank sum test (54). The results presented in Table 4 demonstrate that EquiPNAS is statistically significantly better than the competing methods at 95% confidence level with p-values < 0.05 for both ROC-AUC and PR-AUC metrics across all test sets.

Figure 2 presents nine representative examples from the test datasets comparing the protein–DNA and protein–RNA binding site predictions using EquiPNAS against the second-best

Table 4. Statistical significance test between EquiPNAS and the top-performing methods using AlphaFold2 predicted structural models as input on the test datasets by randomly sampling 70% of the targets for each of the test sets and repeating the sampling process 10 times

	Datasets	Methods	ROC-AUC	PR-AUC
Protein–DNA	Test_129	GraphBind	0.9128 ± 0.008929352	0.492 ± 0.031184042
		P-value	2.22591E-06	7.07626E-10
		GraphSite	0.9219 ± 0.005363457	0.5165 ± 0.022122136
		P-value	1.3961E-09	7.92445E-08
		EquiPNAS	**0.9387** ± 0.004877385	**0.569** ± 0.0264281
	Test_181	GraphBind	0.8916 ± 0.006003703	0.3102 ± 0.017706245
		P-value	8.63327E-08	7.16361E-09
		GraphSite	0.8964 ± 0.006292853	0.3286 ± 0.018124262
		P-value	2.25585E-07	7.9832E-07
		EquiPNAS	**0.9159** ± 0.00395671	**0.3717** ± 0.018372987
Protein–RNA	Test_117	GraphBind	0.7942 ± 0.006250333	0.2019 ± 0.009573691
		P-value	2.3402E-11	1.44E-10
		EquiPNAS	**0.8856** ± 0.006221825	**0.3118** ± 0.013003

The means and the standard deviations of ROC-AUC and PR-AUC are reported. Values in bold represent the best performance in terms of means.

predictors: three from protein–DNA Test_129 (Figure 2A), three from protein–DNA Test_181 (Figure 2B) and three from protein–RNA Test_117 (Figure 2C). The first two examples represent two human protein–DNA interactions: Transcription of Homo sapiens, Mus musculus (PDB ID: 5nj8, chain A), and Hydrolase/DNA of Homo sapiens, DNA launch vector pDE-GFP2 (PDB ID: 5t4i, chain B) as shown in Figure 2A. GraphSite fails to predict the vast majority of protein–DNA binding sites as reflected in its low F1-score, Matthew's Correlation Coefficient (MCC) and PR-AUC in these two targets. In contrast, EquiPNAS achieves reasonably accurate prediction, with a remarkable gain of 0.506 and 0.417 points in F1-score, 0.458 and 0.357 points in MCC, and 0.381 and 0.381 points in PR-AUC, respectively. The third example, Splicing of Caenorhabditis elegans, synthetic construct (PDB ID: 5tkz, chain A) shows inaccurate binding site prediction by GraphSite, resulting in predicting five (out of total 89 residues), which is noticeably high compared to the size of the protein. EquiPNAS accurately predicts these binding sites, with only one (out of total 89 residues) false positive. GraphSite also generates inaccurate predictions for DNA binding protein/DNA in Escherichia coli (PDB ID: 6nua, chain A), with 28 (out of total 227 residues) false positives; whereas EquiPNAS achieves a much better overall prediction performance with only three (out of total 227 residues) false positives. Interestingly, EquiPNAS attains perfect prediction with both ROC-AUC and PR-AUC values of 1.0, as well as an F1-score and MCC of approximately 0.93 for a smaller target (73 residues), a transcription protein in Mycobacterium tuberculosis (PDB ID: 7kuf chain A). In contrast, GraphSite's prediction is contaminated by several false positives, resulting in F1-score and MCC values of less than 0.65. Additionally, for an RNA binding protein/DNA in *Homo sapiens* (PDB ID: 7csz, chain A), our method still outperforms GraphSite, with a performance gain of 0.27 points in PR-AUC, 0.187 points in F1-score, and 0.216 points in MCC, whereas GraphSite fails to identify majority of binding site residues, particularly for DNA chain C, resulting in a high number of false negatives. The RNA binding protein example in Danio rerio and Caenorhabditis elegans (PDB ID: 6fq3, chain A) provides a remarkable demonstration of the superior performance of EquiPNAS in predicting protein–RNA binding sites, as compared to the closest competing method GraphBind. While GraphBind fails to accurately detect any binding site, with PR-AUC, F1-score and MCC of 0.024, 0 and −0.017, respectively, EquiPNAS performs reasonably accurate predictions with much better PR-AUC, F1-score, and MCC of 0.732, 0.545 and 0.555, respectively. Furthermore, EquiPNAS shows highly accurate prediction for the transcription factor in Saccharomyces cerevisiae (PDB ID: 5o1y, chain A), exceeding GraphBind by 0.351 points in F1-score, 0.382 points in MCC and 0.303 points in PR-AUC. Additionally, in comparison to EquiPNAS, GraphBind exhibits suboptimal performance due to both false positive and false negative predictions for the binding sites of OXIDOREDUCTASE/RNA in Escherichia coli (PDB ID: 5hr7, chain B).

In the above experiments, all methods use AlphaFold2 predicted structural models as input with EquiPNAS consistently delivering improved performance for both protein–DNA and protein–RNA binding site prediction tasks. However, structure-aware protein–nucleic acid binding site prediction methods traditionally rely on experimentally solved structures as input. Intuitively, using experimental structures as input, whenever available, should lead to better performance than using predicted structural models as input. Consequently, a natural question to ask is: How much performance decline do these methods suffer from when switching from experimental input to prediction? Not surprisingly, as shown in Supplementary Tables 3 and 4, using experimental input leads to better accuracy in almost all cases. Promisingly, the performance decline of EquiPNAS when switching from experimental input to AlphaFold2 prediction is much smaller compared to other methods. For instance, EquiPNAS loses only ∼2.3% of PR-AUC points when using AlphaFold2 predictions as input instead of experimental ones for protein–DNA binding site prediction, whereas GraphBind experiences a higher PR-AUC drop of 4.4–6.9% PR-AUC points. EquiPNAS also demonstrates robustness in protein–RNA binding site prediction with a negligible drop in ROC-AUC (0.1%) when using AlphaFold2 predictions as input, whereas GraphBind shows a much higher ROC-AUC drop (7.7%). That is, EquiPNAS exhibits a minor performance decline when switching from experimental input to prediction while outperforming both GraphBind and GraphSite regardless of the use of predicted or experimental structures, demonstrating its robustness and generalizability and enabling accurate prediction of protein–DNA and protein–RNA binding sites at scale using AlphaFold2 predicted structural models.

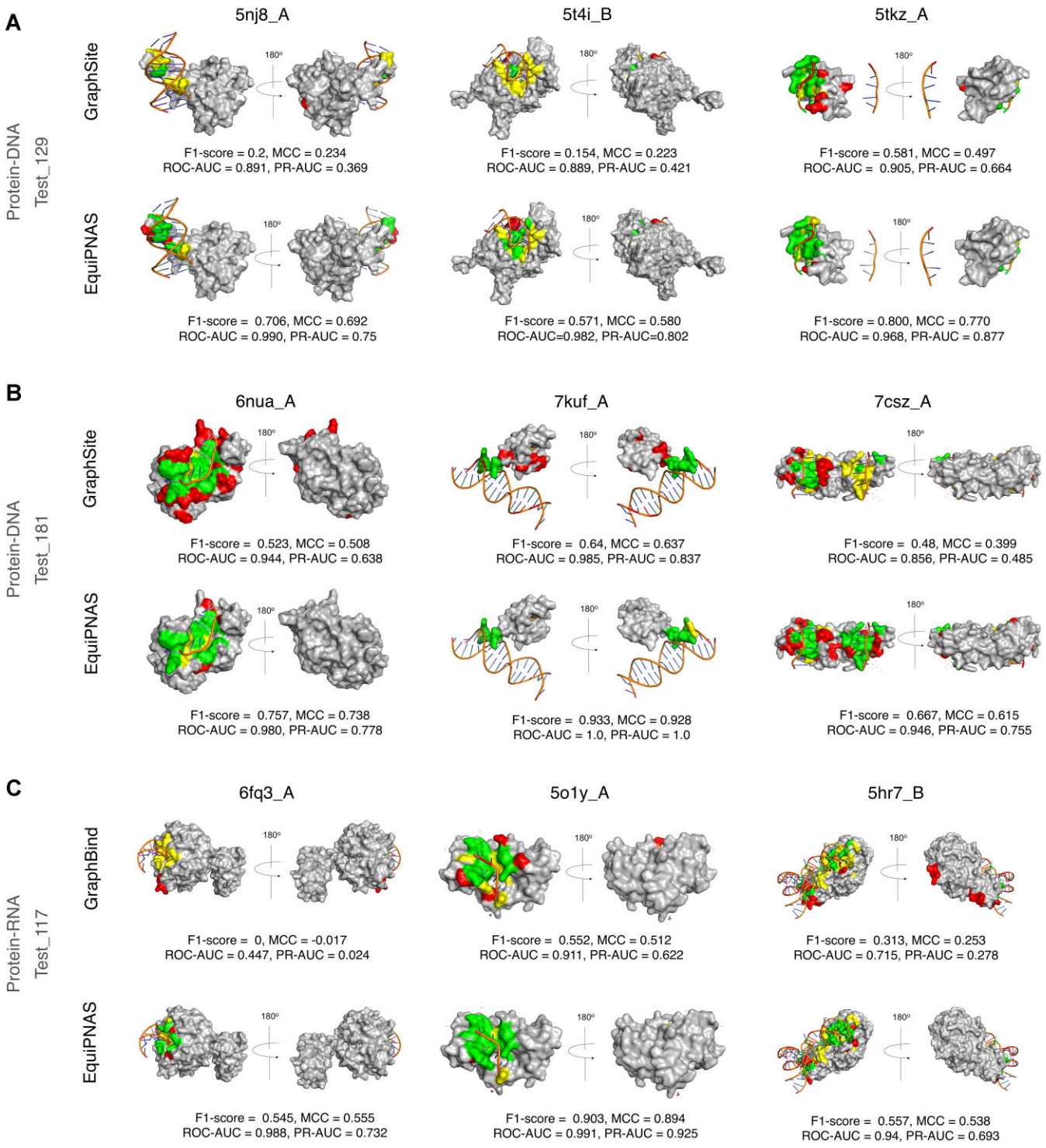


Figure 2. Representative examples of protein–DNA and protein–RNA binding site predictions using EquiPNAS and the closest competing methods compared to the experimental observation. For targets from the Test_129 (**A**) and Test_181 (**B**) sets, protein–DNA binding site prediction using GraphSite versus EquiPNAS are shown. For targets from the Test_117 set (**C**), protein–RNA binding site prediction using GraphBind versus EquiPNAS are shown. True Positive (TP), False Positive (FP), and False Negative (FN) binding sites are represented in green, red, and yellow, respectively.

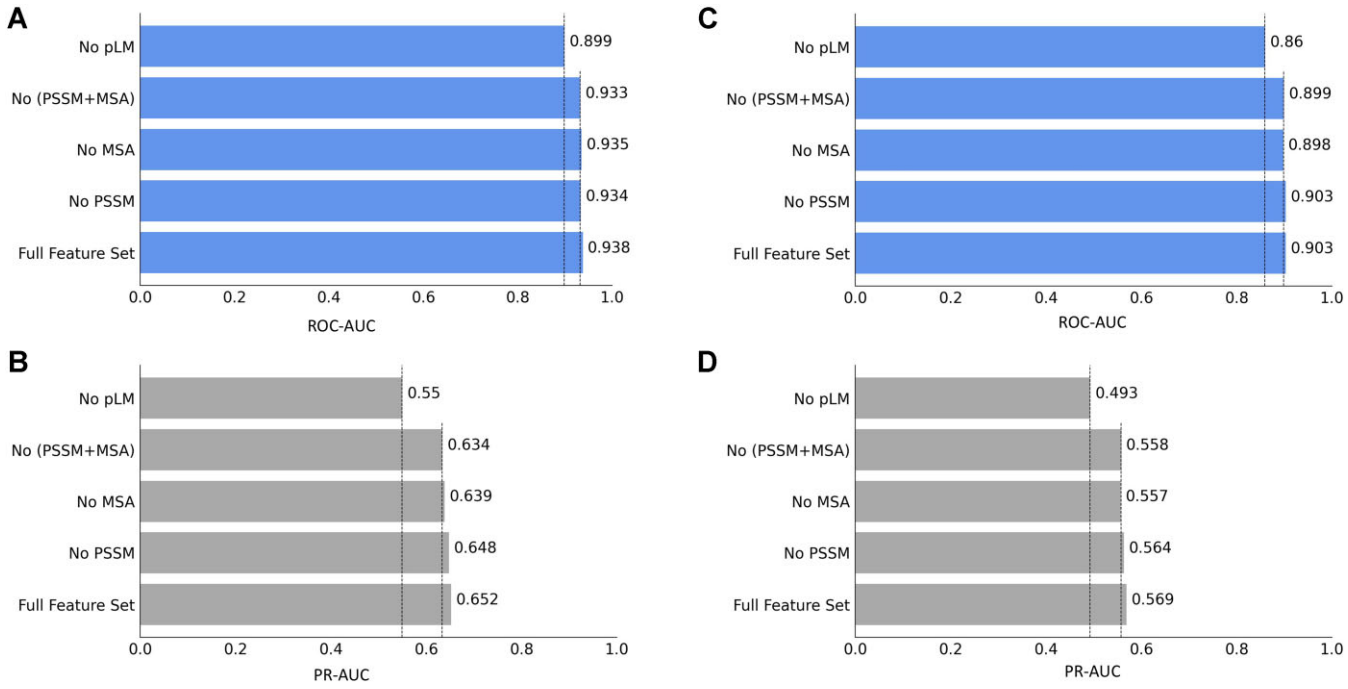


Figure 3. Feature ablation study. For protein–DNA binding site prediction, bar charts representing the performance of the ablated variants in terms of (**A**) ROC-AUC and (**B**) PR-AUC obtained using 5-fold cross validation are shown. For protein–RNA binding site prediction, bar charts representing the performance of the ablated variants in terms of (**C**) ROC-AUC and (**D**) PR-AUC obtained using 5-fold cross validation are shown.

In the context of large-scale protein–nucleic acid binding site prediction using AlphaFold2 predicted structural models, a related question is: Is there any relationship between the self-estimated accuracy of AlphaFold2 predicted structural models and the accuracy of EquiPNAS binding site prediction? We examine the self-estimated accuracy of AlphaFold2 predicted structural models using the AlphaFold2 predicted local distance difference test (pLDDT) and the ROC-AUC and PR-AUC of EquiPNAS binding site prediction resulting from the predicted structure. Using a pLDDT threshold of 0.85, we divide the targets in the test sets into two roughly equal groups: moderate confidence predictions with pLDDT values $\leq$0.85 and high confidence predictions with pLDDT values $>$0.85. Supplementary Figure 1 shows the ROC-AUC and PR-AUC distributions for the two groups. Across the test datasets, high confidence predictions lead to better ROC-AUC and PR-AUC values compared to moderate confidence predictions, with the ROC-AUC and PR-AUC distributions resulting from the high confidence predictions skewed towards higher accuracy binding site prediction. Furthermore, we observe a noticeable difference in binding site prediction accuracy in terms of mean ROC-AUC and PR-AUC values resulting from the moderate confidence predictions versus the high confidence predictions (see Supplementary Table 5), indicating that the self-estimated accuracy of AlphaFold2 predicted structural models can inform the accuracy of EquiPNAS binding site prediction in the absence of any experimental information in that highly confident AlphaFold2 predictions tend to yield more accurate binding site prediction.

Ablation study

Contribution of the pLM embeddings

EquiPNAS utilizes pLM embeddings from the pretrained ESM-2 model (30) as part of the sequence-based features. To evaluate the relative contribution of the protein language model-based features compared to the evolutionary features such as PSSM and MSA, we conduct a feature ablation study by excluding protein language model-based features or the evolutionary features from the full-fledged EquiPNAS feature set. Figure 3 displays the 5-fold cross-validation performance of the ablated variants of EquiPNAS in terms of ROC-AUC and PR-AUC values for protein–DNA and protein–RNA binding site prediction. The results demonstrate that excluding pretrained protein language model-based features (no pLM) results in the worst performance with a relative PR-AUC drop of 18.5% (Figure 3B) and 15.4% (Figure 3D) for protein–DNA and protein–RNA binding site predictions, respectively. Such a significant performance drop highlights the importance of using pLM embeddings for our prediction. In contrast, we observe only minor performance drops when one or both evolutionary features were discarded. Even discarding both the evolutionary features (no (PSSM + MSA)) results in a relative PR-AUC drop of only 2.8% and 2% for protein–DNA and protein–RNA binding site predictions, respectively. Overall, compared to the relatively minor but positive contribution of evolutionary features, protein language model-based features have a major contribution to the improved performance of the new EquiPNAS model.

The ESM-2 offers a range of pretrained pLMs with varying scale ranging from 8 million to 15 billion parameters including esm2_t6_8M_UR50D, esm2_t12_35M_UR50D, esm2_t30_150M_UR50D, esm2_t33_650M_UR50D, esm2_t36_3B_UR50D and esm2_t48_15B_UR50D-trained. The largest pLM esm2_t48_15B_UR50D with 15 billion parameters serves as the default choice for the pLM embeddings in our EquiPNAS method. To assess the impact of the scale of pretrained pLMs on performance, we retrain five separate protein–DNA and protein–RNA binding site prediction models on the full training set after replacing the pLM

Table 5. Protein-DNA and protein–RNA binding site prediction performance using different ESM-2 pLMs with lower number of parameters (esm2_t6_8M_UR50D, esm2_t12_35M_UR50D, esm2_t30_150M_UR50D, esm2_t33_650M_UR50D and esm2_t36_3B_UR50D) compared to the default choice of the pLM used in EquiPNAS (esm2_t48_15B_UR50D) with 15 billion parameters

	Datasets	Models	ROC-AUC	PR-AUC
Protein–DNA	Test_129	esm2_t6_8M_UR50D	0.921	0.504
		esm2_t12_35M_UR50D	0.923	0.507
		esm2_t30_150M_UR50D	0.928	0.539
		esm2_t33_650M_UR50D	0.933	0.543
		esm2_t36_3B_UR50D	0.935	0.531
		EquiPNAS (esm2_t48_15B_UR50D)	**0.940**	**0.569**
	Test_181	esm2_t6_8M_UR50D	0.897	0.332
		esm2_t12_35M_UR50D	0.901	0.339
		esm2_t30_150M_UR50D	0.910	0.359
		esm2_t33_650M_UR50D	0.912	0.362
		esm2_t36_3B_UR50D	0.908	0.352
		EquiPNAS (esm2_t48_15B_UR50D)	**0.918**	**0.384**
Protein–RNA	Test_117	esm2_t6_8M_UR50D	0.856	0.285
		esm2_t12_35M_UR50D	0.862	0.299
		esm2_t30_150M_UR50D	0.863	0.297
		esm2_t33_650M_UR50D	0.869	0.309
		esm2_t36_3B_UR50D	0.874	0.303
		EquiPNAS (esm2_t48_15B_UR50D)	**0.886**	**0.320**

Values in bold represent the best performance.

Table 6. Protein–DNA and protein–RNA binding site prediction performance of EquiPNAS variant trained without any evolutionary information (w/o MSA + PSSM) against the top-performing methods on the test datasets using AlphaFold2 predicted structural models as input

	Datasets	Methods	ROC-AUC	PR-AUC
Protein–DNA	Test_129	GraphBind*	0.916	0.497
		GraphSite*	0.934	**0.544**
		EquiPNAS w/o (MSA + PSSM)	**0.936**	**0.544**
	Test_181	GraphBind*	0.893	0.317
		GraphSite*	**0.917**	**0.369**
		EquiPNAS w/o (MSA + PSSM)	**0.917**	0.364
Protein–RNA	Test_117	GraphBind	0.793	0.204
		EquiPNAS w/o (MSA + PSSM)	**0.877**	**0.299**

Note: * Results are obtained directly from the published work of GraphSite.
Values in bold represent the best performance.

embeddings from the default esm2_t48_15B_UR50D choice with other ESM-2 pLMs with lower number of parameters including esm2_t6_8M_UR50D, esm2_t12_35M_UR50D, esm2_t30_150M_UR50D, esm2_t33_650M_UR50D, and esm2_t36_3B_UR50D models. Table 5 reports the performance of these alternative models in comparison to EquiPNAS (utilizing esm2_t48_15B_UR50D) for both protein–DNA and protein–RNA test sets. The results demonstrate that models trained with the pLM having the lowest number of parameters (esm2_t6_8M_UR50D) perform the poorest in both protein–DNA and protein–RNA binding site prediction tasks, with a 12.2–15.7% lower PR-AUC values compared to EquiPNAS in the test datasets. With the increase in number of parameters of the ESM-2 pLMs, test set performance tends to improve. EquiPNAS leveraging the largest pLM esm2_t48_15B_UR50D with 15 billion parameters consistently achieves the best performance across all test sets, justifying our choice of the ESM-2 pLM. A recent method called GeoBind (55), which exploits protein molecular surfaces for protein–nucleic acid binding site prediction using geometric deep learning, attains state-of-the-art performance by extracting molecular surfaces computed from experimental structures coupled with evolutionary information in the form of MSA or pLM embeddings to replace MSA. The published work of GeoBind, trained on the same training set used in our method, reports its performance for protein–DNA (on Test_129) and protein–RNA (on Test_117) binding site prediction tasks using experimental structures as input. In a head-to-head comparison with GeoBind on the identical set of test targets, our method EquiPNAS consistently outperforms GeoBind in both protein–DNA and protein–RNA binding site prediction tasks (see Supplementary Table 6). For example, EquiPNAS using experimental structures as input attains higher ROC-AUC of 0.943 (and 0.887) than GeoBind having an ROC-AUC of 0.940 (and 0.874) for protein–DNA (and protein–RNA) binding site prediction tasks. Once again, EquiPNAS exhibits remarkable robustness by attaining comparable or even better accuracy with predicted structural models from AlphaFold2 than what GeoBind can achieve even with experimental structures. That is, EquiPNAS is robust and more accurate compared to GeoBind.

Recognizing the major contribution of pLM features compared to the relatively minor impact of the evolutionary features, we investigate the performance of our method utilizing the pLM embeddings, but without using any evolutionary information. Specifically, we discard the PSSM and MSA features and retrain our method on the full training set, and evaluate the performance on the test sets for both protein–DNA and protein–RNA binding site prediction tasks. As reported in Table 6, We find that for protein–DNA binding site prediction,

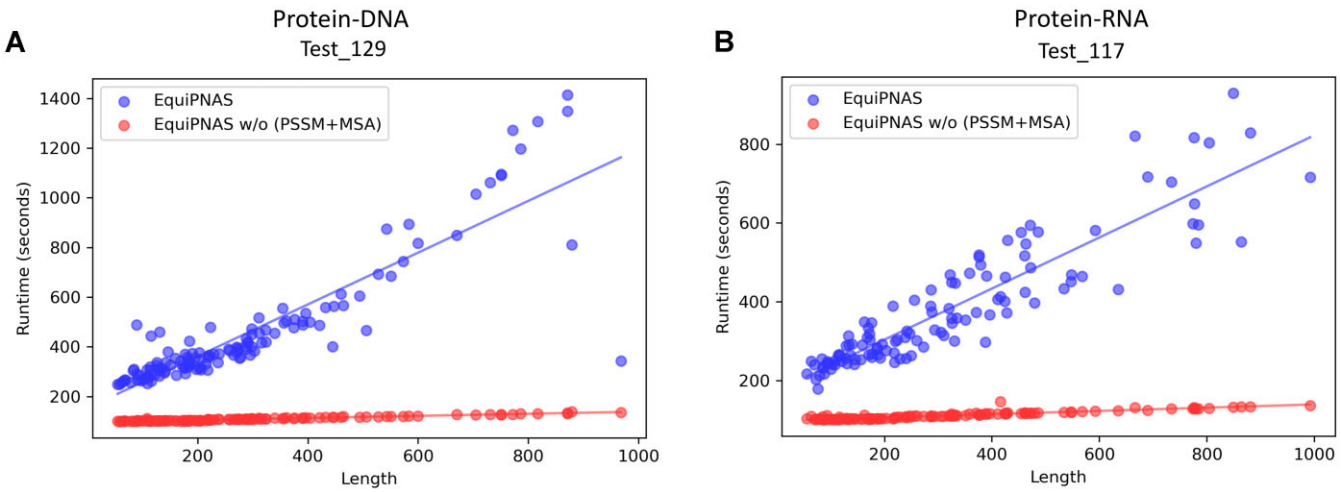


Figure 4. The running time of the full-fledged version of EquiPNAS and its variant trained without any evolutionary information on (**A**) protein–DNA (Test_129) and (**B**) protein–RNA (Test_117) binding site prediction. For each target, input protein length versus runtime (in seconds) are shown.

EquiPNAS without PSSM or MSA (denoted by 'EquiPNAS w/o (PSSM + MSA)') outperforms GraphBind, and performs comparably to GraphSite; with only a slight performance decline compared to the full-fledged version of EquiPNAS. For example, in Test_129, EquiPNAS w/o (PSSM + MSA) achieves a ROC-AUC of 0.936 and a PR-AUC of 0.544, which is comparable to GraphSite (ROC-AUC of 0.934 and PR-AUC of 0.544) and much higher than GraphBind (ROC-AUC of 0.916 and PR-AUC of 0.497). We observed a similar trend in Test_181. In contrast, the state-of-the-art GraphSite experiences a noticeable performance drop without using any evolutionary features. As reported in the published work of GraphSite, PR-AUC drops from 0.544 down to 0.452 without using its MSA-derived features (-AF2 Single). For protein–RNA binding site prediction (Test_117), EquiPNAS w/o (PSSM + MSA) achieves a ROC-AUC of 0.877 and a PR-AUC of 0.299, which is noticeably better than GraphBind (ROC-AUC of 0.793 and PR-AUC of 0.204). Collectively, the results demonstrate the robustness of EquiPNAS over the state-of-the-art methods in that EquiPNAS is able to significantly reduce the dependence on the availability of evolutionary information which is not always abundant such as with orphan proteins or rapidly evolving proteins. Even without using any evolutionary information, and thus at a much lower computational overhead required for MSA and PSSM feature generation, our method performs comparably (in the case of protein–DNA), even superior (in the case of protein–RNA) to the full-fledged state-of-the-art protein–DNA and protein–RNA binding site prediction methods. In summary, EquiPNAS enables us to build generalizable and scalable models.

We further analyze the running time of the full-fledged version of EquiPNAS against its variant, 'EquiPNAS w/o (PSSM + MSA)', that utilizes the pLM embeddings but without any evolutionary information. As shown in Figure 4, the running time of EquiPNAS is clearly dependent on the length of the input protein, whereas the variant trained without any evolutionary information w/o (PSSM + MSA) exhibits no such trend and yields a near-constant running time regardless of the protein length. With an average running time of approximately 110 s, EquiPNAS w/o (PSSM + MSA) attains a speed boost of around 3–4 times compared to the full-fledged EquiPNAS version. That is, bypassing the evolutionary features leads to orders of magnitude speedup in running time.

Contribution of equivariance

EquiPNAS delivers robust and improved performance across various datasets and predictive modeling scenarios. In order to understand the reasons behind such improved performance and verify that it is connected to the equivariant nature of the model, we perform an ablation study by isolating the effect of the equivariant graph convolutions used in EquiPNAS. In particular, we train a family of baseline graph neural networks for protein–DNA and protein–RNA binding site prediction tasks after turning off the coordinate updates of the equivariant graph convolution layers, thus making it an invariant network. Both the equivariant (the full-fledged version of EquiPNAS) and invariant counterparts are trained on the same training datasets using the same set of input features and hyperparameters as the full-fledged version of EquiPNAS. Figure 5 shows the performance of the equivariant and invariant networks using both experimentally determined (native) and AlphaFold2 predicted structures. The results demonstrate that equivariant networks used in the full-fledged version of EquiPNAS consistently outperform the invariant baseline networks regardless of the use of predicted or native structures as input. Strikingly, the invariant baseline models even using the native structures perform worse than the equivariant models using the AlphaFold2 predicted structures, let alone the equivariant models using the experimental structures. For instance, in the Test_129 set, the baseline invariant model attains ROC-AUC (and PR-AUC) of 0.938 (and 0.565) using the native structures, whereas the equivariant model attains ROC-AUC (and PR-AUC) of 0.940 (and 0.569) using AlphaFold2 predicted structures, and 0.943 (0.582) using native structures. A similar trend is also overserved in test sets Test_181 and Test_117. Overall, the results highlight the performance contribution and remarkable robustness of the equivariant networks used in EquiPNAS, attaining better accuracy with AlphaFold2 predicted structural models than what an invariant counterpart can achieve even with experimental structures for both protein–DNA and protein–RNA binding site prediction tasks.

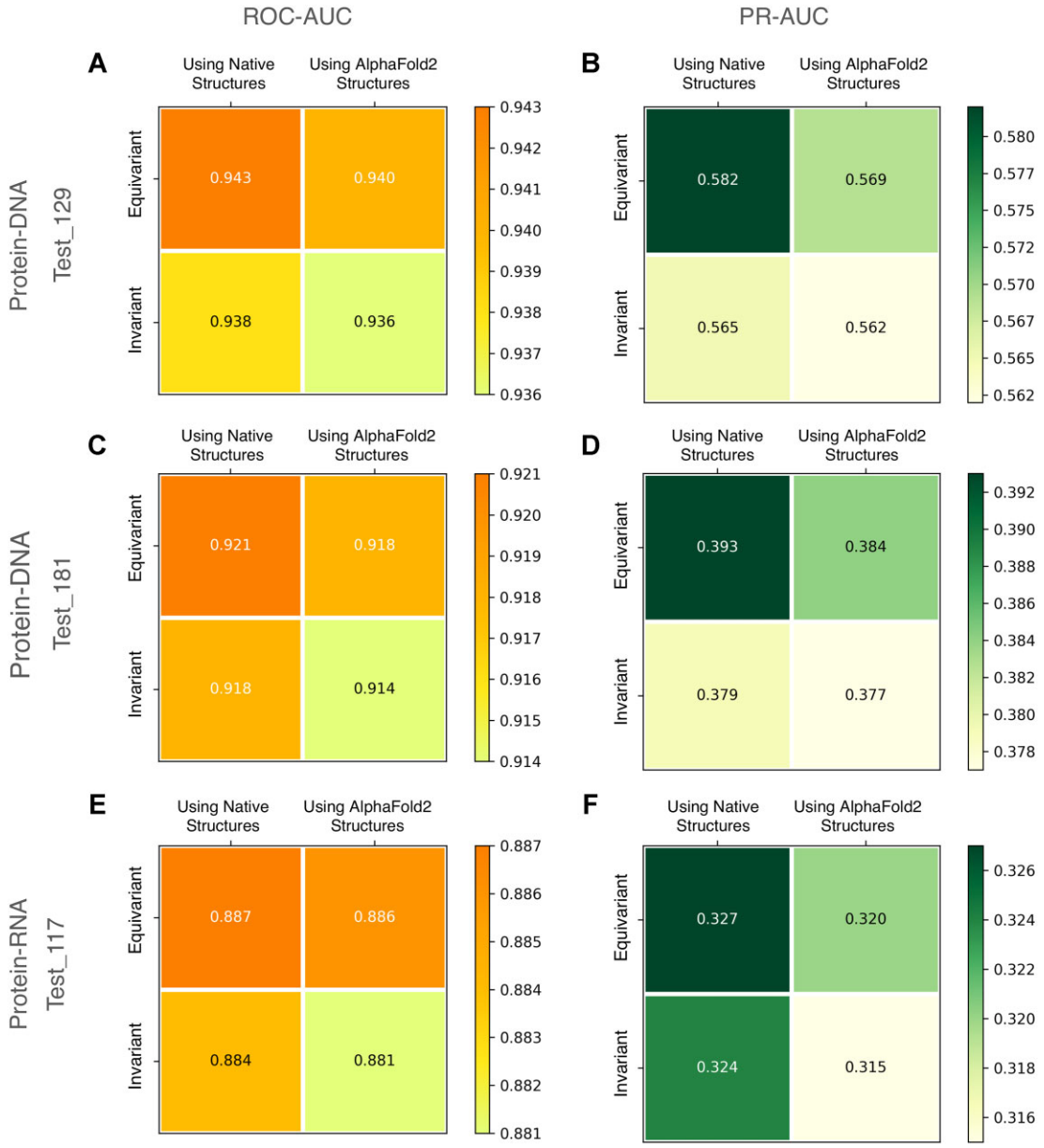


Figure 5. The performance of equivariant networks used in the full-fledged version of EquiPNAS compared against the invariant baseline networks using both experimental (native) and AlphaFold2 predicted structures as input. ROC-AUC and PR-AUC for protein–DNA test set Test_129 are presented in (**A, B**); ROC-AUC and PR-AUC for protein–DNA test set Test_181 are presented in (**C, D**); ROC-AUC and PR-AUC for protein–RNA test set Test_117 are presented in (**E, F**).

Discussion

This work presents EquiPNAS, a new pLM-informed equivariant deep graph neural network framework for accurate protein–nucleic acid binding site prediction. We demonstrate that EquiPNAS consistently outperforms the state-of-the-art methods on both protein–DNA and protein–RNA binding site prediction tasks. A major contribution of our work is the successful utilization of protein language model (pLM) embeddings, a previously unexplored avenue in the context of protein–DNA and protein–RNA binding site predictions. Our ablation study reveals that the pLM embeddings are sufficiently powerful that can dramatically reduce the dependence on the availability of evolutionary information which is not always abundant such as with orphan proteins or rapidly evolving proteins, enabling us to build generalizable models. Moreover, despite being trained on experimental structures as input, our method exhibits remarkable robustness and performance resilience by attaining high predictive accuracy even when AlphaFold2 predicted structural models are used as input, dramatically enhancing the scalability of protein–nucleic acid binding site prediction without compromising on accuracy. Through controlled experiments, we directly verify that the symmetry-aware nature of the E(3) equivariant graph-based framework is a major driving force behind the improved performance of EquiPNAS, particularly when predicted structures are used as input.

While this work focuses on partner-independent protein–nucleic acid binding site prediction, that is, predicting the binding sites based only upon the surface of an isolated

protein without any prior knowledge about the interacting nucleic acid partner; incorporating additional information regarding the DNA or RNA molecules interacting with the protein may lead even more accurate binding sites prediction. Beyond the realm of binding site prediction, a promising direction for future work is to develop accurate, robust, and scalable computational approaches for protein–DNA or protein–RNA complex structure modeling, capturing protein–DNA and protein–RNA interactions at the atomic level. In this regard, the predicted protein–nucleic acid binding sites can serve as additional restraints, alongside physics- and/or knowledge-guided force fields, to facilitate more efficient and accurate protein–DNA or protein–RNA complex structure modeling. The predicted binding site information can complement and supplement the existing force fields as an additional scoring term to efficiently navigate the conformational space accessible to protein–nucleic acid complexes, leading to improved predictive modeling.

Data availability

The raw data used in this study, including the datasets for train, test and validation are collected from publicly available sources and freely available at http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/ and https://github.com/biomed-AI/GraphSite.

Code availability. An open-source software implementation of EquiPNAS, licensed under the GNU General Public License v3, is freely available at https://github.com/Bhattacharya-Lab/EquiPNAS.

Supplementary data

Supplementary Data are available at NAR Online.

Funding

National Institute of General Medical Sciences [R35GM138146 to D.B.]; National Science Foundation [DBI2208679 to D.B.]. Funding for open access charge: National Institute of General Medical Sciences [R35GM138146 to D.B.]

Conflict of interest statement

None declared.

References

1. Hirota,K., Miyoshi,T., Kugou,K., Hoffman,C.S., Shibata,T. and Ohta,K. (2008) Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature*, **456**, 130–134.
2. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.
3. Zhang,J. and Kurgan,L. (2019) SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics*, **35**, i343–i353.
4. Zhao,H., Yang,Y. and Zhou,Y. (2010) Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. *Bioinformatics*, **26**, 1857–1863.
5. Konc,J., Hodošček,M., Ogrizek,M., Trykowska Konc,J. and Janežič,D. (2013) Structure-based function prediction of uncharacterized protein using binding sites comparison. *PLoS Comput. Biol.*, **9**, e1003341.
6. Schmidtke,P. and Barril,X. (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.*, **53**, 5858–5867.
7. 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**, 930–936.
8. Zhang,J., Chen,Q. and Liu,B. (2021) NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. *Briefings Bioinf.*, **22**, bbaa397.
9. 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. *J. Chem. Inf. Model.*, **59**, 3057–3071.
10. Zhang,J., Ghadermarzi,S., Katuwawala,A. and Kurgan,L. (2021) DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences. *Briefings Bioinf.*, **22**, bbab336.
11. Walia,R.R., Xue,L.C., Wilkins,K., El-Manzalawy,Y., Dobbs,D. and Honavar,V. (2014) RNABindRPlus: A predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS One*, **9**, e97725.
12. Armon,A., Graur,D. and Ben-Tal,N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.
13. Hu,J., Li,Y., Zhang,M., Yang,X., Shen,H.-B. and Yu,D.-J. (2016) Predicting protein–DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **14**, 1389–1398.
14. 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 Trans. Comput. Biol. Bioinf.*, **10**, 994–1008.
15. 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 Res.*, **46**, W438–W442.
16. 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 Struct. Funct. Bioinf.*, **81**, 1885–1899.
17. Jiménez,J., Doerr,S., Martínez-Rosell,G., Rose,A.S. and De Fabritiis,G. (2017) DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, **33**, 3036–3042.
18. Li,S., Yamashita,K., Amada,K.M. and Standley,D.M. (2014) Quantifying sequence and structural features of protein–RNA interactions. *Nucleic Acids Res.*, **42**, 10086–10098.
19. 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. *Nat. Commun.*, **10**, 4941.
20. 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 Res.*, **49**, e51.
21. 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 Bioinf.*, **23**, bbab564.
22. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
23. 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**, 583–589.
24. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2021) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
25. Elnaggar,A., Heinzinger,M., Dallago,C., Rihawi,G., Wang,Y., Jones,L., Gibbs,T., Feher,T., Angerer,C., Steinegger,M., *et al.* (2020) ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv doi: https://arxiv.org/abs/2007.06225, 04 May 2021, preprint: not peer reviewed.
26. Brandes,N., Ofer,D., Peleg,Y., Rappoport,N. and Linial,M. (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, **38**, 2102–2110.
27. Ferruz,N., Schmidt,S. and Höcker,B. (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.*, **13**, 4348.
28. Chowdhury,R., Bouatta,N., Biswas,S., Floristean,C., Kharkar,A., Roy,K., Rochereau,C., Ahdritz,G., Zhang,J., Church,G.M., *et al.* (2022) Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.*, **40**, 1617–1623.
29. Rives,A., Meier,J., Sercu,T., Goyal,S., Lin,Z., Liu,J., Guo,D., Ott,M., Zitnick,C.L., Ma,J., *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2016239118.
30. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
31. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,L. and Polosukhin,I. (2017) Attention is all you need. arXiv doi: https://arxiv.org/abs/1706.03762, 02 August 2023, preprint: not peer reviewed.
32. Madani,A., Krause,B., Greene,E.R., Subramanian,S., Mohr,B.P., Holton,J.M., Olmos,J.L., Xiong,C., Sun,Z.Z., Socher,R., *et al.* (2023) Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.*, **41**, 1099–1106.
33. Horne,J. and Shukla,D. (2022) Recent advances in machine learning variant effect prediction tools for protein engineering. *Ind. Eng. Chem. Res.*, **61**, 6235–6245.
34. Garcia Satorras,V., Hoogeboom,E. and Welling,M. (2021) E(n) equivariant graph neural networks. arXiv doi: https://arxiv.org/abs/2102.09844, 16 February 2022, preprint: not peer reviewed.
35. Roche,R., Moussad,B., Shuvo,M.H. and Bhattacharya,D. (2023) E (3) equivariant graph neural networks for robust and accurate protein–protein interaction site prediction. *PLoS Comput. Biol.*, **19**, e1011435.
36. Moussad,B., Roche,R. and Bhattacharya,D. (2023) The transformative power of transformers in protein structure prediction. *Proc. Natl. Acad. Sci. U.S.A.*, **120**, e2303499120.
37. Hie,B.L., Shanker,V.R., Xu,D., Bruun,T.U., Weidenbacher,P.A., Tang,S., Wu,W., Pak,J.E. and Kim,P.S. (2023) Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-023-01763-2.
38. Wu,F., Wu,L., Radev,D., Xu,J. and Li,S.Z. (2023) Integration of pre-trained protein language models into geometric deep learning networks. *Commun. Biol.*, **6**, 876.
39. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
40. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H. and the UniProt Consortium (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
41. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S. and Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nat. Methods*, **19**, 679–682.
42. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
43. Jing,B., Eismann,S., Suriana,P., Townshend,R.J.L. and Dror,R. (2020) Learning from protein structure with geometric vector perceptrons. arXiv doi: https://arxiv.org/abs/2009.01411, 16 May 2021, preprint: not peer reviewed.
44. Paszke,A., Gross,S., Massa,F., Lerer,A., Bradbury,J., Chanan,G., Killeen,T., Lin,Z., Gimelshein,N., Antiga,L., *et al.* (2019) PyTorch: an imperative style, high-performance deep learning library. arXiv doi: https://doi.org/10.48550/arXiv.1912.01703, 03 December 2019, preprint: not peer reviewed.
45. Wang,M., Zheng,D., Ye,Z., Gan,Q., Li,M., Song,X., Zhou,J., Ma,C., Yu,L., Gai,Y., *et al.* (2019) Deep graph library: a graph-centric, highly-performant package for graph neural networks. arXiv doi: https://doi.org/10.48550/arXiv.1909.01315, 25 August 2020, preprint: not peer reviewed.
46. Loshchilov,I. and Hutter,F. (2016) SGDR: stochastic gradient descent with warm restarts. arXiv doi: https://doi.org/10.48550/arXiv.1608.03983, 03 May 2017, preprint: not peer reviewed.
47. Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. arXiv doi: https://doi.org/10.48550/arXiv.1412.6980, 30 January 2017, preprint: not peer reviewed.
48. Yang,J., Roy,A. and Zhang,Y. (2012) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.
49. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
50. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
51. Remmert,M., Biegert,A., Hauser,A. and Söding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
52. Cho,K., van Merrienboer,B., Gulcehre,C., Bahdanau,D., Bougares,F., Schwenk,H. and Bengio,Y. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv doi: https://doi.org/10.48550/arXiv.1406.1078, 03 September 2014, preprint: not peer reviewed.
53. Anderson,T.W. and Darling,D.A. (1952) Asymptotic theory of certain "Goodness of Fit" criteria based on stochastic processes. *Ann. Math. Stat.*, **23**, 193–212.
54. Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics Bull.*, **1**, 80–83.
55. Li,P. and Liu,Z.-P. (2023) GeoBind: segmentation of nucleic acid binding interface on protein surface with geometric deep learning. *Nucleic Acids Res.*, **51**, e60.

Received: September 13, 2023. **Revised:** December 22, 2023. **Editorial Decision:** January 6, 2024. **Accepted:** January 11, 2024