Interpretable Protein-DNA Interactions Captured by Structure-based Optimization

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

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

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

E-mail: Xingcheng_Lin@ncsu.edu

Abstract

Sequence-specific DNA recognition underlies essential processes in gene regulation, yet predictive methods for simultaneous prediction of genome-wide DNA recognition sites and their binding affinity remain lacking. Here, we present IDEA, an interpretable residue-level biophysical model capable of predicting binding sites and strengths of DNA-binding proteins across the genome. By leveraging the sequence-structure relationship from known protein-DNA complexes, IDEA learns an energy model enabling direct interpretation of physicochemical interactions among individual amino acids and nucleotides. Using transcription factors as examples, we demonstrate that this energy model accurately predicts genomic DNA recognition sites and their binding strengths. Additionally, the IDEA model is integrated into a coarse-grained simulation framework that accurately captures the absolute protein-DNA binding free energies. Overall, IDEA provides an integrated computational platform alleviating 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 expressions are 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 performed.[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 protein-DNA binding specificity. Numerous methods have emerged to enable predictions of binding sites and affinities of DNA-binding proteins.[23–29] These methods often utilized machine-learning-based training to extract sequence preference information from DNA or protein by utilizing experimental high-throughput (HT) assays,[23–26] which relies on the availability and quality of experimental binding assays. Additionally, many approaches employ deep neural networks,[27–29] 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 transcription factors.[30]

Nowadays, over 5000 protein-DNA 3D structures, including TF-DNA complexes, have been published.[31,32] 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 of protein-DNA interactions. The very robustness of evolution[33–36] offers an approach to extract the sequence-structure relation embedded in these available complexes, which learns an interpretable binding energy landscape that governs the recognition processes of those DNA-binding proteins.

Here, we introduce the Interpretable protein-DNA Energy Associative (IDEA) model, a predictive model that learns protein-DNA physicochemical interactions from available biophysical structures and their associated sequences (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 DNA-binding sites on a genome-wide scale. Notably, IDEA learns the interaction matrix between each amino acid and nucleotide, allowing for direct interpretation of the "molecular grammar" governing the binding specificities of DNA binding proteins. This interpretable energy model is further integrated into a simulation framework, facilitating mechanistic studies of various biomolecular functions involving protein-DNA interactions.

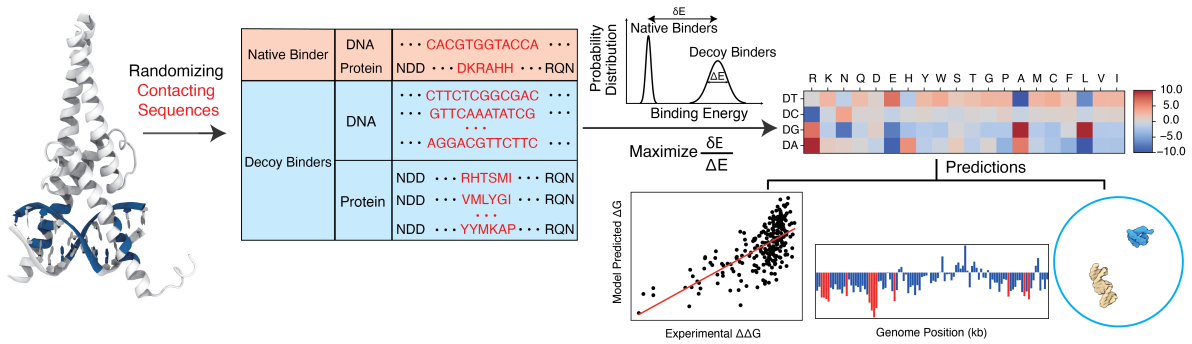

Figure 1: **Overview of the IDEA protocol.** The protein-DNA complex, represented by the human MAX-DNA complex structure (PDB ID: 1HLO), was used for training. The sequence of the protein and DNA that form close contacts (highlighted in blue) were included in the training dataset. In addition, a series of synthetic decoy sequences were generated by randomizing the contacting protein and DNA sequences. The amino-acid-nucleotide energy model was optimized by maximizing the binding energy gap between the native binder and the decoy binders. The optimized energy model can be used for multiple predictive applications, including evaluating the binding energy of protein-DNA pairs with different sequences, genome-wide prediction of DNA binding sites by transcription factors or other DNA-binding proteins, and constructing a sequence-specific residue-resolution simulation model for dynamic simulations.

Results

Residue-Level Protein-DNA Energy Model for Predicting Protein-DNA Recognition Specificities

IDEA is a coarse-grained biophysical model at the residue resolution for investigating the binding interactions between protein and DNA (Fig. 1). It leverages 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 interface. The model was trained using available protein-DNA complexes curated from existing database.[31,37] Unlike existing deep-learning-based protein-DNA binding prediction models, IDEA aims to learn a physicochemical-based energy model that quantitatively characterizes sequences-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 given protein-DNA pairs

based on their structures and sequences. Additionally, it enables genome-wide prediction of DNA binding sites by a given protein, such as a transcription factor. Finally, the learned energy model is integrated into a simulation framework that can be used to investigate the molecular mechanisms underlying various DNA-templated activities, such as initiation of gene transcription,[38] epigenetic regulation of gene expression,[4] and DNA replication.[39] Further details of the optimization protocol are provided in Methods Section *Energy Model Optimization*.

IDEA Accurately Predicts Transcription Factor Binding Specificity

We first examine the predictive accuracy of IDEA by comparing its predicted transcription factor binding affinity with experimental measurements.[13,14] We focused on the MAX transcription factor (TF), a basic Helix-loop-helix (bHLH) TF with the most comprehensive 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.[13,40] Among them, the 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, a consensus MAX DNA binding sequence, and can thus serve 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 0.67, Fig. 2A). Including additional human-associated MAX-DNA complex structures and their associated sequences slightly improves the correlation (Pearson Correlation 0.68, Fig. S1), albeit not significantly, likely due to the structural similarity of 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.[26,41] Inspired by these approaches, we developed an enhanced predictive protocol that integrates additional experimental binding data if available (see the Method section: Enhanced Modeling Prediction with SELEX Data). Encouragingly, when

training the model with the SELEX-seq data[26] of MAX TF, IDEA gained a significant predictive improvement in reproducing the MITOMI measurement (Pearson Correlation 0.79, Fig. 2B).

To examine IDEA's predictive accuracy across other DNA-binding proteins, we applied it to calculate the binding affinity of additional TFs with available MITOMI measurements.[40] IDEA's predicted binding affinity of Pho4 protein, another bHLH transcription factor, correlates well with the experimental measurement (Pearson Correlation 0.6, Fig. S2A) by training on the only available protein-DNA structure (PDB ID: 1A0A). We further evaluated IDEA's predictive performance on Zinc-finger Zif268 protein, which has a different structure than the bHLH TFs and thus requires a different atom to represent the DNA nucleotides (Table S1). Due to a lack of experimental data for all the possible DNA sequence combinations, we focused on testing the model's predictions on point-mutated DNA sequences, which was known to be a challenging task due to their minor deviation from the native sequence. Despite the challenge, IDEA achieves accurate predictions, achieving a Pearson Correlation of 0.57 (Fig. S2B). We further expanded the training dataset to include all the 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 0.63, Fig. S3).

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 (Fig. 2C). Notably, DNA Cytosine (DC) exhibited strong

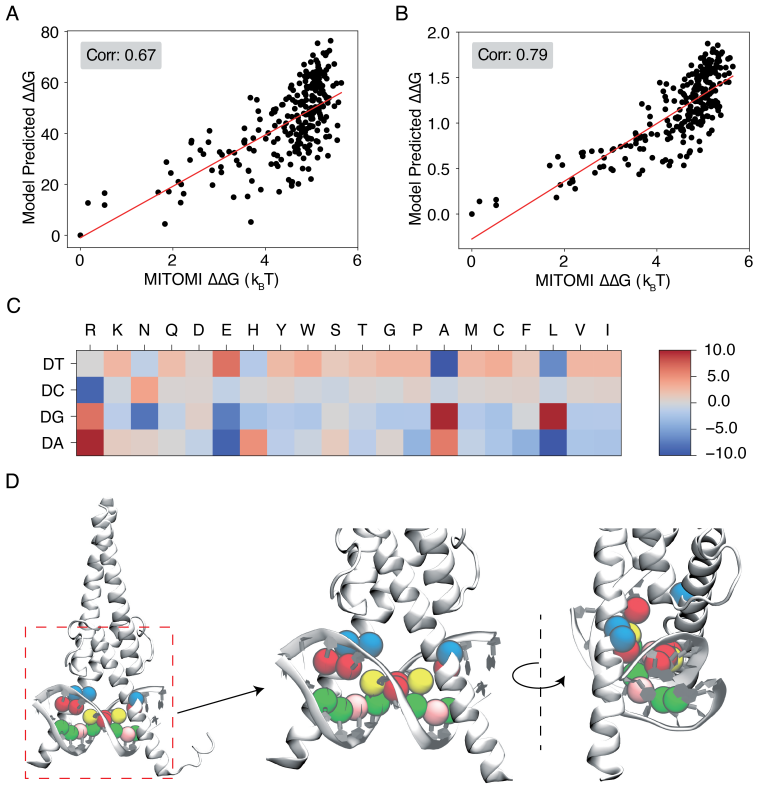

Figure 2: **Results for MAX-based predictions.** (A) The binding energy calculated by IDEA, learned using one MAX-DNA complex (PDB ID: 1HLO), correlates well with experimentally measured MAX-DNA binding energy.[13] $\Delta\Delta G$ represents the changes in binding energy relative to the protein binding to the wild-type DNA. (B) Including SELEX-measured MAX-binding data in the training set further improves the modeling prediction. (C) The learned energy model reveals an amino-acid-nucleotide interaction pattern governing MAX-DNA recognition. (D) The 3D structure of the MAX-DNA complex (zoomed in with different views) illustrates important amino acid-nucleotide contacts located on the protein-DNA interface, where multiple DNA Cytosines (red spheres) form close contacts with Arginine (blue spheres). The predicted binding energies and optimized energy model are presented in reduced units, as explained in Methods.

interactions with Arginine (R), consistent with the fact that the E-box region (CACGTG) frequently attracts the positively charged residues of bHLH TFs.[42] Importantly, Arginine was in close contact with Cytosine in the crystal structure (Fig. 2D), 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 glutamic acid (E) and all DNA nucleotides, consistent with their negative charges (Fig. S4). 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 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 binding specificities 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 MIT-OMI 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 Correlation coefficients larger than 0.5 (Fig. 3A).

The transferability of IDEA within the same CATH superfamily can be understood from the similar structural interfaces among different DNA-binding proteins, which determines a similar learned energy model. For example, the Pho4 protein (PDB ID: 1A0A) shares a highly similar DNA-binding interface with the MAX protein (PDB ID: 1HLO) (Fig. 3B), despite having a 33.41% probability of being homologous. Consequently, the energy model derived from the Pho4-DNA complex (Fig. 3C) exhibits a similar amino-acid-nucleotide interactive pattern as that learned from the MAX-DNA complex (Fig. 2C).

Genome-wide Identification of 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 map-

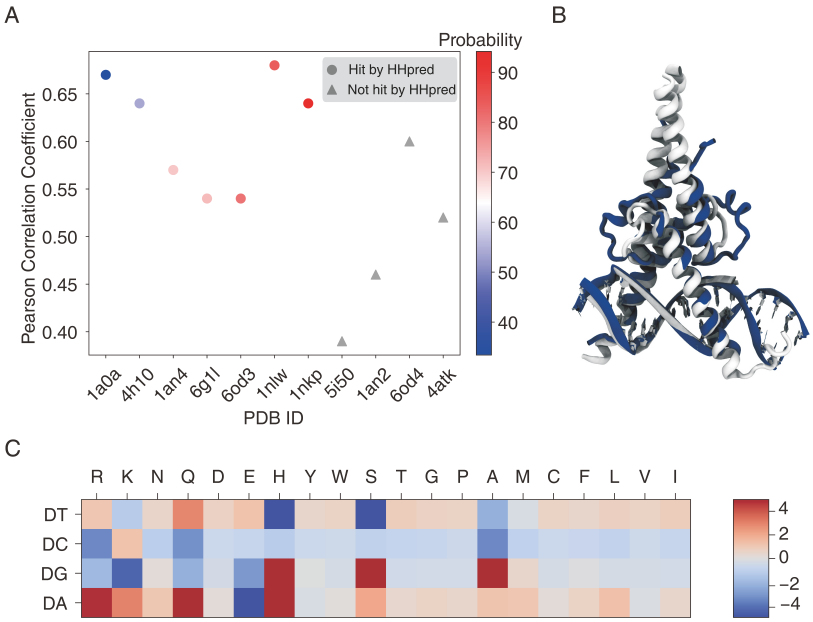

Figure 3: **IDEA prediction shows transferability across the same CATH superfamily.** (A) The predicted MAX binding specificity, when trained on other protein-DNA complexes within the same protein CATH superfamily, correlates with experimental measurement. The proteins are ordered by their probability of being homologous to the MAX protein, determined using HHpred.[43] Training with a homologous protein (hit by HHpred) usually leads to better predictive performance (Pearson Correlation coefficient > 0.5). (B) Structural alignment between 1HLO (white) and 1A0A (blue), two protein-DNA complexes in the same CATH Helix-loop-helix superfamily. The alignment was performed based on the E-box region of the DNA.[44] (C) The learned 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 Methods.

ping of genome-wide protein-DNA binding locations, such as ChIP–seq,[9,17,45] DAP-seq,[10] and FAIRE-seq,[46] 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 facilitate the genome-wide prediction of protein-DNA binding sites given a DNA sequence at a base-pair resolution. To assess the predictive accuracy of this protocol, we utilized publicly available MAX-based ChIP-seq binding data

for GM12878 lymphoblastoid cell lines.[47] As the experimental measurements were conducted at the 420 base pairs resolution, we averaged our modeling prediction at 500 base pairs. Due to computational constraints, we focused our prediction on the 1 Mb region of chromosome 1, which has the largest number of MAX-binding sites (156000 kb - 157000 kb, representative window shown in Fig. 4A top). Lower binding energy corresponds to stronger binding affinity between the protein and DNA, indicating higher probable binding sites. The predicted binding energies were further normalized by the binding energies of the randomized decoy sequences (see the Method 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 accurately identifies the majority of the MAX binding sites, achieving an AUC score of 0.81 (Fig. 4B). Therefore, IDEA provides an accurate initial prediction of the genome-wide binding sites of DNA-binding proteins based only on DNA sequences before more costly experiments are conducted. Furthermore, IDEA holds the potential to identify the binding sites beyond the experimental resolution and can aid in the interpretation of other sequencing techniques to uncover additional factors influencing genome-wide DNA binding, such as chromatin accessibility.[48,49]

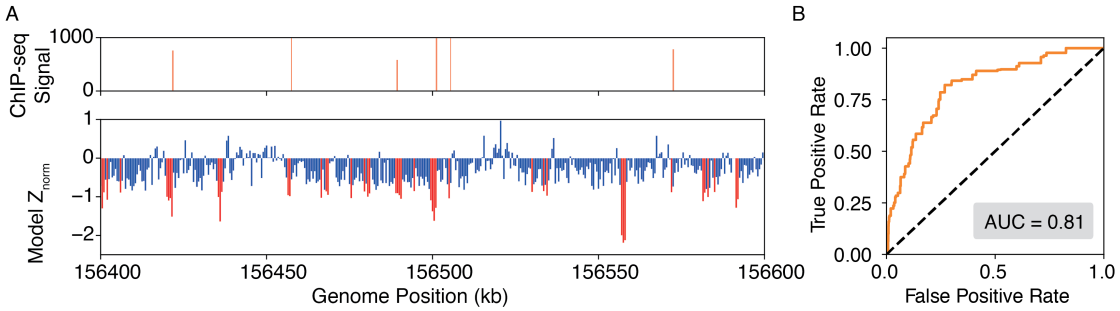

Figure 4: **IDEA accurately identifies genome-wide protein-binding sites.** (A) The IDEA-predicted MAX-transcription factor binding sites on the Lymphoblastoid cells chromosome (bottom) correlate well with ChIP-seq measurements (top), shown for a representative window of the 1Mb region of chromosome 1 where ChIP-seq signals are densest. Highly probable binding sites detected by IDEA are represented as red peaks. The predicted normalized Z scores were averaged over a 500 bp window to match the experimental resolution (420 bp). (B) The predicted binding sites of the 1Mb region of chromosome 1 were used to construct a ROC curve, reaching an AUC score of 0.81.

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.[50–54] Notably, the structure-based protein model[55] and 3-Site-Per-Nucleotide (3SPN) DNA model[56] 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,[57] including nucleosomal DNA unwrapping,[58–60] interactions between nucleosomes,[61,62] large chromatin organizations,[62–64] and its modulation by chromatin factors.[65–67] Although both the protein and DNA force fields have been systematically optimized,[56,68–71] 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 Method 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 targeted 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.[72] We applied IDEA to learn an energy model based on these 9 protein-DNA complexes, which leads to a strong correlation between the modeling predicted binding affinity and the experimental measurements (Pearson Correlation coefficient 0.72, Fig. 5A). The learned energy model was incorporated into the simulation model, which was used to simulate the binding process between the protein and

DNA target using umbrella sampling techniques.[73] The predicted binding free energy[74] (Fig. S5) from our simulation shows a strong correlation (Pearson Correlation 0.79) and a near-perfect fit with the experimental measurement (Fig. 5B), greatly improved over the previous model, which depicts protein-DNA interactions by non-sequence-specific homogeneous electrostatics interactions (Pearson Correlation 0.6, Fig.S6). 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 energy given the protein and DNA structures with physical units.

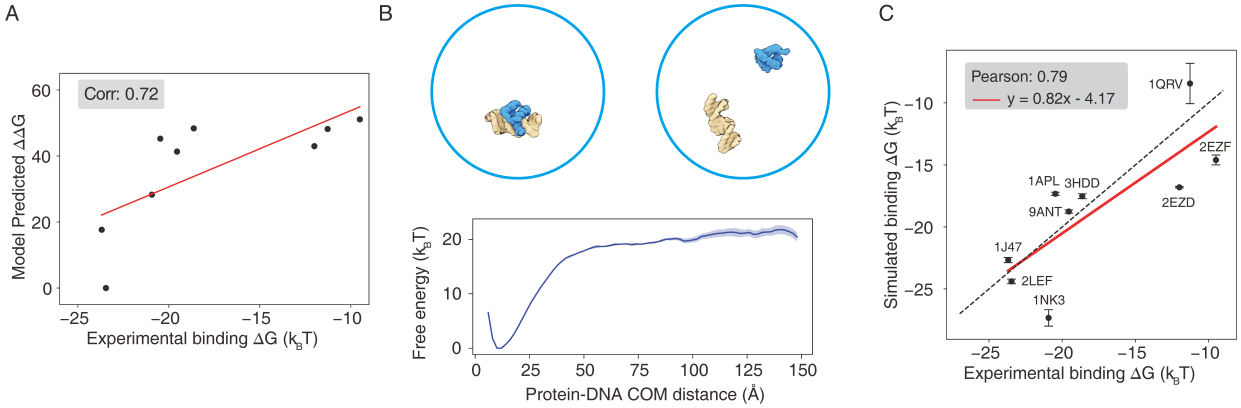

Figure 5: **Enhanced protein-DNA simulation model by incorporating IDEA-optimized energy model.** (A) The prediction of the protein-DNA binding affinity for 9 protein-DNA complexes using the IDEA-learned energy model shows a strong correlation with the experimental measurements.[72] $\Delta\Delta G$ represents the changes in binding energy relative to the protein-DNA complex with the lowest predicted binding energy. The predicted binding energies are presented in reduced units, as explained in Methods. (B) Illustration of an example protein-DNA complex structure (PDB ID: 9ANT) and the coarse-grained simulation used to evaluate protein-DNA binding free energy. A typical free energy profile was extracted from the simulation, using the center of mass (COM) distance as the collective variable. The shaded region represents the standard deviation of the mean. Representative structures of bound and unbound proteins 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 non-specific attraction between protein and DNA (Figure S6). The predicted binding energies are presented in physical units. Error bars represent the standard deviation of the mean.

Discussion and Conclusion

The protein-DNA interaction landscape has evolved to enable specific targeting of proteins towards their functional target sites, which underlie essential processes in controlling gene expression. Such interaction specifics are determined by the physicochemical interactions between amino acids and nucleotides. By harnessing available sequences and structural data of available protein-DNA complexes, we introduce IDEA, a data-driven method that learns a system-specific energy model that enables high-throughput *in silico* predictions of protein-DNA binding specifics and can be scaled up to the genome level. IDEA achieves accurate *de novo* prediction by using only protein-DNA complex structures and their associated sequences but can be further enhanced with additional binding assay measurements, such as the SELEX data.[15,16,21] Despite significant progress in genome-wide sequencing techniques,[9,10,17,46] determining the binding specificities of DNA-binding biomolecules remains time-consuming and expensive. Therefore, IDEA presents a cost-effective alternative for making the initial prediction 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 training data. Prior efforts have applied deep learning techniques to predict the interactions between a given protein and DNA based on their sequences. These approaches usually required a large amount of sequencing data.[24–26] By leveraging the sequence-structure relationship inherent in protein-DNA complex structures, IDEA achieves comparable or better predictive accuracy while using only a few protein-DNA complexes available within the same protein superfamily. Moreover, the predictive power of IDEA can be further enhanced by incorporating available experimental data, as demonstrated in our prediction of MAX-DNA binding specificities (Fig. 2A,B).

Another highlight of IDEA is that it presents an interpretable amino acid-nucleotide interaction energy model of given protein-DNA complexes. Unlike other machine-learning approaches, this optimized energy model not only predicts binding affinities and binding sites of the DNA-binding proteins but also provides an interpretable representation of the

interaction specifics between each type of amino acid and nucleotide. Additionally, we integrated this physicochemical-based energy model into a simulation model, thereby improving the characterization of protein-DNA binding dynamics. Therefore, IDEA-based simulation enables investigations into dynamic interactions among various proteins and DNA, facilitating molecular-based simulations to understand the physical mechanisms underlying many DNA-binding processes, such as transcription, epigenetic regulations, and their modulation by sequence mutations such as single-nucleotide polymorphisms (SNPs).[75,76]

Since the IDEA-optimized energy model serves as a "molecular grammar" guiding the protein-DNA interaction specifics, it can be expanded to include additional variants of amino acids and nucleotides. With recent advancement in sequencing and structural characterization,[77,78] as well as deep-learning-based structural predictions,[79–81] thousands of structural and sequence data 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 amino acid and nucleotide variants, such as methylated DNA[82] and post-translationally modified amino acids.[83] Such strategies will facilitate studies on the functional relevance of epigenetic variants, such as those caused by exposure to environmental hazards,[84] and their structural impact on the human genome.[85–87]

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. Specifically, we represented each protein amino acid by extracting 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 at the nitrogenous base. The choice between these two DNA atom types depends on the number of Cα atoms surrounding that DNA atom. We hypothesized that DNA atoms with more surrounding Cα atoms would contain more

information on the protein-DNA interactions and chose the DNA atom type with a larger sum of the number of Cα atoms within 8Å, 9Å, and 10Ådistances from the DNA atoms (Table S1). We consistently applied the same selection rule across all complexes included in the same training dataset based on the majority of selections.

Energy Model Optimization

Training Protocol

We included the native protein and DNA sequences from the experimentally determined protein-DNA structural interface 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 Fig. 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) from the strong binders. The IDEA protocol optimizes a pair-wise amino-acid-nucleotide energy model 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.[33,35,51,88,89] Specifically, the binding energies for strong binders (E_{strong}) and their corresponding decoy weak binders (E_{decoy}) are calculated using the following expression:

$$E_{\text{binding}} = \sum_{i \in \text{protein}, j \in \text{DNA}} \gamma_{i,j}(a_i, a_j)\Theta_{i,j} \tag{1}$$

where $\Theta_{i,j}$ utilized a switching function that captures an effective interaction range between each amino acid-nucleotide pair within the protein-DNA complex:

$$\Theta_{i,j} = 0.5 \cdot (\tanh(\kappa \cdot (r_{i,j} - r_{\min})) \cdot \tanh(\kappa \cdot (r_{\max} - r_{i,j}))) + 0.5 \tag{2}$$

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.

IDEA optimizes γ, a protein-DNA energy model represented by a 20-by-4 interaction matrix. Each entry $\gamma_{i,j}(a_i, a_j)$ describes the pairwise interaction between an amino acid i and a nucleotide j. In practice, the parameters $\gamma_{i,j}(a_i, a_i)$ were optimized to maximize $\delta E/\Delta E$, where $\delta E = \langle E_{\text{decoy}} \rangle - \langle E_{\text{strong}} \rangle$, representing the energy gap between strong and decoy binders. Mathematically, δE can be represented as $A\gamma$. The standard deviation of the decoy binding energies ΔE can be calculated as $\Delta E = \sqrt{\gamma^T B \gamma}$, where:

$$A = \langle \phi_{\text{decoy}} \rangle - \langle \phi_{\text{strong}} \rangle \tag{3}$$

$$B = \left\langle \phi_{\text{decoy}} \phi_{\text{decoy}}^T \right\rangle - \langle \phi_{\text{decoy}} \rangle \left\langle \phi_{\text{decoy}}^T \right\rangle \tag{4}$$

here, ϕ takes the function form of E_{binding} and summarizes the total number of contacts between each type of amino acid and nucleotide in the given protein-DNA complexes. The vector A thus specifies the difference in interaction strengths for each pair of amino acids and nucleotides between the strong and decoy binders, with a dimension of 1×300. Additionally, matrix B is a 300×300 covariance matrix that contains information about which types of interactions tend to co-occur in the decoy binders. Optimizing the function $\delta E/\Delta E$ with respect to the parameters $\gamma_{i,j}(a_i, a_j)$ corresponds to maximizing the ratio $A^T\gamma/\sqrt{\gamma^T B \gamma}$. This maximization is effectively attained by maximizing the functional objective $R(\gamma)$, defined as:

$$R(\gamma) = A^T\gamma - \lambda\sqrt{\gamma^T B \gamma} \tag{5}$$

where λ is a Lagrange multiplier. Setting the derivative $\frac{\partial R(\gamma)}{\partial \gamma}$ equal to zero leads to $\gamma \propto B^{-1}A$. Here, γ represents a 300×1 vector that encodes the learned binding strengths across different amino acid-nucleotide interactions. 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 of our optimization reported in this paper, N was set to 70 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 Fig. 2C. Due to the presence of an undetermined scaling factor after the model optimization, the predicted energies are presented in reduced units. Only when the optimized energy is integrated into a simulation model, can IDEA-based simulations predict binding energies in physical units (Fig. 5C).

Enhanced Modeling Prediction with SELEX Data

When additional experimental binding data, such as SELEX,[26] is available, we expanded 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 energy with the expression $E = -\ln(\text{Affinity})$, with Affinity being the normalized affinity generated from the SELEX package. We then calculated the ϕ value by utilizing the protein-DNA complex structure threaded with the SELEX-provided DNA sequences, which can be used together with the converted E values to compute the γ energy model based on the equation 1:

$$\gamma = \phi^{-1}E \tag{6}$$

here γ is the enhanced protein-DNA energy model represented as a 300×1 vector.

For predicting the MAX-DNA binding specificity, we utilized the Human MAX SELEX data deposited in the European Nucleotide Archive database (accession no. PRJEB25690).[26,90] The R/Bioconductor package SELEX version 1.30.0[91] 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 energy to compute the γ energy model.

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,[92] which include a total of 11 protein-DNA structures with the E-box sequence (CACGTC), 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 and their associated sequences, predicting the MAX-DNA binding affinity in the MITOMI dataset.[13] To examine the connection between the predictive outcome and the probability that the training protein is homologous to MAX, HHpred[93] was utilized to search for MAX's homologous proteins, using the sequence of 1HLO as the query (Fig. 3A).

Processing of ChIP-seq Data

To evaluate the genome-wide predictive capabilities of our protocol, we selected the MAX ChIP-seq data from the human GM12878 chromosome 1 (chr1) for testing (accession number CM000663.2).[94] Notably, the genomic segment from 156,000,000 bp to 157,000,000 bp exhibited the densest ChIP-seq signals, containing a total of 33 binding sites. We thus use this 1 Mb region to test our modeling predictions. To predict the binding sites of a MAX

transcription factor, we utilized the strand of the 11-bp DNA sequence closest to the associated MAX protein in the protein-DNA complex structure (PDB ID: 1HLO) as a reference and scanned it through the entire DNA region base by base. This process generated a total of $1,000,000 - 11 + 1 = 999,990$ binder sequences. We then predicted the binding energies of these target sequences using the trained MAX-TF energy model (Fig. 2), which were normalized into a Z score using the calculated binding energies of the randomized decoy sequences with

$$Z_{norm} = \frac{E - \langle E_{decoy} \rangle}{\mathrm{SD}(E_{decoy})} \tag{7}$$

where $\langle\ \rangle$ denotes the average over all the decoy sequences, and SD denotes the standard deviation. Since the whole genome sequence has evolved to favor MAX binding, most of the normalized Z scores are negative (Fig. 4A). The ROC curve was constructed by considering the lower binding energy as representing stronger binding affinity.

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,[55] 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,[56] 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.[55,56,59,61–64,66,67,70,95]

Detailed descriptions of this protein-DNA force field have been documented in previous studies.[58,63,64,67] Specifically, we followed Ref. 56 for simulating DNA-DNA interactions, and Ref. 55 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:

$$U_{\text{elec}}(r) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon r}e^{-r/l_D} \tag{8}$$

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 electric constant in vacuum, 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:

$$U_{\text{exclude}}(r) = 4\epsilon(\frac{\sigma}{r})^{12} \tag{9}$$

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

$$V_{\text{PD}}(r) = \frac{W}{2}(1 + \tanh[\eta(r_0 - r)]) \tag{10}$$

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, W used the normalized IDEA-learned energy model based on 9 protein-DNA complexes (Fig. 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 S2. Additional details on the protein-DNA simulation are provided in the Supplementary Materials.

Acknowledgement

This work was supported by the startup funding from North Carolina State University. This research was also funded, in part, by the NC State Genetics and Genomics Academy. We thank Dr. Keith Weninger, Dr. Faruck Morcos, and Dr. Qin Zhou for very helpful discussions. We also thank Dr. Sebastian Maerkl for directing us to the MITOMI experimental

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 specificities

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

(26) 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*.

(27) 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.

(28) 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.

(29) 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.

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

(31) Sagendorf, J. M.; Markarian, N.; Berman, H. M.; Rohs, R. DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Research* **2019**, *48*, D277–D287.

(32) 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.

(33) 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.

(34) 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.

(35) Schafer, N. P.; Kim, B. L.; Zheng, W.; Wolynes, P. G. Learning to fold proteins using energy landscape theory. *Israel journal of chemistry* **2014**, *54*, 1311–1337.

(36) 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.

(37) 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.

(38) Petrenko, N.; Jin, Y.; Dong, L.; Wong, K. H.; Struhl, K. Requirements for RNA polymerase II preinitiation complex formation in vivo. *eLife* **2019**, *8*, e43654.

(39) Marchal, C.; Sima, J.; Gilbert, D. M. Control of DNA replication timing in the 3D genome. *Nat Rev Mol Cell Biol* **2019**, *20*, 721–737.

(40) 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.

(41) 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.

(42) 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.

(43) 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.

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

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

(46) 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.

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

(48) 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.

(49) 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.

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

(51) 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.

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

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

(54) 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.

(55) 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.

(56) 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.

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

(58) 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.

(59) 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.

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

(61) 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.

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

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

(64) Liu, S.; Lin, X.; Zhang, B. Chromatin fiber breaks into clutches under tension and crowding. *Nucleic Acids Research* **2022**, gkac725.

(65) 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.

(66) 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.

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

(68) 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.

(69) 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.

(70) 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.

(71) 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.

(72) 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.

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

(74) 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.

(75) 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.

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

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

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

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

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

(81) 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.

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

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

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

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

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

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

(88) 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.

(89) 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.

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

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

(92) 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.

(93) 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.

(94) National Center for Biotechnology Information (NCBI) GenBank. https://www.ncbi.nlm.nih.gov/genbank/.

(95) 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*.