Structural Biology and Molecular Biophysics

Interpretable Protein-DNA Interactions Captured by Structure-Sequence Optimization

Reviewed Preprint
v1 • February 7, 2025
Not revised

Yafan Zhang, Irene Silvernail, Zhuyang Lin, Xingcheng Lin

Bioinformatics Research Center, North Carolina State University, Raleigh, United States • Department of Physics, North Carolina State University, Raleigh, United States

eLife Assessment

This **valuable** work presents an interpretable protein-DNA Energy Associative (IDEA) model for predicting binding sites and affinities of DNA-binding proteins. The study provides a detailed description of the method, making it reproducible. However, the generalizability of the prediction model presents certain concerns, and the supporting evidence appears **incomplete**. Nonetheless, with a thorough re-examination of the training and testing procedures, this model can be widely applicable for predicting genome-wide protein-DNA binding sites.

https://doi.org/10.7554/eLife.105565.1.sa3

Abstract

Sequence-specific DNA recognition underlies essential processes in gene regulation, yet methods for simultaneous prediction 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 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. [23–30] 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] 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.[31]

Nowadays, over 5000 protein-DNA 3D structures, including TF-DNA complexes, have been published.[32,33] 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 [34–37] 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 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 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 dynamics.

Results

Residue-Level Protein-DNA Energy Model for Predicting ProteinDNA 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.[32,38] 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 the prediction of genomic 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,[39] epigenetic regulation of gene expression,[4] and DNA replication.[40] Further details of the optimization protocol are provided in Methods Section *Energy Model Optimization*.

IDEA Accurately Predicts Protein-DNA Binding Specificity

We first examine the predictive accuracy of IDEA by comparing its predicted TF-DNA binding affinity with experimental measurements.[13,14] We focused on the MAX 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.[13,41] 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 (Ebox) 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.[27,42] 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 gained a significant predictive improvement in reproducing the MITOMI measurement (Pearson Correlation 0.79, Fig. S2).

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-

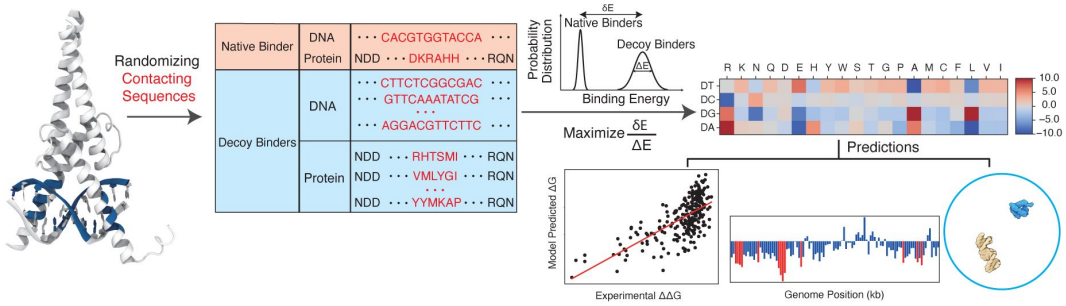

Figure 1

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) in the structure were included in the training dataset. In addition, a series of synthetic decoy sequences were 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.

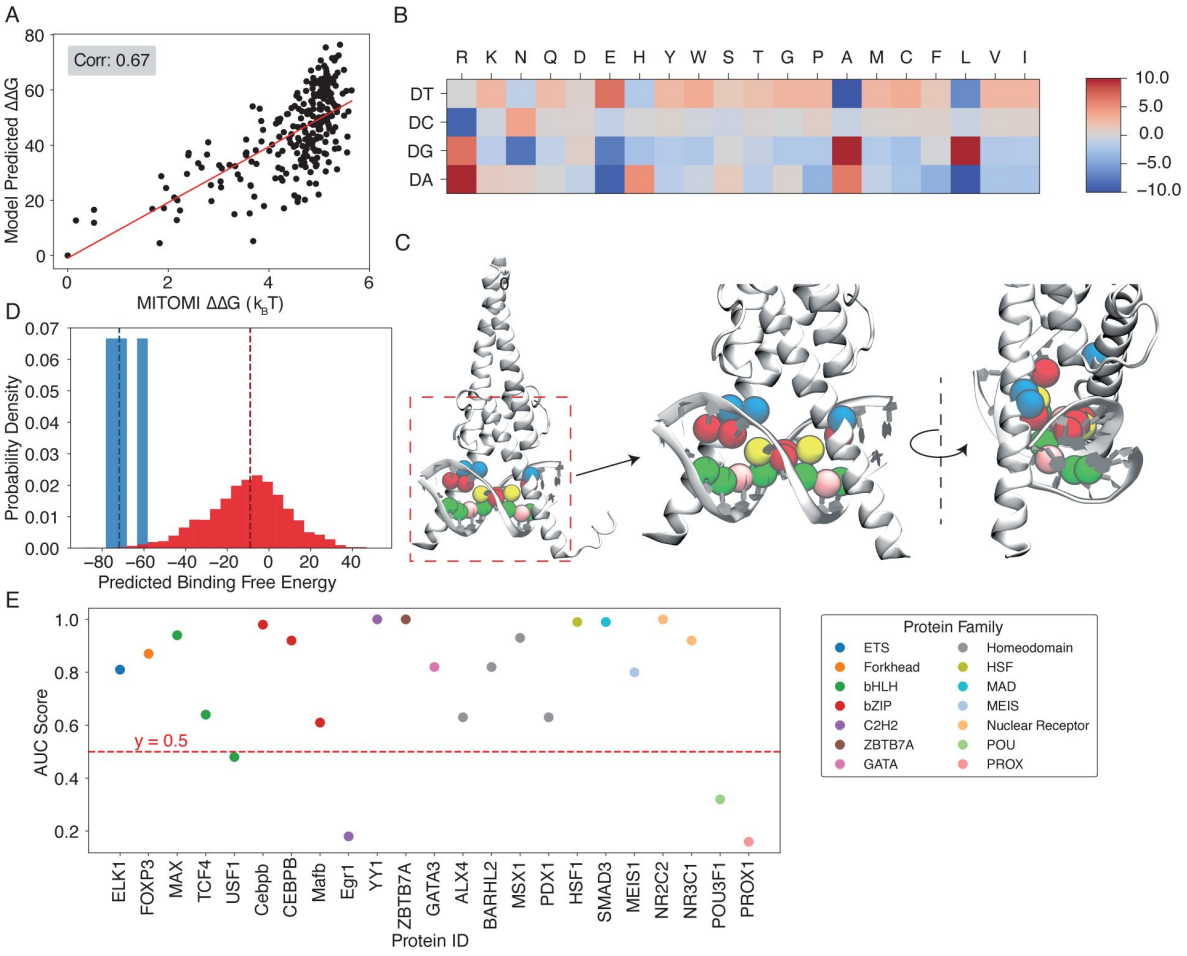

Figure 2

Results for MAX-based predictions.

(A) The binding free energy calculated by IDEA, trained using one MAX–DNA complex (PDB ID: 1HLO), correlates well with experimentally measured MAX–DNA binding free energy.[13] $\Delta\Delta G$ represents the changes in binding free energy relative to that of the wild-type protein–DNA complex. (B) The optimized energy model reveals an amino acid–nucleotide interaction pattern governing MAX–DNA recognition. The predicted binding free energies and optimized energy model are presented in reduced units, as explained in the Methods. (C) The 3D structure of the MAX–DNA complex (zoomed in with different views) highlights important amino acid–nucleotide contacts at the protein–DNA interface, where several DNA cytosines (red spheres) form close contacts with arginine (blue spheres). (D) Probability density distribution of strong and weak binders for the protein ZBTB7A. The mean of each distribution is marked with a dashed line. (E) The AUC score for each protein–DNA pair, calculated based on the predicted probability distributions.

DNA complexes (**Fig. 2B**). Notably, DNA Cytosine (DC) exhibited strong interactions with Arginine (R), consistent with the fact that the E-box region (CACGTG) frequently attracts the positively charged residues of bHLH TFs.[43] Importantly, Arginine was in close contact with Cytosine in the crystal structure (**Fig. 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 glutamic acid (E) and all DNA nucleotides, consistent with their negative charges (Fig. 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 experimentally determined structures. 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 Fig. S4). Receiver Operating Characteristic (ROC) analysis was performed to calculate the Area Under the Curve (AUC) score for these predictions. Further details are provided in Methods Section *Evaluation of IDEA prediction using HT-SELEX data.* Our analysis shows that for 80% of the proteins across 14 protein families, IDEA successfully differentiates strong binders from weak ones, with an AUC score greater than 0.5. We further performed 10-fold cross-validation on the binding affinities of the protein-DNA pairs in this dataset and found that IDEA outperforms state-of-theart protein-DNA models for cases with available experimentally determined protein-DNA complex structures (Fig. S5)

We also applied IDEA to calculate the binding affinity of additional TFs with available MITOMI measurements.[41] IDEA's predicted binding affinity of Pho4 protein, another bHLH transcription factor, correlates well with the experimental measurement (Pearson Correlation 0.6, Fig. S6A) by training on the only available protein-DNA structure (PDB ID: 1A0A). We further evaluated IDEA's predictive performance for the zinc-finger protein Zif268, which has a different structure from the bHLH TFs and thus requires a different atom to represent the DNA nucleotides (Table S1). Due to limited experimental data for all the possible DNA sequence combinations, we focused on testing IDEA's predictions on point-mutated DNA sequences, which was known to be a challenging task due to their minor deviation from the wild-type sequence. Despite the challenge, IDEA achieves accurate predictions, with a Pearson Correlation of 0.57 (Fig. S6B). 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. S7).

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 proteinDNA 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 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 Correlation coefficients larger than 0.5 (**Fig. 3A**).

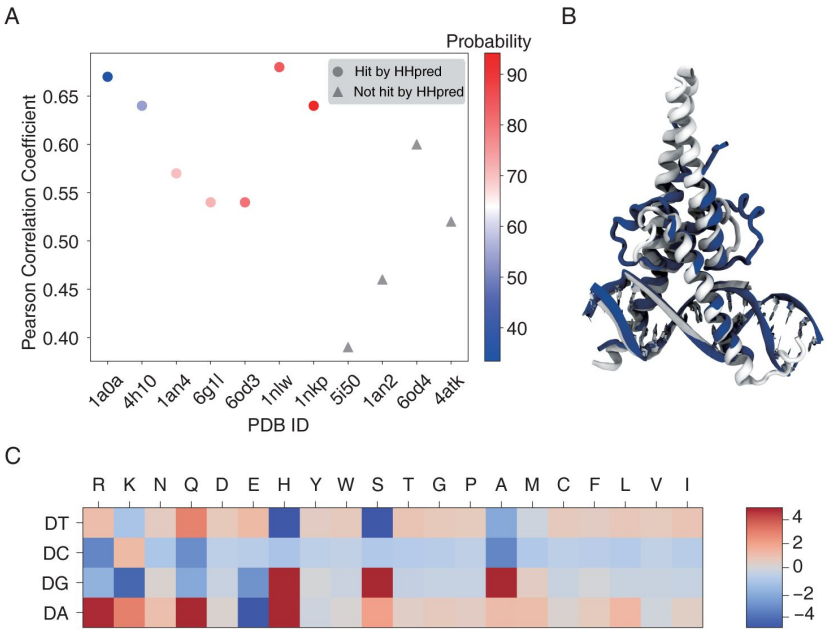

Figure 3

IDEA prediction shows transferability within the same CATH superfamily.

(A) The predicted MAX binding specificity, 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.[44] 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.[45] (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.

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. 2B**).

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, [9,17,46] DAP-seq,[10] and FAIREseq,[47] 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.[48] As the experimental measurements were conducted at the 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 **Fig. 4A** top). Lower predicted binding free energy corresponds to stronger protein-DNA binding affinity, indicating higher probability binding sites.

The predicted binding free energies were further normalized by those of 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 (**Fig. 4B**). We also tested IDEA's predictive performance on two additional typical DNA-binding proteins and successfully identified their genomic binding sites (Fig. S8). 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.[49–51]

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.[52–56] Notably, the structure-based protein model[57] and 3-Site-Per-Nucleotide (3SPN) DNA model[58] 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,[59] including nucleosomal DNA unwrapping,[60–62] interactions between nucleosomes,[63,64] large chromatin organizations,[64–66] and its modulation by chromatin factors.[67–69] Although both the protein and DNA force fields have been systematically optimized,[58,70–73]

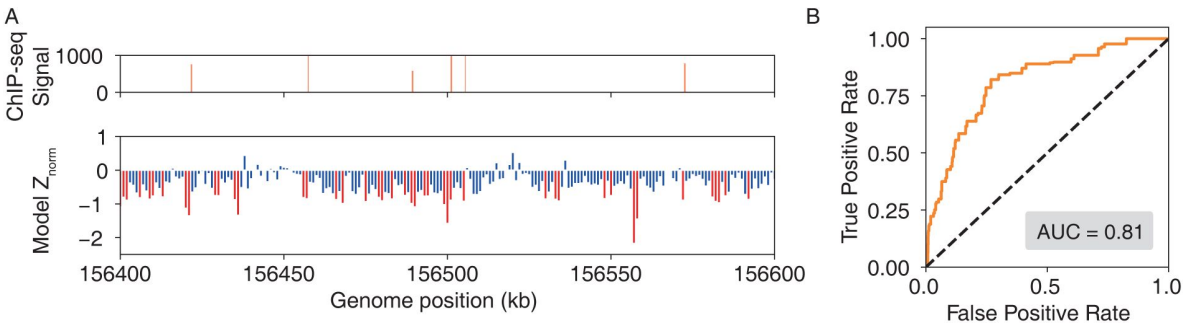

Figure 4

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.

the non-bonded interactions between protein and DNA were primarily guided by Debye-Hückel treatment of electrostatic interactions, which did not consider sequencespecific 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.[74] 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.[75] The predicted binding free energy[76] (Fig. S9) 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. S10). 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.

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 datadriven method that optimizes a system-specific energy model. This model enables highthroughput *in silico* predictions of protein-DNA binding specificities 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 (Figs. S2 and S5). Despite significant progress in genome-wide sequencing techniques,[9,10,17,47] determining the binding specificities 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.

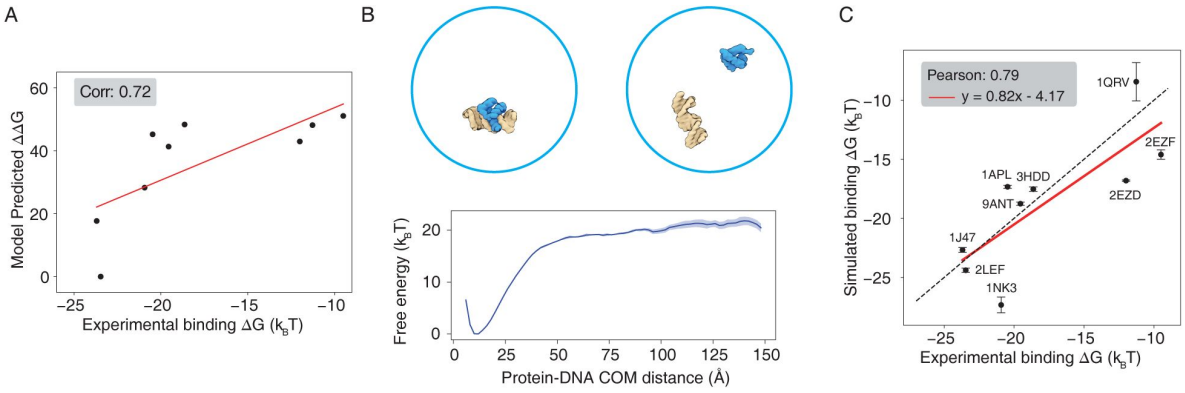

Figure 5

Enhanced protein-DNA simulation model by incorporating IDEAoptimized 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.[74] $\Delta\Delta G$ represents the changes in binding free energy relative to the protein-DNA complex with the lowest predicted binding free energy. The predicted binding free 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 (Fig. S9). The predicted binding free energies are presented in physical units. Error bars represent the standard deviation of the mean.

In addition, despite the complicated *in vivo* environment, IDEA enables predictions of genomic binding sites and shows good agreement with the ChIP-seq measurements (**Fig. 4**, S8). 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 amino acid-nucleotide interaction energy model for given protein-DNA complexes. Unlike other machine-learning approaches, IDEA's optimized energy model not only predicts binding affinities and binding sites of DNA-binding proteins but also provides an interpretable representation of interaction specifics between each type of amino acid and nucleotide. Additionally, we integrated this physicochemical-based energy model into a simulation framework, 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). [77,78]

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,[79,80] as well as deep-learning-based structural predictions,[81–83] 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 [84] and post-translationally modified amino acids.[85] Such strategies will facilitate studies on the functional relevance of epigenetic variants, such as those caused by exposure to environmental hazards,[86] and their structural impact on the human genome.[87–89]

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 S1). This selection rule was applied consistently across all complexes in the training dataset to maintain uniform criteria.

Energy Model Optimization

Training Protocol

We fused the native protein and DNA sequences from the experimentally determined proteinDNA 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 ≤ 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 pairwise 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.[34,36,53,90,91] Specifically, the binding energies for strong binders (E_{strong}) and their corresponding decoy weak binders (E_{decoy}) are calculated using the following expression:

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

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:

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

here, $r_{\min}$ = −8 Å and $r_{\max}$ = 8 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:

$$A = \langle \phi_{\text{decoy}} \rangle - \langle \phi_{\text{strong}} \rangle \quad (3)$$

$$B = \langle \phi_{\text{decoy}} \phi_{\text{decoy}}^T \rangle - \langle \phi_{\text{decoy}} \rangle \langle \phi_{\text{decoy}}^T \rangle \quad (4)$$

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:

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

where λ is a Lagrange multiplier. Setting the derivative $\frac{\partial R(\gamma)}{}$ 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 N th 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. 2B**. 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 free energies in physical units (**Fig. 5C**).

Enhanced Modeling Prediction with SELEX Data

When additional experimental binding data, such as SELEX,[27] 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 free 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).[27,92] The R/Bioconductor package SELEX version 1.30.0[93] 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 energy to compute the γ energy model.

Evaluation of IDEA Predictions Using HT-SELEX data

The processed HT-SELEX data used in this study is 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 (Fig. S4). A well-separated distribution between strong and weak binders indicates a successful prediction. Additionally, we calculated the AUC score based on ROC analysis of these probability distributions. An AUC score of 1.0 means IDEA can fully separate the strongand weak-binder distributions, while an AUC score of 0.5 indicates no separation.

We further performed 10-fold cross-validation on this dataset, using the protocol described in 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 set and the remaining fold as the test set. 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 our model's predictive performance. Our results are

compared with the 1mer and 1mer+shape methods from[26] (Fig. S5). our comparative analysis shows IDEA achieved higher predictive accuracy than the existing state-of-the-art sequence-based protein-DNA binding predictors for those protein-DNA complexes that have available experimentally resolved structures.

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,[94] 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[95] 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 predictive capabilities of our protocol on specific chromosomal segments, we selected ChIP-seq data for three transcription factors—MAX, EGR1, and CTCF. The data were sourced from the ENCODE project, with accession numbers ENCFF361EVH (MAX), ENCSR026GSW (EGR1), and ENCFF960ZGP (CTCF).

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

To predict binding sites, we scanned the corresponding DNA regions base by base using the DNA sequence closest to each protein in the relevant protein-DNA complex structures. For MAX, we used the 11-bp sequence from the PDB structure 1HLO, generating 1, 000, 000− 11 + 1 = 999, 990 sequences. For EGR1, we used the 10-bp sequence from PDB 1AAY, producing 1, 000, 000 − 10 + 1 = 999, 991 sequences. For CTCF, no single PDB structure covers the entire protein,[96] so two PDB structures were used: 8SSS (ZF1-ZF7, 23-bp) and 8SSQ (ZnFs 3-11, 35-bp), generating 1, 000, 000 −23 + 1 = 999, 978 and 1, 000, 000 −35 + 1 = 999, 966 sequences, respectively.

Predicted binding free energies for each transcription factor were normalized into Z-scores using randomized decoy sequences:

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

where ⟨ ⟩ 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 (**Fig. 4**, S8).

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,[57] 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, [58] 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.[57,58,61,63–66,68,69,72,97]

Detailed descriptions of this protein-DNA force field have been documented in previous studies.[60,65,66,69] Specifically, we followed Ref. 58 for simulating DNA-DNA interactions, and Ref. 57 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 ε = 78.0 is the dielectric constant of the bulk solvent. q_i and q_j correspond to the charges of the two particles. ε_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 σ = 4.0Å:

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

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 η = 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.

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.

Acknowledgements

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. 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.

Additional files

Supplementary Information

References

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

(2) Orphanides G., Reinberg D 2002) **A Unified Theory of Gene Expression** *Cell* **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 2000) **Genome-Wide Location and Function of DNA Binding Proteins** *Science* **290**:2306–2309

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

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

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

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

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

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

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

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

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

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

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

(15) Ogawa N., Biggin M. D., Deplancke B., Gheldof N. 2012) **Gene Regulatory Networks: Methods and Protocols** Totowa, NJ: Humana Press :51–63

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

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

(18) Bulyk M. L., Huang X., Choo Y., Church G. M 2001) **Exploring the DNA-binding specificities of zinc fingers with DNA microarrays** *Proc. Natl. Acad. Sci. U.S.A* **98**:7158–7163

(19) Gord^an, R.; Shen, N.; Dror, I.; Zhou, T.; Horton, J.; Rohs, R.; Bulyk, M 2013) **Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape** *Cell Reports* **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., Gordan R. 2020) **DNA mismatches reveal conformational penalties in protein–DNA recognition** *Nature* **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. 2010) **; others Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities** *Genome research* **20**:861–873

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

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

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

(25) Zhou T., Shen N., Yang L., Abe N., Horton J., Mann R. S., Bussemaker H. J., Gordân R., Rohs R. 2015) **Quantitative modeling of transcription factor binding specificities using DNA shape** *Proceedings of the National Academy of Sciences* **112**:4654–4659

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

(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 2018) **Accurate and sensitive quantification of protein-DNA binding affinity** *Proc. Natl. Acad. Sci. U.S.A* **115**

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

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

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

Bioinformatics **20**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(54) Bascom D., Schlick G. 2018) **Nuclear Architecture and Dynamics** Elsevier :123–147

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

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

(57) Clementi C., Nymeyer H., Onuchic J. N 2000) **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* **298**:937–953

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(73) Freeman G. S., Lequieu J. P., Hinckley D. M., Whitmer J. K., De Pablo J. J 2014) **DNA Shape Dominates Sequence Affinity in Nucleosome Formation** *Phys. Rev. Lett* **113**

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

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

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

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

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

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

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

(81) Abramson J., et al. 2024) **Accurate structure prediction of biomolecular interactions with AlphaFold 3** *Nature* **630**:493–500

(82) Tunyasuvunakool K., et al. 2021) **Highly accurate protein structure prediction for the human proteome** *Nature* **596**:590–596

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

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

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

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

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

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

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

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

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

(92) 2024) **European Nucleotide Archive European Nucleotide Archive** https://www.ebi.ac.uk/ena

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

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

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

(96) Yang J., Horton J. R., Liu B., Corces V. G., Blumenthal R. M., Zhang X., Cheng X 2023) **Structures of CTCF–DNA complexes including all 11 zinc fingers** *Nucleic acids research* **51**:8447–8462

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

Author information

Yafan Zhang
Bioinformatics Research Center, North Carolina State University, Raleigh, United States

Irene Silvernail
Department of Physics, North Carolina State University, Raleigh, United States

Zhuyang Lin
Bioinformatics Research Center, North Carolina State University, Raleigh, United States

Xingcheng Lin
Bioinformatics Research Center, North Carolina State University, Raleigh, United States, Department of Physics, North Carolina State University, Raleigh, United States
ORCID iD: 0000-0002-9378-6174

For correspondence: Lin@ncsu.edu

Editors

Reviewing Editor
Anna Panchenko
Queen's University, Kingston, Canada

Senior Editor
Qiang Cui
Boston University, Boston, United States of America

Reviewer #1 (Public review):

Summary:

This work presents an Interpretable protein-DNA Energy Associative (IDEA) model for predicting binding sites and affinities of DNA-binding proteins. Experimental results demonstrate that such an energy model can predict DNA recognition sites and their binding strengths across various protein families and can capture the absolute protein-DNA binding free energies.

Strengths:

(1) The IDEA model integrates both structural and sequence information, although such an integration is not completely original.

(2) The IDEA predictions seem to have agreement with experimental data such as ChIP-seq measurements.

Weaknesses:

(1) The authors claim that the binding free energy calculated by IDEA, trained using one MAX-DNA complex, correlates well with experimentally measured MAX-DNA binding free energy (Figure 2) based on the reported Pearson Correlation of 0.67. However, the scatter plot in Figure 2A exhibits distinct clustering of the points and thus the linear fit to the data (red line) may not be ideal. As such. the use of the Pearson correlation coefficient that measures linear correlation between two sets of data may not be appropriate and may provide misleading results for non-linear relationships.

(2) In the same vein, the linear Pearson Correlation analysis performed in Figure 5A and the conclusion drawn may be misleading.

(3) The authors included the sequences of the protein and DNA residues that form close contacts in the structure in the training dataset, whereas a series of synthetic decoy sequences were generated by randomizing the contacting residues in both the protein and DNA sequences. In particular, synthetic decoy binders were generated by randomizing either the DNA (1000 sequences) or protein sequences (10,000 sequences) from the strong binders. However, the justification for such randomization and how it might impact the model's generalizability and transferability remain unclear.

(4) The authors performed Receiver Operating Characteristic (ROC) analysis and reported the Area Under the Curve (AUC) scores in order to quantitate the successful identification of the strong binders by IDEA. It would be beneficial to analyze the precision-recall (PR) curve and report the PRAUC metric which could be more robust.

https://doi.org/10.7554/eLife.105565.1.sa2

Reviewer #2 (Public review):

Summary:

Zhang et al. present a methodology to model protein-DNA interactions via learning an optimizable energy model, taking into account a representative bound structure for the system and binding data. The methodology is sound and interesting. They apply this model for predicting binding affinity data and binding sites in vivo. However, the manuscript lacks discussion of/comparison with state-of-the-art and evidence of broad applicability. The interpretability aspect is weak, yet over-emphasized.

Strengths:

The manuscript is well organized with good visualizations and is easy to follow. The methodology is discussed in detail. The IDEA energy model seems like an interesting way to study a protein-DNA system in the context of a given structure and binding data. The authors show that an IDEA model trained on one system can be transferred to other structurally similar systems. The authors show good performance in discriminating between binding-vs-decoy sequences for various systems, and binding affinity prediction. The authors also show evidence of the ability to predict genome-wide binding sites.

Weaknesses:

An energy-based model that needs to be optimized for specific systems is inherently an uncomfortable idea. Is this kind of energy model superior to something like Rosetta-based energy models, which are generally applicable? Or is it superior to family-specific knowledge-based models? It is not clear.

Prediction of binding affinity is a well-studied domain and many competitors exist, some of which are well-used. However, no quantitative comparison to such methods is presented. To understand the scope of the presented method, IDEA, the authors should discuss/compare with such methods (e.g. PMID 35606422).

The term "interpretable" has been used lavishly in the manuscript while providing little evidence on the matter. The only evidence shown is the family-specific residue-nucleotide interaction/energy matrix and speculations on how these values are biologically sensible. Recent works already present more biophysical, fine-grained, and sometimes family-independent interpretability (e.g. PMID 39103447, 36656856, 38352411, etc.). The authors should put into context the scope of the interpretability of IDEA among such works.

The manuscript disregards subtle yet important differences in commonly used terminology in the field. For example, the authors use the term "specificity" and "affinity" almost interchangeably (for example, the caption for Figure 3A uses "specificity" although the Methods text describes the prediction as about "affinity"). If the authors are looking to predict specificity, IDEA needs to be put in the context of the corresponding state-of-the-art (PMID 36123148, 39103447, 38867914, 36124796, etc).

It is not clear how much the learned energy model is dependent on the structural model used for a specific system/family. It would be interesting to see the differences in learned model based on different representative PDB structures used. Similarly, the supplementary figures show a lack of discriminative power for proteins like PDX1 (homeodomain family), POU, etc. Can the authors shed some light on why such different performances?

It is also not clear if IDEA's prediction for reverse complement sequences is the same for a given sequence. If so, how is this property being modelled? Either this description is lacking or I missed it.

https://doi.org/10.7554/eLife.105565.1.sa1

Reviewer #3 (Public review):

Summary:

Protein-DNA interactions and sequence readout represent a challenging and rapidly evolving field of study. Recognizing the complexity of this task, the authors have developed a compact and elegant model. They have applied well-established approaches to address a difficult problem, effectively enhancing the information extracted from sparse contact maps by integrating artificial sequences decoy set and available experimental data. This has resulted in the creation of a practical tool that can be adapted for use with other proteins.

Strengths:

(1) The authors integrate sparse information with available experimental data to construct a model whose utility extends beyond the limited set of structures used for training.

(2) A comprehensive methods section is included, ensuring that the work can be reproduced. Additionally, the authors have shared their model as a GitHub project, reflecting their commitment to transparency of research.

Weaknesses:

(1) The coarse-graining procedure appears artificial, if not confusing, given that full-atom crystal structures provide more detailed information about residue-residue contacts. While the selection procedure for distance threshold values is explained, the overall motivation for

adopting this approach remains unclear. Furthermore, since this model is later employed as an empirical potential for molecular modeling, the use of P and C5 atoms raises concerns, as the interactions in 3SPN are modeled between Cα and the nucleic base, represented by its center of mass rather than P or C5 atoms.

(2) Although the authors use a standard set of metrics to assess model quality and predictive power, some ΔΔG predictions compared to MITOMI-derived ΔΔG values appear nonlinear, which casts doubt on the interpretation of the correlation coefficient.

(3) The discussion section lacks information about the model's limitations and a comprehensive comparison with other models. Additionally, differences in model performance across various proteins and their respective predictive powers are not addressed.

https://doi.org/10.7554/eLife.105565.1.sa0

Author response:

> ***Reviewer 1:***
>
> *Summary: This work presents an Interpretable protein-DNA Energy Associative (IDEA) model for predicting binding sites and affinities of DNA-binding proteins. Experimental results demonstrate that such an energy model can predict DNA recognition sites and their binding strengths across various protein families and can capture the absolute protein-DNA binding free energies.*

We appreciate the reviewer's careful assessment of the paper, and we thank the reviewer for the insightful suggestions and comments.

> *Strengths:*
>
> *(1) The IDEA model integrates both structural and sequence information, although such an integration is not completely original. (2) The IDEA predictions seem to have agreement with experimental data such as ChIP-seq measurements.*

We appreciate the reviewer's comments on the strength of the paper.

> *Weaknesses:*
>
> *(1) The authors claim that the binding free energy calculated by IDEA, trained using one MAX-DNA complex, correlates well with experimentally measured MAX-DNA binding free energy (Figure 2) based on the reported Pearson Correlation of 0.67. However, the scatter plot in Figure 2A exhibits distinct clustering of the points and thus the linear fit to the data (red line) may not be ideal. As such. the use of the Pearson correlation coefficient that measures linear correlation between two sets of data may not be appropriate and may provide misleading results for non-linear relationships.*

We thank the reviewer for the insightful comments and agree that the linear fit between our predictions and the experimental data may not be ideal. The primary utility of the IDEA model is for assessing the relative binding affinities of different DNA sequences. To further support this, we plan to conduct additional statistical analyses that are independent of the linear correlation assumption but instead focus on the ranked order of DNA sequence binding affinities.

> *(2) In the same vein, the linear Pearson Correlation analysis performed in Figure 5A and the conclusion drawn may be misleading.*

We thank the reviewer for the insightful comments. We will perform the same analysis for Figure 5A as detailed in our response to the previous comments.

> *(3) The authors included the sequences of the protein and DNA residues that form close contacts in the structure in the training dataset, whereas a series of synthetic decoy sequences were generated by randomizing the contacting residues in both the protein and DNA sequences. In particular, synthetic decoy binders were generated by randomizing either the DNA (1000 sequences) or protein sequences (10,000 sequences) from the strong binders. However, the justification for such randomization and how it might impact the model's generalizability and transferability remain unclear.*

We thank the reviewer for the insightful comments. We will perform additional analyses to assess the robustness of our model predictions with respect to the number of randomized decoys. Additionally, we will examine how randomization would potentially affect the model's generalizability and transferability.

> *(4) The authors performed Receiver Operating Characteristic (ROC) analysis and reported the Area Under the Curve (AUC) scores in order to quantitate the successful identification of the strong binders by IDEA. It would be beneficial to analyze the precision-recall (PR) curve and report the PRAUC metric which could be more robust.*

We agree with Reviewer 1 that more statistical metrics should be used to evaluate our model's performance. We will include a more robust approach, such as PRAUC, to evaluate our model.

> ***Reviewer 2:***
>
> *Summary:*
>
> *Zhang et al. present a methodology to model protein-DNA interactions via learning an optimizable energy model, taking into account a representative bound structure for the system and binding data. The methodology is sound and interesting. They apply this model for predicting binding affinity data and binding sites in vivo. However, the manuscript lacks discussion of/comparison with state-of-the-art and evidence of broad applicability. The interpretability aspect is weak, yet over-emphasized.*

We appreciate the reviewer's excellent summary of the paper, and we thank the reviewer for the insightful suggestions and comments.

> *Strengths:*
>
> *The manuscript is well organized with good visualizations and is easy to follow. The methodology is discussed in detail. The IDEA energy model seems like an interesting way to study a protein-DNA system in the context of a given structure and binding data. The authors show that an IDEA model trained on one system can be transferred to other structurally similar systems. The authors show good performance in discriminating between binding-vs-decoy sequences for various systems, and binding affinity prediction. The authors also show evidence of the ability to predict genome-wide binding sites.*

We appreciate the reviewer's strong assessment of the strengths of this paper.

> *Weaknesses:*
>
> *An energy-based model that needs to be optimized for specific systems is inherently an uncomfortable idea. Is this kind of energy model superior to something like Rosetta-based energy models, which are generally applicable? Or is it superior to family-specific knowledge-based models? It is not clear.*

We thank the reviewer for the insightful comments. We will include predictions by generic protein-DNA energy models, such as the Rosetta-based energy model or family-specific knowledge-based model, to compare with our model performance.

> *Prediction of binding affinity is a well-studied domain and many competitors exist, some of which are well-used. However, no quantitative comparison to such methods is presented. To understand the scope of the presented method, IDEA, the authors should discuss/compare with such methods (e.g. PMID 35606422).*

We thank the reviewer for the insightful comments. In our initial submission, Figure S5 presents a comparison between our model's prediction and those of an existing method using 10-fold cross-validation. We agree a more comprehensive comparison with other methods is needed and will include a discussion and comparison of the IDEA model's performance with additional state-of-the-art models.

> *The term "interpretable" has been used lavishly in the manuscript while providing little evidence on the matter. The only evidence shown is the family-specific residue-nucleotide interaction/energy matrix and speculations on how these values are biologically sensible. Recent works already present more biophysical, fine-grained, and sometimes family-independent interpretability (e.g. PMID 39103447, 36656856, 38352411, etc.). The authors should put into context the scope of the interpretability of IDEA among such works.*

We agree that "interpretability" should be discussed in a relevant context. We will discuss the scope of IDEA interoperability within the context of recent works, including those suggested by the reviewers.

> *The manuscript disregards subtle yet important differences in commonly used terminology in the field. For example, the authors use the term "specificity" and "affinity" almost interchangeably (for example, the caption for Figure 3A uses "specificity" although the Methods text describes the prediction as about "affinity"). If the authors are looking to predict specificity, IDEA needs to be put in the context of the corresponding state-of-the-art (PMID 36123148, 39103447, 38867914, 36124796, etc).*

We really appreciate the reviewer for pointing out our conflation of "specificity" and "affinity" in the manuscript. To clarify, IDEA's primary function is to predict the binding affinities of protein-DNA pairs in a sequence-specific manner. The acquired binding affinities of target DNA sequences can then be used to assess the specific binding motifs. We will revise our text to clarify this point.

> *It is not clear how much the learned energy model is dependent on the structural model used for a specific system/family. It would be interesting to see the differences in learned model based on different representative PDB structures used. Similarly, the supplementary figures show a lack of discriminative power for proteins like PDX1 (homeodomain family), POU, etc. Can the authors shed some light on why such different performances?*

We thank the reviewer for the insightful comments and agree that the familyspecific energy model could provide insight into the model predictions. We will examine different energy models based on the protein family, and especially investigate whether they can explain the lack of discriminative power for certain proteins.

> *It is also not clear if IDEA's prediction for reverse complement sequences is the same for a given sequence. If so, how is this property being modelled? Either this description is lacking or I missed it.*

We thank the reviewer for the insightful comments. The IDEA model treats reverse complementary sequences separately. We will provide additional details on how these sequences are modeled.

> ***Reviewer 3:***
>
> *Summary:*
>
> *Protein-DNA interactions and sequence readout represent a challenging and rapidly evolving field of study. Recognizing the complexity of this task, the authors have developed a compact and elegant model. They have applied well-established approaches to address a difficult problem, effectively enhancing the information extracted from sparse contact maps by integrating artificial sequences decoy set and available experimental data. This has resulted in the creation of a practical tool that can be adapted for use with other proteins.*

We appreciate the reviewer's excellent summary of the paper, and we thank the reviewer for the insightful suggestions and comments.

> *Strengths:*
>
> *(1) The authors integrate sparse information with available experimental data to construct a model whose utility extends beyond the limited set of structures used for training. (2) A comprehensive methods section is included, ensuring that the work can be reproduced. Additionally, the authors have shared their model as a GitHub project, reflecting their commitment to transparency of research.*

We appreciate the reviewer's strong assessment of the strengths of this paper.

> *Weaknesses:*
>
> *(1) The coarse-graining procedure appears artificial, if not confusing, given that full-atom crystal structures provide more detailed information about residue-residue contacts. While the selection procedure for distance threshold values is explained, the overall motivation for adopting this approach remains unclear. Furthermore, since this model is later employed as an empirical potential for molecular modeling, the use of P and C5 atoms raises concerns, as the interactions in 3SPN are modeled between C_α and the nucleic base, represented by its center of mass rather than P or C5 atoms.*

We appreciate the reviewer's insightful comments. The selection of P and C5 atoms will augment our model prediction, but the prediction is robust without this selection scheme. We will provide more details on the motivation behind this selection.

Regarding the simulation model, we acknowledge a potential disconnection between the coarse-grained level of the 3SPN model (3 coarse-grained sites per nucleotide) and the data-driven model (1 coarse-grained site per nucleotide). The selection of nucleic bases for

molecular interactions in the 3SPN model follows the PI's previous work [PMID: 34057467] and its code implementation. We will test the simulation model by incorporating interactions between Cff and P atoms. In the future, we will work on implementing IDEA model output for 1-bead-per-nucleotide DNA simulation models.

> *(2) Although the authors use a standard set of metrics to assess model quality and predictive power, some ΔΔG predictions compared to MITOMI-derived ΔΔG values appear nonlinear, which casts doubt on the interpretation of the correlation coefficient.*

We thank the reviewer for the insightful comments and agree that the linear fit between our model's prediction and the experimental data may not be ideal. The primary utility of the IDEA model is for assessing the relative binding affinities of different DNA sequences. To this end, we plan to perform additional statistical analyses that are independent of the linear correlation assumption but instead focus on the ranked order of DNA sequence binding affinities.

> *(3) The discussion section lacks information about the model's limitations and a comprehensive comparison with other models. Additionally, differences in model performance across various proteins and their respective predictive powers are not addressed.*

We thank the reviewer for the insightful comments and will compare the performance of the IDEA model with state-of-the-art methods. We will also perform detailed analyses of the learned energy models across different proteins and examine their correlation with the model's predictive powers.

https://doi.org/10.7554/eLife.105565.1.sa4