Acta Pharmacologica Sinica

www.nature.com/aps

ARTICLE

Graph-based deep learning approach for high-throughput protein-DNA interaction scoring

Yi-hao Zhao[1], Ying Wang[1], Chao Shen[2], De-jun Jiang[3], Shu-kai Gu[1], Hui-feng Zhao[1], Zi-yi You[1], Ting-jun Hou[1,4]✉ and Yu Kang[1,4]✉

Accurately quantifying protein-DNA interactions (PDIs) is critical for understanding biological processes and facilitating drug design. However, the inherent flexibility of nucleic acids limits the availability of experimentally determined structures of PDI complexes, posing a significant challenge for training reliable scoring functions (SFs). To address this, we developed PDIScore, a novel deep learning-based SF for PDI prediction. PDIScore utilizes a comprehensive graph representation to capture nucleotide flexibility, employs a scalable GraphGPS architecture with BigBird linear global attention to handle large interaction interfaces, and leverages Mixture Density Networks (MDNs) to model residue-nucleotide distance distributions. PDIScore was trained on a self-collected dataset of ~7000 protein-nucleic acid complex structures and validated on three rigorous test sets for evaluating its screening, docking, and ranking capabilities. The results illustrated that PDIScore significantly outperformed existing methods: it achieved the best screening power on the screening set (e.g., $EF_{1\%} = 14.13$, AUROC = 0.82 using AlphaFold3 structures), the highest docking success rate on the docking set (48.94% top1), and superior ranking capability on the ranking set (PCC = 0.50). Case studies demonstrated PDIScore's ability to elucidate biological mechanisms (e.g., adenovirus transcription, SOCS1 regulation) and its interpretability at the nucleotide level for identifying key interaction sites. PDIScore represents a robust, generalizable tool with significant potential for advancing PDI-related research and therapeutic design.

Key words: protein-DNA interactions; machine learning; deep learning; molecular docking; virtual screening

Acta Pharmacologica Sinica (2025) 0:1–13; https://doi.org/10.1038/s41401-025-01688-3

INTRODUCTION

Protein-DNA interactions (PDIs) are fundamental to many biological processes, including DNA replication, RNA transcription, gene repair, and gene regulation [1, 2]. PDIs are also related to various diseases, such as inflammation, cancer and Alzheimer's disease [3–5]. Understanding these interactions offers valuable insights into the mechanisms of life, disease pathways, and drug discovery [6–8]. The former two are typically associated with transcription factors (TFs), such as TATA-box-binding protein (TBP) and early growth response protein 1 (Egr1), which specifically recognize DNA sequences to regulate the transcription of the associated genes [9, 10]. For instance, during the adenovirus replication cycle, TBP binds to the TATA box of the adenovirus major late promoter (AdMLP), thereby activating the transcription of adenovirus RNA [11]. With respect to disease pathways, Egr1 regulates the transcription of suppressor of cytokine signaling-1 (SOCS1), the dysregulation of which can lead to immune defects, including excessive inflammation, autoimmune conditions, and malignancies [12, 13]. As for PDI-related drugs, there are protein-targeted therapies, such as DNA aptamer [14, 15], and DNA-targeted drugs, such as cyclic peptides [16]. Furthermore, synthetic DNAs engineered for high affinity and specificity to proteins are being developed as drugs, diagnostic tools, and antagonists to elucidate the role of specific proteins [17].

Quantitative analysis of PDIs in high-throughput studies is essential for the rational design of synthetic DNA. However, experimental methods such as electrophoretic mobility shift assay [18], isothermal titration calorimetry [19], and surface plasmon resonance [20] face significant challenges in measuring protein-DNA binding affinity quickly and accurately. These challenges arise from the structural end-effects of short DNA and the abundance of nonspecific binding sites of long DNA [18]. Moreover, these methods are typically time-intensive and costly, especially when applied to high-throughput experiments [21]. Therefore, there is a passing need for computational methods that can rapidly estimate protein-DNA binding, including both sequence-based and structure-based methods. The inherent flexibility and diverse conformations of nucleic acids complicate accurate binding affinity predictions using sequence-based methods. Consequently, structure-aware methods are generally expected to yield higher prediction accuracy compared to sequence-based methods [22].

A number of structure-based computational scoring functions (SFs) have been developed for predicting protein-DNA binding affinity [21, 23–28]. However, the inherent flexibility of nucleic

[1]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China; [2]Department of Clinical Pharmacy, the First Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou 310003, China; [3]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410004, China and [4]Zhejiang Provincial Key Laboratory for Intelligent Drug Discovery and Development, Jinhua 321016, China
Correspondence: Ting-jun Hou (tingjunhou@zju.edu.cn) or Yu Kang (yukang@zju.edu.cn)
These authors contributed equally: Yi-hao Zhao, Ying Wang.

Received: 14 July 2025 Accepted: 2 October 2025
Published online: 01 December 2025

acids leads to a scarcity of experimentally determined structures with known binding affinities, posing a significant challenge for existing methods, which typically rely on datasets containing fewer than 500 entries [29]. Traditional approaches, such as MM/GBSA [25], FoldX [23, 24], and ref2015 and dna_gb in PyRosetta [26], generally quantify PDIs as a weighted sum of multiple empirical energy terms. The advent of machine learning (ML) algorithms has facilitated the development of a pool of ML-based SFs (MLSFs) that can learn directly from data, represented by PreDBA [27], emPDBA [28], and SAMPDI-3D [21]. Although these methods demonstrate impressive performance on their internal datasets, the limited size of these datasets and the lack of external validation or systematic evaluation may compromise their reliability [29]. For example, the largest dataset of SAMPDI-3D included 463 mutants for training [29]. Based on our evaluation, most of these SFs struggle with ranking a series of known binders (ranking power), and some can even hardly distinguish active binders from non-binders or weakly active binders (screening power).

The ability to identify native-like binding poses from incorrect ones (docking power) is also an important criterion for evaluating a SF. The aforementioned methods need complex structures for scoring, but the flexibility of nucleic acids often limits the availability of crystal structures, necessitating the use of docking techniques to predict protein-DNA complex structures. Representative docking methods include HDOCK [30], PyDockDNA [31], and HADDOCK [32, 33]. Meanwhile, artificial intelligence (AI)-based structure prediction methods like AlphaFold [34] have demonstrated remarkable success in protein structure prediction, and its latest version, AlphaFold3 [35], can even predict protein-DNA complexes directly from sequences. These methods have internal SFs to rank and filter multiple docking poses. The capability of SFs to distinguish native-like binding poses from incorrect ones is crucial for accurate structure prediction and subsequent binding affinity prediction.

To develop a more reliable SF, we curated an extensive dataset of ~7000 protein-nucleic acid complex structures sourced from Protein Data Bank (PDB) [36]. To the best of our knowledge, such a large-scale structure dataset has never been reported, and it can be served as a pioneering source for training predictive models, thereby advancing the study of PDIs. To comprehensively evaluate existing PDI SFs, we also compiled three validation sets, including (1) a screening set that involves 5 systems and a total of ~28000 protein-DNA complexes with the binding affinities determined by the same research group using the same experimental method [37], (2) a published docking benchmark with 47 unbound protein-DNA complex structures [38], and (3) a ranking set that contains 9 systems and ~200 protein-DNA complexes, with measured binding affinities that cover both protein mutations and DNA mutations [21, 39–42]. Due to the known binding affinities of each ligand in the screening set, it can also be utilized to assess the ranking power. Relying on these datasets, we presented a deep learning (DL)-based approach named PDIScore for PDI prediction. Our model architecture comprises four primary modules: graph representation, feature extraction, concatenation, and mixture density network (MDN). Given the higher conformational flexibility of nucleic acids compared to proteins, we used a wider array of structural descriptors (11 atom distances and 20 dihedral angles) into the graph representation. For feature extraction, we used the general, powerful, scalable (GPS) graph [43] with linear global attention provided by BigBird [44] to replace the fully-connected graph transformer (GT) used in our previous studies [45, 46]. This featurization strategy is scalable to graphs with several thousand nodes, making it particularly suitable for this study. Compared with protein-small molecule interactions, PDIs typically have larger contact areas, which are reflected in larger graphs with more nodes and edges during feature extraction. The MDN modules are utilized to learn the probability density distribution of the distance between each residue and each nucleic acid. Overall, PDIScore is trained on the structure dataset and rigorously tested on the screening, docking and ranking sets, demonstrating its ability to achieve an improved correlation with experimental measurements and distinguish native-like structures from decoys. PDIScore also demonstrates its utility as a rescoring tool for AlphaFold3, offering valuable support for PDI-related drug design.

MATERIALS AND METHODS

Dataset preparation

A total of four datasets, i.e., structure dataset, screening set, docking set, and ranking set, were curated in this study. The structure dataset was used for model training, and its distribution is shown in Fig. S1. The remaining three sets were used as the test sets for model evaluation and are detailed in Fig. S2 and Table S1.

The structure dataset consisted of 7108 protein-nucleic acid complex structures retrieved from PDB (before October 16, 2023). Among them, 532 protein-DNA complexes were labeled with affinity data from the PDBbind database (v2020) [47] to fine-tune the model, forming the affinity dataset. As shown in Fig. S1, the distribution of the structure dataset closely resembled that of the affinity dataset. Approximately 86% of the complexes in the structure dataset and about 97% of the complexes in the affinity dataset contained fewer than 60 nucleotides. Both datasets shared the same top four most prevalent protein types: transferase, transcription, DNA binding protein, and hydrolase.

The screening set contained ~28,000 protein-DNA complexes for five protein targets, including TBP, Ets1, Egr1, Max, and GR, whose wild-type crystal structures bound with DNA had been determined (PDB entries: 1QNE, 2NNY, 1P47, 1AN2, and 1R4R, respectively). For each specific protein target, the binding affinities of equal-length DNA sequences were measured by the same research group using the same bioassay. The mutants with unavailable crystal structures were either obtained by point mutation from the corresponding wild type or predicted by AlphaFold3. For point mutation, the reference crystal structures were first mutated using Chimera [48] based on the DNA sequence information, followed by the minimization of the whole system using Rosetta [49]. For Alphafold3, the sequences of DNA and protein were input to predict the corresponding structures. To ensure a fair comparison with the SF embedded in AlphaFold3, the predicted structures were not processed with energy minimization. The top 1% DNAs ranked by experimental binding data were regarded as the active ligands for a specific protein. The others were regarded as the decoys. The major aim of this set was to test whether the approach could enrich these ligands from the entire library. Additionally, the binding preference for each DNA between Egr1 and Max was available, which could be used as a small-scale reverse screening test to estimate whether the approach could reproduce these preferences.

The docking set was directly retrieved from a published non-redundant protein-DNA docking benchmark, which contained 47 unbound-unbound test cases. The unbound structure was defined as a conformation that existed in the absence of binding partners or within a different complex. The unbound DNA structures were generated using the program 3DNA [50] and the unbound protein structures were determined by X-ray or NMR. The benchmark was redundancy reduced using the program MMseqs2 [51], removing the complexes whose sequences were >40% similar to those in the structure dataset. Following this process, 41 complexes remained, forming the de-redundant dataset.

The ranking set was employed to evaluate the capability of each approach to capture changes in binding free energy (ΔΔG) resulting from different mutations, covering both protein and DNA mutations. We selected systems from the test sets of SAMPDI-3D, each containing more than 15 mutation ΔΔG values. To ensure a fair comparison and provide supplementary data, we also

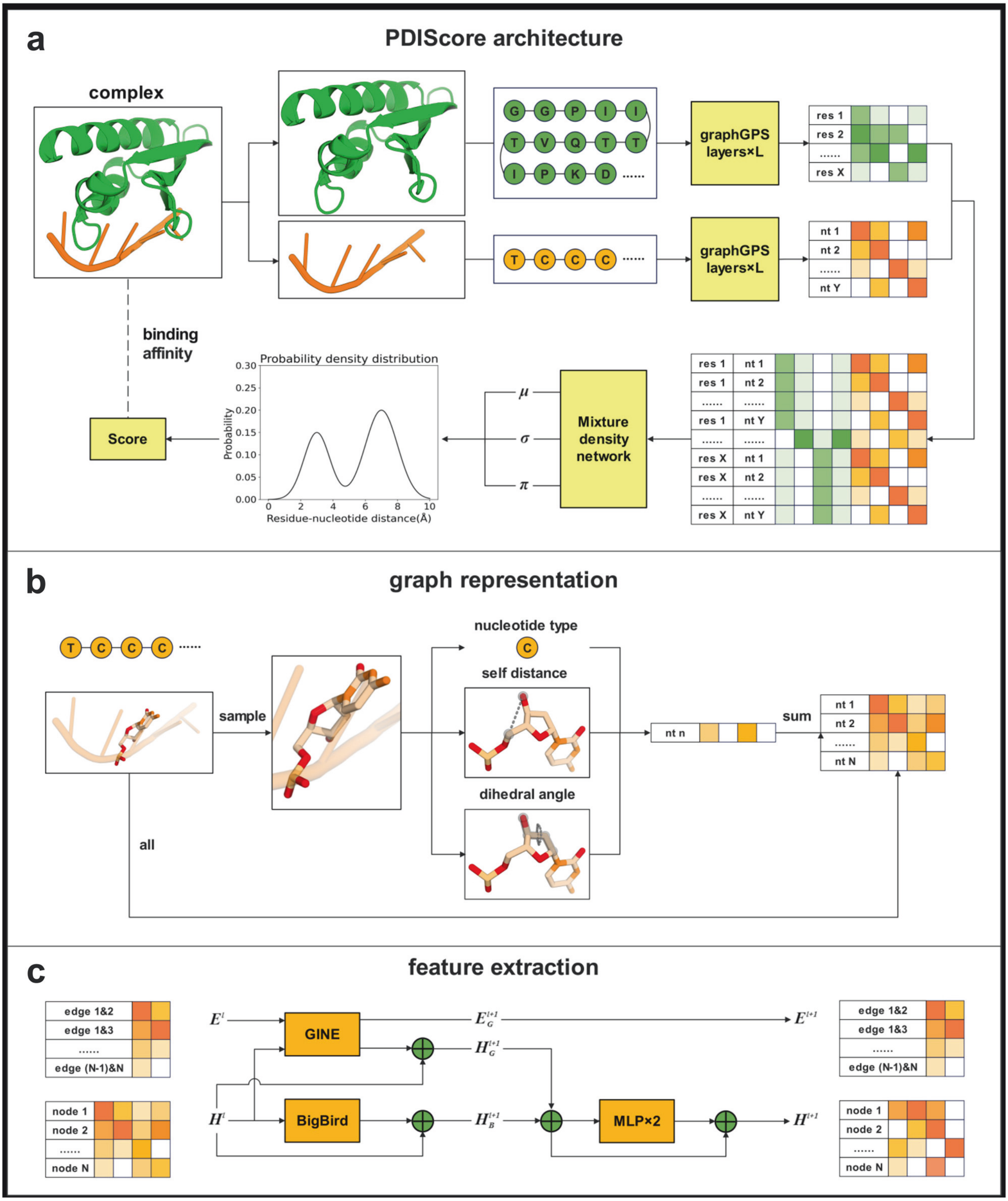

Fig. 1 The model architecture of PDIScore. a Each protein-nucleic acid complex is separated into protein and nucleic acid, which are then fed to independent graphGPS layers at the residue (res) level and the nucleotide (nt) level, respectively. Each graphGPS layer primarily consists of a GINE module and a BigBird module for updating graph representations. The final node representations of protein and nucleic acid are concatenated and then input to an MDN to generate probability density distributions. These distributions are ultimately assembled into a score which indicates the binding affinity of protein-nucleic acid complex. **b** In the graph representation, the node features of nucleic acids comprise nucleotide types, self_distances, and dihedral_angles. **c** In the feature extraction, the node features are aggregated by the output of BigBird module and GINE module, followed by the processing of Multilayer Perceptron (MLP) module.

identified systems that were either entirely or partially absent from the training sets of SAMPDI-3D, thus leaving nine proteins with resolved DNA-bound crystal structures, including CEBPB, MafB, ELK1, ETV5, ERG, Tral, CAP, PadR, and c-Fos-c-Jun (PDB entries: 1GU4, 2WTY, 1DUX, 4UNO, 4IRI, 2A0I, 1RUN, 5×11, and 1FOS, respectively). The first five proteins involved DNA mutations, while the remaining four involved protein mutations. The unknown structures were also predicted by point mutation or AlphaFold3, as mentioned above.

Graph representations

The model architecture of PDIScore is depicted in Fig. 1, which consists of graph representation, feature extraction, concatenation, and MDN modules. The graph representations of proteins were referenced from our previous study, RTMScore, which had demonstrated efficacy in predicting protein-small molecule interactions [46]. The residues located within 10.0 Å radius around the co-crystalized DNA were defined as the binding pocket, while the other residues were deemed irrelevant to binding and thereby removed from the structure. Each pocket was then represented as an undirected graph ($G^P = (N^P, E^P)$), with nodes representing the residues in a pocket and edges representing the interactions between any two residues with a minimum distance of less than 10.0 Å. The features of nodes were based on the residue level, which was demonstrated to be effective in our previous study [46]. For edges, we considered the non-covalent interactions that played important roles to maintain the secondary structures and tertiary structures of the protein. The features of protein graph are outlined in Table 1. Specifically, the node features included amino acid type, self-distances, and dihedral angles for each residue, and the edge features included bonding states, CA-to-CA distances, center-to-center distances, and the maximum and minimum distances between any two residues.

Similarly, the graph ($G^{NA} = (N^{NA}, E^{NA})$) for each nucleic acid was constructed at the nucleotide level. In this graph, nodes represent the nucleotides of the entire nucleic acid, and edges represent the interactions between any two nucleotides whose minimum distances are less than 10.0 Å. Table 2 summarizes the features of each nucleic acid graph. The node features include the types, self-distances, and dihedral angles of each nucleotide, whereas the edge features include bonding states, C5′-to-C5′ distances, center-to-center distances, and the maximum and minimum distances between any two nucleotides. Due to the larger number of atoms in nucleotides compared to residues, there are more distances and dihedral angles to consider. In addition to standard nucleotides, we also included three nonstandard nucleotides present in the structure dataset. For the interatomic distances, we calculated the distances between atomic pairs composed of sugar-phosphate backbone atoms (O3′, C3′, C4′, C5′, O5′, P, C2′, C1′, and O4′). To avoid excessive features while minimizing the loss of structural information, we sampled the distances from the top 10 crystal structures ranked by the number of nucleotides and ignored the distances that were relatively evenly distributed among the nucleotides, such as the minimum distance among all interatomic distances (Fig. S3). When calculating dihedral angles, the atom types varied a lot across different nucleobases. Therefore, the features related to dihedral angles primarily involved sugar-phosphate backbone atoms, with the exception of two atoms from nucleobases (atoms N9 and C4 from purines A, G, I, or atoms N1 and C2 from pyrimidines C, T, U). The features were computed using the MDAnalysis package [52], and the graphs were created with the Deep Graph Library (DGL) package [53].

Feature extraction

The protein and nucleic acid graphs shared the same model architecture but employed independent feature extractors to transform the features into their respective hidden representations. For a graph with node features $n_i \in \mathbb{R}^{1\times d_n}$ for node i and edge features $e_{ij} \in \mathbb{R}^{1\times d_e}$ for the edge between node i and its neighboring node j, these features were first embedded into d-dimensional initial hidden representations by linear transformations:

$$h_i^0 = n_i W_n^0 + b_n^0; e_{ij}^0 = e_{ij} W_e^0 + b_e^0 \tag{1}$$

where $W_n^0 \in \mathbb{R}^{d_n\times d}$, $W_e^0 \in \mathbb{R}^{d_e\times d}$, and b_n^0, $b_e^0 \in \mathbb{R}^{1\times d}$ are the weights and biases of the linear layers, respectively; h_i^0 and e_{ij}^0 are the node and edge features in the first hidden layer, respectively.

The output features were then updated through several layers of graphGPS [43]. At each layer of graphGPS, the features were updated by aggregating the output of a message-passing graph neural networks (MPNN) layer and a global attention layer. The MPNN layer operated by aggregating information from the neighbors of a node and subsequently updating its feature representation accordingly. On the other hand, the global attention layer enabled nodes to consider all other nodes in a graph. This architecture design was particularly vital for overcoming the inherent limitations associated with traditional MPNNs, such as over-smoothing and over-squashing. For convenience, in the subsequent sections, the linear transformations were simplified as O:

$$O(i) = iW + b \tag{2}$$

where $W \in \mathbb{R}^{d_i\times d_o}$ and $b \in \mathbb{R}^{1\times d_o}$ are the weight and bias of the linear layer, respectively; $i \in \mathbb{R}^{1\times d_i}$ is the input data; d_i and d_o are the dimensions of input data and output data, respectively.

Now the BigBird operation of the lth layer could be described as follows:

$$ATTN(h_i^l) = \sum_{a=1}^{A} Softmax\left(\left(h_i^l Q_a\right) \left(H_{N(i)}^l K_a\right)^T \right) \cdot H_{N(i)}^l V_a \tag{3}$$

$$\hat{h}_{Bi}^{l+1} = O_{B1}^l (ATTN(h_i^l) + h_i^l) \tag{4}$$

$$h_{Bi}^{l+1} = O_{B2}^l ((ReLU(\hat{h}_{Bi}^{l+1})) + \hat{h}_{Bi}^{l+1} \tag{5}$$

where $H^l = \{h_1^l, \dots \dots, h_N^l\}$ is the set of h_i^l, so $H^l \in \mathbb{R}^{N\times d}$; $N(i)$ denotes the out-neighbors set of node i; Q_a, $K_a \in \mathbb{R}^{d\times m}$ and $V_a \in \mathbb{R}^{d\times d}$ are the ath weights from A-sets of Query, Key and Value weight matrices, respectively; Softmax represents the softmax operation; *ReLU* is a type of nonlinear activation function; O_{B1}^l and O_{B2}^l are two linear transformations with $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^{1\times d}$ in the lth layer of BigBird; $\hat{h}_{Bi}^{l+1}$ and h_{Bi}^{l+1} are the intermediate and final node representations in the (l+1)th layer of BigBird, respectively. In the lth layer of BigBird, the hidden representation h_i^l was initially processed through an attention mechanism based on multi-head attention, as shown in Eq. (3). Then the output was used to calculate the representation for the next layer h_{Bi}^{l+1} by a two-layer fully connected network.

When the BigBird model was used for global attention layer in graphGPS, the graph isomorphism network with edges (GINE) [54] was employed for MPNN layer. The GINE operation of the lth layer could be described as follows:

$$\hat{h}_{Gi}^{l+1} = \sum_{j\in N(i)} ReLU\left(h_j^l + e_{ji}^l\right) + h_i^l \tag{6}$$

$$h_{Gi}^{l+1} = O_{G2}^l ReLU(O_{G1}^l \hat{h}_{Gi}^{l+1}) \tag{7}$$

where $\hat{h}_{Gi}^{l+1}$ and h_{Gi}^{l+1} are the intermediate and final node representations in the (l+1)th layer of BigBird, respectively; O_{G1}^l and O_{G2}^l are two linear transformations with $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^{1\times d}$ in the lth layer of GINE. In the lth layer of GINE module, the hidden representation h_i^l was initially processed through an

Table 1. Node and edge features of protein graphs.

Features	Size	Description
Nodes		
Type	32	residue type ([“GLY”, “ALA”, “VAL”, “LEU”, “ILE”, “PRO”, “PHE”, “TYR”, “TRP”, “SER”, “THR”, “CYS”, “MET”, “ASN”, “GLN”, “ASP”, “GLU”, “LYS”, “ARG”, “HIS”, “MSE”, “CSO”, “PTR”, “TPO”, “KCX”, “CSD”, “SEP”, “MLY”, “PCA”, “LLP”, “metal”, “other”]) with one hot encoding
Self_distance	5	maximum and minimum distances within any atomic pair, distances between the atomic pairs CA-O, O-N, C-N (multiplied by 0.1)
Dihedral_angle	4	dihedral angles phi, psi, omega, and chi1 (multiplied by 0.01)
Edges		
Whether_connected	1	whether two residues are connected
CA_distance	1	distance between the CA atoms of two residues (multiplied by 0.1)
Center_distance	1	distance between the centers of two residues (multiplied by 0.1)
Maximum_distance	2	maximum and minimum distances between two residues (multiplied by 0.1)

Table 2. Node and edge features of nucleic acid graphs.

Features	Size	Description
Nodes		
Type	12	nucleotide type ([“A”, “G”, “C”, “U”, “I”, “N”, “DA”, “DG”, DC”, “DT”, “DU”, “DI”]) with one hot encoding
Self_distance	11	maximum distance within any atomic pair, distances between the atomic pairs O3′-C5′, O3′-O5′, O3′-P, C3′-O5′, C3′-P, C4′-P, O3′-C1′, O3′-O4′, C5′-C1′, C5′-C2′ (multiplied by 0.1)
Dihedral_angle	20	dihedral angles O5′(b)-P(b)-O3′-C3′, P(b)-O3′-C3′-C4′, O3′-C3′-C4′-C5′, C3′-C4′-C5′-O5′, C4′-C5′-O5′-P, C5′-O5′-P-O3′(f), P(f)-O3′-C3′-C2′, O3′-C3′-C2′-C1′, C3′-C2′-C1′-O4′, C2′-C1′-O4′-C4′, C1′-O4′-C4′-C5′, O4′-C4′-C5′-O5′, O3′-C3′-C4′-O4′, C5′-C4′-C3′-C2′, C3′-C4′-O4′-C1′, C4′-C3′-C2′-C1′, C4′-O4′-C1′-N9(N1), C3′-C2′-C1′-N9(N1), O4′-C1′-N9(N1)-C4(C2), C2′-C1′-N9(N1)-C4(C2) (multiplied by 0.01)
Edges		
Whether_connected	1	whether two nucleotides are connected
C5′_distance	1	distance between the atoms C5′ of two nucleotides (multiplied by 0.1)
Center_distance	1	distance between the centers of two nucleotides (multiplied by 0.1)
Maximum_distance	2	maximum and minimum distances between two nucleotides (multiplied by 0.1)

aggregation function, which considered the features of neighboring nodes and the attributions of the edges connecting these nodes, as shown in Eq. (6). Then the output of aggregation function was used to calculate the representation for the next layer h_{Gi}^{l+1} by two linear transformations. The edge features e_{ij}^{l} were used but not updated in GINE layers, also it should be noted that the BigBird layers did not employ or update edge features, which meant $e_{ij}^{l+1} = e_{ij}^{l}$.

In graphGPS, the features were updated by aggregating the output of the BigBird module and GINE module. The graphGPS operation of the lth layer could be described as follows:

$$\begin{aligned}\hat{h}_i^{l+1} &= BatchNorm\left(Dropout\left(\hat{h}_{Bi}^{l+1}\right) + h_i^l\right)\\ &\quad + BatchNorm(Dropout\left(\hat{h}_{Gi}^{l+1}\right) + h_i^l)\end{aligned} \tag{8}$$

$$h_i^{l+1} = Dropout(O_{GPS2}^{l}(Dropout(ReLU(O_{GPS1}^{l}\hat{h}_i^{l+1})))) + \hat{h}_i^{l+1} \tag{9}$$

$$e_{ij}^{l+1} = e_{ij}^{l} \tag{10}$$

where $\hat{h}_i^{l+1}$ and h_i^{l+1} are the intermediate and final node representations in the (l+1)th layer of graphGPS, respectively; O_{GPS1}^{l} and O_{GPS2}^{l} are two linear transformations with $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^{1\times d}$ in the lth layer of graphGPS; Dropout denotes the dropout operation; BatchNorm denotes the batch normalization operation. The node features of BigBird $\hat{h}_{Bi}^{l+1}$ and GINE $\hat{h}_{Gi}^{l+1}$ were both processed through the dropout and batch normalization operations, and subsequently aggregated to form the intermediate representations $\hat{h}_i^{l+1}$, which was then used to calculate the final node representations h_i^{l+1} by two linear transformations with the dropout operations. As for the edge features, they were only passed to the GINE module and were not updated within it. Therefore, the edge representations remained unchanged during the graphGPS operations, as shown in Eq. (10). The outputs h_i^{l+1} and e_{ij}^{l+1} were then fed to the next graphGPS layer until the final graphGPS layer, when the output was input into the feature concatenation module.

Feature concatenation and MDN

In the feature concatenation module, the extracted node representations for protein and nucleic acid were concatenated in pairs to form the input for MDN. The related operations could be described as follows:

$$h_{x,y} = Dropout\left(ELU\left(BatchNorm\left(O_{MDN1}Cat\left(\left[h_x^P, h_y^{NA}\right]\right)\right)\right)\right) \tag{11}$$

$$\mu_{x,y} = ELU\left(O_{MDN2}h_{x,y}\right) + 1 \tag{12}$$

$$\sigma_{x,y} = ELU\left(O_{MDN3}h_{x,y}\right) + 1.1 \tag{13}$$

$$\pi_{x,y} = Softmax\left(O_{MDN4}h_{x,y}\right) \tag{14}$$

where h_x^P and h_y^{NA} are the extracted node representations for protein and nucleic acid, respectively; O_{MDN1} is the linear transformation with $W \in \mathbb{R}^{2d\times d_{x,y}}$ and $b \in \mathbb{R}^{1\times d_{x,y}}$; O_{MDN2}, O_{MDN3}, and O_{MDN4} are three linear transformations with $W \in \mathbb{R}^{d_{x,y}\times d_g}$ and $b \in \mathbb{R}^{1\times d_g}$; *Cat* represents the concatenation operation; Softmax represents the softmax operation; ELU is a type of activation function; the averages $\mu_{x,y}$, standard deviations $\sigma_{x,y}$, and mixing coefficients $\pi_{x,y}$ are the three components that determine the probability density distribution of the distance between protein node *x* and nucleic acid node *y*. The probability density distributions of the distances could be learned from the concatenated features within the MDN.

Model training

Following the parameterization strategy in our previous study [45], we trained the model with two steps: normal training and fine-tuning. During the initial training phase, *α* was set to 0, and the prepared structure set was randomly divided into two subsets: a training set of 6508 complexes and a validation set of 600 complexes. In the subsequent fine-tuning phase, *α* was adjusted to 0.5, and the affinity dataset was also randomly split into a training set of 482 complexes and a validation set of 50.

The loss function was defined as follows:

$$\mathcal{L} = \mathcal{L}_{MDN} + 0.001 \times (\mathcal{L}_{nt} + \mathcal{L}_{bt}) + \alpha\mathcal{L}_{PCC} \tag{15}$$

$$\mathcal{L}_{MDN} = -\log P\left(d_{x,y}|h_x^P, h_y^{NA}\right) = -\log\sum_{n=1}^{d_g}\pi_{x,y,n}N\left(d_{x,y}|\mu_{x,y,n}, \sigma_{x,y,n}\right) \tag{16}$$

$$\mathcal{L}_{nt} = \sum_{y=1}^{Y} Ent\left(O_{type1}h_y^{NA}, N_{y,type}^{NA}\right) \tag{17}$$

$$\mathcal{L}_{bt} = \sum_{yi=1}^{Y}\sum_{yj\in N(yi)}^{Y} Ent\left(O_{type2}Cat\left(h_{yi}^{NA}, h_{yj}^{NA}\right), E_{yij,type}^{NA}\right) \tag{18}$$

$$\mathcal{L}_{PCC} = Cor(BA_{comp}, BA_{expt}) \tag{19}$$

$$BA_{comp} = \sum_{x=1}^{X}\sum_{y=1}^{Y}\log P\left(d_{x,y}|h_x^P, h_y^{NA}\right) \tag{20}$$

where the total loss $\mathcal{L}$ is composed of MDN loss $\mathcal{L}_{\mathrm{MDN}}$, cross-entropy losses $\mathcal{L}_{nt}$ and $\mathcal{L}_{bt}$, and correlation loss $\mathcal{L}_{\mathrm{PCC}}$. Here, *α* is the weight of $\mathcal{L}_{\mathrm{PCC}}$, $d_{x,y}$ is the minimum distance between *x*th residue and *y*th nucleotide, O_{type1} is the linear transformation with $W \in \mathbb{R}^{d\times d_{nt}}$ and $b \in \mathbb{R}^{1\times d_{nt}}$, O_{type2} is the linear transformation with $W \in \mathbb{R}^{2d\times d_{bt}}$ and $b \in \mathbb{R}^{1\times d_{bt}}$, $N_{y,\mathrm{type}}^{\mathrm{NA}}$ represents the type of *y*th nucleotide, $E_{yij,\mathrm{type}}^{\mathrm{NA}}$ represents the type of bond between *yi*th nucleotide and *yj*th nucleotide, *Ent* stands for the cross-entropy operation, *Cor* represents the Pearson correlation coefficient (PCC) operation, and $\mathrm{BA}_{\mathrm{comp}}$ and $\mathrm{BA}_{\mathrm{expt}}$ represent the computed and experimental binding affinities, respectively. The MDN loss $\mathcal{L}_{\mathrm{MDN}}$ is calculated by summing the negative log-likelihood values of all potential residue-nucleotide pairs. The model's capabilities to predict nucleotide types and their bond types could be reflected by the cross-entropy losses $\mathcal{L}_{nt}$ and $\mathcal{L}_{bt}$, respectively. Lastly, the PCC between the computed and experimental binding affinities of a batch of protein-nucleic acid complexes was calculated and shared as the correlation loss $\mathcal{L}_{\mathrm{PCC}}$.

Model evaluation

The model, along with other SFs, were evaluated on the screening, docking and ranking sets. In the screening set, the DNA ligand pools of five protein targets were used for the forward screening and ranking tasks. Among the five protein targets, two targets shared the same DNA ligand pool, and the binding preferences of each ligand were used for the reverse screening task. The forward screening power was measured using the area under the receiver operating characteristic curve (AUROC) and the enrichment factor (EF). EF was calculated as the average percentage of true binders among the top-scoring candidates (1%, 5%, or 10%). The reverse screening power was measured using AUROC. Given the known binding affinities of ligands in the screening set, the ranking power was also measured using PCC and Spearman correlation coefficient (SCC). As for the docking set, the docking power was measured using the success rate (SR), defined as a prediction being successful if at least one of the top-ranked poses had a root-mean-square deviation (RMSD) value of less than 6.0 Å from the native pose. Both DNA RMSD(D) and protein RMSD(P) were required to satisfy the condition. The atom P from DNA and atom CA from protein were used to calculate RMSD(D) and RMSD(P), respectively. To enhance the breadth of our analysis, the ranking set included four DNA targets alongside additional five protein targets. The ranking power for these nine targets was also evaluated using PCC and SCC.

In addition to PDIScore, several other SFs were included as the baselines for comparison. In the screening set, we used three classical SFs, including two version of PyRosetta (ref2015, dna_gb) and FoldX, and two ML-based SFs (PreDBA and emPDBA) for point mutation predictions, and used the internal SF of Alphafold3, PyRosetta (ref2015, dna_gb), and FoldX to re-score the structures obtained through Alphafold3. In the docking set, we utilized three docking methods: HDOCK, PyDockDNA, HADDOCK, along with the AI-based method AlphaFold3, as pose prediction baselines. Of note, AlphaFold3 was limited to 5 predictions and pyDockDNA to 100, while the other approaches could output more than 100 predictions. Considering that a number of 5 might be too smaller for certain programs, except AlphaFold3 that generated only 5 predictions, the number of predicted docking poses for other programs was uniformly set to 100 with all the other default parameter settings. In HADDOCK, residues or nucleotides with a minimum distance of less than 10.0 Å from their binding partner were defined as active residues or nucleotides. Additionally, three classical SFs, including two versions of PyRosetta (ref2015 and dna_gb) and FoldX, were employed as the baselines for pose rescoring. In the ranking set, five classical SFs were used for point mutation predictions, including MM/GBSA based on two different force fields (bsc1 and OL21), two versions of PyRosetta (ref2015, dna_gb), FoldX, and one MLSF (SAMPDI-3D), while four approaches, i.e., the internal SF of Alphafold3, PyRosetta (ref2015 and dna_gb), and FoldX, were employed for Alphafold3 predictions.

RESULTS AND DISCUSSION

PDIScore achieves the best performance on the screening set

In the screening task, the DNA ligand pools of five protein targets were used to test whether the approach could enrich these ligands from the entire library. Additionally, the binding preference for each DNA between Egr1 and Max was available, which could be used as a small-scale reverse screening test to estimate whether the approach could reproduce these preferences. In this study, forward screening refers to screening DNAs against given protein targets, whereas reverse screening involves screening proteins for given DNA targets. For forward screening, the average metrics were calculated across five protein targets, each with varying numbers of DNA ligands. Maintaining a 1:99 ratio of active ligands to decoys for each target system, consistent with the typical small-molecule screening datasets [55], we evaluated the performance of PDIScore. Either point mutation or alphafold3 was used to predict the complex structures for those mutants with

Table 3. Screening and ranking powers of SFs on the screening set.

Methods	Screening				Ranking	
	EF(↑)		AUROC(↑)		PCC(↑)	SCC(↑)
	$EF_{1\%}$	$EF_{5\%}$	Forward	Reverse		
Point Mutation						
PyRosetta#ref2015	6.27	2.37	0.58	0.79	0.31	0.34
PyRosetta#dna_gb	5.64	3.07	0.59	0.80	0.31	0.33
FoldX	0.31	3.71	0.60	0.85	0.30	0.26
PreDBA	0.00	1.92	0.38	0.37	−0.12	−0.08
emPDBA	1.13	0.42	0.60	0.66	0.18	0.22
PDIScore	**6.89**	**4.26**	**0.71**	**0.88**	**0.49**	**0.44**
AlphaFold3						
AlphaFold3	1.13	2.35	0.67	0.72	0.28	0.33
PyRosetta#ref2015	3.33	4.15	0.78	0.62	0.30	0.47
PyRosetta#dna_gb	7.50	5.68	0.80	0.62	0.31	0.49
FoldX	7.62	4.80	0.81	0.72	0.45	0.52
PDIScore	**14.13**	**8.15**	**0.82**	**0.84**	**0.65**	**0.62**

Values in bold represent the optimal performances for each metric, and the same applies to all subsequent tables.

unavailable crystal structures. When using point mutation, as shown in Table 3, our model achieved better performance ($EF_{5\%} = 4.26$, AUROC = 0.71) than all baseline methods (e.g., FoldX: $EF_{5\%} = 3.71$, AUROC = 0.60), highlighting its enhanced capability to discriminate the active ligands from a large pole of inactive or weakly active ligands for a specific target. Out of the five protein targets, the DNA ligand pools for Max and Egr1 were identical. Given the known binding preferences of these DNA ligands between Max and Egr1, these targets were also suitable for reverse screening. Among the DNA ligands, 307 had a stronger binding affinity toward Max, while 1419 preferred Egr1. The AUROC values for reverse screening are shown in Table 3 and Fig. 2. PDIScore achieved the highest AUROC value (0.88), better than FoldX (0.85) and PyRosetta (0.80), demonstrating its potential to discriminate both active DNAs based on specific proteins and active proteins based on specific DNAs. Besides the screening power, we also calculated the linear correlation between the predicted and experimentally measured binding affinities, shown in Table 3 and Table S2. PDIScore achieved the top performance with a PCC value of 0.49 and an SCC value of 0.44, whereas the second-ranked PyRosetta achieved a PCC value of 0.31 and an SCC value of 0.33. The leading metric values indicated that PDIScore could serve as a potential tool for PDI-related screening.

The latest version of AlphaFold, AlphaFold3, had been updated to predict protein-DNA complexes. Here, we evaluated the screening power based on the AlphaFold3 predictions on this screening set (Table 3 and Table S3). The internal SF of AlphaFold3 ($EF_{5\%} = 2.35$, AUROC = 0.67) exhibited comparable performance to PyRosetta#ref2015 ($EF_{5\%} = 2.37$, AUROC = 0.58), PyRosetta#dna_gb ($EF_{5\%} = 3.07$, AUROC = 0.59), and FoldX ($EF_{5\%} = 3.71$, AUROC = 0.60), but fell short of PDIScore ($EF_{5\%} = 4.26$, AUROC = 0.71). These metric values were based on the structures predicted by point mutation. To ensure a fair comparison, the four SFs (PyRosetta#ref2015, PyRosetta#dna_gb, FoldX, and PDIScore) were also applied to the structures predicted by AlphaFold3. All these SFs improved the metric values, with PDIScore delivering the best results ($EF_{5\%} = 8.15$, AUROC = 0.82). Additionally, PDIScore demonstrated the top performance with a PCC value of 0.65 and an SCC value of 0.62 based on the AlphaFold3 predictions (Table 3). The enhanced performance indicates that PDIScore could serve as a reliable rescoring tool for AlphaFold3 in PDI-related screening.

The metric values of PDIScore based on the AlphaFold3 predictions were higher than those based on the point mutation predictions. Therefore, we compared the accuracy between the AlphaFold3 and point mutation predictions for three cases (Fig. S4) from the TBP target with crystal structures (PDB IDs: 1QNB, 1QN3, and 6UEP). For point mutation, the RMSD(D) of 6UEP (1.07 Å) was higher than those of 1QNB (0.34 Å) and 1QN3 (0.33 Å), potentially due to an unmatched base pair (C-C) in 6UEP. According to the complementary base pairing principle, matched base pairs (A-T and C-G) maintained a constant width in the DNA double helix, while unmatched base pairs caused deviations from this uniform width. In point mutation, when mutating DNA structures using Chimera [48], the nucleotide base was replaced by the specified type while the backbone remained unchanged. However, unmatched base pairs may cause backbone variations, which could not be predicted by Chimera [48]. This issue was alleviated using AlphaFold3, achieving an RMSD(D) of 0.30 Å for 6UEP. Similarly, for the other matched cases, AlphaFold3 predictions exhibited lower RMSD(D) values: 0.20 Å for 1QNB and 0.27 Å for 1QN3, compared to point mutation predictions. We hypothesized that AlphaFold3 might yield more accurate predictions for the screening set than point mutation, thereby facilitating more effective screening under the same SF.

PDIScore improves the success rates on the docking set

The tested methods relied on crystal structures, which were not always accessible. To solve this issue, some structure prediction methods were proposed and their performance was assessed on the unbound protein-DNA docking benchmark in this study. Up to 100 predictions were set to be generated by those docking methods for a single task, while AlphaFold3 could only provide 5 predictions through an online service. A docking was considered successful if at least one of the top-ranked predictions exhibited RMSD values below 6.0 Å for both the DNA and the protein, relative to their native poses. For comparison, the success rates of the top 1 and top 5 predictions are shown in Table 4. Among the three docking methods, HADDOCK achieved the highest top 1 success rate (14.89%) and top 5 success rate (40.43%). Among all structure prediction methods, AlphaFold3 achieved the highest top 1 success rate (42.55%) and top 5 success rate (48.94%), showing its strong modeling capability. It should be noted that AlphaFold3's training set included bound structures. However, the performance evaluation of SFs might be more sensitive to the similarity between the test set and the training set of the SFs rather than to the structure prediction methods, such as AlphaFold3. Therefore, we removed the samples from the docking benchmark with sequences that were >40% similar to those in the training set. We then aggregated the predictions from HADDOCK and AlphaFold3 for re-scoring, and the re-scoring success rates of SFs were also shown in Table 4. Compared to HADDOCK and AlphaFold3, FoldX did not improve the success rate for either the top 1 (27.66%) or the top 5 (36.17%) predictions. The use of PyRosetta#dna_gb resulted in a decrease in the top 1 success rate (38.30%) but an increase in the top 5 success rate (51.06%). Employing PDIScore, however, improved both the top 1 (48.94%) and top 5 (57.45%) success rates. It achieved the lowest RMSD(D) (10.15 Å) for the top 1 prediction (Fig. 3b). When removing cases similar to the training set (Table S4), PDIScore retained the highest top 1 (51.22%) and top 5 (58.54%) success rates, demonstrating that the superior docking power of PDIScore was built on its generalization capability.

In Fig. 3, we presented two cases (PDB IDs: 1TRO and 1DIZ) to explain why PDIScore could improve success rates. The RMSD between prediction and crystal structure was calculated as the metric with a prior alignment between the predicted and crystal structures. Both RMSD(D) and RMSD(P) < 6 Å were considered

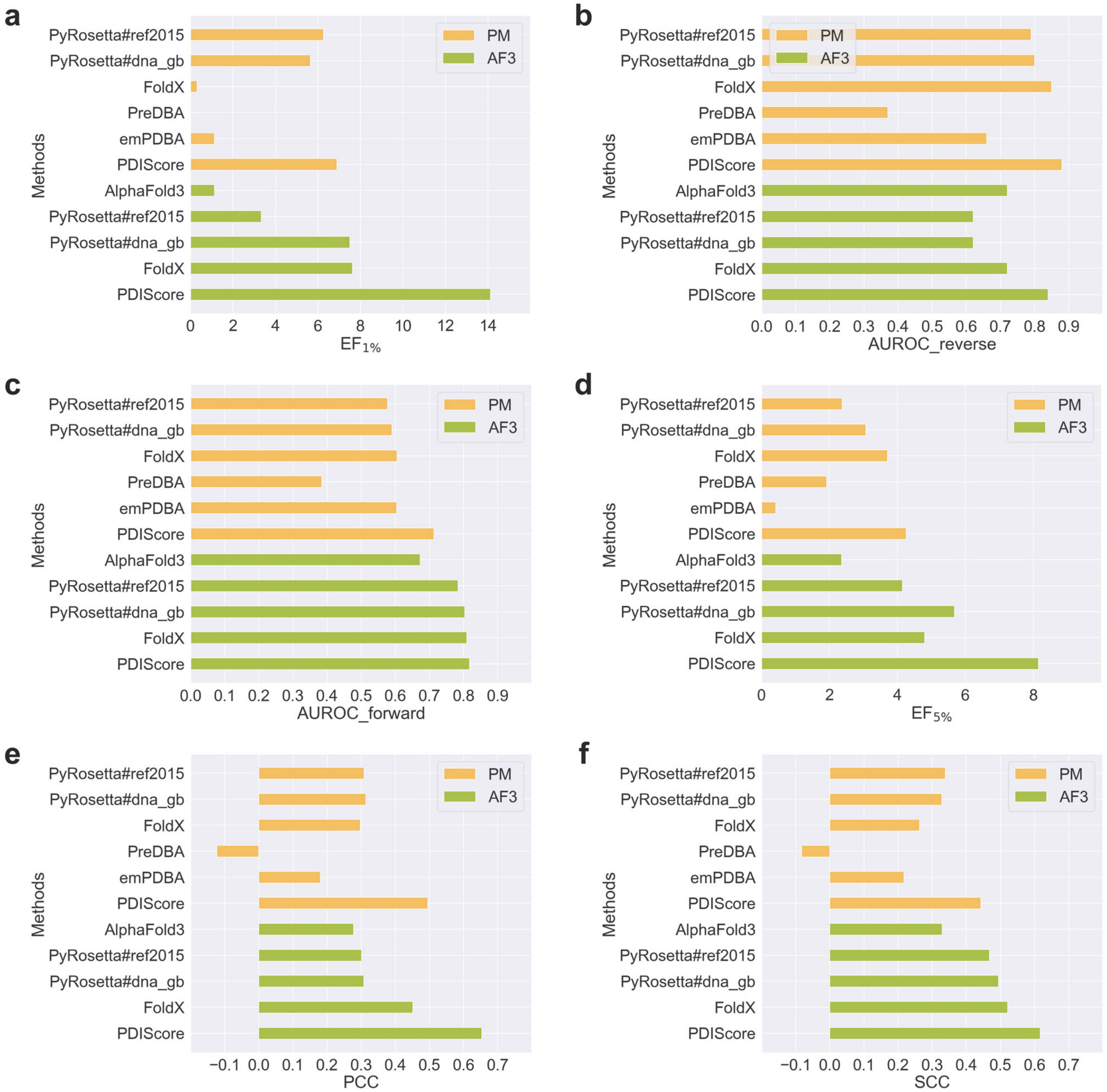

Fig. 2 Screening and ranking powers of SFs on the screening set. Performances of SFs on the screening set based on point mutation (PM, orange) predictions and AlphaFold3 (AF3, green) predictions. The screening powers are in terms of **a** $EF_{1\%}$, **b** $EF_{5\%}$, **c**, **d** AUROC of forward and reverse screening, **e** PCC, and **f** SCC.

successful. The lowest RMSD model of the top 5 predictions is shown in Fig. S5. For PDB 1TRO, the docking methods failed to yield the accurate prediction, whereas AlphaFold3 could generate native-like predictions (RMSD = 1.46 Å). In this case, the HADDOCK predictions could be regarded as decoys in the mixed prediction pool, while PDIScore successfully ranked AlphaFold3 predictions within the top 5, demonstrating its ability to distinguish native-like predictions from decoys. Regarding PDB 1DIZ, AlphaFold3 was unable to obtain the accurate prediction, resulting in a prediction with separate DNA and protein (RMSD = 27.10 Å). The top 5 predictions from docking methods also failed to meet the standard. However, PDIScore could provide the successful model (RMSD = 5.05 Å), showing the advantage of combining and re-scoring the results from different structure prediction methods.

PDIScore retains the best performance on the ranking set

PDIScore excelled in the reverse screening to identify the binding preferences between Egr1 and Max, showing its capability to distinguish active proteins for specific DNA targets. Considering the number of protein ligand was only 2 for each DNA target, we expanded this to over 15 protein ligands. We also selected five protein targets from a previously studied DNA mutant dataset [21]. Both point mutation and alphafold3 were used to predict the complex structures. When using point mutation, we introduced three additional SFs, MM/GBSA#bsc1, MM/GBSA#OL21, and SAMPDI-3D, while excluding two SFs, PreDBA and emPDBA, compared to the screening set. In the screening set, MM/GBSA was not tested due to its substantial computational demands (Table S5), and SAMPDI-3D was not tested because it was only applicable to single-point mutation. In the ranking set, PreDBA and emPDBA were not involved because they struggled with certain targets, yielding identical scores for different ligands. This issue might be caused by the low weights assigned to nucleotide types in the SFs of PreDBA and emPDBA. The PCC and SCC between the predicted and experimentally measured binding affinities were calculated as the evaluation metrics in the ranking set.

As shown in Table 5 and Fig. 4. SAMPDI-3D showed acceptable performance (PCC = 0.41, SCC = 0.37), but PDIScore stood out with superior performance (PCC = 0.50, SCC = 0.48), demonstrating its robustness in ranking different protein mutants. The dataset could be divided into five protein targets and four DNA targets, with the corresponding metrics shown in Table S6–8. MM/GBSA#bsc1 exhibited better performance on the DNA targets (PCC = 0.32, SCC = 0.18) than on the protein targets (PCC = −0.14, SCC = −0.18). Similarly, MM/GBSA#OL21 also showed superior performance on the DNA targets (PCC = 0.31, SCC = 0.25) compared to the protein targets (PCC = 0.14, SCC = 0.08). This discrepancy may be attributed to the choice of force fields; specifically, in MM/GBSA, the same protein force field (ff14SB) was used, while different DNA force fields (bsc1 and OL21) were applied. When calculating DDGs induced by protein mutations, the DNA targets remained the same and so the change of the DNA force field did not affect the final results (indicated by the identical PCC values and similar SCC values). Protein-related interactions were mainly modeled by the protein force field and so the effect of protein mutations was determined by this protein force field. The MM/GBSA results suggested that the protein force field was acceptable but the DNA force fields were inadequate. Notably, among the DNA force fields, OL21 outperformed bsc1 in this assessment. On the contrary, PDIScore performed better on the protein targets (PCC = 0.57, SCC = 0.58) than on the DNA targets (PCC = 0.41, SCC = 0.35). If we removed the redundancy of the training set through the sequence similarity for both proteins and DNAs, 49% of DNAs and only 15% of proteins were retained, indicating that the model learned a broader range of DNA structural information than protein information during training, potentially explaining why PDIScore was better at ranking the DNA ligands for the protein targets.

Table 4. Docking powers of SFs on the docking benchmark.

Methods	Top1 SR(↑)	Top5 SR(↑)	RMSD(D)[a](↓)	RMSD(P)(↓)
HDOCK	0.00	12.77	20.52	2.09
PyDockDNA	2.13	14.89	17.69	1.70
HADDOCK	14.89	40.43	13.80	1.87
AlphaFold3	42.55	48.94	13.86	2.27
PyRosetta#ref2015	23.40	40.43	14.48	2.04
PyRosetta#dna_gb	38.30	51.06	11.79	**1.41**
FoldX	27.66	36.17	15.00	2.04
PDIScore	**48.94**	**57.45**	**10.15**	1.72

[a] The units of RMSD(D) and RMSD(P) are both Å.

For the ranking based on AlphaFold3 predictions, we also evaluated the internal SF of AlphaFold3, PyRosetta#ref2015,

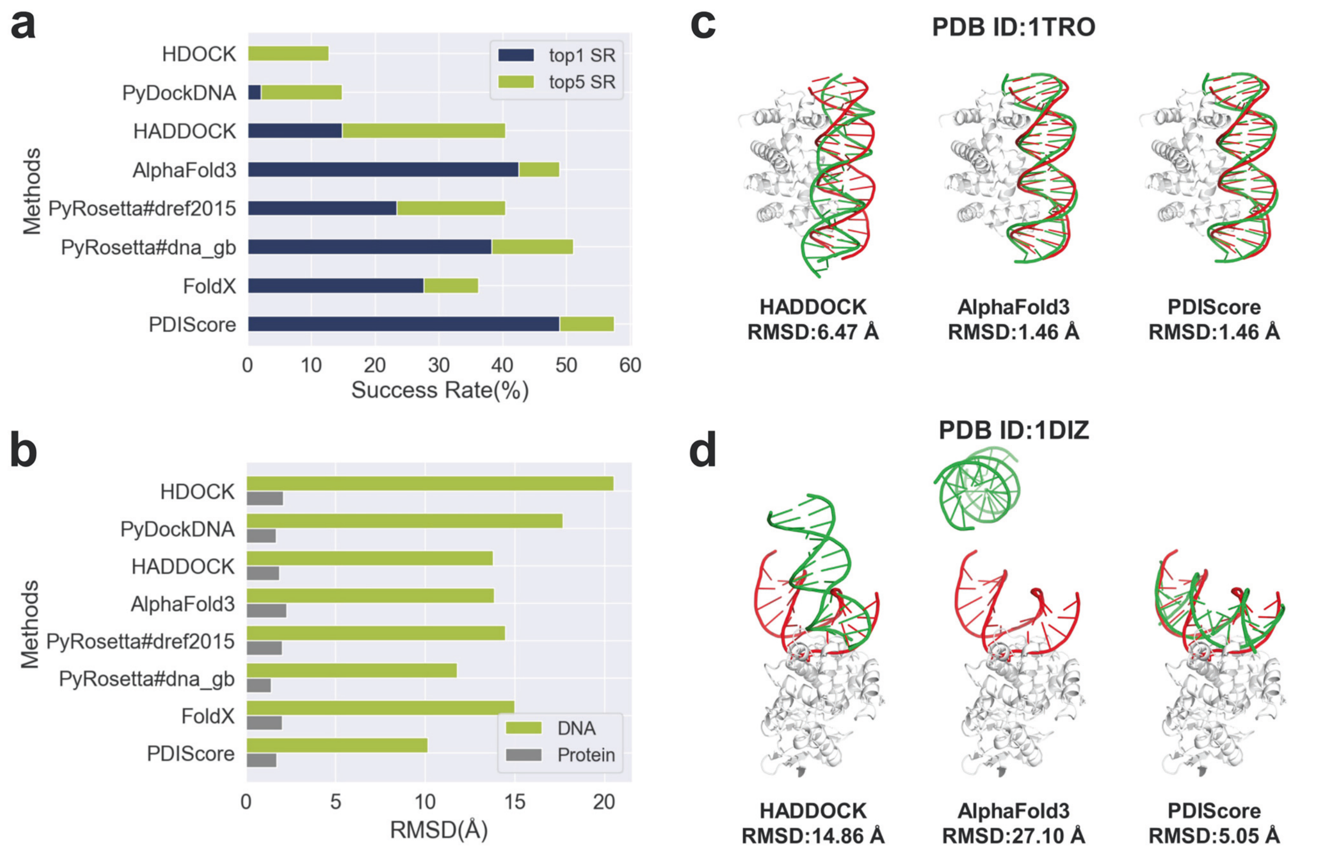

Fig. 3 Docking powers of SFs on the docking benchmark with two cases. a Docking power in terms of the top 1 and top 5 success rates of the tested methods on the docking benchmark. The first four SFs are the internal SFs of the structure prediction methods (HDOCK, PyDockDNA, HADDOCK, and AlphaFold3) tested with their corresponding prediction methods, while the others are external SFs tested with AlphaFold3. **b** The average RMSDs for DNA and protein of the top 1 prediction. Performance of three methods in terms of the lowest RMSD(D) model of the top 5 predictions, including two cases: **c** PDB 1TRO and **d** PDB 1DIZ. The native DNA structures are colored in red and the predicted DNA structures are colored in green.

Table 5. Ranking powers of SFs on the ranking dataset.

Methods	Ranking	
	PCC(↑)	SCC(↑)
Point Mutation		
MM/GBSA#bsc1	0.06	−0.02
MM/GBSA#OL21	0.21	0.16
PyRosetta#ref2015	0.30	0.29
PyRosetta#dna_gb	0.41	0.30
FoldX	0.27	0.22
SAMPDI-3D	0.41	0.37
PDIScore	**0.50**	**0.48**
AlphaFold3		
AlphaFold3	0.12	0.06
PyRosetta#ref2015	0.16	0.16
PyRosetta#dna_gb	0.16	0.17
FoldX	0.09	0.01
PDIScore	**0.37**	**0.32**

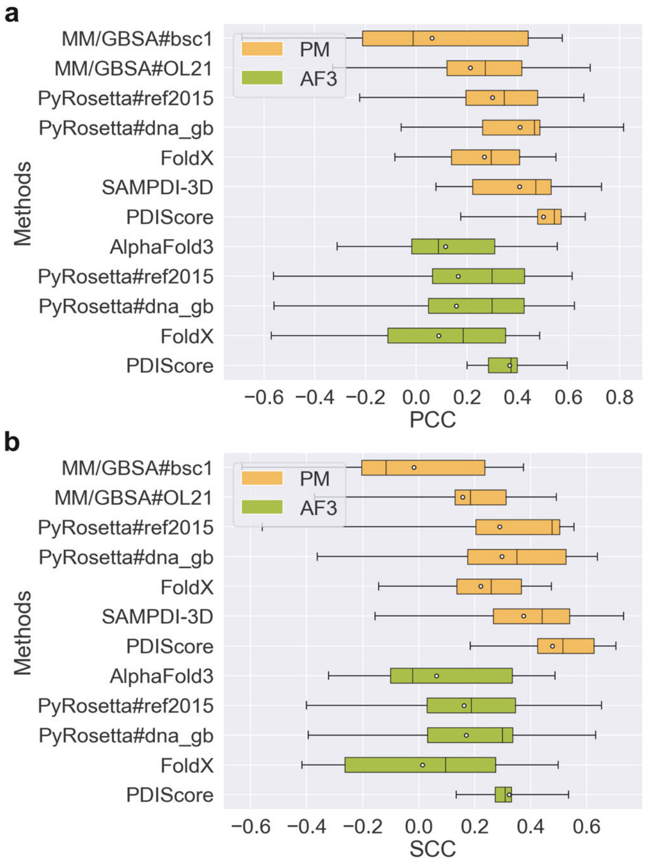

Fig. 4 Ranking powers of SFs on the ranking dataset. Ranking power of SFs based on the point mutation (PM, orange) predictions and AlphaFold3 (AF3, green) predictions, in terms of **a** PCC and **b** SCC.

PyRosetta#dna_gb, FoldX, and PDIScore. PDIScore achieved the best performance (PCC = 0.37, SCC = 0.32), showing its potential as an effective rescoring tool for AlphaFold3. However, for PDIScore, the metric values based on AlphaFold3 predictions were lower than those based on point mutation predictions, in contrast to the trend observed in the screening set. Therefore, we calculated the RMSD(D) between AlphaFold3 predictions and crystal structures for each system (PDB IDs: 1QNE, 2NNY, 1P47, 1AN2, 1R4R, 1GU4, 2WTY, 1DUX, 4UNO, 4IRI, 2A0I, 1RUN, 5×11, 1FOS). On average, the RMSD(D) of the screening systems (0.61 Å) was lower than that of the ranking systems (1.26 Å), suggesting that AlphaFold3 produced more accurate predictions for the screening set. As shown in Fig. S6, the systems with high RMSD(D) tended to show low PCC. The systems with RMSD(D) ≤ 0.8 Å exhibited an average PCC of 0.55, while those with RMSD(D) > 0.8 Å showed an average PCC of 0.32, highlighting the importance of prediction accuracy in achieving reliable scoring performance.

Quantifying PDIs sheds light on the mechanisms of life and disease pathways

According to the screening test, SFs for quantifying PDIs should be beneficial to the discovery of PDI-related drugs. However, to our knowledge, the lack of open-source or accessible relevant datasets has hindered validation in this area. To show the importance of PDI predictions, we presented two examples related to the mechanisms of life and disease pathways. The first case occurs in the adenovirus infection cycle, where AdMLP is essential for gene transcription. The TATA box within AdMLP serves as the TBP binding site, and its mutants cause changes in transcriptional activity [56]. Using the wild-type (wt) box as the template, a mutation from TATA to TCTA reduced specific transcription to 49% of the wt level, while a mutation to TATC decreased it to 21%, as reported. We also found the binding affinities of three sequences to TBP from another report [37], as shown in Table S9. The mutation from TATA to TCTA and TATC led to a reduction in binding affinity to 53% and 37%, respectively. The rank of binding affinities and transcriptional activities among the three sequences remained in the same order, suggesting that the PDI strength relates with the transcriptional activity, which is crucial for virus assembly and maturation. By quantifying PDIs with PDIScore, the predicted order consistently aligned with experimental measurements, demonstrating the capability of PDIScore to recognize the crucial promoter sequences and the possible impacts of their mutations.

The second case is the regulation of SOCS1 by Egr1 [57]. The loss of SOCS1 in tumor cells could upregulate PD-L1 expression and suppress the antitumor response mediated by cytotoxic T lymphocytes (CTLs), thereby increasing tumor aggressivity (Fig. 5). Two Egr1 binding sites are present within the SOCS1 promoter. Mutating these sites lead to a decrease in the transcriptional activity of the promoter [12], as shown in Table S10. The importance of the Egr1 binding sites was further investigated by testing the S1 and S2 mutants within the SOCS1 promoter using a luciferase assay. The observed reduction in luciferase activity in the promoter mutants confirmed the critical role of these binding sites. According to the results of PDIScore, S1wt and S2wt achieved the highest scores. The mutation of either S1 or S2 led to lower scores compared to the wild type, and the mutation of both S1 and S2 resulted in the lowest score. The predictions aligned with the experimental results, suggesting that the promoter mutations could be a possible cause of disease during SOCS1 regulation.

PDIScore is interpretable at the nucleotide level

Besides the interaction between protein and DNA, we might sometimes be curious about the key nucleotides or residues in PDI, as this information could provide guidance in the design of DNAs or proteins. The predicted scores of PDIScore could be decomposed into the contributions from individual nucleotides in a DNA or residues in a protein pocket, since PDIScore was built at the nucleotide-residue level, summing all nucleotide-residue distance likelihood values to obtain the final score. Taking DNA ligands for TBP as an example, the contributions of individual nucleotides in DNAs are shown in Fig. 6, and the darker color of the nucleotide indicated the greater contribution to the total score. The DNA

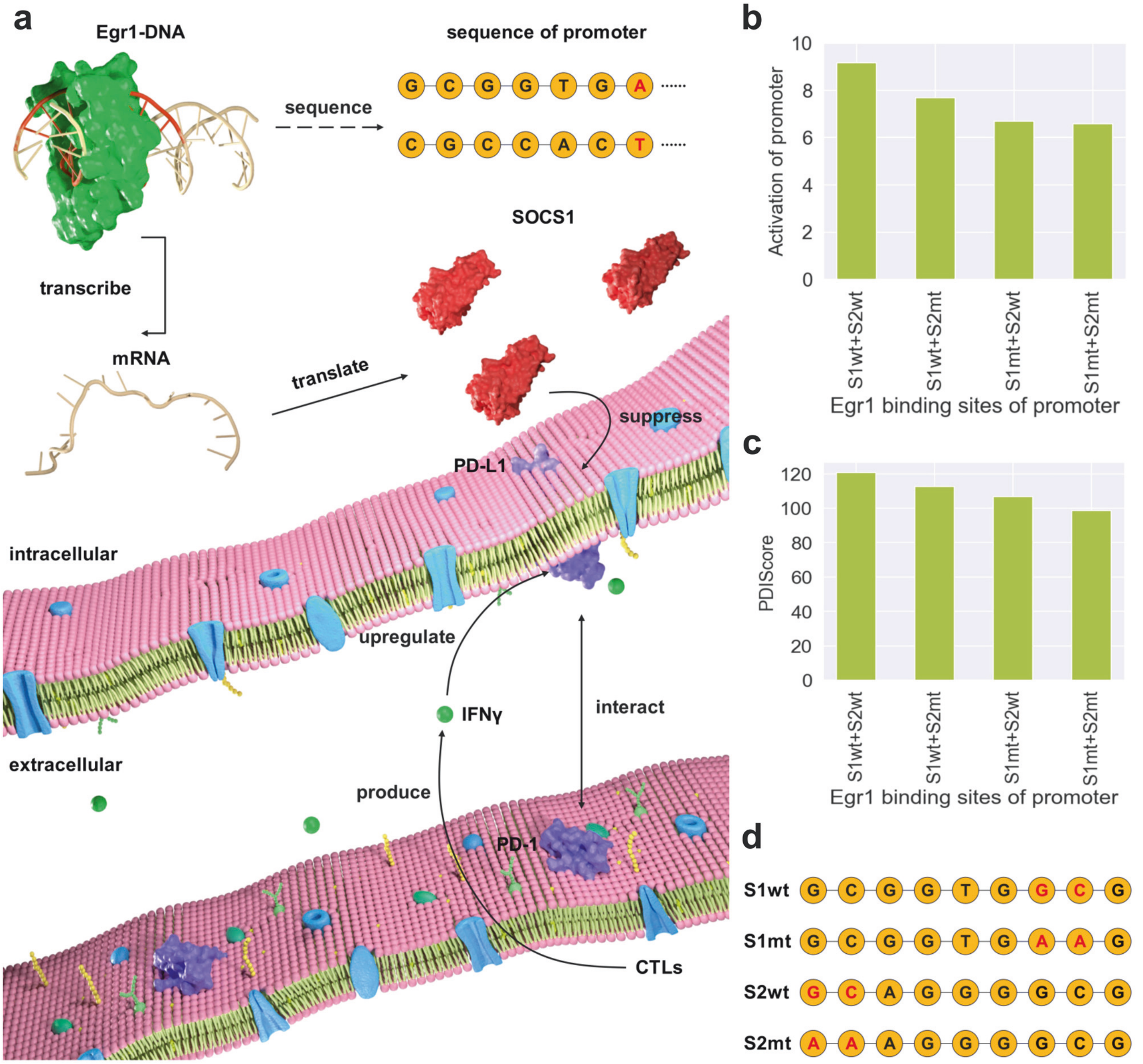

Fig. 5 PDIScore predicts the transcriptional activity of the promoter in the regulation of SOCS1 by Egr1. a The regulation of SOCS1 by Egr1 and the role of SOCS1 in the immune system. During the regulation process, Egr1 identifies and binds to its specific binding site, often characterized by a core consensus sequence like GCGGTGGCG. This binding triggers the transcription of target genes, resulting in the production of mRNA, which is subsequently translated into the SOCS1 protein. In the immune system, SOCS1 acts as a crucial regulator and inhibits adaptive immunosuppression mediated by IFNγ-induced PD-L1 expression. **b**, **c** Transcriptional activity and binding affinity predicted by PDIScore for the wild-type (wt) and mutant (mt) promoters. **d** The S1 and S2 sequences of the wt and mt.

ligands were grouped into four sets, each with a different nucleotide base pair at the 7th position: AT, TA, CG, and GA, while the other bases remained the same. In each group, the AT paired DNA was experimentally detected to have the highest binding affinity, suggesting that the AT pair at the 7th position might be the key nucleotide pair. We also counted the top10 DNAs out of 19131 based on binding affinity and found that the 7th position was AT pair in each top DNA, supporting the 7AT pair as the key nucleotides. In the heat map of nucleotide contribution, the color of the 7AT pair was the darkest in each group, and the colors of other positions were similar. This indicated that the importance of 7AT pair could be detected by the decomposition of PDIScore. The complex structures were also shown in Fig. 6, where the nucleotides were dyed with the corresponding colors from the heat map. It could be observed that the 7th nucleotide pair was located in the protein pocket, and the base part was in full contact with the protein surface, which might explain the significant change in binding affinity when the 7th position was mutated. Overall, PDIScore could provide a nucleotide level perspective to identify the key nucleotides in the design of DNA ligands.

CONCLUSIONS

In this work, we report PDIScore, a new DL approach that could predict the binding strength between protein and DNA. Our approach is composed of graph representation, feature extraction, concatenation, and MDN modules. We employ a comprehensive set of structural descriptors, including 11 atom distances and 20 dihedral angles for graph representation, in order to capture the subtle conformations of nucleotides. Considering the large

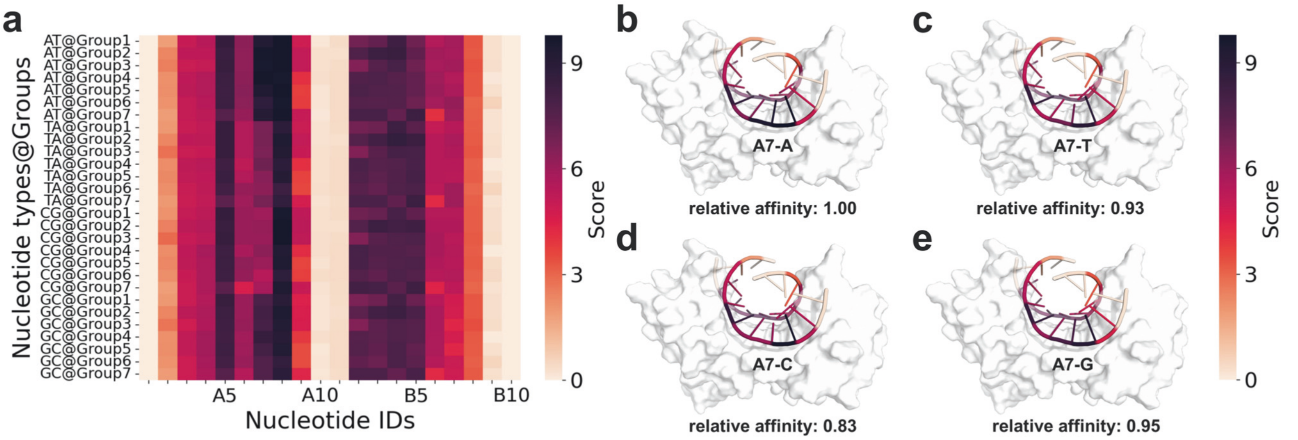

Fig. 6 The decomposed contributions of PDIScore at the nucleotide level. a In the heat map, the darker color of the nucleotide indicates the greater contribution to the total score. Each nucleotide ID is composed of chain ID (A or B) and the nucleotide order. Among the seven DNA groups, the nucleotide base pair is mutated at the 7th position, while the other base pairs remain identical within the same group. **b**–**e** In complex structures, each nucleotide pair is tinted according to its corresponding color on the heat map. The four complexes come from group 1, varying in the 7th nucleotide base pairs: **b** AT, **c** TA, **d** CG, and **e** GC.

contact areas of PDIs, which are reflected as large graphs in feature extraction, we adapt graphGPS and BigBird to handle graphs comprising several thousand nodes. The MDN modules are proven to effectively learn the probability density distribution of the distance between each residue and each nucleic acid.

PDIScore, in turn, outperforms all existing methods on various datasets, whether in terms of screening, docking or ranking power. In the screening set, PDIScore based on the AlphaFold3 predictions achieves an optimal $EF_{1\%}$ value of 14.13% and the best PCC value of 0.65. In the docking set, PDIScore reaches the highest top1 success rate of 48.94% compared with HDOCK, PyDockDNA, HADDOCK, AlphaFold3, PyRosetta#ref2015, PyRosetta#dna_gb, and FoldX. As for the ranking set, PDIScore based on the point mutation predictions achieves the best PCC value of 0.50. In the screening and ranking sets, the performance of AlphaFold3 is enhanced by PDIScore, showing its utility as a rescoring tool for AlphaFold3. As for interpretability, PDIScore is built at the residue/nucleotide level, so its predicted scores can be naturally decomposed into the contributions of residue-nucleotide pairs, identifying the key nucleotides in our case study.

Given its various capabilities and wider application range, PDIScore shows promising potential for PDI-related drug design. Its scoring capability can be used in screening task, saving experimental time and costs, and its docking capability enables it to be embedded into an ensemble strategy for structure prediction. As for the application range, on the one hand, PDIScore can be used for the screening of DNA drugs, such as DNA aptamers for the protein targets; on the other hand, it can be used for the screening of protein drugs or peptide drugs, such as cyclic peptides for the DNA targets. Considering the scarcity of open-source related datasets, we look forward to conducting more validation tasks in real scenarios in the future. Additionally, the current calculations are based on static conformations, while the interactions between proteins and DNAs are dynamic processes. The lack of consideration for flexibility might be a limitation of PDIScore. Introducing dynamic structures into the training set can expand the dataset, mitigate distributional imbalance, and potentially improve the performance of PDIScore in future research.

DATA AVAILABILITY

The datasets are available at https://doi.org/10.5281/zenodo.13764615. The code is available at https://github.com/roger-yh-zhao/pdiscore.

ACKNOWLEDGEMENTS

This work was financially supported by National Natural Science Foundation of China (22220102001, 22307112, 92370130, 22303081) and Postdoctoral Fellowship Program of CPSF (GZB20230648).

AUTHOR CONTRIBUTIONS

TJH and YK designed the research study. YHZ developed the method and wrote the code. YHZ, YW, CS, DJJ, SKG, HFZ, and ZYY performed the analysis. YHZ and TJH wrote the paper. All authors read and approved the manuscript.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41401-025-01688-3.

Competing interests: The authors declare no competing interests.

REFERENCES

1. Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNA-binding transcription factors across the tree of life. Nucleic Acids Res. 2010;38:7364–77.
2. Markodimitraki CM, Rang FJ, Rooijers K, de Vries SS, Chialastri A, de Luca KL, et al. Simultaneous quantification of protein-DNA interactions and transcriptomes in single cells with scDam&T-seq. Nat Protoc. 2020;15:1922–53.
3. Yu L, Liu P. Cytosolic DNA sensing by cGAS: regulation, function, and human diseases. Signal Transduct Target Ther. 2021;6:170.
4. Wells A, Heckerman D, Torkamani A, Yin L, Sebat J, Ren B, et al. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nat Commun. 2019;10:5241.
5. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, Subtypes PT, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun. 2020;11:728.
6. Guo AD, Yan KN, Hu H, Zhai L, Hu TF, Su H, et al. Spatiotemporal and global profiling of DNA-protein interactions enables discovery of low-affinity transcription factors. Nat Chem. 2023;15:803–14.
7. Chan LL, Pineda M, Heeres JT, Hergenrother PJ, Cunningham BT. A general method for discovering inhibitors of protein-DNA interactions using photonic crystal biosensors. ACS Chem Biol. 2008;3:437–48.
8. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
9. Nikolov DB, Chen H, Halay ED, Usheva AA, Hisatake K, Lee DK, et al. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. Nature. 1995;377:119–28.
10. Hashimoto H, Olanrewaju YO, Zheng Y, Wilson GG, Zhang X, Cheng X. Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. Genes Dev. 2014;28:2304–13.
11. Kays AR, Schepartz A. Virtually unidirectional binding of TBP to the AdMLP TATA box within the quaternary complex with TFIIA and TFIIB. Chem Biol. 2000;7:601–10.

12. Mostecki J, Showalter BM, Rothman PB. Early growth response-1 regulates lipopolysaccharide-induced suppressor of cytokine signaling-1 transcription. J Biol Chem. 2005;280:2596–605.
13. Sobah ML, Liongue C, Ward AC. SOCS proteins in immunity, inflammatory diseases, and immune-related cancer. Front Med. 2021;8:727987.
14. Sefah K, Shangguan D, Xiong X, O'Donoghue MB, Tan W. Development of DNA aptamers using Cell-SELEX. Nat Protoc. 2010;5:1169–85.
15. Huynh L, Chen A. Design of a protein-targeted DNA aptamer using atomistic simulation. J Biomol Struct Dyn. 2023;41:672–80.
16. Wang X, Wang Y, Cao A, Luo Q, Chen D, Zhao W, et al. Development of cyclopeptide inhibitors of cGAS targeting protein-DNA interaction and phase separation. Nat Commun. 2023;14:6132.
17. Kimoto M, Yamashige R, Matsunaga K, Yokoyama S, Hirao I. Generation of high-affinity DNA aptamers using an expanded genetic alphabet. Nat Biotechnol. 2013;31:453–7.
18. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat Protoc. 2007;2:1849–61.
19. Hadzi S, Lah J. Analysis of protein-DNA interactions using isothermal titration calorimetry: successes and failures. Methods Mol Biol. 2022;2516:239–57.
20. Majka J, Speck C. Analysis of protein-DNA interactions using surface plasmon resonance. Adv Biochem Eng Biotechnol. 2007;104:13–36.
21. Townshend RJL, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, et al. Geometric deep learning of RNA structure. Science. 2021;373:1047–51.
22. Roche R, Moussad B, Shuvo MH, Tarafder S, Bhattacharya D. EquiPNAS: improved protein-nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. Nucleic Acids Res. 2024;52:e27.
23. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019;35:4168–9.
24. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33:W382–8.
25. Wang E, Sun H, Wang J, Wang Z, Liu H, Zhang JZH, et al. End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. Chem Rev. 2019;119:9478–508.
26. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics. 2010;26:689–91.
27. Yang W, Deng L. PreDBA: a heterogeneous ensemble approach for predicting protein-DNA binding affinity. Sci Rep. 2020;10:1278.
28. Yang S, Gong W, Zhou T, Sun X, Chen L, Zhou W, et al. emPDBA: protein-DNA binding affinity prediction by combining features from binding partners and interface learned with ensemble regression model. Brief Bioinform. 2023;24:bbad192.
29. Zhang X, Mei LC, Gao YY, Hao GF, Song BA. Web tools support predicting protein-nucleic acid complexes stability with affinity changes. WIREs RNA. 2023;14:e1781.
30. Yan Y, Zhang D, Zhou P, Li B, Huang SY. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. 2017;45:W365–73.
31. Rodriguez-Lumbreras LA, Jimenez-Garcia B, Gimenez-Santamarina S, Fernandez-Recio J. pyDockDNA: a new web server for energy-based protein-DNA docking and scoring. Front Mol Biosci. 2022;9:988996.
32. van Zundert GCP, Rodrigues J, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. J Mol Biol. 2016;428:720–5.
33. Honorato RV, Koukos PI, Jimenez-Garcia B, Tsaregorodtsev A, Verlato M, Giachetti A, et al. Structural biology in the clouds: the WeNMR-EOSC ecosystem. Front Mol Biosci. 2021;8:729513.
34. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
35. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
36. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.
37. Afek A, Shi H, Rangadurai A, Sahay H, Senitzki A, Xhani S, et al. DNA mismatches reveal conformational penalties in protein-DNA recognition. Nature. 2020;587:291–6.
38. van Dijk M, Bonvin AM. A protein-DNA docking benchmark. Nucleic Acids Res. 2008;36:e88.
39. Larkin C, Datta S, Harley MJ, Anderson BJ, Ebie A, Hargreaves V, et al. Inter- and intramolecular determinants of the specificity of single-stranded DNA binding and cleavage by the F factor relaxase. Structure. 2005;13:1533–44.
40. Li J, Lee JC. Modulation of allosteric behavior through adjustment of the differential stability of the two interacting domains in *E. coli* cAMP receptor protein. Biophys Chem. 2011;159:210–6.
41. Yu S, Maillard RA, Gribenko AV, Lee JC. The N-terminal capping propensities of the D-helix modulate the allosteric activation of the Escherichia coli cAMP receptor protein. J Biol Chem. 2012;287:39402–11.
42. Seldeen KL, Deegan BJ, Bhat V, Mikles DC, McDonald CB, Farooq A. Energetic coupling along an allosteric communication channel drives the binding of Jun-Fos heterodimeric transcription factor to DNA. FEBS J. 2011;278:2090–104.
43. Rampášek L, Galkin M, Dwivedi VP, Luu AT, Wolf G, Beaini D. Recipe for a general, powerful, scalable graph transformer. NeurIPS. 2022;35:14501–15.
44. Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, et al. Big bird: transformers for longer sequences. NeurIPS. 2020;33:17283–97.
45. Shen C, Zhang X, Hsieh CY, Deng Y, Wang D, Xu L, et al. A generalized protein-ligand scoring framework with balanced scoring, docking, ranking and screening powers. Chem Sci. 2023;14:8129–46.
46. Shen C, Zhang X, Deng Y, Gao J, Wang D, Xu L, et al. Boosting protein-ligand binding pose prediction and virtual screening based on residue-atom distance likelihood potential and graph transformer. J Med Chem. 2022;65:10691–706.
47. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem. 2004;47:2977–80.
48. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera-a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.
49. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–74.
50. Lu XJ, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 2003;31:5108–21.
51. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
52. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem. 2011;32:2319–27.
53. Wang M, Xu L, Zheng D, Gan Q, Gai Y, Ye Z, et al. Deep graph library: towards efficient and scalable deep learning on graphs. Preprint at. https://doi.org/10.48550/arXiv.1909.01315 (2019).
54. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. In *presented in part at the International Conference on Learning Representations (ICLR)*, (ICLR, 2020).
55. Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0-a public library of challenging docking benchmark sets. J Chem Inf Model. 2013;53:1447–62.
56. Concino M, Goldman RA, Caruthers MH, Weinmann R. Point mutations of the adenovirus major late promoter with different transcriptional efficiencies in vitro. J Biol Chem. 1983;258:8493–6.
57. Ilangumaran S, Gui Y, Shukla A, Ramanathan S. SOCS1 expression in cancer cells: potential roles in promoting antitumor immunity. Front Immunol. 2024;15:1362224.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.