PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Physicochemical models of protein–DNA binding with standard and modified base pairs

Tsu-Pei Chiu[a], Satyanarayan Rao[a,1], and Remo Rohs[a,b,c,d,2]

Edited by Barry Honig, Columbia University, New York, NY; received April 1, 2022; accepted December 12, 2022

DNA-binding proteins play important roles in various cellular processes, but the mechanisms by which proteins recognize genomic target sites remain incompletely understood. Functional groups at the edges of the base pairs (bp) exposed in the DNA grooves represent physicochemical signatures. As these signatures enable proteins to form specific contacts between protein residues and bp, their study can provide mechanistic insights into protein–DNA binding. Existing experimental methods, such as X-ray crystallography, can reveal such mechanisms based on physicochemical interactions between proteins and their DNA target sites. However, the low throughput of structural biology methods limits mechanistic insights for selection of many genomic sites. High-throughput binding assays enable prediction of potential target sites by determining relative binding affinities of a protein to massive numbers of DNA sequences. Many currently available computational methods are based on the sequence of standard Watson–Crick bp. They assume that the contribution of overall binding affinity is independent for each base pair, or alternatively include dinucleotides or short *k*-mers. These methods cannot directly expand to physicochemical contacts, and they are not suitable to apply to DNA modifications or non-Watson–Crick bp. These variations include DNA methylation, and synthetic or mismatched bp. The proposed method, DeepRec, can predict relative binding affinities as function of physicochemical signatures and the effect of DNA methylation or other chemical modifications on binding. Sequence-based modeling methods are in comparison a coarse-grain description and cannot achieve such insights. Our chemistry-based modeling framework provides a path towards understanding genome function at a mechanistic level.

transcription factor | binding specificity | readout mode | quantitative modeling | deep learning

Significance

The Watson–Crick sequence model enables a simplified representation of DNA, wherein four letters, A, C, G, and T, describe the chemical identities and orientations of all possible nucleotide pairs. In this coarse-grained model, each letter describes an assembly of over 60 atoms. However, the atomic composition of a nucleotide pair can be altered by chemical modifications, different base-pairing geometries, or mismatches. As only a few atoms contribute to binding specificity, we propose that compared to a sequence model, a chemistry-based model that directly encodes protein–DNA contacts may more robustly capture the chemical variations of DNA. We introduce models that directly and precisely represent physicochemical readout, which is, importantly, not restricted to standard Watson–Crick base pairs.

DNA-binding proteins selectively bind to their genomic binding sites and regulate various cellular processes. This selective binding occurs when the DNA-binding domain of a gene regulatory protein, a transcription factor (TF), recognizes its binding site by reading physicochemical signatures at the base-pair (bp) edges (Fig. 1*A*). These physicochemical signatures, consisting of hydrogen bond (H-bond) acceptors, H-bond donors, methyl groups, and nonpolar hydrogen atoms, are exposed on the surface of the DNA major and minor grooves (Fig. 1*B*) and allow protein residues to form a series of chemical contacts, including H-bonds, water-mediated H-bonds, and hydrophobic interactions (1).

Structural information from TF–DNA complexes provides insight into specific mechanisms used by a TF to recognize DNA targets (2–5). One example for highly specific H-bonds occurs with arginine, which recognizes guanine through a bidentate interaction—forming two adjacent H-bonds—often contributing significantly to TF–DNA binding specificity (4, 6). In addition, protein residues can employ hydrophobic interactions to differentiate nucleotides, for example thymine versus cytosine (7–9). In some cases, structural deviations from a B-form double helix increase accessibility of DNA physicochemical signatures, enabling a TF to establish an optimized set of H-bonds or hydrophobic interactions that can determine DNA-binding specificity (10–12). Nevertheless, there is a paucity of experimentally determined TF structures in complex with DNA target sites. Available structural information for TF–DNA complexes is typically limited to complexes where a protein or its DNA-binding domain binds to a single DNA sequence.

In the last decade, several high-throughput binding technologies have been developed to enable sampling of TF–DNA binding. These technologies quantitatively measure the binding affinities of one TF against thousands or even millions of different DNA sequences in vitro (13–21). These methods provide an alternative path to infer TF–DNA binding without requiring time-consuming structural biology experiments. Concurrently, several DNA motif discovery methods have been established for modeling TF–DNA binding preferences using these high-throughput experimental data. Methods based on position weight matrices (PWMs) assume that each nucleotide at a corresponding position

Author contributions: T.P.C. and R.R. designed research; T.P.C. performed research; T.P.C. and S.R. contributed new reagents/analytic tools; T.P.C. analyzed data; and T.P.C. and R.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]Present address: Department of Biochemistry and Molecular Genetics, University of Colorado Denver School of Medicine, Aurora, CO 80045.

[2]To whom correspondence may be addressed. Email: rohs@usc.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2205796120/-/DCSupplemental.

Published January 19, 2023.

independently contributes to the overall binding affinity (22, 23). To describe interdependent contributions between nucleotides, more complex modeling methods have been developed (24–26). Nevertheless, none of these methods directly infers chemical interactions of amino acids in DNA grooves that are essential for binding specificity.

Physicochemical signatures of conventional DNA bases (Fig. 1*C*) can be altered by modifying a chemical group. These signatures constitute an important layer of reprogrammable information in DNA (Fig. 1*D*). For example, the chemical signatures of 5mCpG dinucleotides where cytosine is methylated at the 5-carbon position can be recognized by the methyl-CpG-binding domain protein that recruits histone deacetylases and promotes local chromatin condensation to regulate transcription (27). Non-Watson–Crick bp, such as Hoogsteen (6, 28) (Fig. 1*E*), synthetic (29) (Fig. 1*F*), and mismatched (30) (Fig. 1*G*) bp, represent unique physicochemical signatures in DNA. These unique bp introduce new layers of complexity and possibly influence TF–DNA recognition. For example, mismatched DNA can be recognized by a specific class of TFs that acts as a repair barrier to increase the mutation rate and thereby regulate cellular replication and repair processes (30–32). Mismatched bp have recently also been reported to play a role in the CRISPR-Cas9 gene editing system (33). Investigating how these non-standard physicochemical signatures affect binding specificity is an important step toward understanding the binding mechanisms. However, existing methods are difficult to apply to DNA modifications or non-Watson–Crick bp, due to a potential overfitting problem as a result of one-hot encoding with expanded alphabets representing such non-standard bp.

Here, we introduce DeepRec (Deep Recognition for TF–DNA binding), a deep-learning-based method that integrates two convolutional neural network (CNN) modules for extracting the pattern of physicochemical signatures in the major and minor grooves of DNA. Each CNN module extracts nonlinear spatial context among physicochemical signatures of bp to mine potential insights beyond DNA sequence. We use a grid hyperparameter search to find a combination of hyperparameters, which yields an optimal model to minimize a predefined loss function on a given dataset. To reduce the error introduced by an individual predictive model, we performed ensemble training with multiple random seeds and average the contribution of each physicochemical signature. DeepRec integrates a forward perturbation-based interpretative approach that highlights the important physicochemical signatures for deciphering binding mechanisms. This method aims to reveal important physicochemical patterns recognized by TFs and further explain biological insights that cannot be elucidated by sequence-based models. Such a chemistry-based approach is necessary, given the increasing evidence for the biological importance of various chemical modifications of DNA in gene regulation, cellular function, and disease.

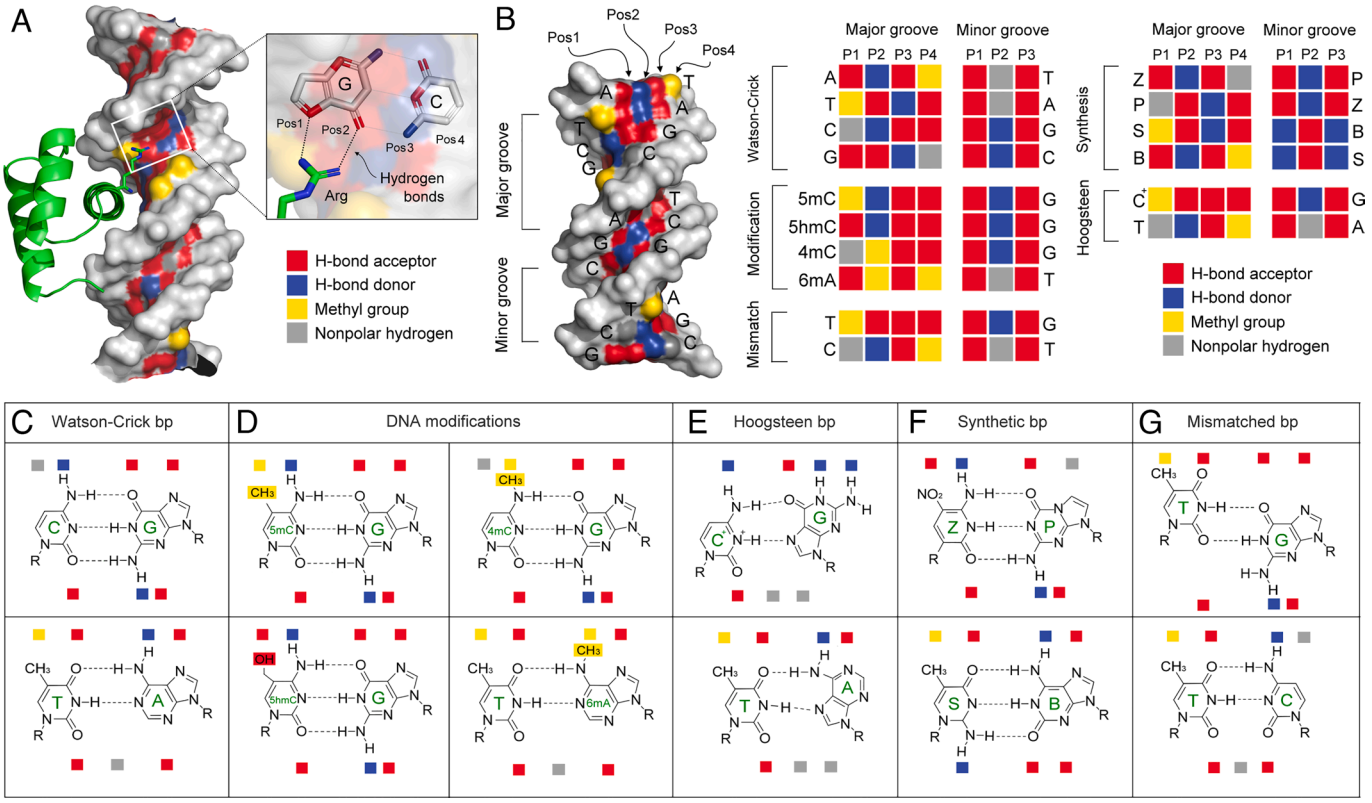

Fig. 1. Physicochemical signatures of base pairs (bp) characterized by a set of chemical groups at bp edges in the major or minor groove. (*A*) Schematic showing that the DNA binding domain of a TF recognizes DNA physicochemical signatures, including an H-bond acceptor (red), H-bond donor (blue), methyl group (yellow), and nonpolar hydrogen (gray). (*B*) Schematic view demonstrating the standard and expanded physicochemical signatures in DNA major and minor grooves. Labels 'Pos1-4' (major groove) and 'Pos1-3' (minor groove) indicate positions of physicochemical groups at various bp. In addition to (*C*) standard Watson–Crick bp signatures, additional unique signatures are introduced by (*D*) DNA modifications, (*E*) Hoogsteen bp, (*F*) synthetic nucleotides, and (*G*) mismatched bp. 5-methylcytosine (5mC)/G and 5-hydroxymethylcytosine (5hmC)/G modifications carry unique signatures in the major groove by an additional methyl or hydroxymethyl group on the 5-position of cytosine, respectively. Methylations on the 4-position of the cytosine pyrimidine ring (N4-methylcytosine, 4mC) and on the 6-position of the adenine purine ring (N6-methyladenine, 6mA) create alternate unique signatures in the major grove. Hoogsteen bp flip their purines and exhibit different physicochemical patterns in the major or minor groove. Hachimoji DNA is composed of four synthetic bases (P, Z, S, B). The Z/P bp has a unique pattern in the major groove, whereas the S/B bp has a unique pattern in the minor groove. Mismatched bp can change the number of H-bonds in the base-pairing geometry. For example, the T/G bp increases by one H-bond, and the T/C bp loses one H-bond acceptor.

Results and Discussion

Description of the DeepRec Framework. The framework of DeepRec consists of two parallel CNN modules to extract binding patterns from physicochemical signatures in the DNA major and minor groove, respectively (Fig. 2*A*). A joint layer combines patterns extracted from the two CNN modules, and a hidden layer further discovers higher-level binding patterns (Fig. 2*A*). The model is capable of characterizing the relative affinities of all binding sites by building a predictive model based on high-throughput experimental data.

We introduce a new encoding method that has a contact-based representation in the context of TF–DNA recognition in comparison to the widely adopted one-hot encoding method for DNA sequences (A, C, T, and G). Our method encodes physicochemical signatures in the major and minor grooves with respect to their defined positions into two-dimensional features or images that contain geometric relationships among physicochemical signatures (Fig. 1*B*) (1). The CNN modules keep the original structure of the input and extract the spatial context of the physicochemical signatures to identify binding patterns. This information might be neglected by sequence-based models that consider DNA as a one-dimensional string of letters, likely assuming that the features of DNA bp are independent of each other. The hidden layer models the higher-order interactions between patterns implicitly. In this way, our method is able to mine spatial and nonlinear information from the physicochemical space.

Compared to the one-hot sequence-encoding scheme, our method provides a fundamental description of the actual molecular interactions between TFs and DNA, and it retains the dependency between bp based on the physicochemical signatures at their major and minor groove edges. For example, the only difference between cytosine and 5-methylcytosine (5mC) is a methyl group at one major-groove position (Fig. 1*B*). One-hot sequence encoding assumes that all nucleotides are independent, which might lead to loss of information about nucleotide interdependence. Specifically, the dependency of unmethylated, fully methylated, and hemimethylated CpG bp steps varies in different degrees. Such variable dependency cannot be represented by independent letters in a one-hot sequence-encoding scheme.

Another advantage of our new encoding scheme is that the encoding of physicochemical signatures does not need to be expanded when introducing diverse nucleotide types, such as those from chemical DNA modifications, or from Hoogsteen (6, 28), synthetic (29), or mismatched bp (30). By contrast, the one-hot sequence-encoding scheme must expand the feature dimension, and a massive yet sparse encoding matrix is more likely to lead to an overfitting issue when modeling the data.

We use deep learning techniques to infer model parameters. Our training pipeline alleviates the need for manual parameter adjustment by automatically tuning several calibration parameters through threefold cross-validation (Fig. 2*B*). With the tuned parameters, we perform ensemble training with multiple random seeds and filter out low-performance models (Fig. 2*C*). DeepRec utilizes a forward perturbation-based approach to calculate the binding difference between the presence and absence of a chemical group at a given position, regardless of nucleotide sequence (Fig. 2*D*). To visualize the binding preferences of an individual TF, DeepRec introduces a new visualization of physicochemical signatures using physicochemical energy logos (Fig. 2*E*). The package is available on GitHub (https://github.com/TsuPeiChiu/DeepRec).

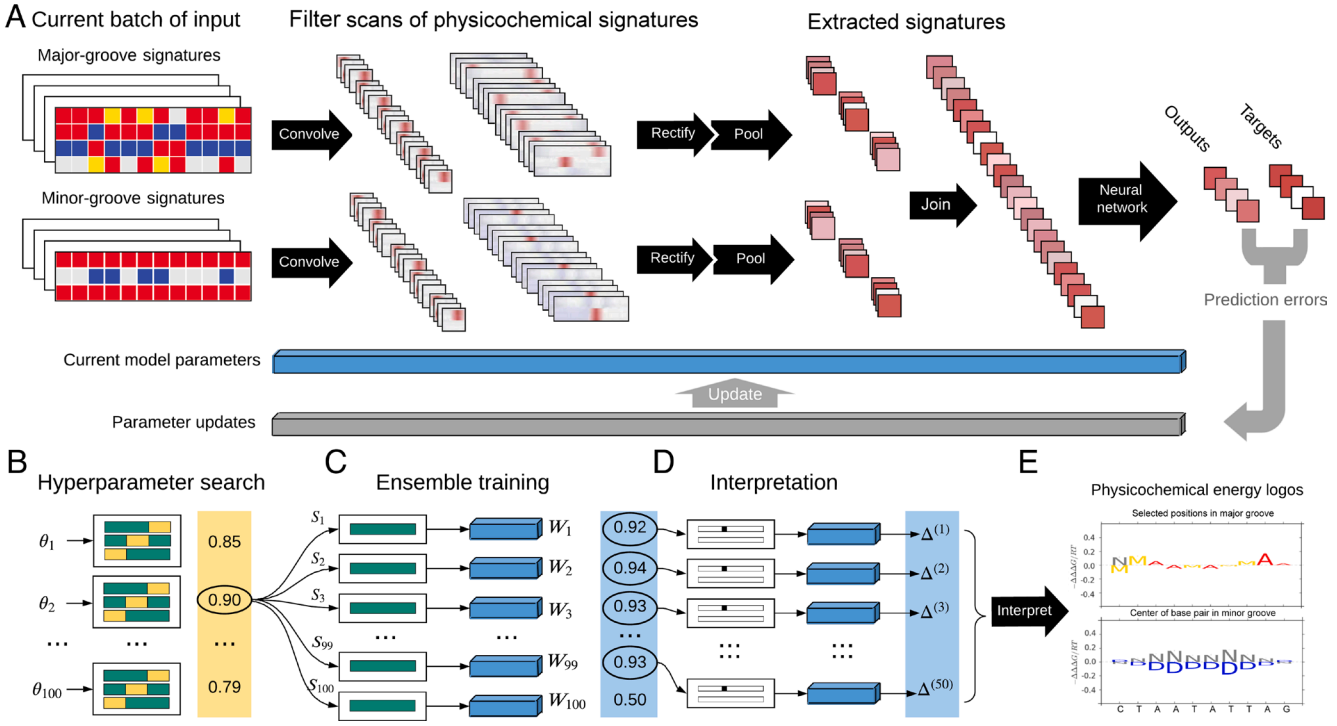

Fig. 2. Workflow of the DeepRec prediction framework. (*A*) The stages of convolution, rectifying, pooling, and neural networks predict binding affinities based on major and minor groove positions (Pos) for each input sample. During the training phase, back-propagation stages simultaneously update all trainable parameters to improve prediction accuracy. The entire process starts from (*B*) a hyperparameter search that randomly chooses 100 parameter combinations in hyperparameter space θ and performs threefold cross-validation on the training set. The parameter combination with the highest average r^2 is used for (*C*) ensemble training with 100 random seeds (S_1 to S_{100}) and results in 100 models (W_1 to W_{100}). Models of the top 50 performers in r^2 are tested on held-out validation sets and used for (*D*) interpretation. The delta value (Δ) is the gap between relative binding affinities when a specific physicochemical signature is or is not present. (*E*) The average of delta values (Δ_1 to Δ_{50}) over the 50 models is used for calculation of the physicochemical energy logos.

In this study, we demonstrated and validated our physicochemical encoding scheme of DNA for the TFs MAX, MEF2B, p53, ATF4, and C/EBPβ. We selected these TFs based on the primary consideration that they have representative binding mechanisms, and they each have available high-quality high-throughput experimental data and corresponding co-crystal structures. In terms of binding mechanisms, MAX binds to the DNA major groove with several types of H-bonding (34). MEF2B is known for its essential minor groove binding (35). p53 recognizes its DNA binding site as a protein tetramer and through an interplay of base and shape readout (6). Finally, ATF4 binds to a CpG dinucleotide through the mechanism of thymine mimicry (36). All these systems have SELEX-seq experimental data that characterize a full range of TF-binding affinities for model training and corresponding co-crystal structures in the Protein Data Bank (PDB) for validation (37).

DeepRec Predicts DNA Contacts in TF-DNA Binding. Experimentally determined structures of TF–DNA complexes provide critical insights into binding mechanisms. By observing the number and geometry of H-bonds as well as hydrophobic contacts between protein residues and bp, one can understand how proteins use their unique readout mechanisms to achieve DNA-binding specificity. However, co-crystal structures are available for relatively few TFs and are typically limited to complexes in which a protein or its DNA-binding domain binds to a single DNA sequence. Moreover, crystal-packing contacts near the binding site, variations of side-chain rotamers, and the presence of low-electron-density polar hydrogens can limit attempts to identify a binding contact. DeepRec leverages large data generated from high-throughput binding assays and enables the prediction of physicochemical readout from sequencing data. In this way, DeepRec has the potential to confirm and reveal unknown binding mechanisms without the requirement of solving a structure.

We first targeted the widely studied human helix–loop–helix (bHLH) protein MAX, which preferentially recognizes a subset of enhancer-box (E-box) sequences (38). We trained DeepRec on MAX SELEX-seq data (39), interpreted the model using DNA physicochemical energy logos (Fig. 3*A*), and compared predicted logos with sequence logos (Fig. 3*B*) and TF contacts within a MAX–DNA co-crystal structure (PDB ID 5EYO) (34) (Fig. 3 *C*–*F*). We observed two prominent H-bond acceptors (denoted as 'A's) at physicochemical signature position 3 (Pos3) and Pos4 in the major groove of the C_{-3}/G_{-3} bp (or at Pos1 and Pos2 of the G_3/C_3 bp) in the E-box of the physicochemical energy logos (Fig. 3*A*). The logos showed a positive average change in binding free energy $-\Delta\Delta\Delta G$ (see definitions in *SI Appendix*). Consistent with the co-crystal structure of the MAX–DNA complex, His28 forms one H-bond with either O_6 or N_7 of the 3′ guanine (G_{-3}), or forms a bifurcated H-bond (with the two H-bonds sharing a donor) (1) (Fig. 3*C*). We observed no clear signal in energy logos at Pos2 in the major groove of the C_{-3}/G_{-3} bp (or Pos3 of G_3/C_3) (Fig. 3*A*). In the MAX–DNA co-crystal structure (Fig. 3*C*), the H-bond acceptor of Glu32 might be occupied by donors of Arg35, preventing bonds with the donor of the 5′ cytosine (C_{-3} or C_3).

Another notable H-bond acceptor was found in the physicochemical energy logos at Pos3 in the major groove of the A_{-2}/T_{-2} bp (or Pos2 of the T_2/A_2 bp) (Fig. 3*A*). This observation is consistent with the co-crystal structure, in which Arg36 donates one or two H-bonds to O_4 of the 3′ thymine (T_{-2} or T_2) (Fig. 3*D*). One

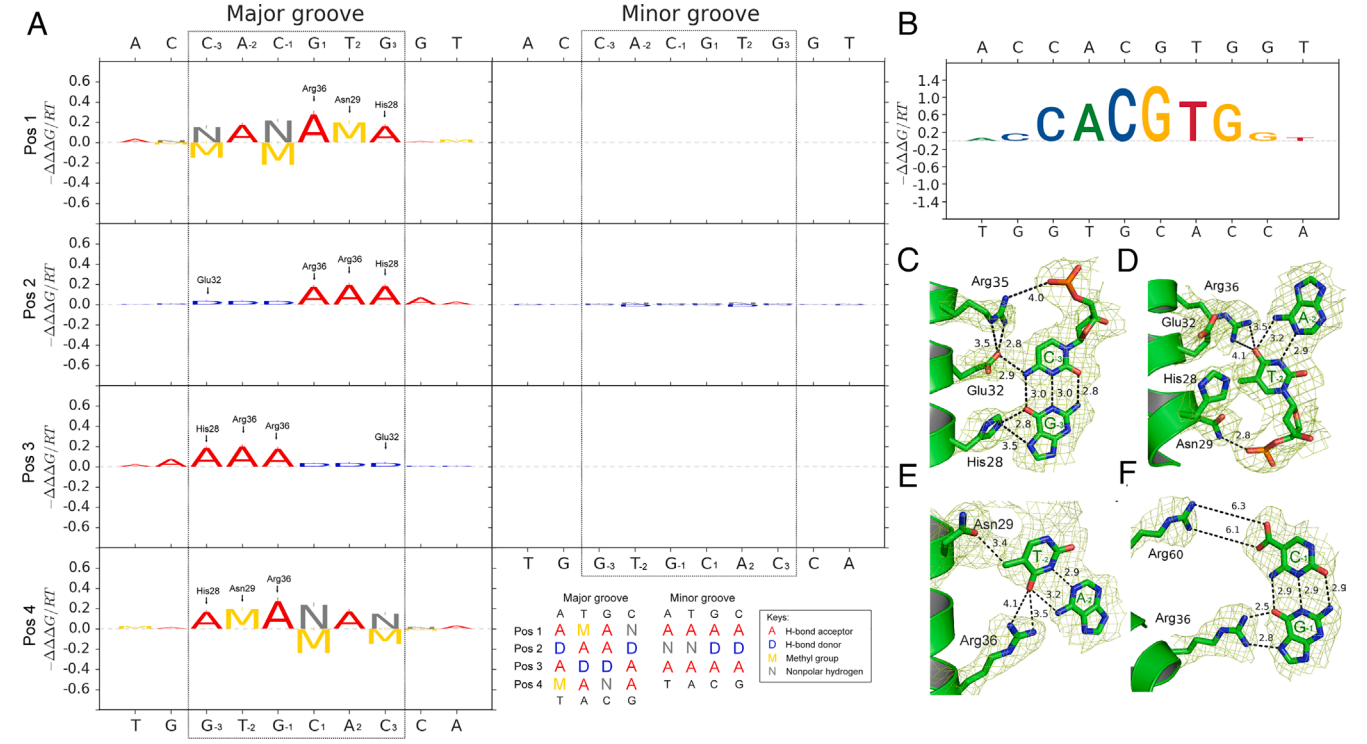

Fig. 3. DNA physicochemical energy logos and co-crystal structure of the MAX–5caC complex (PDB ID 5EYO). (*A*) Physicochemical energy logos were generated by DeepRec based on MAX SELEX-seq data. The physicochemical signature position (Pos) and the nucleotide position of the bp are indicated by the name of the MAX protein residue when a contact between the protein and DNA was found, using the DNAproDB method (26). Pos1-4 refer to physicochemical signature positions 1-4, respectively, of the bp. The dashed box highlights the MAX binding site, and the sequence on top of these panels indicates the input sequence. In the energy logos, 'M' represents the thymine methyl group, since the model was trained on unmethylated data. (*B*) Sequence logos obtained from DeepRec sequence model. (*C*) Interactions between MAX residues and the C_{-3}/G_{-3} bp in the major groove. Numbers indicate distances between atoms in Å. (*D* and *E*) Interactions between MAX residues and the A_{-2}/T_{-2} bp in the major groove. (*F*) Interactions between MAX residues and the C_{-1}/G_{-1} bp in the major groove.

notable methyl-group was found at Pos4 in the major groove of the A_{-2}/T_{-2} bp (or Pos1 of the T_2/A_2 bp). In the co-crystal structure, Asn29 formed a van der Waals (vdW) interaction with the methyl-group of the 3′ thymine (T_{-2} or T_2) (Fig. 3*E*). Two additional prominent H-bond acceptors were found at Pos3 and Pos4 in the major groove of the C_{-1}/G_{-1} bp (or Pos1 and Pos2 of the G_1/C_1 bp) (Fig. 3*A*). In the co-crystal structure (Fig. 3*F*), Arg36 interacts with the central 5′-CpG-3′ dinucleotide by donating two hydrogen atoms to the O_6 and N_7 atoms of the 3′ guanine (G_{-1} or G_1), forming a bidentate H-bond (two H-bonds with different donor and acceptor atoms). This geometry conveys a high degree of specificity (40). Similar results were obtained when we cross-validated our method with data from the microfluidics-based SMiLE-seq platform (21) (*SI Appendix*, Fig. S1). To further prove that our method is generalizable across different experimental platforms, we predicted relative binding affinities measured by SMiLE-seq with the model trained by the SELEX-seq data. The results showed a good correlation (*SI Appendix*, Fig. S2). As a control, we shuffled the relative binding affinities of the training data, and no signals were detected from the predictive model (*SI Appendix*, Fig. S3). Together, these results showed that DeepRec is capable of predicting binding mechanisms with respect to the co-crystal structure.

DeepRec Predicts TF Binding Preference in the Context of DNA Grooves. Physicochemical signatures in the DNA major groove are more diverse and unique than those in the minor groove, thereby conveying a higher degree of TF-binding specificity (1). For example, MAX only recognizes signatures in the major groove to achieve its binding specificity, as evidenced by the physicochemical energy logos (Fig. 3*A*) and the MAX–DNA co-crystal structure (PDB ID 5EYO). Nevertheless, TF-binding specificity can be achieved through a complex recognition process involving both major and minor groove readout (1). Therefore, we asked whether DeepRec is able to predict binding mechanisms that involve physicochemical signatures in the major and minor grooves.

We next studied myocyte enhancer factor 2B (MEF2B), a member of the MEF2 family that plays vital roles in the development and functioning of neuronal and muscle cells. Minor groove contacts have been shown to be important for TF binding for this family (41). We trained DeepRec on MEF2B SELEX-seq data (41), interpreted the model using DNA physicochemical energy logos (Fig. 4*A*), and compared predicted logos with the sequence logos (Fig. 4*B*) and the TF–DNA contacts of a MEF2B–DNA co-crystal structure (PDB ID 1N6J) (42) (Fig. 4 *C*–*F*). One prominent H-bond acceptor at Pos4 in the major groove of the T_{-4}/A_{-4} bp (or Pos1 in the major groove of the A_4/T_4 bp) was observed in the physicochemical energy logos (Fig. 4*A*). Consistent with the MEF2B–DNA co-crystal structure, Lys23 donates one H-bond to N_7 of the 3′ adenine (A_{-4} or A_4) (Fig. 4*D*). Intriguingly, the major groove signals mainly occurred at the 3-bp half sites (5′-CTAW$_4$TAG-3′). In contrast, notable minor groove signals were observed within the central W_4 region of the binding site (Fig. 4 *A*, *Right*).

Compared to the physicochemical energy logos of MAX (Fig. 3 *A*, *Right*), MEF2B logos showed stronger signals in the minor groove, with positive nonpolar hydrogens ('N's) and negative H-bond donors ('D's) (Fig. 4 *A*, *Right*). In the MEF2B–DNA co-crystal structure (Fig. 4*E*), the Gly2-Arg3 conformation inserts into the minor groove: Arg3 makes electrostatic interactions with the phosphodiester backbone, whereas Gly2 makes vdW or hydrophobic interactions with nonpolar atoms. The energy logos demonstrated long-range interactions in the minor

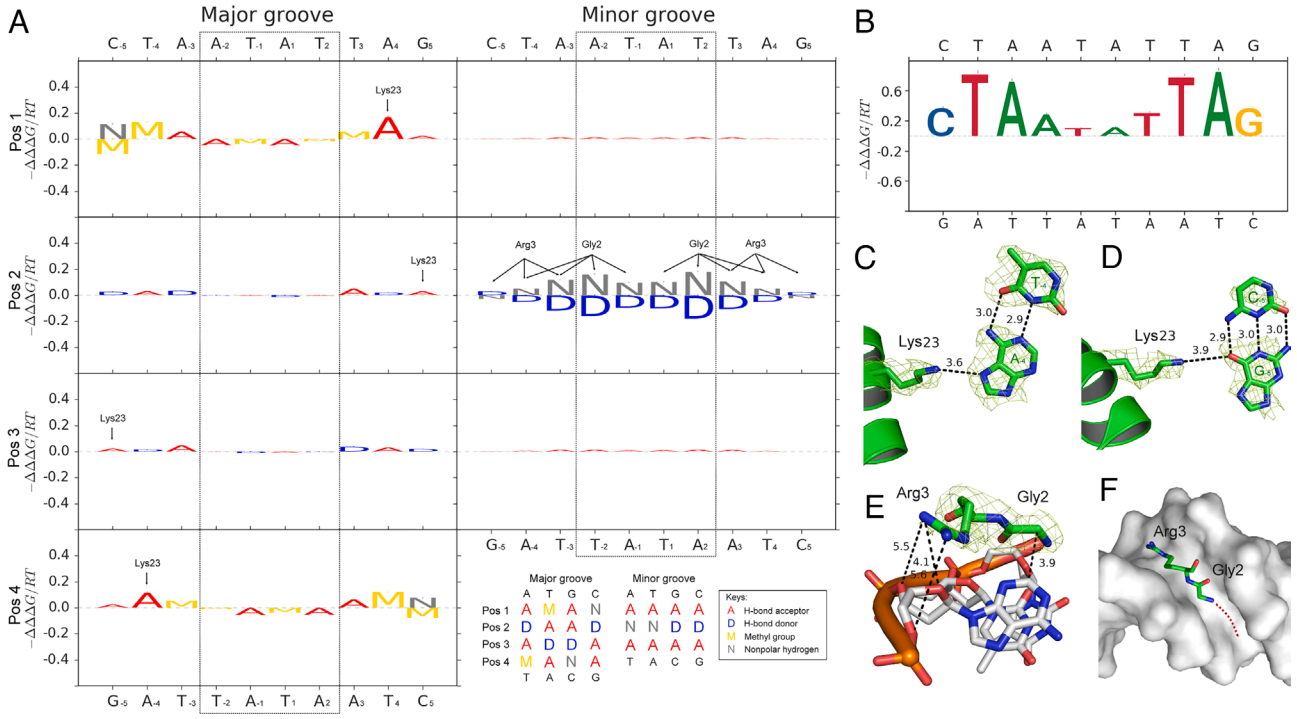

Fig. 4. DNA physicochemical energy logos and co-crystal structure of MEF2B–DNA complex (PDB ID 1N6J). (*A*) Physicochemical energy logos were generated by DeepRec based on MEF2B SELEX-seq data. The position with respect to the physicochemical signature position (Pos) and the nucleotide position of the bp are indicated by the name of the MEF2B protein residue when a contact between the protein and DNA was found, using the DNAproDB method (26). Pos1-4 refer to physicochemical signature positions 1-4, respectively, of the bp. The dashed box highlights the central core-binding site. (*B*) Sequence logos obtained from DeepRec sequence model. (*B*) Interactions between MEF2B residues and the C_{-5}/G_{-5} bp in the major groove. Numbers indicate distances between atoms in Å. (*C*) Interactions between MEF2B residues and T_{-4}/A_{-4} bp in the major groove. (*D*) Interactions between MEF2B residues and the T_{-5}/A_{-5} bp in the major groove. (*E*) Interactions between MEF2B residues and atoms in the minor groove. (*F*) "Met-Gly-Arg" conformation in the minor groove.

groove (Fig. 4 *A*, *Right*), suggesting the potential presence of another nonpolar residue such as Met, which can form highly specific interactions for minor-groove recognition (Fig. 4*F*).

DeepRec Predicts the Geometry of H-Bonding and Hydrophobic Interactions. DNA-binding specificity is achieved through H-bond contacts and hydrophobic interactions. While the number of contacts formed between protein residues and DNA bases provides binding specificity, the uniqueness of the geometry of H-bonds and the hydrophobic interaction conveys a higher degree of specificity for protein recognition (1). For example, the H-bond geometry of bidentate and bifurcated H-bonds provides a higher degree of binding specificity than single H-bonds (43). In a complex example, bidentate H-bonds combined with a hydrophobic interaction form a triad geometry that conveys an even higher degree in binding specificity. For instance, a methyl-Arg-G triad forms during recognition of TpG and methyl-CpG dinucleotides in double-stranded DNA (44).

The human tumor suppressor p53 binds as a tetramer to two dimeric sites (5′- RRRCWWGYYY -3′) separated by a spacer of 0-13 bp (45). Each p53 monomer uses a methyl-Arg-G triad to recognize the TpG dinucleotide (44). We trained DeepRec on p53 SELEX-seq data (39), interpreted the model using DNA physicochemical energy logos (Fig. 5*A*), and compared predicted logos with the sequence logos and TF–DNA contacts of the p53–DNA co-crystal structure (PDB ID 3Q06) (1) (Fig. 5 *B*–*D*). We observed two prominent H-bond acceptors at Pos1 and Pos2 in the major groove of the G_2/C_2 bp in the physicochemical energy logos (Fig. 5*A*). This observation is consistent with the p53–DNA co-crystal structure, in which Arg280 donates two H-donors to N_6 and O_7 of the 3′ guanine, forming bidentate H-bonds (Fig. 5*D*).

On the other hand, a prominent methyl group was found at Pos1 in the major groove of the T_1/A_1 bp (Fig. 5*A*), in agreement with the co-crystal structure wherein the methyl group of thymine stabilizes the Arg280 positioning through a vdW contact (Fig. 5*D*). The bidentate H-bond and the hydrophobic interaction formed a so-called "methyl-Arg-G triad," providing unique specificity. Intriguingly, three consecutive 'A's on the side are found at Pos1 in the major groove of G_{-5}, G_{-4}, and A_{-3}, near H-donor Arg280 and Lys120, suggesting why p53 prefers the RRR triplet at the binding-site edges (Fig. 5*C*). The results showed that DeepRec is capable of detecting the binding geometry for p53–DNA recognition.

DeepRec Predicts Impact of DNA Modifications on Protein–DNA Binding. DNA modifications play key roles in gene regulation (36, 46), but their effects on TF-binding are not completely understood. Several studies have adopted structural biology and low-throughput binding methods (e.g., X-ray crystallography and electrophoretic mobility shift assays) on TFs in complexes with modified DNA to explain the effects of chemical modifications at atomic resolution (34, 47). Recently, high-throughput methods, such as EpiSELEX-seq (36), methyl-HT-SELEX (46), and methyl-Spec-seq (48), have been developed to determine the effects of the CpG modification on TF-binding using a large pool of

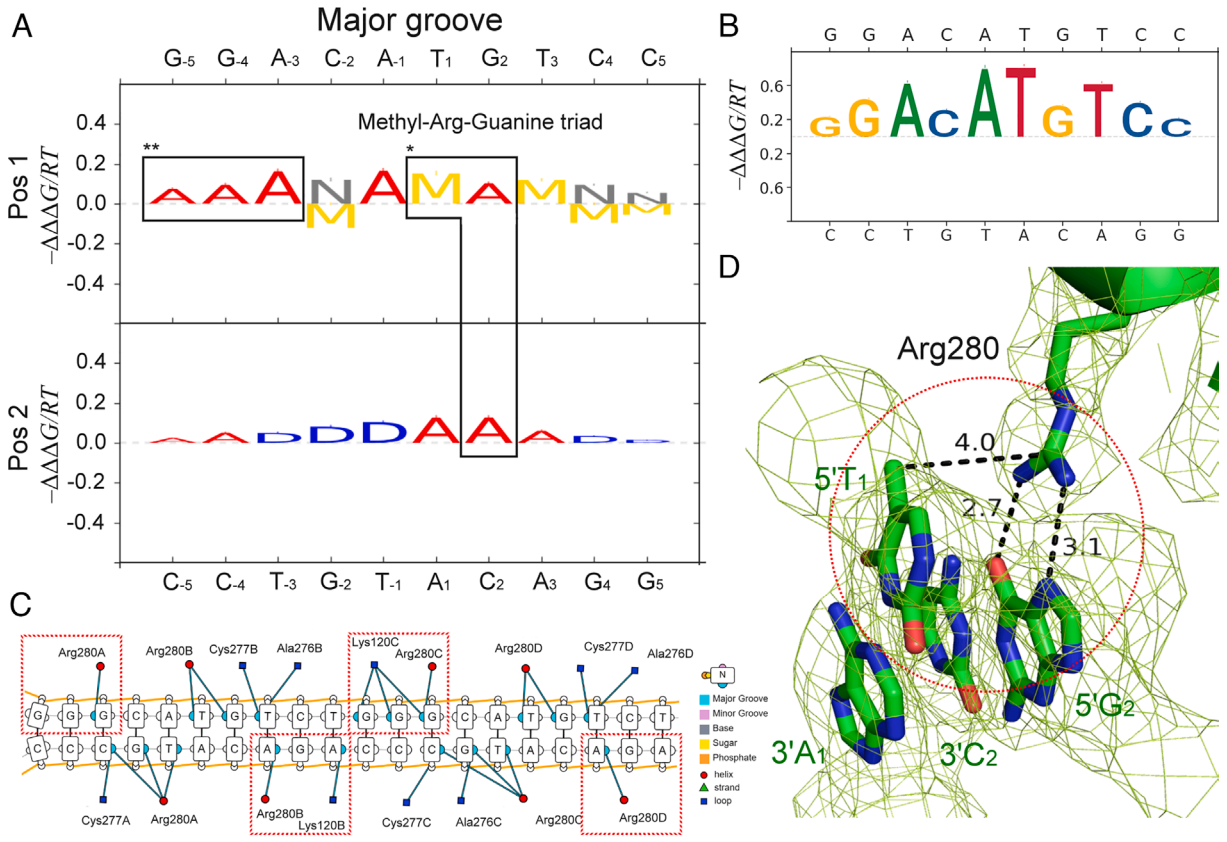

Fig. 5. Selected DNA physicochemical energy logos and co-crystal structure of p53–DNA complex structure (PDB ID 3Q06). (*A*) Preferred physicochemical signature position (Pos) indicates preference for the thymine methyl group and two guanine acceptors in the major groove of the p53 response element. Pos1-2 refer to physicochemical signature positions 1-2, respectively, of the bp. Positions marked with * represent Arg280 contacts corresponding to (*D*) the p53–DNA co-crystal structure. (*B*) Sequence logos obtained from DeepRec sequence model. (*C*) TF–DNA contacts in red dashed-boxes show preference of purine bases due to H-bonds between the acceptors and the donors provided from arginine and lysine. Contact map is obtained from DNAproDB. Contact map is consistent with the geometrical pattern shown in positions marked with ** in panel (*A*).

chemically modified oligonucleotides. These high-throughput methods are powerful tools but lack insights into structural readout mechanisms.

Here, we asked whether we could predict the effects of methylation on TF binding using high-throughput binding data and explain possible binding mechanisms. We trained DeepRec on EpiSELEX-seq data (36) of the human bZIP proteins ATF4 and C/EBPβ using a similar process as was used in previous systems. In this case, we combined methylated (Lib-M) and unmethylated (Lib-U) DNA fragments generated from EpiSELEX-seq (36) as input. We modeled the systems without expanding the sequence alphabets by instead describing physicochemical signatures (Fig. 1 *B* and *D*). Predictions for ATF4 showed a decrease in binding affinity when a CpG bp step is present at the center of the binding site (at position −1/+1), and an increase in binding affinity when the sequences contained a CpG dinucleotide at the flank of the motif (at position −4/−3 or +3/+4) (Fig. 6*A* and *SI Appendix*, Fig. S3*A*). In contrast, predictions for C/EBPβ showed a weak or no preference for 5mCpG at the center of the binding site (Fig. 6*B* and *SI Appendix*, Fig. S3*B*) (49). The methylation effect was precisely identified in low-affinity binding sequences for ATF4 (Fig. 6*A* and *SI Appendix*, Fig. S3*E*). These results are consistent with results from previous studies (*SI Appendix*, Fig. S3 *C–F*) (36, 46, 49, 50).

Next, we trained DeepRec on SELEX-seq data (39) for ATF4 and C/EBPβ. We asked whether we could learn the methylation effect from unmethylated DNA. Predictions for ATF4 showed a positive effect on binding affinity when the sequences contained a CpG bp step in the flank, demonstrating the 'thymine mimicry' that could possibly be learned from the methyl group of thymine

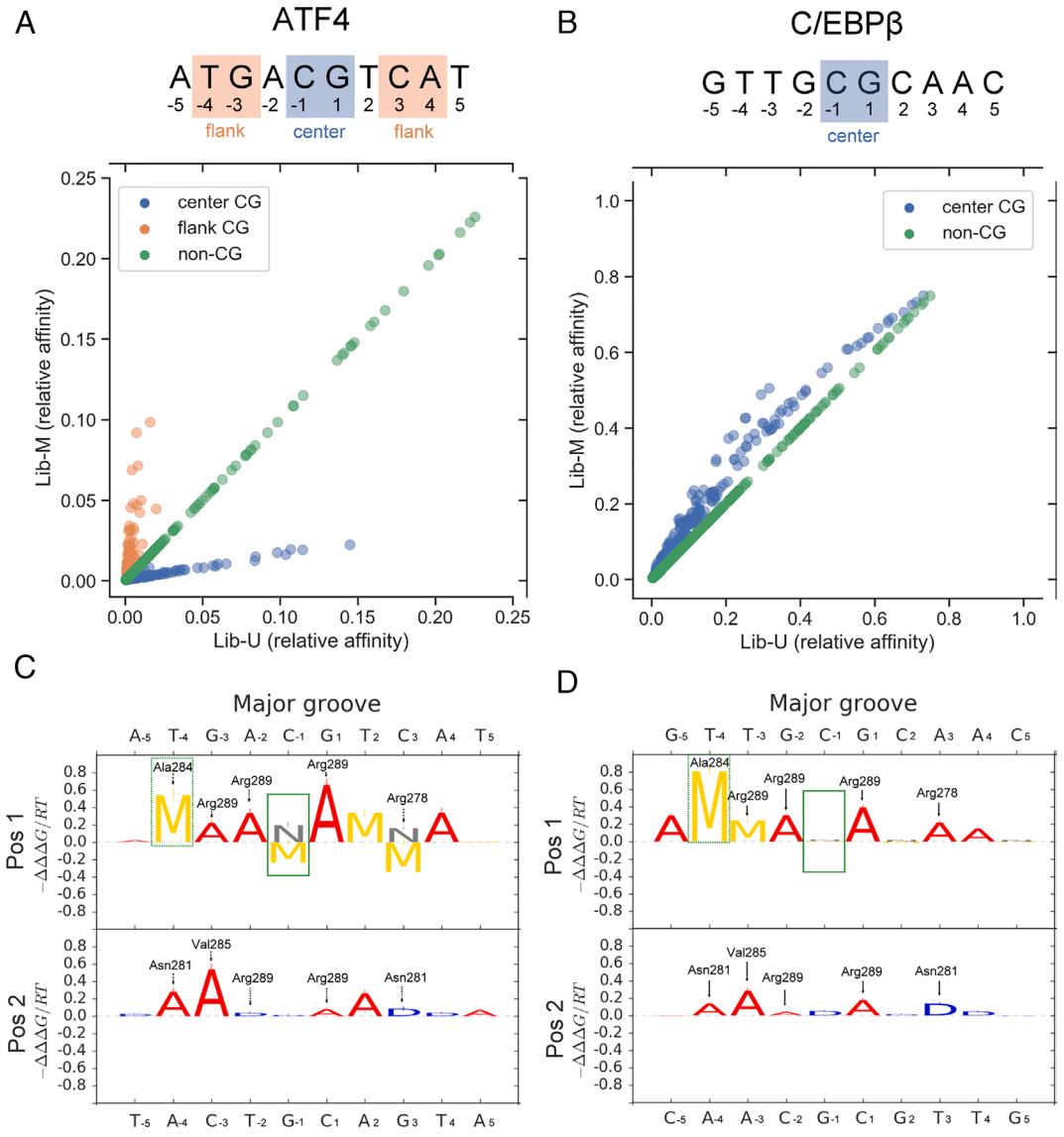

Fig. 6. Predicting and measuring methylation sensitivity for ATF4 and C/EBPβ using DeepRec on EpiSELEX data and validating the contacts with C/EBPβ complex structure (PDB ID 6MG1 and 2E42). Symmetric consensus motifs and comparison of relative affinities of 10-mer sequences between a methylated sequence library (Lib-M) versus an unmethylated sequence library (Lib-U) predicted by DeepRec are shown for (*A*) ATF4 and (*B*) C/EBPβ. For ATF4 (*A*), non-CpG sequences (green) show the same binding affinities in both libraries because no methylation is involved. Flanking CpG-containing sequences are preferred in Lib-M, whereas central CpG-containing sequences are preferred in Lib-U. For C/EBPβ (*B*), central CpG-containing sequences are slightly preferred in Lib-M. DNA physicochemical energy logos are shown for (*C*) ATF4 and (*D*) C/EBPβ. Corresponding contacts from co-crystal structures (PDB ID 6MG1) are indicated by residue name and number. Due to the unavailability of a ATF4 co-crystal structure, the C/EBPβ co-crystal structure was compared between ATF4 logos and contacts marked by a dashed line. Green dashed-boxes highlight hydrophobic interactions between the methyl group of the thymine and the hydrophobic side chains that are shared by ATF4 and C/EBPβ. Green solid boxes indicate a negative effect of methylation on central CpG for ATF4 but a slight positive effect for C/EBPβ.

in the unmethylated data (*SI Appendix*, Fig. S3 *G* and *I*). However, the model trained for C/EBPβ did not capture the positive contribution toward binding from the thymine in the central TpG dinucleotide (*SI Appendix*, Fig. S3 *H* and *J*), which implied that the thymine mimicry event might not be selected in this case. This comparison allowed us to study the possible mechanisms of the methylation effect on binding.

The resulting physicochemical logos supported the previously identified molecular basis for methylation effects on binding (36). The predictions showed a positive 'M' at Pos1 in the major groove of T_{-4}/A_{-4} (Pos4 in the major groove of A_4/T_4) for ATF4 (green dashed boxes in Fig. 6 *C* and *D*), demonstrating the positive effect of methylation on binding affinity. Based on TF–DNA contacts in a C/EBPβ co-crystal structure (PDB ID 6MG2) (51), Ala284 interacts with the methyl group of a thymine at position −4 of the binding site, making a vdW contact Fig. 6*E*. A negative 'M' is found at Pos1 in the major groove of C_{-1}/G_{-1} (Pos4 in the major groove of G_1/C_1) for ATF4 (Fig. 6*C*), showing a negative effect on binding affinity. However, there is no obvious signal of 'M' detected at Pos1 in the major groove of C_{-1}/G_{-1} (or at Pos4 in the major groove of G_1/C_1) for C/EBPβ (Fig. 6*D*). In the co-crystal structure Fig. 6*H*, the 'M' of an 5mC might interact with the carbon Arg289 to form a vdW interaction, which may explain how C/EBPβ interacts with methylated DNA Fig. 6*H*. However, why ATF4 has less preference for methylated DNA is still unknown.

Conclusions

Investigating physicochemical readout signatures in DNA that are important for protein binding is an approach to uncover TF–DNA readout mechanisms. However, current experimental methods are limited in their ability to provide sufficient numbers of structures to explain the entirety of TF–DNA binding mechanisms. Until now, existing methods have been incapable of mining important physicochemical signatures by leveraging large data. Here, we describe DeepRec, a deep learning framework capable of mining the importance of physicochemical readout signatures for TF–DNA binding specificity in the context of binding free energy. DeepRec accurately predicts possible binding contacts for several TFs across TF families at physicochemical resolution, and indicates corresponding forces (e.g., H-bonds, hydrophobic interactions, etc.) that current sequence-based modeling methods cannot identify. Furthermore, DeepRec detects physicochemical signatures and binding geometries that can contribute to highly specific TF–DNA binding, such as bifurcated and bidentate H-bonds, methyl-Arg-G triads, and long-range-DNA minor groove interaction patterns.

Strikingly, DeepRec is capable of revealing effects of chemical DNA modifications on learning on a dataset that combines methylated and unmethylated data. Compared with results trained on unmethylated data alone, one can imply the possible binding mechanism of a methylation event. Because DeepRec can mine lesser coarse-grained information than DNA sequence, we envision that this method can be easily expanded to studies beyond the concept of DNA sequence towards physicochemical modeling. For example, DeepRec might be employed to investigate effects of Hoogsteen bp (52) observed in p53 (6), as well as effects of synthetic bp (29), mismatched bp (30), or other modified bp, such as 4mC and 6mA (53), in the context of TF–DNA binding.

Future work can improve the DeepRec approach. First, we currently limit the number of physicochemical signature positions in major and minor grooves to four and three, respectively; however, in some cases this definition might be vague. For example, the number of physicochemical signature positions in the minor groove could be two or three. In addition, we encode physicochemical signatures as discrete data and consider O and N as the same type of physicochemical signatures with equal physicochemical property strength. However, considering the strength of the acceptor from a hydroxymethyl group, O, and N would be different. To address this issue, we should introduce a function to describe the distribution of strength for different chemical groups rather than use a discrete one-hot encoding representation. Furthermore, DeepRec requires large data containing enough variance to be able to train thousands of parameters. Some data, such as HT-SELEX, may not be ideal for our method. Compared to our end-to-end method, a pipeline training from a sequence model or a conversion from PWM to obtain physicochemical logos might neglect the important spatial context of physicochemical signatures. The content-switch process would introduce more arbitrary parameters. DeepRec trains models with shape features implicitly; therefore, the impact of three-dimensional structure could not be highlighted by our method. Finally, DeepRec might not be able to predict all contacts occurring in the co-crystal structure, or vice versa. To improve the prediction, we could co-train our model with existing crystal structures to fine-tune the predictions.

Materials and Methods

Deep Learning Framework DeepRec. Deep neural networks are a type of artificial neural networks comprised of multiple layers between input and output layers. Each layer consists of a number of neurons, which receive input from a set of previous layer neurons. This sequential layer-by-layer structure executes a sequence of functional transformations to model complex nonlinear relationships between predictive features and response variables.

We developed a multimodule deep-learning framework capable of mining important patterns in multimodal systems. The architecture includes convolutional and down-sampling layers for each module to extract features from input data, a joint layer combining features retrieved from different sources, and a hidden layer that further integrates features to discover higher-level patterns. Based on this framework, we developed DeepRec (Deep Recognition of TF–DNA binding), which integrates physicochemical features of DNA in the major and minor grooves, followed by a perturbation-based forward-propagation approach to interpret the resulting model (Fig. 2). This method aims to discover important physicochemical readout signatures recognized by TFs and to explain biological insights that cannot be revealed by sequence-based models.

DNA Physicochemical Signatures and Feature Encoding. Physicochemical signatures at the edges of bp in the DNA major or minor grooves underlie the ability of TFs to recognize bp through H-bonds or hydrophobic contacts, as shown in Fig. 1. For a given DNA sequence, we encode the corresponding physicochemical signatures using a binary representation for H-bond acceptor, H-bond donor, thymine methyl group, and nonpolar hydrogen. A detailed description about the encoding method is provided in *SI Appendix*. The encoding method can be extended to non-Watson–Crick bp, including Hoogsteen, synthetic, and mismatched bp, without increasing the feature dimension. In contrast, the sequence-based model introduces entirely new letters of the sequence alphabet when using the one-hot encoding method, which would increase the dimension of input features by making them sparse, which might result in an overfitting issue. Using a different letter also implies independence, for instance of a methylated cytosine from cytosine despite the largely overlapping chemical characteristics of C/G and 5mC/G bp.

Hyperparameter Search. The hyperparameter search begins by sampling 100 sets of random calibration parameters. *SI Appendix*, Table S1 lists the sampling used for each parameter in a set. The calibration phase evaluates the quality of each parameter set by threefold cross-validation on the training set. Each model is trained on a different two-thirds of the data, and its performance is evaluated on the held-out one-third. Calibration parameters are scored by averaging the three r^2 values of the validation dataset (Fig. 2*B*).

Ensemble Training. Once the best calibration parameters have been identified, we trained 100 new models using 80% of the training data (a single fold) with different random seeds. Resulting models are stored and used to reduce the variance of predictions and generalization errors that generally happen in neural networks. The best 0.5 quantile of models is selected and is returned by the entire pipeline for interpretation or further analysis.

Model Interpretation. DeepRec utilizes a perturbation-based forward propagation approach that nullifies a physicochemical signature at each defined position of a bp, one at a time, and then quantifies its impact on binding free energy. The binding free-energy difference is calculated between the presence and absence of the signature at each corresponding physicochemical signature position. To visualize the detailed binding preferences of an individual TF, DeepRec introduces a new visualization, coined the 'DNA physicochemical energy logo' (e.g., Fig. 3*A*). In these logos, letters are used to represent DNA physicochemical features ('A' for H-bond acceptor, 'D' for H-bond donor, 'M' for methyl group, and 'N' for nonpolar hydrogen). The logos describe the binding preference in each DNA groove (major and minor) and physicochemical group position 1-4 (Pos1-4) at single-nucleotide resolution. The height of each letter indicates the average change in binding free energy ($-\Delta\Delta G$) resulting from the comparison of the reference probe to its mutants with a nullified physicochemical signature in ensemble models. The vertical bar on each letter specifies the standard error of the mean, which measures how far the sample mean is likely to be separated from the true population mean.

This method is expanded to handle DNA modifications by considering the addition or replacement of a specific physicochemical signature. For example, the 'N' signature at Pos1 of the C/T bp can be swapped with an 'M' signature in the major groove to represent methylation at position 5 of the cytosine, 5mC. Next, the change in binding free energy can be measured by comparing the reference probe with nullification of the physicochemical signature.

Data, Materials, and Software Availability. Previously published data were used for this work (PRJEB25690; SRP073361; GSE116401; GSE98652).

ACKNOWLEDGMENTS. We thank all current members of the Rohs laboratory for valuable input. This work was supported by the NIH (grant R35GM130376 to R.R.) and the Human Frontier Science Program (grant RGP0021/2018 to R.R.).

Author affiliations: [a]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089; [b]Department of Chemistry, University of Southern California, Los Angeles, CA 90089; [c]Department of Physics and Astronomy, University of Southern California, Los Angeles, CA 90089; and [d]Department of Computer Science, University of Southern California, Los Angeles, CA 90089

1. R. Rohs *et al.*, Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
2. J. M. Sagendorf, N. Markarian, H. M. Berman, R. Rohs, DNAproDB: An expanded database and web-based tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res.* **48**, D277–D287 (2020).
3. R. Joshi *et al.*, Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131**, 530–543 (2007).
4. C. W. Garvie, C. Wolberger, Recognition of specific DNA sequences. *Mol. Cell* **8**, 937–946 (2001).
5. N. M. Luscombe, S. E. Austin, H. M. Berman, J. M. Thornton, An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**, REVIEWS001 (2000).
6. M. Kitayner *et al.*, Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.* **17**, 423–429 (2010).
7. S. C. Harrison, A. K. Aggarwal, DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* **59**, 933–969 (1990).
8. A. K. Aggarwal, D. W. Rodgers, M. Drottar, M. Ptashne, S. C. Harrison, Recognition of a DNA operator by the repressor of phage 434: A view at high resolution. *Science* **242**, 899–907 (1988).
9. C. Wolberger, Y. C. Dong, M. Ptashne, S. C. Harrison, Structure of a phage 434 Cro/DNA complex. *Nature* **335**, 789–795 (1988).
10. R. S. Hegde, S. R. Grossman, L. A. Laimins, P. B. Sigler, Crystal structure at 1.7 A of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **359**, 505–512 (1992).
11. Y. Kim, J. H. Geiger, S. Hahn, P. B. Sigler, Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512–520 (1993).
12. J. L. Kim, D. B. Nikolov, S. K. Burley, Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365**, 520–527 (1993).
13. Y. Zhao, D. Granas, G. D. Stormo, Inferring binding energies from selected binding sites. *PLoS Comput. Bio.l* **5**, e1000590 (2009).
14. A. Jolma *et al.*, Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
15. M. Slattery *et al.*, Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270–1282 (2011).
16. M. F. Berger *et al.*, Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
17. C. L. Warren *et al.*, Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 867–872 (2006).
18. R. Gordân *et al.*, Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* **3**, 1093–1104 (2013).
19. R. Nutiu *et al.*, Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
20. D. D. Le *et al.*, Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3702–E3711 (2018).
21. A. Isakova *et al.*, SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods* **14**, 316–322 (2017).
22. G. D. Stormo, DNA binding sites: Representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
23. G. D. Stormo, Y. Zhao, Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.* **11**, 751–760 (2010).
24. M. T. Weirauch *et al.*, Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**, 126–134 (2013).
25. T. Zhou *et al.*, Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4654–4659 (2015).
26. B. Alipanahi, A. Delong, M. T. Weirauch, B. J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
27. D. Schübeler, Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
28. D. Golovenko *et al.*, New insights into the role of DNA shape on its recognition by p53 proteins. *Structure* **26**, 1237–1250.e6 (2018).
29. S. Hoshika *et al.*, Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science* **363**, 884–887 (2019).
30. A. Afek *et al.*, DNA mismatches reveal conformational penalties in protein-DNA recognition. *Nature* **587**, 291–296 (2020).
31. M. A. M. Reijns *et al.*, Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502–506 (2015).
32. R. Sabarinathan, L. Mularoni, J. Deu-Pons, A. Gonzalez-Perez, N. López-Bigas, Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
33. M. Pacesa *et al.*, Structural basis for Cas9 off-target activity. *Cell* **185**, 4067–4081.e21 (2022).
34. D. Wang *et al.*, MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic Acids Res.* **45**, 2396–2407 (2017).
35. X. Lei *et al.*, Crystal structure of apo MEF2B reveals new insights in DNA binding and cofactor interaction. *Biochemistry* **57**, 4047–4051 (2018).
36. J. F. Kribelbauer *et al.*, Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. *Cell Rep.* **19**, 2383–2395 (2017).
37. H. M. Berman *et al.*, The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
38. P. Brownlie *et al.*, The crystal structure of an intact human Max-DNA complex: New insights into mechanisms of transcriptional control. *Structure* **5**, 509–520 (1997).
39. C. Rastogi *et al.*, Accurate and sensitive quantification of protein-DNA binding affinity. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3692–E3701 (2018).
40. N. M. Luscombe, R. A. Laskowski, J. M. Thornton, Amino acid-base interactions: A three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29**, 2860–2874 (2001).
41. A. C. Dantas Machado *et al.*, Landscape of DNA binding signatures of myocyte enhancer factor-2B reveals a unique interplay of base and shape readout. *Nucleic Acids Res.* **48**, 8529–8544 (2020).
42. A. Han *et al.*, Sequence-specific recruitment of transcriptional co-repressor Cabin1 by myocyte enhancer factor-2. *Nature* **422**, 730–734 (2003).
43. S. A. Coulocheri, D. G. Pigis, K. A. Papavassiliou, A. G. Papavassiliou, Hydrogen bonds in protein-DNA complexes: Where geometry meets plasticity. *Biochimie* **89**, 1291–1303 (2007).
44. Y. Liu, X. Zhang, R. M. Blumenthal, X. Cheng, A common mode of recognition for methylated CpG. *Trends Biochem. Sci.* **38**, 177–183 (2013).
45. W. S. El-Deiry, S. E. Kern, J. A. Pietenpol, K. W. Kinzler, B. Vogelstein, Definition of a consensus binding site for p53. *Nat. Genet.* **1**, 45–49 (1992).
46. Y. Yin *et al.*, Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
47. X. Liu, E. R. Weikum, D. Tilo, C. Vinson, E. A. Ortlund, Structural basis for glucocorticoid receptor recognition of both unmodified and methylated binding sites, precursors of a modern recognition element. *Nucleic Acids Res.* **49**, 8923–8933 (2021).
48. Z. Zuo, B. Roy, Y. K. Chang, D. Granas, G. D. Stormo, Measuring quantitative effects of methylation on transcription factor-DNA binding affinity. *Sci. Adv.* **3**, eaao1799 (2017).
49. R. Chatterjee *et al.*, High-resolution genome-wide DNA methylation maps of mouse primary female dermal fibroblasts and keratinocytes. *Epigenetics chromatin* **7**, 35 (2014).
50. H. Zhu, G. Wang, J. Qian, Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).
51. J. Yang *et al.*, Structural basis for effects of CpA modifications on C/EBPβ binding of DNA. *Nucleic Acids Res.* **47**, 1774–1785 (2019).
52. K. Hoogsteen, Crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.* **16**, 907–916 (1963).
53. E.-A. Raiber, R. Hardisty, P. van Delft, S. Balasubramanian, Mapping and elucidating the function of modified bases in DNA. *Nat. Rev. Chem.* **1**, 0069 (2017).