Interpretable Protein-DNA Interactions Captured by Structure-Sequence Optimization

Yafan Zhang,[†] Irene Silvernail,[‡] Zhuyang Lin,[†] and Xingcheng Lin[*,†,‡]

†*Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27607*

‡*Department of Physics, North Carolina State University, Raleigh, NC 27607*

E-mail: Xingcheng_Lin@ncsu.edu

Abstract

Sequence-specific DNA recognition underlies essential processes in gene regulation, yet methods for simultaneous predictions of genomic DNA recognition sites and their binding affinity remain lacking. Here, we present the Interpretable protein-DNA Energy Associative (IDEA) model, a residue-level, interpretable biophysical model capable of predicting binding sites and affinities of DNA-binding proteins. By fusing structures and sequences of known protein-DNA complexes into an optimized energy model, IDEA enables direct interpretation of physicochemical interactions among individual amino acids and nucleotides. We demonstrate that this energy model can accurately predict DNA recognition sites and their binding strengths across various protein families. Additionally, the IDEA model is integrated into a coarse-grained simulation framework that quantitatively captures the absolute protein-DNA binding free energies. Overall, IDEA provides an integrated computational platform that alleviates experimental costs and biases in assessing DNA recognition and can be utilized for mechanistic studies of various DNA recognition processes.

Introduction

Gene regulation is essential for controlling the timing and extent of gene expression in cells. It guides important processes from development to disease response.[1] Gene expression is tightly controlled by various DNA-binding proteins, including transcription factors (TFs) and epigenetic regulators.[2–4] TFs initiate gene expression by binding to specific DNA sequences, during which time a primary RNA transcript is synthesized from a gene's DNA.[5] On the other hand, epigenetic regulators control gene expression by binding to specific chromatin regions, which spread post-translational modifications that modulate 3D genome organization.[6] Therefore, characterizing the DNA-interaction specificities of DNA binding proteins is critical for understanding the molecular mechanisms underlying many DNA-templated processes.[7]

To establish a comprehensive understanding of DNA binding processes, various experimental technologies have been developed and utilized.[8–16] These methods, such as Chromatin Immunoprecipitation followed by sequencing (ChIP-Seq),[8,9,17] Protein-binding microarray (PBM),[18–20] and Systematic Evolution of Ligands by Exponential Enrichment (SELEX),[15,16,21] have proven invaluable for measuring DNA-specific protein recognition. Nonetheless, due to the need for protein-specific antibodies,[17] as well as the cost and intrinsic bias associated with these experiments,[22] high-throughput measurement of large numbers of protein-DNA variants remains challenging.

Computational methods complement experimental efforts by providing the initial filter for assessing sequence-specific protein-DNA binding affinity. Numerous methods have emerged to enable predictions of binding sites and affinities of DNA-binding proteins.[23–31] These methods often utilized machine-learning-based training to extract sequence preference information from DNA or protein by utilizing experimental high-throughput (HT) assays,[23–27,31] which rely on the availability and quality of experimental binding assays. Additionally, many approaches employ deep neural networks,[28–30] which could obscure the interpretation of interaction patterns governing protein-DNA binding specificities. Understanding these

patterns, however, is crucial for elucidating the molecular mechanisms underlying various DNA-recognition processes, such as those seen in TFs.[32]

Nowadays, over 5000 protein-DNA 3D structures, including TF-DNA complexes, have been published.[33,34] These data provide invaluable resources for understanding the physicochemical properties of protein-DNA binding patterns, extending beyond mere sequence information. Utilization of these data enables the training of a model for predicting the binding affinities or specificities of protein-DNA interactions. Emerging deep learning,[35] probabilistic[36] and chemistry-based models[37] are beginning to learn from protein–DNA co-crystal structures. Their predictions are biophysically meaningful and can be applied to understand molecular interactions at a mechanistic level, thus holding great potential to guide experimental design and synthetic biology applications. The very robustness of evolution[38–41] provides an opportunity to extract the sequence-structure relationships embedded in existing complexes. Guided by this principle, we can learn an interpretable binding energy landscape that governs the recognition processes of DNA-binding proteins.

Here, we introduce the Interpretable protein-DNA Energy Associative (IDEA) model, a predictive model that learns protein-DNA physicochemical interactions by fusing available biophysical structures and their associated sequences into an optimized energy model (Figure 1). We show that the model can be used to accurately predict the sequence-specific DNA binding affinities of DNA-binding proteins and is transferrable across the same protein superfamily. Moreover, the model can be enhanced by incorporating experimental binding data and can be generalized to enable base-pair resolution predictions of genomic DNA-binding sites. Notably, IDEA learns a family-specific interaction matrix that quantifies energetic interactions between each amino acid and nucleotide, allowing for a direct interpretation of the "molecular grammar" governing sequence-specific protein-DNA binding affinities. This interpretable energy model is further integrated into a simulation framework, enabling mechanistic studies of various biomolecular functions involving protein-DNA dynamics.

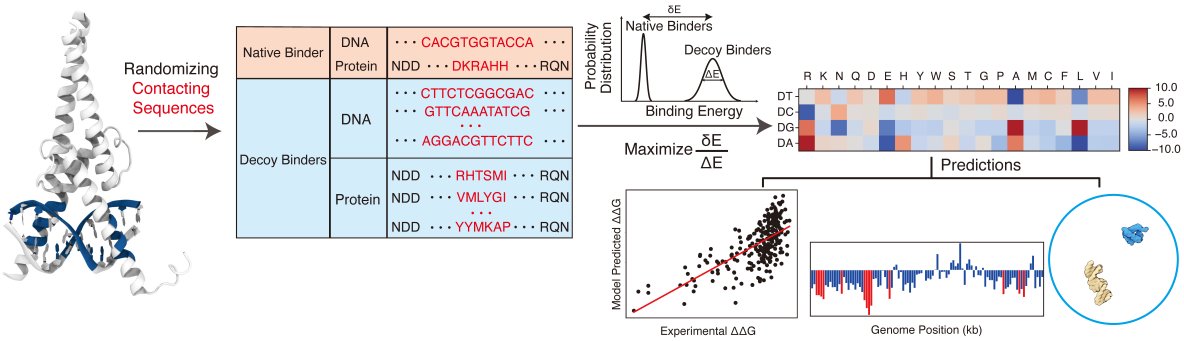

Figure 1: **Overview of the IDEA protocol.** The protein–DNA complex, represented by the human MAX–DNA complex structure (PDB ID: 1HLO), was used for training the IDEA model. The sequences of the protein and DNA residues that form close contacts (highlighted in blue on the structure) in the structure were included in the training dataset. In addition, a series of synthetic decoy sequences was generated by randomizing the contacting residues in both the protein and DNA sequences. The amino acid–nucleotide energy model was then optimized by maximizing the ratio of the binding energy gap (δE) between protein and DNA in the native complex and the decoy complexes relative to the energy variance (ΔE). The optimized energy model can be used for multiple predictive applications, including the evaluation of binding free energies for various protein–DNA sequence pairs, prediction of genomic DNA binding sites by transcription factors or other DNA-binding proteins, and integration into a sequence-specific, residue-resolution simulation framework for dynamic simulations.

Results

Predictive Protein-DNA Energy Model at Residue Resolution

IDEA is a coarse-grained biophysical model at the residue resolution for investigating protein-DNA binding interactions (Figure 1). It integrates both structures and corresponding sequences of known protein-DNA complexes to learn an interpretable energy model based on the interacting amino acids and nucleotides at the protein-DNA binding interface. The model is trained using available protein-DNA complexes curated from existing databases.[33,42] Unlike existing deep-learning-based protein-DNA binding prediction models, IDEA aims to learn a physicochemical-based energy model that quantitatively characterizes sequence-specific interactions between amino acids and nucleotides, thereby interpreting the "molecular grammar" driving the binding energetics of protein-DNA interactions. The optimized energy model can be used to predict the binding affinity of any given protein-DNA pair based on its structures and sequences. Additionally, it enables the prediction of genomic DNA binding sites by a given protein, such as a transcription factor. Finally, the learned energy model can be incorporated into a simulation framework to study the dynamics of DNA-binding processes, revealing mechanistic insights into various DNA-templated processes. Further details of the optimization protocol are provided in the Methods Section *Energy Model Optimization*.

IDEA Accurately Predicts Sequence-Specific Protein-DNA Binding Affinity

We first examine the predictive accuracy of IDEA by comparing its predicted TF-DNA binding affinities with experimental measurements.[13,14] We focused on the MAX TF, a basic Helix-loop-helix (bHLH) TF with the most comprehensive available experimental binding data. The binding affinity of MAX to various DNA sequences has been quantified by multiple experimental platforms, including different SELEX variants[15,16,21] and microfluidic-based MITOMI assay.[13,43] Among them, MITOMI quantitatively measured the binding affinities

of MAX TF to a comprehensive set of 255 DNA sequences with mutations in the enhancer box (E-box) motif, the consensus MAX DNA binding sequence. This dataset serves as a reference for us to benchmark our model predictions. Our *de novo* prediction, based on one MAX crystal structure (PDB ID: 1HLO), correlates well with the experimental values (Pearson correlation coefficient 0.67) and correctly ranks the binding affinities of those DNA sequences (Spearman's rank correlation coefficient 0.65, Figure 2A). Including additional human-associated MAX-DNA complex structures and their associated sequences in the model training slightly improves the predictions (Pearson correlation coefficient 0.68, Spearman's rank correlation coefficient 0.65, Figure S1), albeit not significantly, likely due to the high structural similarity of the protein-DNA interfaces across all MAX proteins. Prior works have proven it instrumental in incorporating experimental binding data to improve predicted protein-nucleic acid binding affinities.[24,27,44] Inspired by these approaches, we developed an enhanced predictive protocol that integrates additional experimental binding data if available (see the Methods Section *Enhanced Modeling Prediction with SELEX Data*). Encouragingly, when training the model with the SELEX-seq data[27] of MAX TF, IDEA showed a significant improvement in reproducing the MITOMI measurement (Pearson correlation coefficient 0.79, Spearman's rank correlation coefficient 0.76, Figure S2).

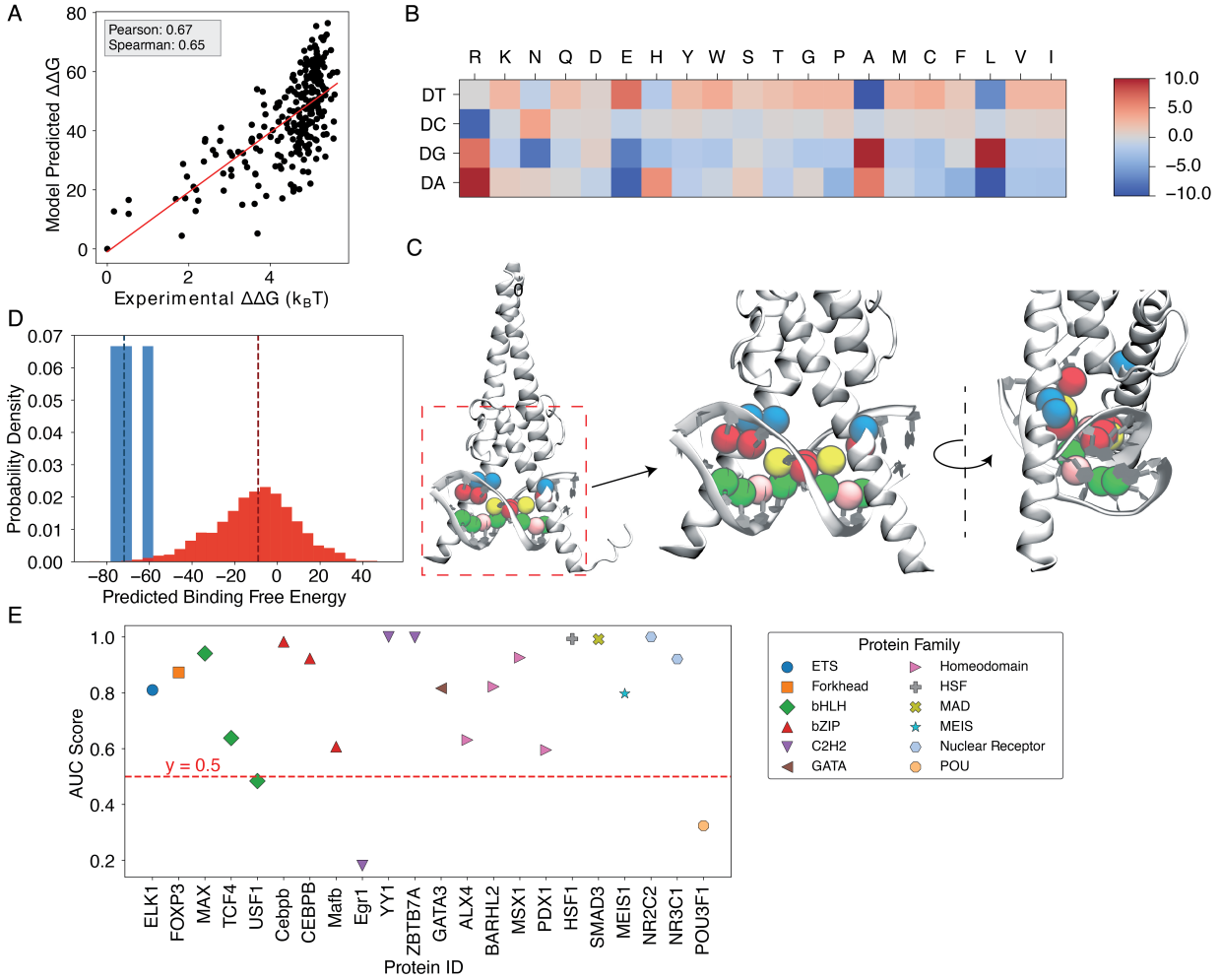

Figure 2: **Results for MAX-based predictions.** (A) The binding free energies calculated by IDEA, trained using a single MAX–DNA complex (PDB ID: 1HLO), correlate well with experimentally measured MAX–DNA binding free energies.[13] $\Delta\Delta G$ represents the changes in binding free energy relative to that of the wild-type protein–DNA complex. (B) The heatmap, derived from the optimized energy model, illustrates key amino acid–nucleotide interactions governing MAX–DNA recognition, showing pairwise interaction energies between 20 amino acids and the four DNA bases—DA (deoxyadenosine), DT (deoxythymidine), DC (deoxycytidine), and DG (deoxyguanosine). Both the predicted binding free energies and the optimized energy model are expressed in reduced units, as explained in the Methods Section *Training Protocol*. Each cell represents the optimized energy contribution, where blue indicates more favorable (lower) energy values, and red indicates less favorable (higher) values. (C) The 3D structure of the MAX–DNA complex (zoomed in with different views) highlights key amino acid–nucleotide contacts at the protein–DNA interface. Notably, several DNA deoxycytidines (red spheres) form close contacts with arginines (blue spheres). Additional nucleotide color coding: adenine (yellow spheres), guanine (green spheres), thymine (pink spheres). (D) Probability density distributions of predicted binding free energies for strong (blue) and weak (red) binders of the protein ZBTB7A. The median of each distribution is marked with a dashed line. (E) Summary of AUC scores for protein–DNA pairs across 12 protein families, calculated based on the predicted probability distributions of binding free energies.

IDEA Decodes the Molecular Grammar Governing Protein-DNA Interactions

IDEA protocol learns the physicochemical interactions that determine protein-DNA interactions by utilizing the sequence-structure relationship embedded in the protein-DNA experimental structures. Such a physicochemical interaction pattern can be interpreted directly from the learned energy model. To illustrate this, we focused on the energy model learned from the MAX-DNA complexes (Figure 2B). Notably, DNA deoxycytidine (DC) exhibited strong interactions with protein arginine (R), consistent with the fact that the E-box region (CACGTG) frequently attracts the positively charged residues of bHLH TFs.[45] Importantly, arginine was in close contact with deoxycytidine in the crystal structure (Figure 2C), thus consistent with the strong DC-R interactions shown in the learned energy model. Upon integrating the SELEX data into our model training, we found that the improved energy model shows additional unfavorable interactions between protein glutamic acid (E) and deoxycytidine, consistent with their negative charges (Figure S3). Thus, including more experimental data can boost IDEA's predictive accuracy by refining the amino-acid-nucleotide interacting energy model to better align with physical principles.

IDEA Generalizes across Various Protein Families

To examine IDEA's predictive accuracy across different DNA-binding protein families, we applied it to calculate protein-DNA binding affinities using a comprehensive HT-SELEX dataset.[26] We focused on evaluating the capability of IDEA to distinguish strong binders from weak binders for each protein with an experimentally determined structure. We calculated the probability density distribution of the top and bottom binders identified in the SELEX experiment. A well-separated distribution indicates the successful identification of strong binders by IDEA (Figure 2D and S4). Receiver Operating Characteristic (ROC) and Precision-Recall (PR) analyses were performed to calculate the Area Under the Curve

(AUC) and the precision-recall AUC (PRAUC) scores for these predictions. Further details are provided in the Methods Section *Evaluation of IDEA Prediction Using HT-SELEX Data*. Our analysis shows that IDEA successfully differentiates strong from weak binders for 80% of the 22 proteins across 12 protein families, achieving AUC and balanced PRAUC scores greater than 0.5 (Figure 2E and S5). To benchmark IDEA's performance against other leading methods, we compared its predictions with several popular models, including the sequence-based predictive models ProBound[31] and DeepBind,[24] the family-based energy model rCLAMPS,[36] and the knowledge-based energy model DBD-Hunter.[46] IDEA demonstrates performance comparable to these state-of-the-art approaches, and incorporating sequence features further improves its prediction accuracy (Figure S6, Table S1, and Table S2). We also performed 10-fold cross-validation on the binding affinities of protein–DNA pairs in this dataset and found that IDEA outperforms a recent regression model that considers the shape of DNA with different sequences[26] (Figure S7). Details are provided in the Appendix *Section: Comparison of IDEA predictive performance Using HT-SELEX data*.

We also applied IDEA to predict the binding affinity of additional TFs with available MITOMI measurements.[43] For PHO4, another bHLH transcription factor, IDEA's predictions, trained on the only available protein-DNA structure (PDB ID: 1A0A), correlate well with experimental measurements, showing a Pearson correlation coefficient of 0.60 and a Spearman's rank correlation coefficient of 0.63 (Figures S8A, B). We further evaluated IDEA's predictive performance for the zinc-finger protein Zif268 (see the Methods Section *Structural Modeling of Protein and DNA*). Due to the limited experimental data for all possible DNA sequence combinations, we focused on testing IDEA's predictions on point-mutated DNA sequences. Predictions on point-mutated sequences are known to be a challenging task due to their minor deviations from the wild-type sequence. Despite this, IDEA achieves accurate predictions, with a Pearson correlation coefficient of 0.57 and a Spearman's rank correlation coefficient of 0.60 (Figures S8C, D). We further expanded the training dataset to include all available Zif268 structures and their associated sequences from

the same CATH zinc finger superfamily (CATH ID: 3.30.160.60). Incorporating these additional training data further improves the predictive accuracy (Pearson correlation coefficient 0.63; Spearman's rank correlation coefficient 0.60, Figure S9).

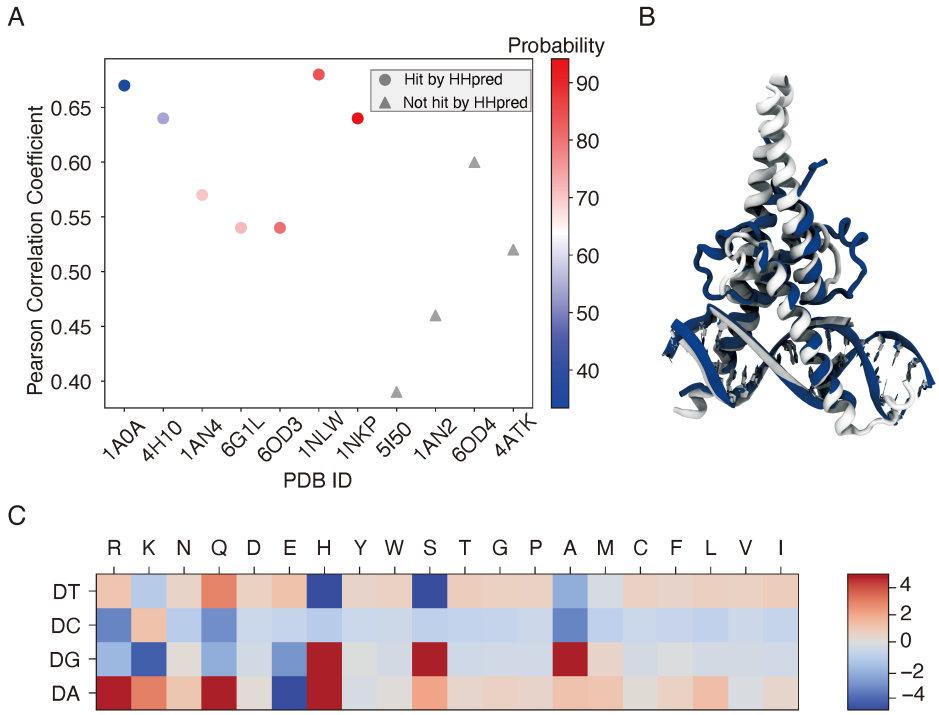

Figure 3: **IDEA prediction shows transferability within the same CATH superfamily.** (A) The predicted MAX binding affinity, trained on other protein-DNA complexes within the same protein CATH superfamily, correlates well with experimental measurement. The proteins are ordered by their probability of being homologous to the MAX protein, determined using HHpred.[47] Training with a homologous protein (determined as a hit by HHpred) usually leads to better predictive performance (Pearson correlation coefficient > 0.5) compared to non-homologous proteins. (B) Structural alignment between 1HLO (white) and 1A0A (blue), two protein-DNA complexes within the same CATH Helix-loop-helix superfamily. The alignment was performed based on the E-box region of the DNA.[48] (C) The optimized energy model for 1A0A, a protein-DNA complex structure of the transcription factor PHO4 and DNA, with 33.41% probability of being homologous to the MAX protein. The optimized energy model is presented in reduced units, as explained in the Methods Section: *Training Protocol*.

IDEA Demonstrates Transferability across Proteins in the Same CATH Superfamily

Since IDEA relies on the sequence-structure relationship of given protein-DNA complexes to reach predictive accuracy, we inquired whether the trained energy model from one protein-DNA complex could be generalized to predict the sequence-specific binding affinities of other complexes. To test this, we assessed the transferability of IDEA predictions across all 11 structurally available protein-DNA complexes within the MAX TF-associated CATH superfamily (CATH ID: 4.10.280.10, Helix-loop-helix DNA-binding domain). We trained IDEA based on each of these 11 complexes and then used the trained model to predict the MAX-based MITOMI binding affinity. Our results show that IDEA generally makes correct predictions of the binding affinity when trained on proteins that are homologous to MAX, with Pearson and Spearman Correlation coefficients larger than 0.5 (Figure 3A and Figure S10).

The transferability of IDEA within the same CATH superfamily can be understood from the similarities in protein-DNA binding interfaces, which determine similar learned energy models. For example, the PHO4 protein (PDB ID: 1A0A) shares a highly similar DNA-binding interface with the MAX protein (PDB ID: 1HLO) (Figure 3B), despite sharing only a 33.41% probability of being homologous. Consequently, the energy model derived from the PHO4-DNA complex (Figure 3C) exhibits a similar amino-acid-nucleotide interactive pattern as that learned from the MAX-DNA complex (Figure 2B). To further evaluate the similarity between the learned energy models and their connection to protein families, we performed principal component analysis (PCA) on the normalized energy models across 24 proteins from 12 protein families.[26] Our analysis (Figure S11) reveals that most of the energy models from the same protein family fall within the same cluster, while those from different protein families exhibit distinct patterns. Moreover, the relative distance between energy models in PCA space reflects the degree of transferability between them. For example, PHO4 (PDB ID: 1A0A) is positioned close to MAX (PDB ID: 1HLO), whereas USF1 (PDB ID: 1AN4) and TCF4 (PDB ID: 6OD3) are farther away. This is consistent with the results in

Figure 3A, where the energy model trained on PHO4 has better transferability than those trained on USF1 or TCF4.

Identification of Genomic Protein-DNA Binding Sites

The genomic locations of DNA binding sites are causally related to major cellular processes.[17] Although multiple techniques have been developed to enable high-throughput mapping of genomic protein-DNA binding locations, such as ChIP–seq,[9,17,49] DAP-seq,[10] and FAIRE-seq,[50] challenges remain for precisely pinpointing protein-binding sites at a base-pair resolution. Furthermore, the applicability of these techniques for different DNA-binding proteins is restricted by the quality of antibody designs.[17] Therefore, a predictive computational framework would significantly reduce the costs and accelerate the identification of genomic binding sites of DNA-binding proteins.

We incorporated IDEA into a protocol to predict the genomic protein-DNA binding sites at the base-pair resolution, given a DNA-binding protein and a genomic DNA sequence. To evaluate the predictive accuracy of this protocol, we utilized publicly available ChIP-seq data for MAX TF binding in GM12878 lymphoblastoid cell lines.[51] As the experimental measurements were conducted at a 420 base pairs resolution, we averaged our modeling prediction over a window spanning 500 base pairs. Since the experimental signals are sparsely distributed across the genome, we focused our prediction on the 1 Mb region of Chromosome 1, which has the densest and most reliable ChIP-seq signals (156000 kb - 157000 kb, representative window shown in Figure 4A top). In our model, lower predicted binding free energy corresponds to stronger protein-DNA binding affinity and thus a higher probability of a binding site. These predicted binding free energies were further normalized against those calculated from the randomized decoy sequences (see the Methods Section *Processing of ChIP-seq Data* for details). The identified MAX-binding sites are marked in red, with a normalized Z score < -0.75. Our prediction successfully identifies the majority of the MAX binding sites, achieving an AUC score of 0.81 (Figure 4B). We also tested IDEA's predictive

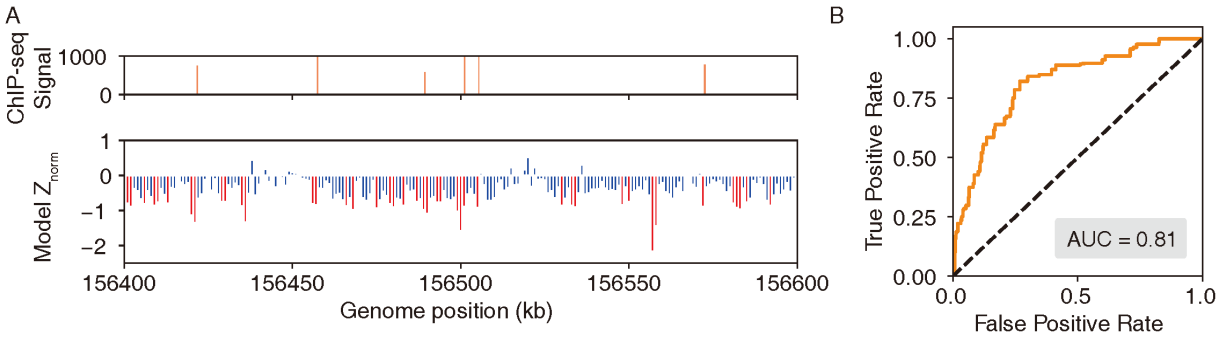

Figure 4: **IDEA accurately identifies genome-wide protein-binding sites.** (A) The IDEA-predicted MAX-transcription factor binding sites on the Lymphoblastoid cells chromosome (bottom) correlate well with ChIP-seq measurements (top), shown for a representative 1Mb region of chromosome 1 where ChIP-seq signals are densest. For visualization purposes, a 1 kb resolution was used to plot the predicted normalized Z scores, with highly probable binding sites represented as red peaks. (B) AUC score for prediction accuracy based on the normalized Z scores, averaged over a 500 bp window to match the experimental resolution of 420 bp.

performance on two additional typical DNA-binding proteins and successfully identified their genomic binding sites (Figure S12). Therefore, IDEA provides an accurate initial prediction of genomic recognition sites of DNA-binding proteins based only on DNA sequences, facilitating more focused experimental efforts. Furthermore, IDEA holds the potential to identify binding sites beyond the experimental resolution and can aid in the interpretation of other sequencing techniques to uncover additional factors that influence genome-wide DNA binding, such as chromatin accessibility.[52–54]

Integration of IDEA into a Residue-Resolution Simulation Model Captures Protein-DNA Binding Dynamics

The trained residue-level model can be incorporated into a simulation model, thus enabling investigations into dynamic interactions between protein and DNA for mechanistic studies. Coarse-grained protein-nucleic acid models have shown strong advantages in investigating large biomolecular systems.[55–59] Notably, the structure-based protein model[60] and 3-Site-Per-Nucleotide (3SPN) DNA model[61] have been developed at the residue resolution to enable efficient sampling of biomolecular systems involving protein and DNA. A combination of these two models has proved instrumental in quantitatively studying multiple large chromatin systems,[62] including nucleosomal DNA unwrapping,[63–65] interactions between nucleosomes,[66,67] large chromatin organizations,[67–69] and its modulation by chromatin factors.[70–72] Although both the protein and DNA force fields have been systematically optimized,[61,73–76] the non-bonded interactions between protein and DNA were primarily guided by Debye-Hückel treatment of electrostatic interactions, which did not consider sequence-specific interactions between individual protein amino acids and DNA nucleotides.

To refine this simulation model by including protein-DNA interaction specificity, we incorporated the IDEA-optimized protein-DNA energy model into the coarse-grained simulation model as a short-range Van der Waals (vdW) interaction in the form of a Tanh function (See Methods Section *Sequence Specific Protein-DNA Simulation Model* for modeling details). This refined simulation model is used to simulate the dynamic interactions between DNA-binding proteins and their target DNA sequences. As a demonstration of the working principle of this modeling approach, we selected 9 TF-DNA complexes whose binding affinities were measured by the same experimental group.[77] We applied IDEA to train an energy model based on these protein-DNA complexes, which yielded a modest correlation between the modeling-predicted binding affinity and experimental measurements (Pearson correlation coefficient 0.46, Spearman's rank correlation coefficient 0.55, Figure 5A). However, upon incorporating the learned energy model into our simulation model and computing

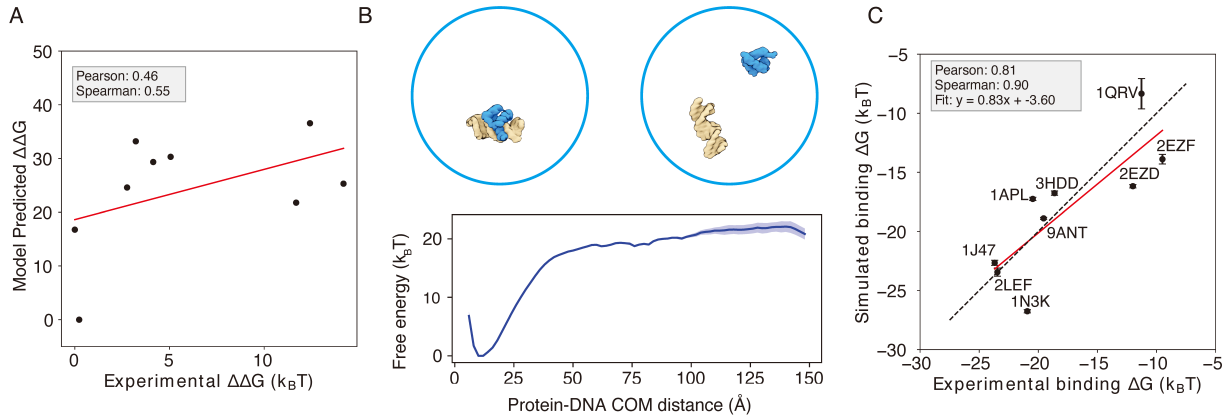

Figure 5: **Enhanced protein-DNA simulation model by incorporating IDEA-optimized energy model.** (A) Predicted protein-DNA binding free energies for 9 protein-DNA complexes using the IDEA-learned energy model correlate with the experimental measurements.[77] $\Delta\Delta G$ represents the changes in binding free energy relative to the protein-DNA complex with the lowest predicted value. The predicted binding free energies are presented in reduced units, as explained in *Methods*. (B) Example of a protein-DNA complex structure (PDB ID: 9ANT) and the coarse-grained simulation used to evaluate protein-DNA binding free energy. A representative free energy profile was extracted from the simulation, using the center of mass (COM) distance between protein and DNA as the collective variable. The shaded region represents the standard deviation of the mean. Representative bound and unbound structures are shown above, with protein in blue and DNA in yellow. (C) Incorporating the IDEA-optimized energy model into the simulation model improves the prediction of protein-DNA binding affinity, compared to the prediction by the previous model with electrostatic interactions and uniform nonspecific attraction between protein and DNA (Figure S14). The predicted binding free energies are presented in physical units. Error bars represent the standard deviation of the mean.

protein-DNA binding free energies through umbrella sampling,[78] we observed a significant improvement in predictive accuracy. The predicted binding free energies[79] (Figure S13) from our simulations correlate strongly with experimental measurements (Pearson correlation coefficient 0.81; Spearman's rank correlation coefficient 0.90) and near-perfectly reproduce the experimental binding affinities (Figure 5B). This prediction also represents a substantial improvement over the previous model, which depicts protein-DNA interactions through non-sequence-specific homogeneous electrostatic interactions (Pearson correlation coefficient 0.60, Spearman's rank correlation coefficient 0.70; Figure S14). Notably, due to an undetermined prefactor in the IDEA training, it can only provide relative binding free energies in reduced units for different protein-DNA complexes. In contrast, the IDEA-incorporated simulation model can predict the absolute binding free energies with physical units.

Discussion

The protein-DNA interaction landscape has evolved to facilitate precise targeting of proteins towards their functional binding sites, which underlie essential processes in controlling gene expression. These interaction specifics are determined by physicochemical interactions between amino acids and nucleotides. By integrating sequences and structural data from available protein-DNA complexes into an interaction matrix, we introduce IDEA, a data-driven method that optimizes a system-specific energy model. This model enables high-throughput *in silico* predictions of protein-DNA binding affinities and can be scaled up to predict genomic binding sites of DNA-binding proteins, such as TFs. IDEA achieves accurate *de novo* predictions using only protein-DNA complex structures and their associated sequences, but its accuracy can be further enhanced by incorporating available experimental data from other binding assay measurements, such as the SELEX data,[15,16,21] achieving accuracy comparable or better than state-of-the-art methods (Figures S2 and S7, Table S1 and S2). Despite significant progress in genome-wide sequencing techniques,[9,10,17,50] determin-

ing sequence-specific binding affinities of DNA-binding biomolecules remains time-consuming and expensive. Therefore, IDEA presents a cost-effective alternative for generating the initial predictions before pursuing further experimental refinement.

A key advantage of IDEA is its incorporation of both structural and sequence information, which greatly reduces the demand for extensive training data. Prior efforts have applied deep learning techniques to predict the interactions between proteins and DNA based on their sequences.[24,25,27] These approaches usually require a large amount of sequencing training data. By leveraging the sequence-structure relationship inherent in protein-DNA complex structures, IDEA achieves accurate *de novo* prediction using only a few protein-DNA complexes within the same protein superfamily.

In addition, despite the complicated *in vivo* environment, IDEA enables predictions of genomic binding sites and shows good agreement with the ChIP-seq measurements (Figure 4, S12). Moreover, IDEA features rapid prediction, typically requiring 1-3 days to predict base-pair resolution genomic binding sites of a 1 Mb DNA region using a single CPU. Furthermore, IDEA's prediction can be trivially parallelized, making it possible to perform genome-wide predictions of DNA recognition sites by a given protein within weeks.

Another highlight of IDEA is its ability to present an interpretable, family-specific amino acid-nucleotide interaction energy model for given protein-DNA complexes. The optimized IDEA energy model can not only predict sequence-specific binding affinities of protein-DNA pairs but also provide a residue-specific interaction matrix that dictates the preferences of amino acid-nucleotide interactions within specific protein families (Figure S11). This interpretable energy matrix would facilitate the discovery of sequence binding motifs for target DNA-binding proteins, complementing both sequence-based[80–82] and structure-based approaches.[35–37,83] Additionally, we integrated this physicochemical-based energy model into a simulation framework, thereby improving the characterization of protein-DNA binding dynamics. IDEA-based simulation enables the investigation into dynamic interactions between various proteins and DNA, facilitating molecular-level understanding of the physical mech-

anisms underlying many DNA-binding processes, such as transcription, epigenetic regulations, and their modulation by sequence variations, such as single-nucleotide polymorphisms (SNPs).[84,85]

Since the IDEA-optimized energy model serves as a "molecular grammar" for guiding protein-DNA interaction specifics, it can be expanded to include additional variants of amino acids and nucleotides. With recent advancements in sequencing and structural characterization,[86,87] as well as deep-learning-based structural predictions,[88–90] thousands of structures and sequences have been solved for these variants. The IDEA procedure can be repeated for these variant-included structures to expand the optimized energy model by including modifications such as methylated DNA nucleotides[91] and post-translationally modified amino acids.[92] Such strategies will facilitate studies on the functional relevance of epigenetic variants, such as those caused by exposure to environmental hazards,[93] and their structural impact on the human genome.[94–96]

Although IDEA has proven successful in many examples, it can be improved in several aspects. The model currently assumes the training and testing sequences share the same protein-DNA structure. While double-stranded DNA is generally rigid, recent studies have shown that sequence-dependent DNA shape contributes to their binding specificity.[25,35,97] To improve predictive accuracy, one could incorporate predicted DNA shapes or structures into the IDEA training protocol. In addition, the model is residue-based and evaluates the binding free energy as the additive sum of contributions from individual amino-acid-nucleotide contacts. This assumption does not account for cooperative effects that may arise from multiple nucleotide changes. A potential refinement could utilize a finer-grained model that includes more atom types within contacting residues and employs a many-body potential to account for such cooperative effects.

Methods

Structural Modeling of Protein and DNA

We utilized a coarse-grained framework to extract structural information of protein and DNA at the residue level (Figure 1). Specifically, each protein amino acid was represented by the coordinates of the Cα atom, and each DNA nucleotide was represented by either the P atom in the phosphate group or the C5 atom on the nucleobase. The choice between these two DNA atom types (C5 or P) depended on their distances to surrounding Cα atoms. We hypothesized that DNA atoms closer to Cα atoms would provide more meaningful information on protein-DNA interactions. Therefore, we chose the DNA atom type (C5 or P) with the largest sum of occurrences within 8 Å, 9 Å, and 10 Å distances from their surrounding Cα atoms (Table S3). This selection rule was applied consistently across all complexes in the training dataset to maintain uniform criteria.

Energy Model Optimization

Training Protocol

IDEA is a residue-based, data-driven biophysical model that predicts sequence-specific protein-DNA binding affinities. The model outputs relative binding free energies between protein and DNA, which can be converted to the relative binding affinity (represented by the dissociation constant K_d) using:[7]

$$\Delta\Delta G_{\text{binding}} = RTln\Delta K_d \tag{1}$$

The model integrates both the sequence and structural information from experimentally determined protein-DNA structural interface into an interaction matrix for model training. These sequences, located at the protein-DNA structural interface (defined as those residues whose representative atoms have a distance $\leq$ 8Å, highlighted in blue in Figure 1 Left), are

hypothesized to have evolved to favor strong amino acid-nucleotide interactions. We collected the sequences from both the strong binders and decoy binders (considered weak binders) for model parameterization. Synthetic decoy binders were generated by randomizing either the DNA (1000 sequences) or protein sequences (10,000 sequences) of the strong binders (see Figure S15 for an assessment of the model's robustness in transferability and generalizability across different numbers of decoys).

IDEA optimizes an amino-acid-nucleotide energy matrix, $\gamma(a, n)$, by maximizing the gap in binding energies between the strong and decoy binders. Similar strategies have been successfully applied in protein folding and protein-protein binding.[38,40,56,98,99] Specifically, the solvent-averaged binding free energies for strong binders (E_{strong}) and their corresponding decoy weak binders (E_{decoy}) are calculated using the following expression:

$$E_{\text{binding}} = \sum_{i \in \text{protein}, j \in \text{DNA}} \gamma(a, n) \Theta_{i,j}(a_i, n_j) \tag{2}$$

Here, $\gamma(a, n)$ is a 20-by-4 residue-type-specific energy matrix encoding residue-type-specific interaction energies between amino acid (a) and nucleotide (n). $\Theta_{i,j}(a_i, n_j)$ utilizes a switching function that captures an effective interaction range between each amino acid-nucleotide pair within the protein-DNA complex:

$$\Theta_{i,j}(a_i, n_j) = 0.5 \cdot (\tanh(\kappa \cdot (r_{i,j} - r_{\min})) \cdot \tanh(\kappa \cdot (r_{\max} - r_{i,j}))) + 0.5 \tag{3}$$

here, $r_{\min} = -8\,\text{Å}$ and $r_{\max} = 8\,\text{Å}$. This defines two residues to be "in contact" if their distance is less than 8 Å. κ (here taken 0.7) is a scaling factor that modulates the steepness of the hyperbolic tangent function.

In practice, the γ matrix was optimized to maximize $\delta E / \Delta E$, where $\delta E = \langle E_{\text{decoy}} \rangle - \langle E_{\text{strong}} \rangle$ represents the average energy gap between strong and decoy binders, and ΔE is the standard deviation of decoy binding energies. Mathematically,

$$\delta E = A(a,n)^T \gamma(a,n) \tag{4}$$

and

$$\Delta E = \sqrt{\gamma(a,n)^T B(a,n) \gamma(a,n)} \tag{5}$$

with

$$A(a,n) = \langle \phi_{\text{decoy}} \rangle - \langle \phi_{\text{strong}} \rangle \tag{6}$$

$$B(a,n) = \left\langle \phi_{\text{decoy}} \phi_{\text{decoy}}^T \right\rangle - \langle \phi_{\text{decoy}} \rangle \left\langle \phi_{\text{decoy}}^T \right\rangle \tag{7}$$

Here, $\phi(a,n) = \sum_{i \in \text{protein}, j \in \text{DNA}} \Theta_{i,j}(a_i, n_j)$ summarizes the total number of contacts between each type of amino acid and nucleotide in the given protein-DNA complexes. The vector $A(a,n)$ quantifies the difference in interaction frequency between the strong and decoy binders, with a dimension of 1×300. The covariance matrix $B(a,n)$ is a 300×300 matrix that quantifies the propensity of co-occurring residue type interactions in the decoy binders. Maximizing the function $\delta E / \Delta E$ with respect to $\gamma(a,n)$ corresponds to maximizing $A^T \gamma / \sqrt{\gamma^T B \gamma}$, which can be achieved by maximizing the functional objective $R(\gamma)$:

$$R(\gamma) = A^T \gamma - \lambda \sqrt{\gamma^T B \gamma} \tag{8}$$

where λ is a Lagrange multiplier. Setting the derivative $\frac{\partial R(\gamma)}{\partial \gamma}$ equal to zero leads to $\gamma \propto B^{-1} A$. A filtering procedure is further applied to reduce the noise arising from the finite sampling of certain types of interactions: The matrix B is first diagonalized such that $B^{-1} = P \Lambda^{-1} P^{-1}$, where P is the matrix of eigenvectors of B, and Λ is a diagonal matrix composed of B's eigenvalues. We retained the principal N modes of B, which are ordered by the descending magnitude of their eigenvalues and replaced the remaining $300 - N$ eigenvalues in Λ with the Nth eigenvalue of B. In all optimizations reported in this work, $N = 70$ was generally a good choice to maximize the utilization of non-zero eigenvalues in the matrix

B. For visualization purposes, the vector γ is reshaped into a 20-by-4 matrix, as shown in Figure 2B.

To predict the binding free energy for a target protein-DNA pair, we substitute the target sequence into a known protein-DNA complex structure and recalculate the target ϕ matrix, which will be combined with the trained γ matrix to compute the target protein-DNA binding free energies using Eq. 2. Due to the presence of an undetermined scaling factor after the model optimization, the predicted free energies are presented in reduced units. Despite this, the relative free energies can be used to accurately rank binding affinities of target protein-DNA pairs. We utilize this computed solvent-averaged free energy to approximate the protein-DNA binding free energy in Equation 1. When the optimized energy is integrated into a simulation model, the IDEA-based simulations can predict the absolute binding free energies in physical units (Figure 5C).

Treatment of Complementary DNA Sequences

To replace the DNA sequence in the protein-DNA complex structure with a target sequence, IDEA uses sequence identity to determine whether the target sequence belongs to the forward or reverse strand of the DNA in the protein-DNA structure. The more similar strand is selected and replaced with the target sequence. As the orientations of test sequences are specified from 5' to 3' in all datasets used in this study (e.g., processed MITOMI, HT-SELEX, and ChIP-seq data), this approach ensures that the target sequences are replaced and evaluated correctly. In cases where sequence orientation is not provided (though this was not an issue in this study), we recommend replacing both the forward and reverse strands with the target sequence separately and evaluating the corresponding protein–DNA binding free energies. Since strong binders are likely to dominate the experimental signals, the higher predicted binding affinity, with stronger binding free energies, should be taken as the model's final prediction.

Enhanced Modeling Prediction with SELEX Data

When additional experimental binding data, such as SELEX,[27] is available, we extended IDEA to incorporate this data for enhancing the model's predictive capabilities. SELEX data provides binding affinities between given transcription factors and DNA sequences, which can be converted into dimensionless binding free energy using Eq. 1, with K_d being the normalized affinity generated from the SELEX package. We then calculated the ϕ value by utilizing the protein-DNA complex structure, replacing the native DNA sequence with the SELEX-provided sequences. These ϕ values can be used together with the converted $\Delta\Delta G$ values to compute the γ energy model based on the equation 2:

$$\gamma = \phi^{-1}\Delta\Delta G \tag{9}$$

here γ is the enhanced protein-DNA energy model represented as a 300×1 vector. We estimated γ by applying ridge regression to the linear system in Equation 9, using the previously computed ϕ matrix and the SELEX-derived $\Delta\Delta G$ values. A regularization parameter of $\alpha = 0.01$ was consistently applied throughout all computations to prevent overfitting.

For predicting the MAX-DNA binding specificity, we utilized the Human MAX SELEX data deposited in the European Nucleotide Archive database (accession number PRJEB25690).[27,100] The R/Bioconductor package SELEX version 1.30.0[101] was used to determine the observed R1 count for all 10-mers. Among them, ACCACGTGGT is the motif with the highest frequency, which aligns with the DNA sequence in the MAX-DNA PDB structure (PDB ID: 1HLO), whose full DNA sequence is CACCACGTGGT. Consequently, we used this motif as the reference to align the remaining SELEX-measured 10-mer sequences. A total of 255 10-mer sequences, corresponding to the DNA variants measured from the MITOMI experiment,[13] were selected. We threaded the PDB structure with these sequences to construct the ϕ and utilized the SELEX-converted binding free energies to compute the γ energy model.

Evaluation of IDEA Predictions Using HT-SELEX data

The processed HT-SELEX data used in this study are from,[26] which contains processed DNA binding sequence motifs and their normalized copy number detected from the HT-SELEX experiment, referred to as M-word scores. A higher M-word score corresponds to a stronger binding sequence motif detected in the HT-SELEX experiments. Among all the proteins in the data, 23 have experimentally determined protein-DNA complex structures. We predicted the binding free energies of all the processed sequences of these proteins using our protocol. For evaluating IDEA predictions, we classified motifs with M-word scores above 0.9 as strong binders (label 1) and those with M-word scores below 0.3 as weak binders (label 0). In cases where no sequences had an M-word score below 0.3, we used alternative cutoffs of 0.4 or 0.5 to ensure that at least three weak-binding sequences were included. These labeled strong and weak binders were then used as the ground truth for evaluating IDEA's predictions. We assessed IDEA's predictive accuracy by calculating the probability density distributions of the predicted binding free energies (Figure S4). A well-separated distribution between strong and weak binders indicates a successful prediction. Additionally, we calculated the AUC score and balanced PRAUC score based on the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) analyses of these probability distributions. More details about the evaluation metrics and comparisons with popular existing models are provided in the Appendix *Section: Comparison of IDEA predictive performance Using HT-SELEX data*

Selection of CATH Superfamily for Testing Model Transferability

The MAX protein is characterized by a Helix-loop-helix DNA-binding domain and belongs to the CATH superfamily 4.10.280.10,[102] which includes a total of 11 protein-DNA structures with the E-box sequence (CACGTG), similar to the MAX-binding DNA. We hypothesized that proteins within the same superfamily exhibit common residue-interaction patterns due to their structural and evolutionary similarities. To test the transferability of our model, we conducted individual training on each one of those 11 protein-DNA complex structures

465 and their associated sequences, predicting the MAX-DNA binding affinity in the MITOMI
466 dataset.[13] To examine the connection between the predictive outcome and the probability
467 that the training protein is homologous to MAX, HHpred[103] was utilized to search for MAX's
468 homologous proteins, using the sequence of 1HLO as the query (Figure 3A).

469 Processing of ChIP-seq Data

470 To evaluate the predictive capabilities of our protocol on specific chromosomal segments, we
471 selected ChIP-seq data for three transcription factors—MAX, EGR1, and CTCF. The data
472 were sourced from the ENCODE project, with accession numbers ENCFF361EVH (MAX),
473 ENCSR026GSW (EGR1), and ENCFF960ZGP (CTCF).

474 For testing, we chose genomic regions with the densest ChIP-seq signals. For MAX
475 and CTCF, we focused on the human GM12878 chromosome 1 (GRCh38) region between
476 156,000,000 bp and 157,000,000 bp, as this 1 Mb segment contains the densest signals for both
477 transcription factors, with 33 MAX binding sites and 70 CTCF binding sites. For EGR1,
478 we used the HepG2 cell line, focusing on the GRCh38 chr8 region between 143,000,000 bp
479 and 144,000,000 bp, which contains 97 peaks.

480 To predict binding sites, we scanned the corresponding DNA regions base by base using
481 the DNA sequence closest to each protein in the relevant protein-DNA complex structures.
482 For MAX, we used the 11-bp sequence from the PDB structure 1HLO, generating $1,000,000-$
483 $11+1=999,990$ sequences. For EGR1, we used the 10-bp sequence from PDB 1AAY,
484 producing $1,000,000-10+1=999,991$ sequences. For CTCF, no single PDB structure
485 covers the entire protein,[104] so two PDB structures were used: 8SSS (ZF1-ZF7, 23-bp) and
486 8SSQ (ZnFs 3-11, 35-bp), generating $1,000,000-23+1=999,978$ and $1,000,000-35+1=$
487 $999,966$ sequences, respectively.

488 Predicted binding free energies for each transcription factor were normalized into Z-scores

using randomized decoy sequences:

$$Z_{\mathrm{norm}} = \frac{E - \langle E_{\mathrm{decoy}} \rangle}{\mathrm{SD}(E_{\mathrm{decoy}})} \tag{10}$$

where $\langle\ \rangle$ is the average over all decoy sequences, and SD is the standard deviation. Given the evolutionary preference for strong binding, most Z-scores were negative, indicating stronger binding. ROC curves were then constructed to assess the predictive performance, with lower predicted Z-scores representing stronger binding affinity (Figure 4, S11).

Sequence Specific Protein-DNA Simulation Model

We utilized a previously developed protein and DNA residue-level coarse-grained model for simulating protein-DNA binding interactions. The protein was modeled using the Cα-based structure-based model,[60] with each amino acid represented by a simulation bead located at the Cα site. The DNA was modeled with the 3-Site-Per-Nucleotide (3SPN) DNA model,[61] with each DNA nucleotide modeled by three coarse-grained sites located at the centers of mass of the phosphate, sugar, and base sites. Both these two models and their combination have been validated to reproduce multiple experimental measurements.[60,61,64,66–69,71,72,75,105]

Detailed descriptions of this protein-DNA force field have been documented in previous studies.[63,68,69,72] Specifically, we followed Ref. 61 for simulating DNA-DNA interactions, and Ref. 60 for simulating the protein-protein interactions. The protein-DNA interactions were modeled with a long-range electrostatic interaction and a short-range sequence-specific interaction. The electrostatic interactions were simulated with the Debye-Hückel treatment to capture the screening effect at different ionic concentrations:

$$U_{\mathrm{elec}}(r) = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{\epsilon r} e^{-r/l_D} \tag{11}$$

where $\epsilon = 78.0$ is the dielectric constant of the bulk solvent. q_{i} and q_{j} correspond to the charges of the two particles. $\epsilon_0 = 1$ is the vacuum electric permittivity, and l_D is the Debye-

Hückel screening length.

The excluded volume effect was modeled using a hard wall potential with an apparent radius of $\sigma = 4.0$Å:

$$U_{\text{exclude}}(r) = 4\epsilon(\frac{\sigma}{r})^{12} \tag{12}$$

An additional sequence-specific protein-DNA interaction was modeled with a Tanh function between the phosphate site of the DNA and the amino acid, with the form:

$$V_{\text{PD}}(r) = \frac{W}{2}(1 + \tanh[\eta(r_0 - r)]) \tag{13}$$

where $r_0 = 8$ Å and $\eta = 0.7$ Å^{-1}. W is a 4×20 matrix that characterizes the interaction energy between each type of amino acid and DNA nucleotide. Here, We used the normalized IDEA-learned energy model based on 9 protein-DNA complexes (Figure 5A). The IDEA-learned energy model was normalized by the sum of the absolute value of each term of the energy model. Detailed parameters are provided in Table S4. Additional details on the protein-DNA simulation are provided in the Supplementary Materials.

Appendix

Details of Molecular Dynamics Simulation

The simulations were carried out using the LAMMPS software[106] with the implementation of the protein-DNA force field (Table S4). We performed umbrella-sampling simulations[78] for comprehensive sampling of the binding and unbinding events to compute the protein-DNA binding affinities. To evaluate the model's performance, we selected 9 protein-DNA complexes with binding affinity measured by the same experimental lab.[77,107–110] Initial simulation configurations were prepared based on the crystal structures of the protein-DNA complexes. The PDB IDs of these complexes are provided in Figs 5C, S13, and S14.

The simulations were conducted under the same experimental conditions with 100 mM

monovalent ions. We used a spring constant of 0.02 kcal/mol/Å^2 based on the center of mass (COM) distances between the protein and DNA. The centers of the umbrella windows were placed on a grid ranging from 0.0 to 140.0Å, with a grid size of 10.0Å, and each umbrella simulation lasted for 22 million steps, with a time step of 2.0 fs. We simulated 6 copies of the simulation with different random seeds to ensure sampling convergence. When reweighting the simulation to compute the free energy[79] (Figs. S13 and S14), we excluded the first 2 million steps of each simulation trajectory. To calculate the error bars from our simulation, we divided each simulation trajectory into three equal-length blocks and independently calculated the reweighted free energies from them. The error bar was then estimated as the standard deviation of the three free energy profiles.

Comparison of IDEA predictive performance Using HT-SELEX data

To benchmark the performance of IDEA against state-of-the-art protein-DNA predictive models, we evaluated its ability to recognize strong binders with the HT-SELEX datasets across 22 proteins from 12 families.[26] Specifically, we compare IDEA with two widely used sequence-based models: ProBound[31] and DeepBind.[24] ProBound has demonstrated superior performance over many other predictive protein-DNA models, including JASPAR 2018,[111] HOCOMOCO,[112] Jolma et al.,[113] and DeepSELEX.[114] To use ProBound, we retrieved the trained binding model for each protein from motifcentral.org and used the GitHub implementation of ProBoundTools to infer the binding scores between protein and target DNA sequences. Except for POU3F1, binding models are available for all proteins. Therefore, we excluded POU3F1 and evaluated the protein-DNA binding affinities for the remaining 21 proteins. To use DeepBind, sequence-specific binding affinities were predicted directly with its web server. The Area Under the Curve (AUC) and the Precision-Recall AUC (PRAUC) scores were used as metrics for comparison. An AUC score of 1.0 indicates a perfect separation between the strong- and weak-binder distributions, while an AUC score of 0.5 indicates

no separation. Because there is a significant imbalance in the number of strong and weak binders from the experimental data,[26] where the strong binders are far fewer than the weak binders, we reweighted the samples to achieve a balanced evaluation, using 0.5 as the baseline for randomized prediction.[115] As summarized in Figure S6, Table S1, and Table S2, IDEA ranked second among the three predictive models. In order to assess the performance of IDEA when augmented with additional protein-DNA binding data, we augmented IDEA using randomly selected half of the HT-SELEX data (see the Methods Section *Enhanced Modeling Prediction with SELEX Data*). The augmented IDEA model achieved the best performance among all the models.

In addition, we compared the performance of IDEA with both general and family-specific knowledge-based energy models. First, we incorporated a knowledge-based generic protein-DNA energy model (DBD-Hunter) learned from the protein-DNA database, reported by Skoinick and coworkers,[46] into our prediction protocol. This model assigns interaction energies to different functional groups within each DNA nucleotide, including phosphate (PP), sugar (SU), pyrimidine (PY), and imidazole (IM) groups. For our comparison, we averaged the energy contributions of these groups within each nucleotide and replaced the IDEA-learned energy model with the DBD-Hunter model to assess its ability to differentiate strong binders from weak binders in the HT-SELEX dataset.[26] Additionally, we compared IDEA with rCLAMPS, a family-specific energy model developed to predict protein-DNA binding specificity in the C2H2 and homeodomain families. rCLAMPS learns a position-dependent amino-acid-nucleotide interaction energy model. To incorporate this model into the binding free energy calculation, we averaged the energy contributions across all occurrences of each amino-acid-nucleotide pair, which resulted in a 20-by-4 residue-type-specific energy matrix. This matrix is structurally analogous to the IDEA-trained energy model and can be directly integrated into the binding free energy calculations. As shown in Figure S6, Table S1, and Table S2, the IDEA model generally outperforms DBD-Hunter and rCLAMPS, demonstrating that it can achieve better predictive accuracy than both generic and family-specific

knowledge-based models.

We also performed 10-fold cross-validation using the same HT-SELEX datasets, following the protocol described in the Methods Section *Enhanced Modeling Prediction with SELEX Data*. For each protein, we divided the entire dataset into 10 equal, randomly assigned folds. In each iteration, we used randomly selected 9 of the 10 folds as the training dataset and the remaining fold as the testing dataset. This process was repeated 10 times so that each fold served as the test set once. We then reported the average R^2 scores across these iterations to evaluate IDEA's predictive performance. Our results are compared with the 1mer and 1mer+shape methods from,[26] the latest regression model that considers the shape of DNA with different sequences (Figure S7). This comparative analysis shows IDEA achieved higher predictive accuracy than the state-of-the-art sequence-based protein-DNA binding predictors for protein-DNA complexes that have available experimentally resolved structures.

Overall, these results demonstrate that IDEA can be used to predict the protein-DNA pairs in the absence of known binding sequence data, thus filling an important gap in protein-DNA predictions when experimental binding sequence data are unavailable.

Analysis of examples where IDEA fails to recognize strong DNA binders

We examine IDEA's capability in identifying strong binders from the HT-SELEX dataset across 12 protein families.[26] The model successfully predicts 18 out of 22 protein-DNA systems, but the performance is reduced in 4 cases. Closer investigations revealed the source of these limitations. In some instances, only the protein from a different organism is available. For example, the PDX1 HT-SELEX data utilized the human PDX1 protein, but no human PDX1–DNA complex structure is available. Therefore, the mouse PDX1–DNA complex structure (PDB ID: 2H1K) was used for model training. Differences between model organisms may reduce predictive accuracy. A similar limitation applies to POU3F1, where an available mouse complex (PDB ID: 4Y60) was used to predict human protein–DNA in-

teractions. Notably, DeepBind,[24] a sequence-based prediction tool, also failed to distinguish strong from weak binders when using the mouse POU3F1 protein (AUC score: 0.457), but this was corrected with the human POU3F1 protein (AUC score: 0.956).

IDEA also fails to fully resolve the binding preference of USF1. A closer examination of the HT-SELEX data reveals a lack of distinction among the sequences, as most sequences, including those with the lowest M-word (binding affinity) scores, contain the DNA-binding E-box sequence CACGTG. Therefore, USF1 represents a challenging example where the experimental data only consists of strong binders with limited variations in binding affinity, which likely results from differences in flanking sequences of the E-box motif.

Egr1 stands as a peculiar example. Whereas IDEA does not effectively distinguish between the strong and weak binders in the current HT-SELEX dataset, its predictions are consistent with other experimental datasets, including binding affinities measured by k-MITOMI[43] (Figure S8A, B), preferred binding sequences from protein-binding microarray, an earlier HT-SELEX experiment, and bacterial one-hybrid data.[7] Therefore, further investigation of the current HT-SELEX data is needed to reconcile these differences.

Supplementary Figures

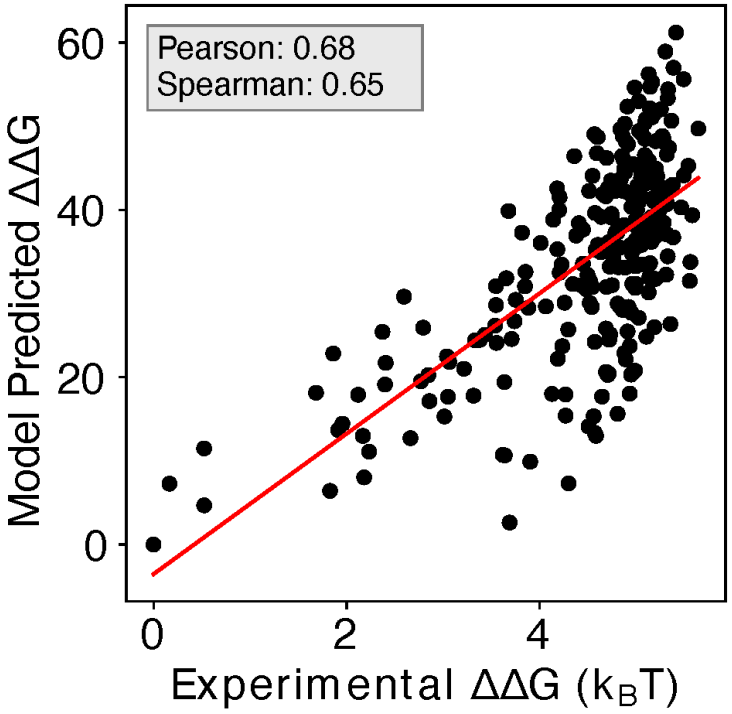

Figure S1: **Including additional human Max-DNA complex structures in IDEA training improves prediction.** Including additional human-associated MAX-DNA complex structures and their associated sequences (PDB IDs: 1HLO, 1NLW, 1NKP) leads to a minor improvement in the correlation between the predicted and experimental binding free energies, with a Pearson correlation coefficient of 0.68 and Spearman's rank correlation coefficient of 0.65. $\Delta\Delta G$ represents the changes in binding free energy relative to the protein binding to the wild-type DNA sequence. The predicted binding free energies are presented in reduced units, as explained in the Methods Section: *Training Protocol*.

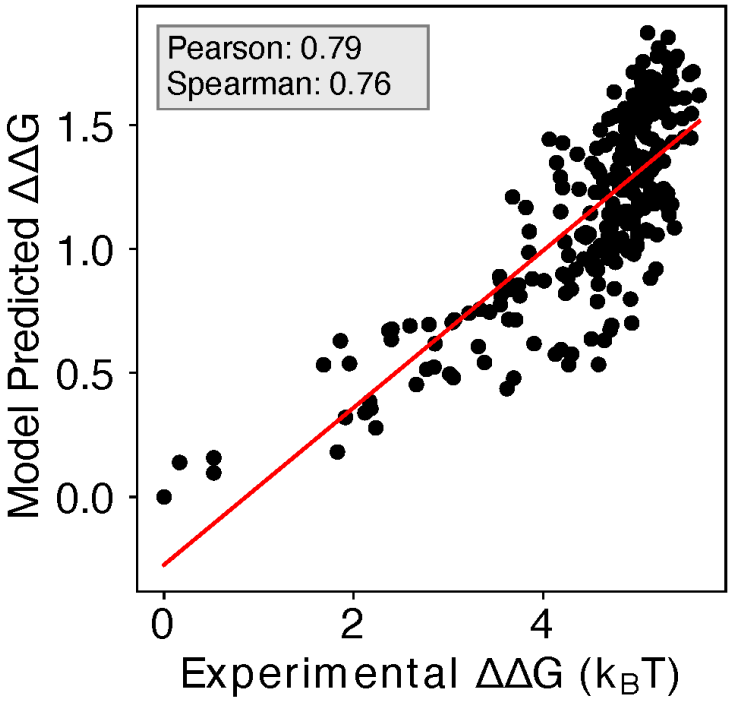

Figure S2: **Enhanced IDEA prediction by integrating SELEX-seq data for MAX transcription factor.** Incorporating SELEX-seq data[27] into IDEA's training protocol significantly improves its predictive accuracy on MITOMI binding measurements[13] for the MAX transcription factor, achieving a Pearson correlation coefficient of 0.79 and a Spearman's rank correlation coefficient of 0.76.

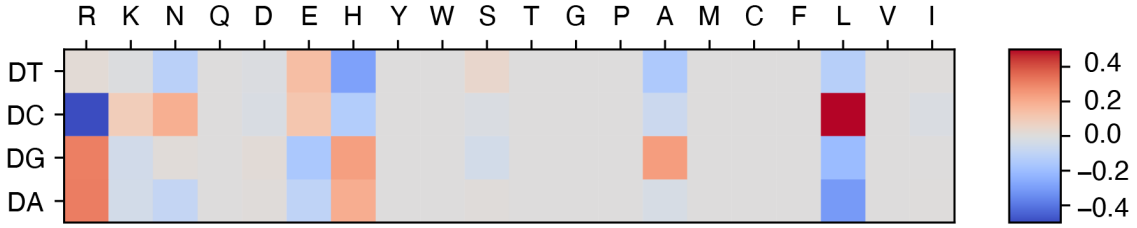

Figure S3: **Refined energy model with SELEX data integration reveals additional physicochemical insights into protein-DNA interactions.** Upon integrating the SELEX data into our model training, we found that the refined energy model reveals additional unfavorable interactions between glutamic acid (E) and deoxycytidine (DC), consistent with their negative charges. The optimized energy model is presented in reduced units, as explained in the Methods Section: *Training Protocol*

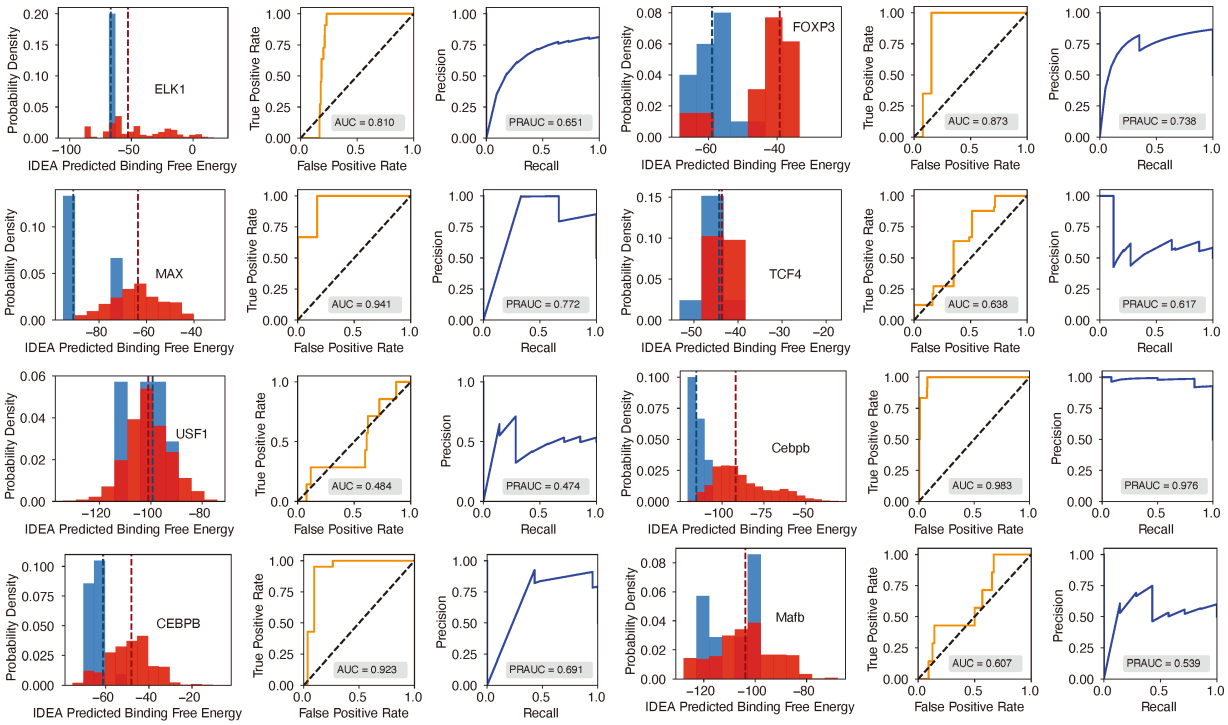

Figure S4: **Evaluation of IDEA's predictive accuracy for distinguishing strong from weak protein-DNA binding interactions (Part 1 of 3).** Probability density distributions of predicted binding free energies for strong (blue) and weak (red) binders identified from the HT-SELEX experiment[26] are shown across multiple protein families. A clear separation between these distributions indicates the model's effectiveness in identifying high-affinity DNA binders. The darker dashed lines within each distribution correspond to the median binding free energy for strong (blue) and weak (red) binders, respectively. For Mafb, the predicted median binding free energies for strong and weak binders are nearly identical, measured at -103.826 and -103.824, respectively. As a result, the dashed lines almost completely overlap, and only one appears visible in the plot. The Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve adjacent to each density plot provide a quantitative assessment of prediction accuracy, with the Area Under the Curve (AUC) and precision-recall AUC (PRAUC) scores displayed in each panel.

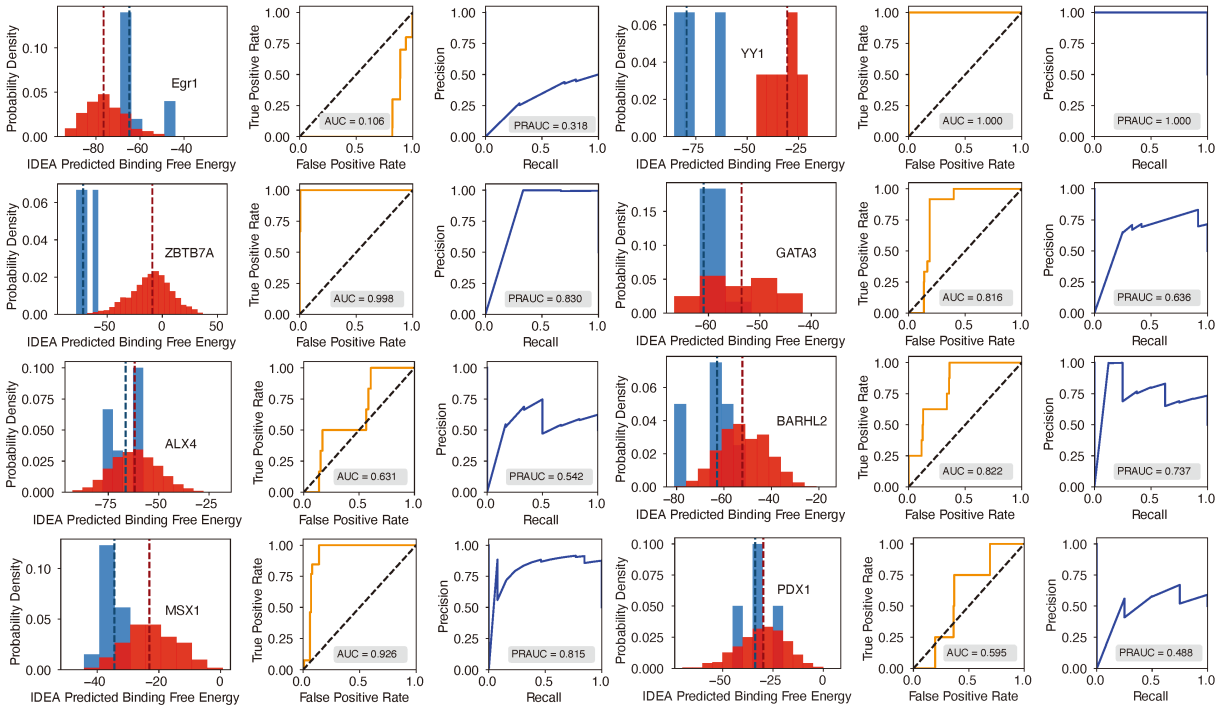

Figure S4: Evaluation of IDEA's predictive accuracy for distinguishing strong from weak protein-DNA binding interactions (Part 2 of 3).

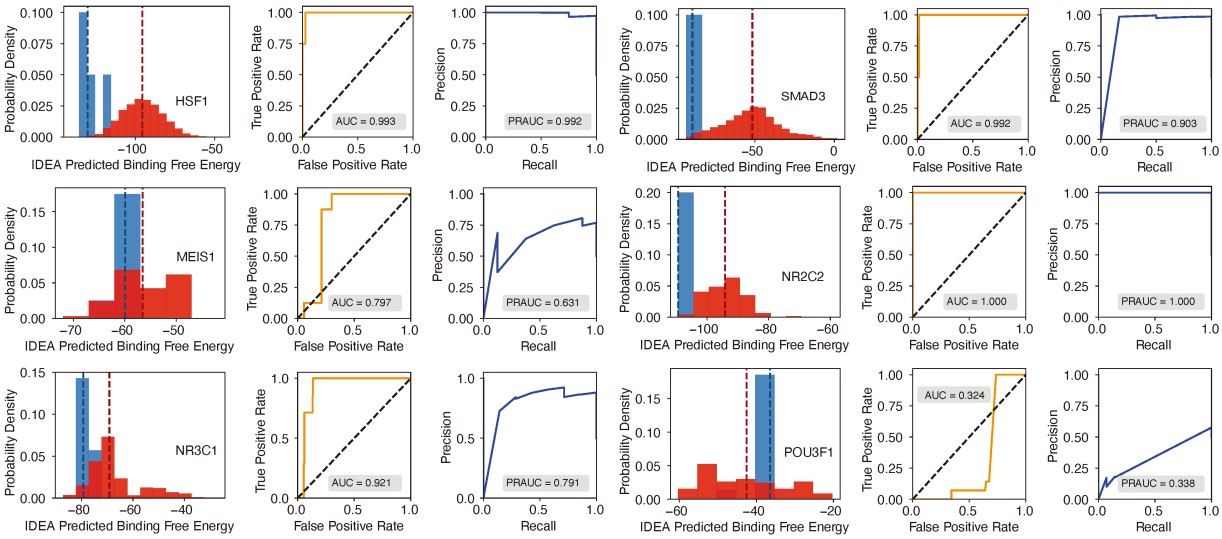

Figure S4: Evaluation of IDEA's predictive accuracy for distinguishing strong from weak protein-DNA binding interactions (Part 3 of 3).

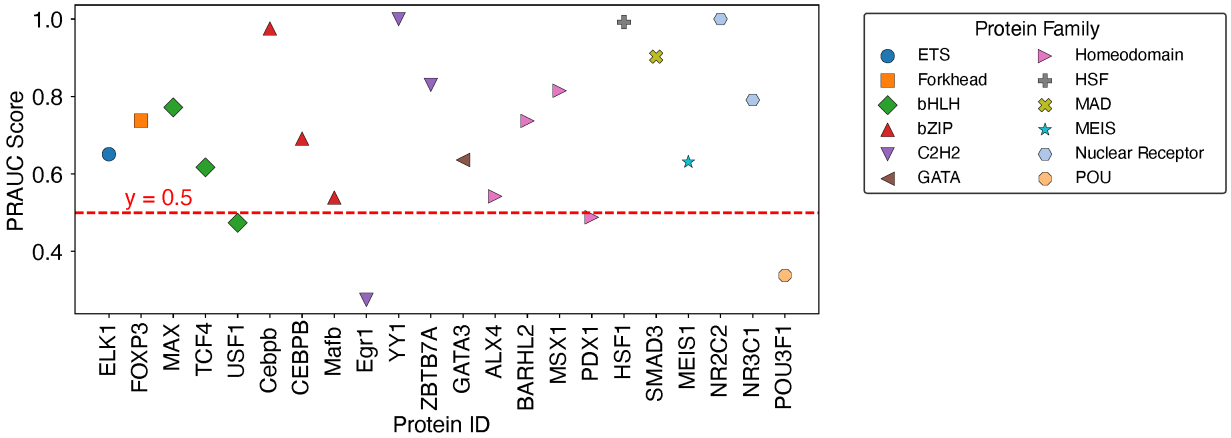

Figure S5: **Summary of balanced PRAUC scores for protein-DNA pairs across 12 protein families.**

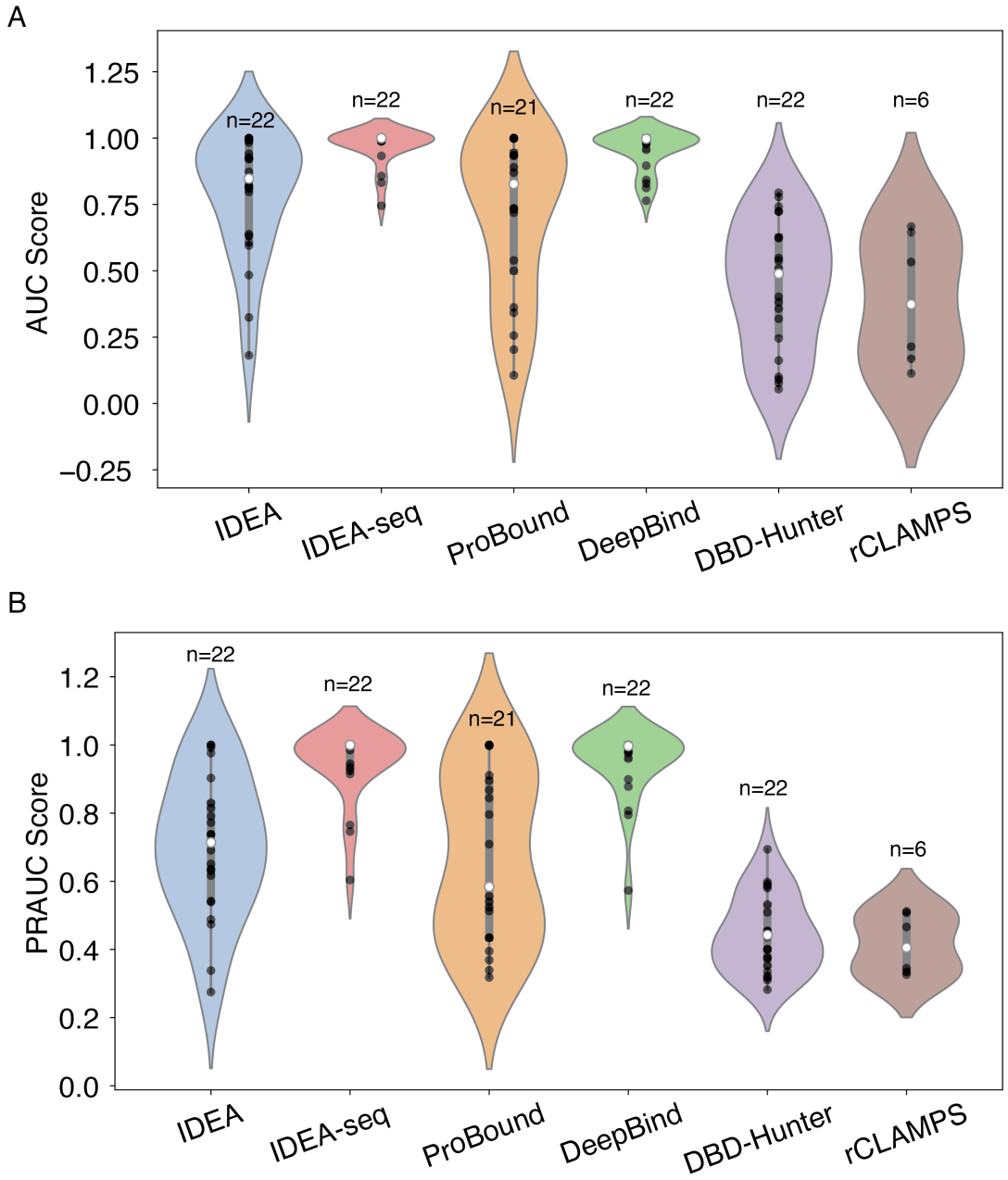

Figure S6: **Performance comparison of IDEA model with other prediction methods.** (A) AUC scores and (B) PRAUC scores for identifying strong binders were evaluated across six different prediction methods: IDEA, IDEA augmented with binding sequences (IDEA-seq), ProBound,[31] DeepBind,[24] the general knowledge-based energy model DBD-Hunter,[46] and family-specific knowledge-based energy model rCLAMPS.[36] Predictive performances were assessed using HT-SELEX datasets spanning 22 proteins from 12 protein families. Each violin plot displays the distribution of scores for individual targets, with the width indicating score density. The thick grey bar represents the interquartile range (first to third quartiles), and the thin line extends to 1.5 times the interquartile range. Individual data points are depicted as scattered black dots, and the white dot represents the median. Sample sizes (n) are labeled above each group.

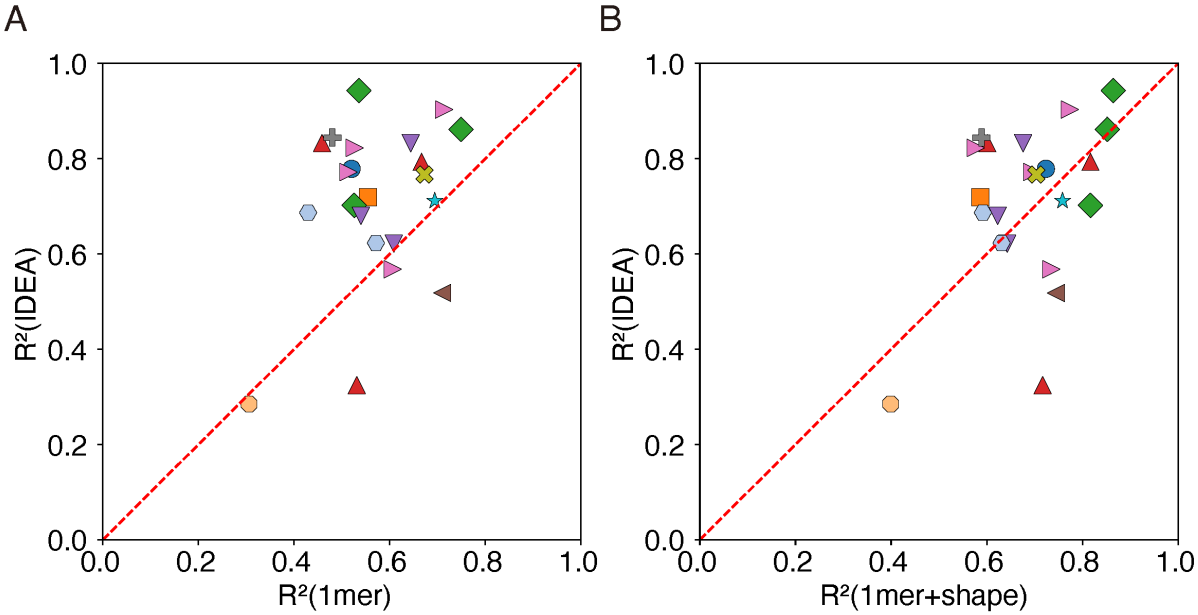

Figure S7: **IDEA outperforms other predictors in cross-validation analysis of protein-DNA binding affinity.** Results from a 10-fold cross-validation indicate that IDEA achieves higher R^2 scores for most of the protein-DNA complexes that have available experimentally resolved structures, compared to the 1mer and 1mer+shape methods reported in.[26] In panel (A), each point represents a protein-DNA complex and compares IDEA's R^2 score with that of the 1mer method, while panel (B) compares IDEA's R^2 score with the 1mer+shape method. Points are color-coded by protein family (see Figure 2E for family annotations).

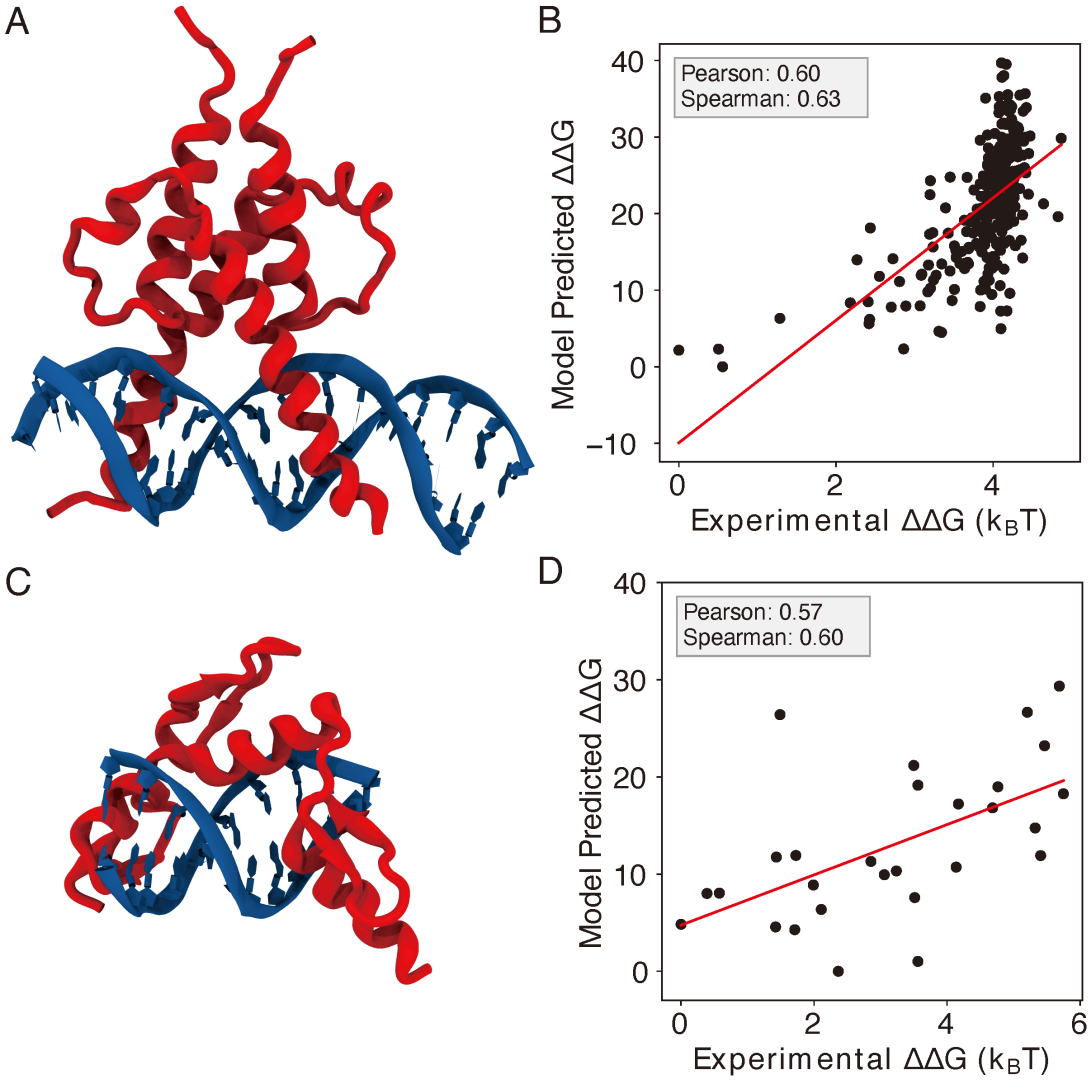

Figure S8: **IDEA correctly predicts the protein-DNA recognition by additional transcription factors.** (A) The 3D structure of basic helix-loop-helix (bHLH) transcription factor PHO4 and its associated DNA (PDB ID: 1A0A). (B) Model-predicted binding free energies of PHO4 correlate with the experimentally determined binding free energies.[13] (C) The 3D structure of zinc finger protein Zif268 and its associated DNA (PDB ID: 1AAY). (D) Model-predicted binding free energies of Zif268 correlate with the experimentally determined binding free energies.[43] $\Delta\Delta G$ represents the changes in binding free energy relative to the protein binding to the wild-type DNA sequence.

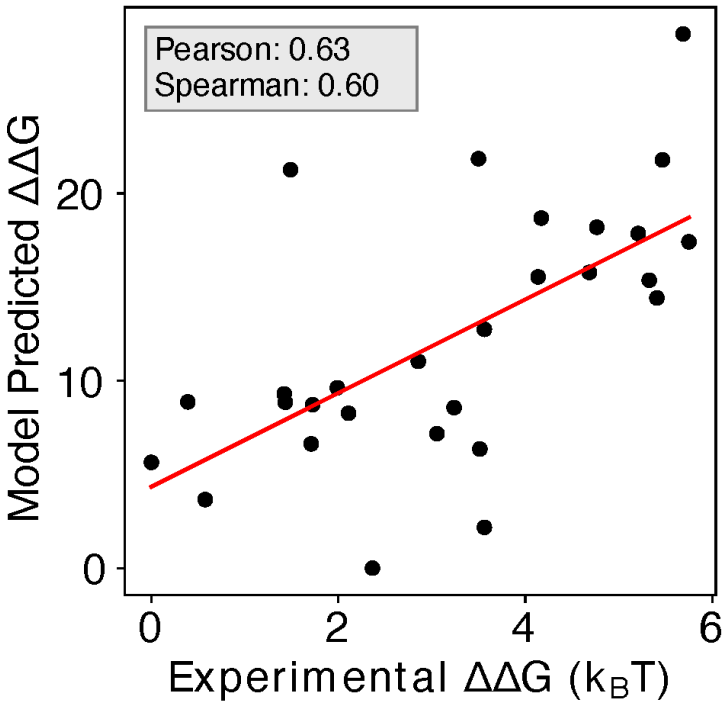

Figure S9: **Enhanced Predictive Accuracy with Inclusion of Zif268 and Related CATH Protein Structures.** Including the structures and associated sequences of 1AAY and other protein-DNA complexes from the same CATH zinc finger superfamily (CATH ID: 3.30.160.60) in the training dataset enhances the predictive accuracy, with a Pearson correlation coefficient of 0.63 and Spearman's rank correlation coefficient of 0.60.

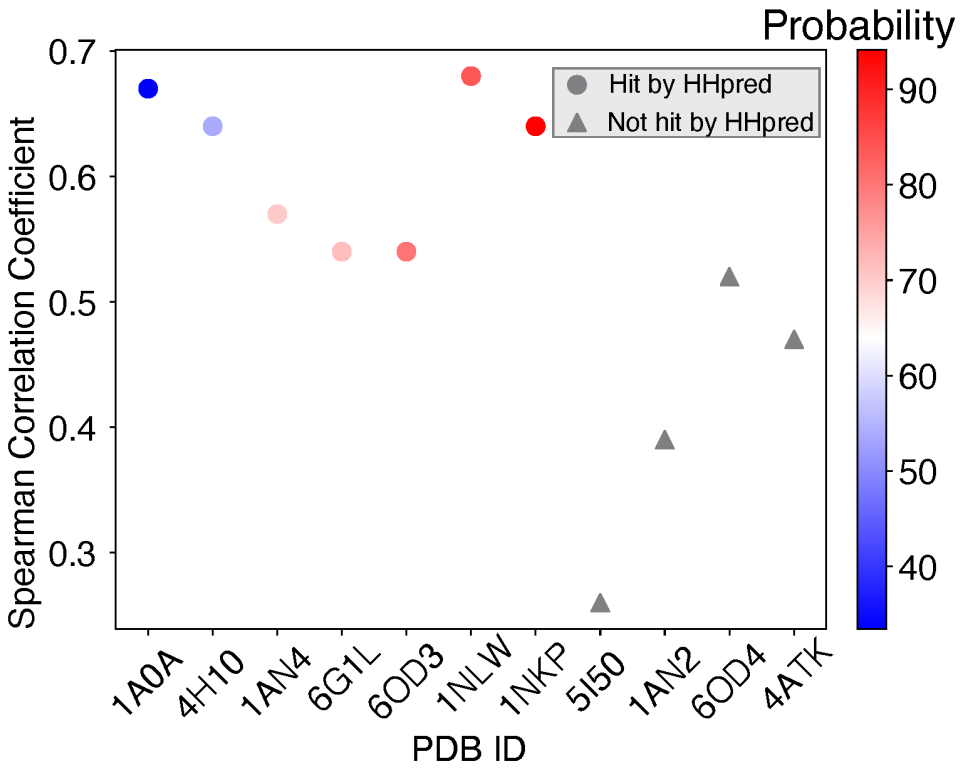

Figure S10: **Spearman's rank correlation coefficients between predicted and experimental MAX binding affinities across training proteins ordered by probability of being homologous to the MAX protein.**

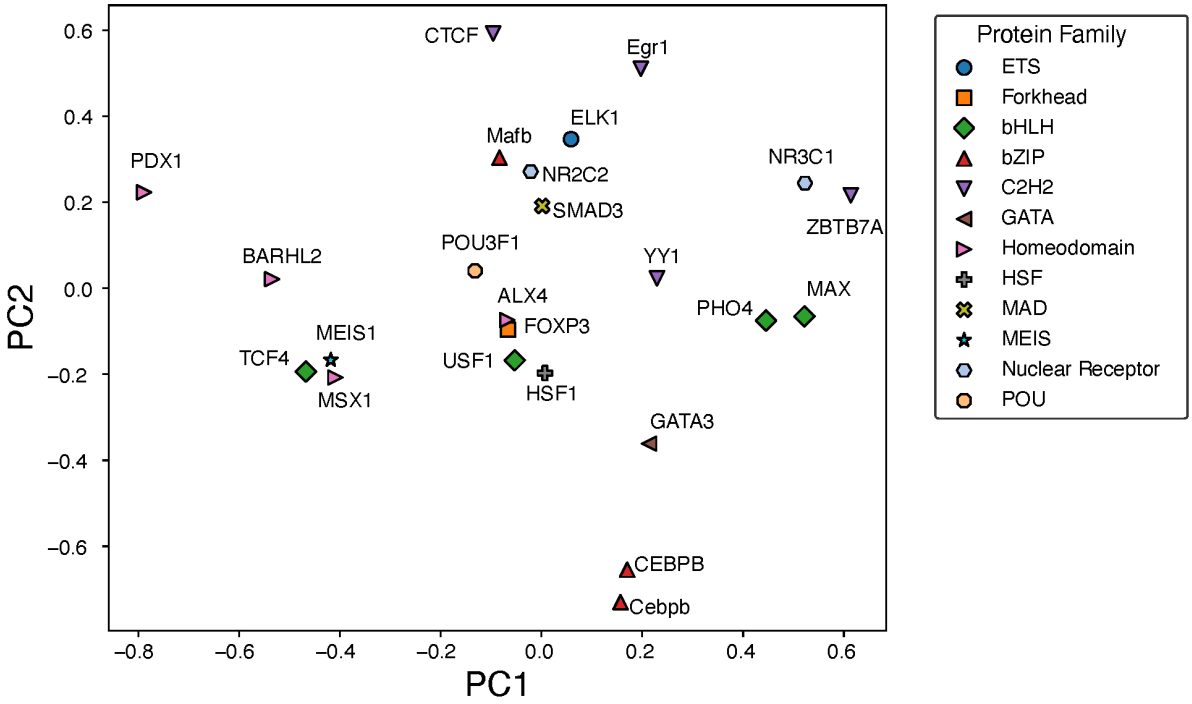

Figure S11: **Principal component analysis of the normalized IDEA-learned energy model across 12 protein families.**

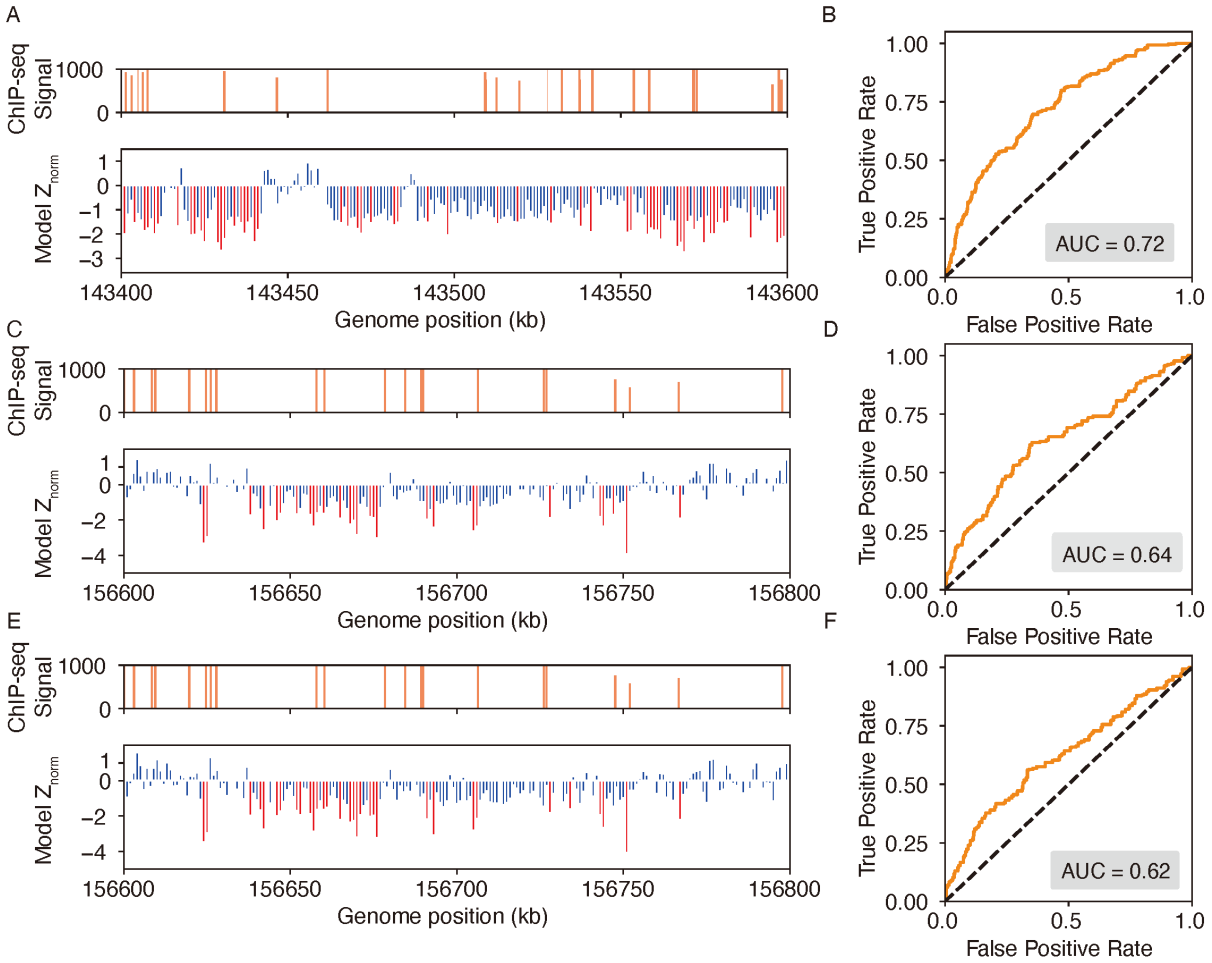

Figure S12: **IDEA accurately identifies genomic binding sites for additional proteins.** (A, B) Predicted binding sites for the EGR1 transcription factor in a 1 Mb region on chromosome 8 of the HepG2 cell line, where ChIP-seq signals are densest. The top plot of (A) shows ChIP-seq signals, while the bottom plot shows IDEA's predicted binding sites, with highly probable sites marked as red peaks. Normalized Z scores were averaged over a 500 bp window to align with the experimental resolution. The ROC curve (B) shows an AUC score of 0.72, reflecting the IDEA's accuracy for EGR1 binding sites prediction. (C–F) Predicted binding sites for the CTCF (CCCTC-Binding factor) protein in the GM12878 cell line on chromosome 1 using two different training structures: 8SSS (C, D) and 8SSQ (E, F). Panels (C) and (E) show IDEA's predictions compared to ChIP-seq signals, with highly probable binding sites highlighted as red peaks. Panels (D) and (F) present ROC curves with AUC scores of 0.64 and 0.62, respectively, indicating IDEA's predictive performance using each training structure.

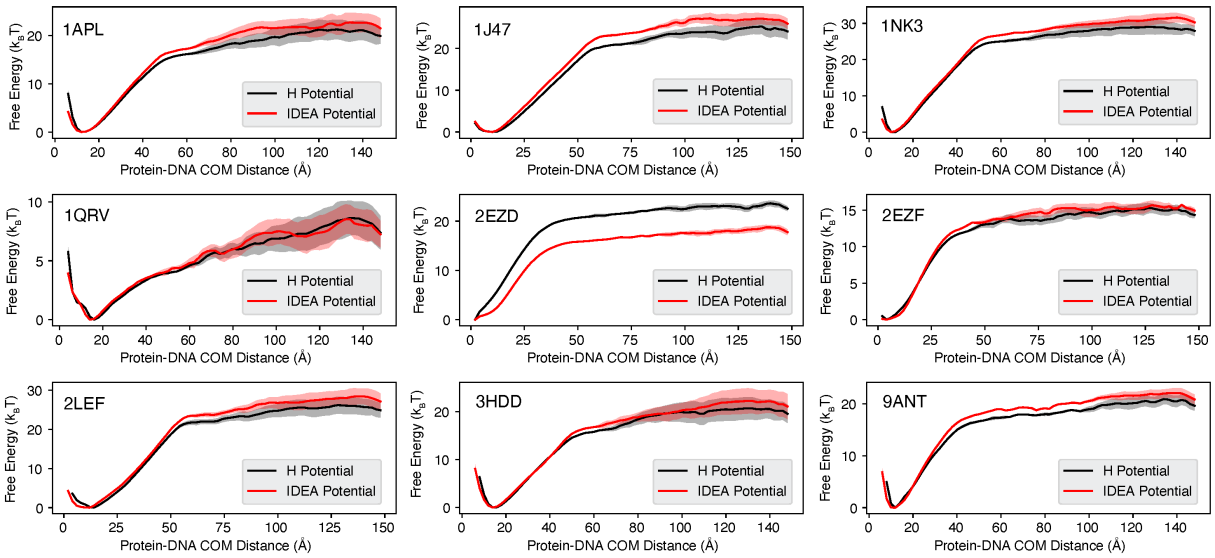

Figure S13: **Binding free energy curves calculated from simulations based on non-sequence-specific homogenous electrostatic potential and IDEA potential models.** The predicted free energy profiles were computed as a function of the Protein-DNA center of mass (COM) distance, comparing the non-sequence-specific homogenous electrostatic potential (H Potential) model (black) and the IDEA potential model (red). The line represents the mean free energy calculated from three equal partitions of the simulation trajectories, and the shaded areas were calculated as the standard deviation.

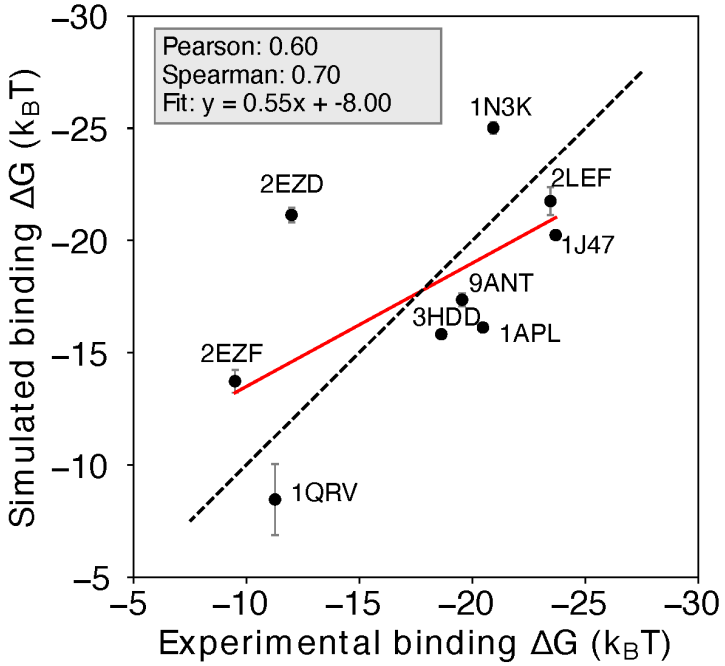

Figure S14: **Prediction of protein-DNA binding free energy with the non-sequence-specific homogeneous electrostatic potential model.** Comparison between the simulation-predicted and experimentally-assessed binding free energies for 9 protein-DNA complexes. The simulation free energies were computed using a non-sequence-specific homogenous electrostatic potential (H potential) model. The predicted binding free energies are presented in physical units. Error bars represent the standard deviation of the mean. See Figure S13 for additional details.

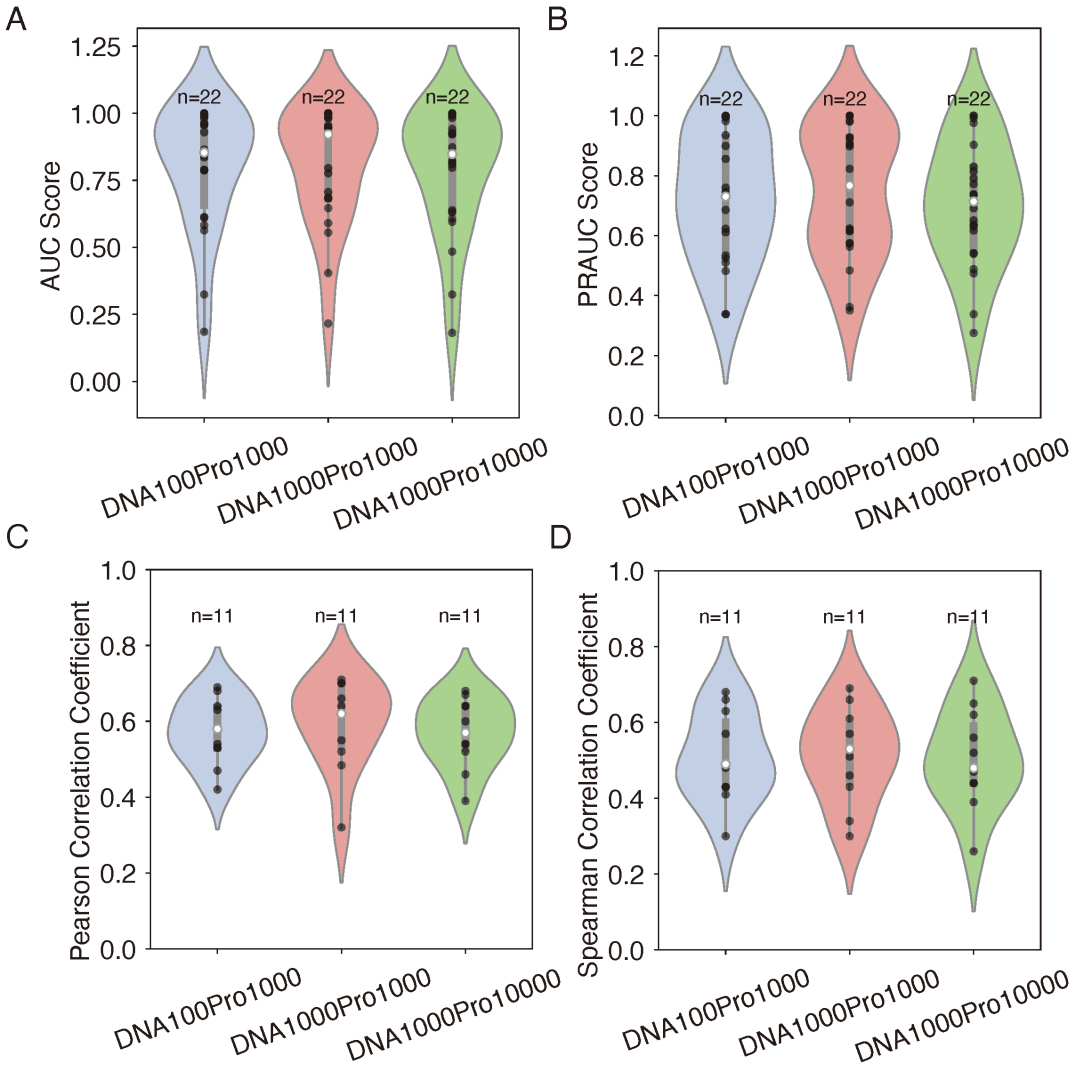

Figure S15: **Effect of the number of decoy sequences on model generalizability and transferrability.** (A-B) Evaluation of IDEA's generalizability across different decoy numbers, corresponding to Figure 2E. Violin plots summarize the distribution of AUC (A) and PRAUC (B) scores for three decoy combinations. (C-D) Analysis of IDEA transferability within the MAX CATH superfamily, corresponding to Figure 3A. Violin plots summarize the distribution of Pearson correlation coefficients (C) and Spearman's rank correlation coefficients (D) between predicted and experimental binding affinities for three decoy combinations. In each violin plot, the thick grey bar represents the interquartile range (first to third quartiles), and the thin line extends to 1.5 times the interquartile range. Individual data are depicted as scattered black points, and the white dot represents the median. Sample sizes (n) are labeled above each group. The three tested decoy combinations include: 100 DNA + 1000 protein decoys (DNA100Pro1000), 1000 DNA + 1000 protein decoys (DNA1000Pro1000), and 1000 DNA + 10000 protein decoys (DNA1000Pro10000). The consistent performance across all decoy combinations demonstrates the robustness of the model's prediction with respect to the number of decoy sequences.

Supplementary Tables

Table S1: Raw AUC scores for distinguishing strong and weak binders across 22 proteins from 12 protein families using six different prediction methods: IDEA, IDEA augmented with binding sequences (IDEA-seq), ProBound,[31] DeepBind,[24] DBD-Hunter[46] and rCLAMPS[36]

ID	IDEA	IDEA-seq	ProBound	DeepBind	DBD-Hunter	rCLAMPS
ELK1	0.810	0.745	0.998	1.000	0.626	–
FOXP3	0.873	1.000	0.342	0.812	0.162	–
MAX	0.941	1.000	0.736	0.995	0.490	–
TCF4	0.638	1.000	0.827	0.842	0.090	–
USF1	0.484	1.000	1.000	0.978	0.623	–
Cebpb	0.983	1.000	0.934	1.000	0.509	–
CEBPB	0.923	1.000	0.362	1.000	0.357	–
Mafb	0.607	0.987	0.944	0.998	0.794	–
Egr1	0.181	0.857	0.106	0.983	0.319	0.214
YY1	1.000	1.000	0.500	1.000	0.778	0.667
ZBTB7A	0.998	1.000	0.890	0.764	0.548	–
GATA3	0.816	0.933	0.870	0.896	0.488	–
ALX4	0.631	1.000	0.998	1.000	0.100	0.168
BARHL2	0.822	0.999	0.934	0.987	0.723	0.533
MSX1	0.926	0.998	0.732	0.829	0.382	0.113
PDX1	0.595	0.832	0.719	0.997	0.539	0.645
HSF1	0.993	1.000	1.000	1.000	0.742	–
SMAD3	0.992	0.999	1.000	1.000	0.402	–
MEIS1	0.797	1.000	0.256	1.000	0.245	–
NR2C2	1.000	1.000	0.203	1.000	0.054	–
NR3C1	0.921	0.999	0.539	1.000	0.724	–
POU3F1	0.324	0.989	–	0.956	0.077	–

Table S2: Raw PRAUC scores for distinguishing strong and weak binders across 22 proteins from 12 protein families using six different prediction methods: IDEA, IDEA augmented with binding sequences (IDEA-seq), ProBound,[31] DeepBind,[24] DBD-Hunter[46] and rCLAMPS[36]

ID	IDEA	IDEA-seq	ProBound	DeepBind	DBD-Hunter	rCLAMPS
ELK1	0.651	0.604	0.997	1.000	0.509	–
FOXP3	0.738	1.000	0.524	0.795	0.335	–
MAX	0.772	1.000	0.556	0.994	0.449	–
TCF4	0.617	1.000	0.796	0.807	0.317	–
USF1	0.474	1.000	1.000	0.977	0.532	–
Cebpb	0.976	1.000	0.868	1.000	0.452	–
CEBPB	0.691	1.000	0.395	1.000	0.377	–
Mafb	0.539	0.985	0.844	0.998	0.694	–
Egr1	0.275	0.746	0.318	0.980	0.373	0.345
YY1	1.000	1.000	0.435	1.000	0.592	0.512
ZBTB7A	0.830	1.000	0.709	0.573	0.449	–
GATA3	0.636	0.925	0.895	0.899	0.436	–
ALX4	0.542	1.000	0.911	1.000	0.322	0.334
BARHL2	0.737	0.934	0.434	0.986	0.598	0.466
MSX1	0.815	0.998	0.584	0.878	0.401	0.326
PDX1	0.488	0.765	0.540	0.996	0.455	0.508
HSF1	0.992	1.000	1.000	1.000	0.580	–
SMAD3	0.903	0.915	1.000	1.000	0.400	–
MEIS1	0.631	1.000	0.369	1.000	0.352	–
NR2C2	1.000	1.000	0.339	1.000	0.310	–
NR3C1	0.791	0.999	0.513	1.000	0.586	–
POU3F1	0.338	0.945	–	0.960	0.282	–

Table S3: This table reports the number of atoms within the cutoff distances from the Cα atoms for all the protein-DNA structures used in this study. For each structure, we selected the DNA atom type (C5 or P) with the largest sum of occurrences within 8 Å, 9 Å, and 10 Å distances from surrounding Cα atoms for modeling. For protein YY1 (PDB ID: 1ubd), where the counts for C5 and P were identical, both models trained on C5 and P achieved an AUC score and PRAUC score of 1.0 for distinguishing strong from weak binders of the HT-SELEX dataset (Figure S4). The R^2 values for 10-fold cross-validation on the HT-SELEX dataset were 0.68 for the C5-trained model and 0.647 for the P-trained model. Figures S4, S5, and S7 use the result from the C5-trained model for YY1.

	8 Å	9 Å	10 Å	Sum
1a0a (C5)	15	22	22	59
1a0a (P)	18	19	20	57
1aay (C5)	11	15	16	42
1aay (P)	14	15	15	44
1an4 (C5)	6	11	16	33
1an4 (P)	16	18	24	58
1apl (C5)	7	11	14	32
1apl (P)	10	14	15	39
1dux (C5)	8	9	11	28
1dux (P)	6	7	12	25
1gu4 (C5)	14	18	20	52
1gu4 (P)	13	16	19	48
1hlo (C5)	11	15	19	45
1hlo (P)	13	14	15	42
1ig7 (C5)	6	8	11	25
1ig7 (P)	9	12	13	34
1j47 (C5)	9	12	16	37
1j47 (P)	16	18	18	52
1nk3 (C5)	5	9	11	25
1nk3 (P)	8	9	10	27
1nkp (C5)	11	16	20	47
1nkp (P)	16	20	20	56
1nlw (C5)	11	16	20	47
1nlw (P)	14	19	20	53

	8 Å	9 Å	10 Å	Sum
1ozj (C5)	9	13	21	43
1ozj (P)	13	18	20	51
1qrv (C5)	5	9	13	27
1qrv (P)	9	9	11	29
1ubd (C5)	14	21	24	59
1ubd (P)	17	19	23	59
2ezd (C5)	6	9	11	26
2ezd (P)	10	11	11	32
2ezf (C5)	7	9	10	26
2ezf (P)	6	7	10	23
2h1k (C5)	6	8	10	24
2h1k (P)	8	11	12	31
2lef (C5)	12	18	21	51
2lef (P)	20	20	21	61
2wty (C5)	13	19	23	55
2wty (P)	17	20	20	57
2xsd (C5)	11	14	14	39
2xsd (P)	10	11	12	33
3hdd (C5)	5	7	10	22
3hdd (P)	7	8	11	26
4hc7 (C5)	5	10	12	27
4hc7 (P)	8	11	13	32
4xrs (C5)	5	11	17	33
4xrs (P)	11	14	15	40
4y60 (C5)	7	11	14	32
4y60 (P)	15	15	17	47

	8 Å	9 Å	10 Å	Sum
5cbx (C5)	8	15	17	40
5cbx (P)	14	14	15	43
6od3 (C5)	5	8	10	23
6od3 (P)	8	11	11	30
7tdw (C5)	6	9	10	25
7tdw (P)	8	10	11	29
7xv6 (C5)	9	16	20	45
7xv6 (P)	14	16	18	48
8e3d (C5)	13	20	23	56
8e3d (P)	20	22	24	66
8k8d (C5)	11	16	18	45
8k8d (P)	9	14	16	39
8osb (C5)	13	22	27	62
8osb (P)	26	30	34	90
8pm7 (C5)	4	8	10	22
8pm7 (P)	9	13	13	35
8ssq (C5)	28	41	46	115
8ssq (P)	34	41	46	121
8sss (C5)	29	39	42	110
8sss (P)	28	35	35	98
9ant (C5)	5	7	11	23
9ant (P)	8	10	12	30

Table S4: Summary of parameters used in our protein-DNA simulation model (unit: kcal/mol).

	PP	SU	A	G	C	T
ALA	0	0	-0.0228	0.0101	0.0199	-0.0061
ARG	0	0	-0.0402	0.0340	-0.0363	-0.0210
ASN	0	0	-0.0163	0.0252	0.0159	-0.0030
ASP	0	0	-0.0001	0.0120	0.0048	0.0101
CYS	0	0	0.0025	0.0107	0.0062	0.0075
GLU	0	0	0.0045	-0.0210	0.0189	-0.0059
GLN	0	0	-0.0138	-0.0060	0.0094	-0.0032
GLY	0	0	0.0027	0.0320	0.0183	-0.0486
HIS	0	0	-0.0030	0.0086	0.0149	-0.0022
ILE	0	0	-0.0188	0.0052	0.0140	0.0099
LEU	0	0	-0.0148	-0.0216	0.0127	0.0121
LYS	0	0	-0.0107	0.0083	-0.0487	0.0115
MET	0	0	-0.0049	0.0085	0.0020	0.0143
PHE	0	0	-0.0015	-0.0117	0.0151	-0.0042
PRO	0	0	0.0009	0.0240	-0.0073	-0.0087
SER	0	0	0.0120	-0.0100	0.0011	-0.0079
THR	0	0	-0.0213	-0.0132	0.0163	0.0066
TRP	0	0	-0.0181	0.0058	0.0172	0.0030
TYR	0	0	-0.0094	0.0016	-0.0117	0.0127
VAL	0	0	-0.0060	0.0102	0.0013	0.0060

Acknowledgement

This work was supported by the startup funding from North Carolina State University. Additionally, this research received partial funding from the NC State Genetics and Genomics Academy and Comparative Medicine Institute. We appreciate Dr. Keith Weninger, Dr. Faruck Morcos, and Dr. Qin Zhou for their insightful discussions and Dr. Sebastian Maerkl for guiding us to the MITOMI experimental data.

Data availability: The implementation of the IDEA model, along with training and prediction examples, is available on our GitHub repository and in the Supplementary data.

References

(1) Lee, T. I.; Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **2013**, *152*, 1237–1251.

(2) Orphanides, G.; Reinberg, D. A Unified Theory of Gene Expression. *Cell* **2002**, *108*, 439–451.

(3) Ren, B.; Robert, F.; Wyrick, J. J.; Aparicio, O.; Jennings, E. G.; Simon, I.; Zeitlinger, J.; Schreiber, J.; Hannett, N.; Kanin, E.; Volkert, T. L.; Wilson, C. J.; Bell, S. P.; Young, R. A. Genome-Wide Location and Function of DNA Binding Proteins. *Science* **2000**, *290*, 2306–2309.

(4) Jaenisch, R.; Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **2003**, *33*, 245–254.

(5) Latchman, D. S. Transcription factors: an overview. *The international journal of biochemistry & cell biology* **1997**, *29*, 1305–1312.

(6) Owen, J. A.; Osmanović, D.; Mirny, L. Design principles of 3D epigenetic memory systems. *Science* **2023**, *382*, eadg3053.

(7) Stormo, G. D.; Zhao, Y. Determining the specificity of protein–DNA interactions. *Nat Rev Genet* **2010**, *11*, 751–760.

(8) Solomon, M. J.; Larsen, P. L.; Varshavsky, A. Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell* **1988**, *53*, 937–947.

(9) Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews genetics* **2009**, *10*, 669–680.

(10) Bartlett, A.; O'Malley, R. C.; Huang, S.-s. C.; Galli, M.; Nery, J. R.; Gallavotti, A.; Ecker, J. R. Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat Protoc* **2017**, *12*, 1659–1672.

(11) Berger, M. F.; Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols* **2009**, *4*, 393–411.

(12) Meng, X.; Brodsky, M. H.; Wolfe, S. A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature biotechnology* **2005**, *23*, 988–994.

(13) Maerkl, S. J.; Quake, S. R. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science* **2007**, *315*, 233–237.

(14) Fordyce, P. M.; Gerber, D.; Tran, D.; Zheng, J.; Li, H.; DeRisi, J. L.; Quake, S. R. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature biotechnology* **2010**, *28*, 970–975.

(15) Ogawa, N.; Biggin, M. D. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. *Gene Regulatory Networks: Methods and Protocols* **2012**, 51–63.

(16) Isakova, A.; Groux, R.; Imbeault, M.; Rainer, P.; Alpern, D.; Dainese, R.; Ambrosini, G.; Trono, D.; Bucher, P.; Deplancke, B. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nature methods* **2017**, *14*, 316–322.

(17) Furey, T. S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* **2012**, *13*, 840–852.

(18) Bulyk, M. L.; Huang, X.; Choo, Y.; Church, G. M. Exploring the DNA-binding speci-

ficities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 7158–7163.

(19) Gordân, R.; Shen, N.; Dror, I.; Zhou, T.; Horton, J.; Rohs, R.; Bulyk, M. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports* **2013**, *3*, 1093–1104.

(20) Afek, A.; Shi, H.; Rangadurai, A.; Sahay, H.; Senitzki, A.; Xhani, S.; Fang, M.; Salinas, R.; Mielko, Z.; Pufall, M. A.; Poon, G. M. K.; Haran, T. E.; Schumacher, M. A.; Al-Hashimi, H. M.; Gordân, R. DNA mismatches reveal conformational penalties in protein–DNA recognition. *Nature* **2020**, *587*, 291–296.

(21) Jolma, A.; Kivioja, T.; Toivonen, J.; Cheng, L.; Wei, G.; Enge, M.; Taipale, M.; Vaquerizas, J. M.; Yan, J.; Sillanpää, M. J.; others Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research* **2010**, *20*, 861–873.

(22) Kohlberger, M.; Gadermaier, G. SELEX: Critical factors and optimization strategies for successful aptamer selection. *Biotechnology and applied biochemistry* **2022**, *69*, 1771–1792.

(23) Weirauch, M. T.; Cote, A.; Norel, R.; Annala, M.; Zhao, Y.; Riley, T. R.; Saez-Rodriguez, J.; Cokelaer, T.; Vedenko, A.; Talukder, S.; others Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology* **2013**, *31*, 126–134.

(24) Alipanahi, B.; Delong, A.; Weirauch, M. T.; Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **2015**, *33*, 831–838.

(25) Zhou, T.; Shen, N.; Yang, L.; Abe, N.; Horton, J.; Mann, R. S.; Bussemaker, H. J.; Gordân, R.; Rohs, R. Quantitative modeling of transcription factor binding specifici-

ties using DNA shape. *Proceedings of the National Academy of Sciences* **2015**, *112*, 4654–4659.

(26) Yang, L.; Orenstein, Y.; Jolma, A.; Yin, Y.; Taipale, J.; Shamir, R.; Rohs, R. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol* **2017**, *13*, 910.

(27) Rastogi, C.; Rube, H. T.; Kribelbauer, J. F.; Crocker, J.; Loker, R. E.; Martini, G. D.; Laptenko, O.; Freed-Pastor, W. A.; Prives, C.; Stern, D. L.; Mann, R. S.; Bussemaker, H. J. Accurate and sensitive quantification of protein-DNA binding affinity. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*.

(28) Roche, R.; Moussad, B.; Shuvo, M. H.; Tarafder, S.; Bhattacharya, D. EquiPNAS: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *Nucleic Acids Research* **2024**, *52*, e27–e27.

(29) Liu, Y.; Tian, B. Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning. *Briefings in Bioinformatics* **2023**, *25*, bbad488.

(30) Nguyen, B. P.; Nguyen, Q. H.; Doan-Ngoc, G.-N.; Nguyen-Vo, T.-H.; Rahardja, S. iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks. *BMC Bioinformatics* **2019**, *20*, 634.

(31) Rube, H. T.; Rastogi, C.; Feng, S.; Kribelbauer, J. F.; Li, A.; Becerra, B.; Melo, L. A. N.; Do, B. V.; Li, X.; Adam, H. H.; Shah, N. H.; Mann, R. S.; Bussemaker, H. J. Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nat Biotechnol* **2022**, *40*, 1520–1527.

(32) Siggers, T.; Gordân, R. Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Research* **2014**, *42*, 2099–2111.

(33) Mitra, R.; Cohen, A. S.; Sagendorf, J. M.; Berman, H. M.; Rohs, R. DNAproDB: an updated database for the automated and interactive analysis of protein-DNA complexes. *Nucleic Acids Res* **2025**, *53*, D396–D402.

(34) Harini, K.; Srivastava, A.; Kulandaisamy, A.; Gromiha, M. M. ProNAB: database for binding affinities of protein-nucleic acid complexes and their mutants. *Nucleic Acids Res* **2022**, *50*, D1528–D1534.

(35) Mitra, R.; Li, J.; Sagendorf, J. M.; Jiang, Y.; Cohen, A. S.; Chiu, T.-P.; Glasscock, C. J.; Rohs, R. Geometric deep learning of protein–DNA binding specificity. *Nat Methods* **2024**, *21*, 1674–1683.

(36) Wetzel, J. L.; Zhang, K.; Singh, M. Learning probabilistic protein-DNA recognition codes from DNA-binding specificities using structural mappings. *Genome Res* **2022**, *32*, 1776–1786.

(37) Chiu, T.-P.; Rao, S.; Rohs, R. Physicochemical models of protein–DNA binding with standard and modified base pairs. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, e2205796120.

(38) Bryngelson, J. D.; Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524–7528.

(39) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* **1997**, *48*, 545–600.

(40) Schafer, N. P.; Kim, B. L.; Zheng, W.; Wolynes, P. G. Learning To Fold Proteins Using Energy Landscape Theory. *Isr J Chem* **2014**, *54*, 1311–1337.

(41) Chu, W.-T.; Yan, Z.; Chu, X.; Zheng, X.; Liu, Z.; Xu, L.; Zhang, K.; Wang, J. Physics of biomolecular recognition and conformational dynamics. *Rep. Prog. Phys.* **2021**, *84*, 126601.

(42) Burley, S. K. et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research* **2022**, *51*, D488–D508.

(43) Geertz, M.; Shore, D.; Maerkl, S. J. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proceedings of the National Academy of Sciences* **2012**, *109*, 16540–16545.

(44) Zhou, Q.; Kunder, N.; De La Paz, J. A.; Lasley, A. E.; Bhat, V. D.; Morcos, F.; Campbell, Z. T. Global pairwise RNA interaction landscapes reveal core features of protein recognition. *Nat Commun* **2018**, *9*, 2511.

(45) Blackwell, T. K.; Kretzner, L.; Blackwood, E. M.; Eisenman, R. N.; Weintraub, H. Sequence-specific DNA binding by the c-Myc protein. *Science* **1990**, *250*, 1149–1151.

(46) Gao, M.; Skolnick, J. DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res* **2008**, *36*, 3978–3992.

(47) Söding, J.; Biegert, A.; Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **2005**, *33*, W244–W248.

(48) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(49) de Souza, N. The ENCODE project. *Nature methods* **2012**, *9*, 1046–1046.

(50) Giresi, P. G.; Kim, J.; McDaniell, R. M.; Iyer, V. R.; Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **2007**, *17*, 877–885.

(51) Zhang, J. et al. An integrative ENCODE resource for cancer genomics. *Nat Commun* **2020**, *11*, 3696.

(52) Boyle, A. P.; Davis, S.; Shulha, H. P.; Meltzer, P.; Margulies, E. H.; Weng, Z.; Furey, T. S.; Crawford, G. E. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **2008**, *132*, 311–322.

(53) Corces, M. R.; Trevino, A. E.; Hamilton, E. G.; Greenside, P. G.; Sinnott-Armstrong, N. A.; Vesuna, S.; Satpathy, A. T.; Rubin, A. J.; Montine, K. S.; Wu, B.; others An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature methods* **2017**, *14*, 959–962.

(54) Granja, J. M.; Corces, M. R.; Pierce, S. E.; Bagdatli, S. T.; Choudhry, H.; Chang, H. Y.; Greenleaf, W. J. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **2021**, *53*, 403–411, Publisher: Nature Publishing Group.

(55) Savelyev, A.; Papoian, G. A. Chemically accurate coarse graining of double-stranded DNA. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 20340–20345.

(56) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* **2012**, *116*, 8494–8503.

(57) D. Bascom, G.; Schlick, T. *Nuclear Architecture and Dynamics*; Elsevier, 2018; pp 123–147.

(58) Tan, C.; Takada, S. Nucleosome allostery in pioneer transcription factor binding. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 20586–20596.

(59) Chakraborty, D.; Mondal, B.; Thirumalai, D. Brewing COFFEE: A Sequence-Specific Coarse-Grained Energy Function for Simulations of DNA-Protein Complexes. *J. Chem. Theory Comput.* **2024**, *20*, 1398–1413.

(60) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *Journal of Molecular Biology* **2000**, *298*, 937–953.

(61) Freeman, G. S.; Hinckley, D. M.; Lequieu, J. P.; Whitmer, J. K.; De Pablo, J. J. Coarse-grained modeling of DNA curvature. *The Journal of Chemical Physics* **2014**, *141*, 165103.

(62) Lin, X.; Qi, Y.; Latham, A. P.; Zhang, B. Multiscale modeling of genome organization with maximum entropy optimization. *J. Chem. Phys.* **2021**, *155*, 010901.

(63) Zhang, B.; Zheng, W.; Papoian, G. A.; Wolynes, P. G. Exploring the Free Energy Landscape of Nucleosomes. *J. Am. Chem. Soc.* **2016**, *138*, 8126–8133.

(64) Lequieu, J.; Córdoba, A.; Schwartz, D. C.; De Pablo, J. J. Tension-Dependent Free Energies of Nucleosome Unwrapping. *ACS Cent. Sci.* **2016**, *2*, 660–666.

(65) Parsons, T.; Zhang, B. Critical role of histone tail entropy in nucleosome unwinding. *The Journal of Chemical Physics* **2019**, *150*, 185103.

(66) Moller, J.; Lequieu, J.; De Pablo, J. J. The Free Energy Landscape of Internucleosome Interactions and Its Relation to Chromatin Fiber Structure. *ACS Cent. Sci.* **2019**, *5*, 341–348.

(67) Lin, X.; Zhang, B. Explicit ion modeling predicts physicochemical interactions for chromatin organization. *eLife* **2024**, *12*, RP90073.

(68) Ding, X.; Lin, X.; Zhang, B. Stability and folding pathways of tetra-nucleosome from six-dimensional free energy surface. *Nat Commun* **2021**, *12*, 1091.

(69) Liu, S.; Lin, X.; Zhang, B. Chromatin fiber breaks into clutches under tension and crowding. *Nucleic Acids Res* **2022**, *50*, 9738–9747.

(70) Watanabe, S.; Mishima, Y.; Shimizu, M.; Suetake, I.; Takada, S. Interactions of HP1 Bound to H3K9me3 Dinucleosome by Molecular Simulations and Biochemical Assays. *Biophysical Journal* **2018**, *114*, 2336–2351.

(71) Leicher, R.; Ge, E. J.; Lin, X.; Reynolds, M. J.; Xie, W.; Walz, T.; Zhang, B.; Muir, T. W.; Liu, S. Single-molecule and in silico dissection of the interaction between Polycomb repressive complex 2 and chromatin. *Proc Natl Acad Sci USA* **2020**, *117*, 30465–30475.

(72) Lin, X.; Leicher, R.; Liu, S.; Zhang, B. Cooperative DNA looping by PRC2 complexes. *Nucleic Acids Research* **2021**, *49*, 6238–6248.

(73) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput Biol* **2016**, *12*, e1004794.

(74) Knotts, T. A.; Rathore, N.; Schwartz, D. C.; De Pablo, J. J. A coarse grain model for DNA. *The Journal of Chemical Physics* **2007**, *126*, 084901.

(75) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; De Pablo, J. J. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *The Journal of Chemical Physics* **2013**, *139*, 144903.

(76) Freeman, G. S.; Lequieu, J. P.; Hinckley, D. M.; Whitmer, J. K.; De Pablo, J. J. DNA

Shape Dominates Sequence Affinity in Nucleosome Formation. *Phys. Rev. Lett.* **2014**, *113*, 168101.

(77) Privalov, P. L.; Dragan, A. I.; Crane-Robinson, C. Interpreting protein/DNA interactions: distinguishing specific from non-specific and electrostatic from non-electrostatic components. *Nucleic Acids Research* **2011**, *39*, 2483–2491.

(78) Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.

(79) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J Comput Chem* **1992**, *13*, 1011–1021.

(80) Mukherjee, S.; Berger, M. F.; Jona, G.; Wang, X. S.; Muzzey, D.; Snyder, M.; Young, R. A.; Bulyk, M. L. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* **2004**, *36*, 1331–1339.

(81) Aizenshtein-Gazit, S.; Orenstein, Y. DeepZF: improved DNA-binding prediction of C2H2-zinc-finger proteins by deep transfer learning. *Bioinformatics* **2022**, *38*, ii62–ii67.

(82) Liu, S.; Gomez-Alcala, P.; Leemans, C.; Glassford, W. J.; Melo, L. A. N.; Lu, X.-J.; Mann, R. S.; Bussemaker, H. J. Predicting the DNA binding specificity of transcription factor mutants using family-level biophysically interpretable machine learning. *bioRxiv* **2025**, 2024.01.24.577115.

(83) Fornes, O.; Meseguer, A.; Aguirre-Plans, J.; Gohl, P.; Bota, P. M.; Molina-Fernández, R.; Bonet, J.; Chinchilla-Hernandez, A.; Pegenaute, F.; Gallego, O.; Fernandez-Fuentes, N.; Oliva, B. Structure-based learning to predict and model protein–DNA interactions and transcription-factor co-operativity in *cis* -regulatory elements. *NAR Genomics and Bioinformatics* **2024**, *6*, lqae068.

(84) Hindorff, L. A.; Sethupathy, P.; Junkins, H. A.; Ramos, E. M.; Mehta, J. P.; Collins, F. S.; Manolio, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **2009**, *106*, 9362–9367.

(85) Lappalainen, T.; Scott, A. J.; Brandt, M.; Hall, I. M. Genomic analysis in the age of human genome sequencing. *Cell* **2019**, *177*, 70–84.

(86) Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **2010**, *11*, 31–46.

(87) Doerr, A. Cryo-electron tomography. *Nat Methods* **2017**, *14*, 34–34.

(88) Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, 1–3.

(89) Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**,

(90) Baek, M.; McHugh, R.; Anishchenko, I.; Jiang, H.; Baker, D.; DiMaio, F. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat Methods* **2024**, *21*, 117–121.

(91) Moore, L. D.; Le, T.; Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **2013**, *38*, 23–38.

(92) Strahl, B. D.; Allis, C. D. The language of covalent histone modifications. *Nature* **2000**, *403*, 41–45.

(93) Jirtle, R. L.; Skinner, M. K. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* **2007**, *8*, 253–262.

(94) Portela, A.; Esteller, M. Epigenetic modifications and human disease. *Nat Biotechnol* **2010**, *28*, 1057–1068.

(95) Allis, C. D.; Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **2016**, *17*, 487–500.

(96) de Goede, O. M. et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **2021**, *184*, 2633–2648.e19.

(97) Li, J.; Chiu, T.-P.; Rohs, R. Predicting DNA structure using a deep learning method. *Nat Commun* **2024**, *15*, 1243.

(98) Lin, X.; George, J. T.; Schafer, N. P.; Ng Chau, K.; Birnbaum, M. E.; Clementi, C.; Onuchic, J. N.; Levine, H. Rapid assessment of T-cell receptor specificity of the immune repertoire. *Nat Comput Sci* **2021**, *1*, 362–373.

(99) Wang, A.; Lin, X.; Chau, K. N.; Onuchic, J. N.; Levine, H.; George, J. T. RACER-m leverages structural features for sparse T cell specificity prediction. *Sci. Adv.* **2024**, *10*, eadl0161.

(100) European Nucleotide Archive European Nucleotide Archive. 2024; https://www.ebi.ac.uk/ena.

(101) Rastogi, C.; Liu, D.; Melo, L.; Bussemaker, H. J. *SELEX: Functions for analyzing SELEX-seq data*; 2022.

(102) Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V. P.; Ashford, P.; Scholes, H. M.; Pang, C. S.; Woodridge, L.; Rauer, C.; Sen, N.; others CATH: increased structural coverage of functional space. *Nucleic acids research* **2021**, *49*, D266–D273.

(103) Gabler, F.; Nam, S.-Z.; Till, S.; Mirdita, M.; Steinegger, M.; Söding, J.; Lupas, A. N.; Alva, V. Protein sequence analysis using the MPI bioinformatics toolkit. *Current Protocols in Bioinformatics* **2020**, *72*, e108.

(104) Yang, J.; Horton, J. R.; Liu, B.; Corces, V. G.; Blumenthal, R. M.; Zhang, X.;

Cheng, X. Structures of CTCF–DNA complexes including all 11 zinc fingers. *Nucleic Acids Research* **2023**, *51*, 8447–8462.

(105) Lequieu, J.; Schwartz, D. C.; De Pablo, J. J. In silico evidence for sequence-dependent nucleosome sliding. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*.

(106) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; In 'T Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **2022**, *271*, 108171.

(107) Dragan, A. I.; Klass, J.; Read, C.; Churchill, M. E.; Crane-Robinson, C.; Privalov, P. L. DNA Binding of a Non-sequence-specific HMG-D Protein is Entropy Driven with a Substantial Non-electrostatic Contribution. *Journal of Molecular Biology* **2003**, *331*, 795–813.

(108) Dragan, A. I.; Read, C. M.; Makeyeva, E. N.; Milgotina, E. I.; Churchill, M. E.; Crane-Robinson, C.; Privalov, P. L. DNA Binding and Bending by HMG Boxes: Energetic Determinants of Specificity. *Journal of Molecular Biology* **2004**, *343*, 371–393.

(109) Dragan, A. I.; Liggins, J. R.; Crane-Robinson, C.; Privalov, P. L. The Energetics of Specific Binding of AT-hooks from HMGA1 to Target DNA. *Journal of Molecular Biology* **2003**, *327*, 393–411.

(110) Dragan, A. I.; Li, Z.; Makeyeva, E. N.; Milgotina, E. I.; Liu, Y.; Crane-Robinson, C.; Privalov, P. L. Forces Driving the Binding of Homeodomains to DNA. *Biochemistry* **2006**, *45*, 141–151.

(111) Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **2018**, *46*, D260–D266.

(112) Kulakovskiy, I. V.; Vorontsov, I. E.; Yevshin, I. S.; Sharipov, R. N.; Fedorova, A. D.; Rumynskiy, E. I.; Medvedeva, Y. A.; Magana-Mora, A.; Bajic, V. B.; Papatsenko, D. A.; Kolpakov, F. A.; Makeev, V. J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **2018**, *46*, D252–D259.

(113) Jolma, A. et al. DNA-Binding Specificities of Human Transcription Factors. *Cell* **2013**, *152*, 327–339.

(114) Asif, M.; Orenstein, Y. DeepSELEX: inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. *Bioinformatics* **2020**, *36*, i634–i642.

(115) Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432.