SUPPORTING INFORMATION

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]

[a] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

[b] Department of Chemistry, University of Southern California, Los Angeles, CA 90089, USA

[c] Department of Physics and Astronomy, University of Southern California, Los Angeles, CA 90089, USA

[d] Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA

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

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

This Supporting Information includes:

Supporting Materials and Methods

Data explanation and preprocessing

Unmethylated DNA

In the last decade, many high-throughput (HT) methods have been developed to quantify transcription factor (TF)-DNA binding. These methods include electrophoretic mobility shift assay (EMSA)-based SELEX-seq (1, 2), bead-based HT-SELEX (3, 4), genomic-context protein-binding microarray(gcPBM) (1), and microfluidics-based SMiLE-seq (5) platforms. Owing to the capability to quantify a wide range of binding affinities and thus fully characterize TF-binding specificities (6), in this study, we trained our DeepRec method mainly on SELEX-seq data for TFs, including the Max homodimer (MAX), myocyte enhancer factor-2B (MEF2B), tumor suppressor protein (p53), activating transcription factor 4 (ATF4), and CCAAT/enhancer-binding protein beta (C/EBPβ). To assess the accuracy of DeepRec predictions, for MAX, we constructed DeepRec on microfluidics-based SMiLE-seq for cross-platform validation.

In this study, sequencing data for MAX derived from the SELEX-seq platform were downloaded from the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/) under study identifier PRJEB25690. Data for MAX derived from the SMiLE-seq platform were downloaded from the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP073361. Data for MEF2B derived from the SELEX-seq platform were downloaded from the Gene Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE116401. Data for p53, ATF4, and C/EBPβ derived from the SELEX-seq platform were downloaded from ENA (https://www.ebi.ac.uk/ena/) under study identifier PRJEB25690.

DeepRec models the relative binding affinities for the first-round (R1) SELEX-seq data. The R1 data covers more optimal information than later selection rounds of the data, which suppresses the low-affinity ligands and includes more experimental errors from the application process. The zero-round (R0) SELEX-seq data are a set of reads sequenced from a random library of DNA ligands. Sequencing data were pre-processed to trim the 3' ends containing adapters and indexing regions. A fifth-order Markov model was generated based on R0 data. Relative binding affinities for oligomers of various length *k* with a defined threshold count were estimated, and the information gain (Kullback-Leibler divergence) associated with two rounds of affinity-based selection (R0 to R1) was calculated based on the SELEX method (7) available at Bioconductor (https://bioconductor.org/packages/SELEX). We chose *k* that optimally captures the DNA binding specificity for each TF, and the refined dataset tables were used to perform DeepRec.

Methylated DNA

Recently, more *in vitro* approaches have been developed to probe the sensitivity of TF binding to methylated DNA, including EMSA-based EpiSELEX-seq (8), methyl-Spec-seq (9), bead-based methyl-SELEX (10), and PBM-based methyl-PBM (11). In this study, we trained our DeepRec approach to model the relative binding affinities on EpiSELEX-seq data for ATF4 and C/EBPβ. Sequencing data were downloaded from ENA (https://www.ebi.ac.uk/ena/) under identifier PRJEB25690.

Each dataset contains a methylated library (Lib-M) and unmethylated library (Lib-U) that can be distinguished by their barcodes (8). Sequencing data for Lib-M and Lib-U were both analyzed using the SELEX tool (7) with the same procedure for analysis of unmethylated DNA (12). Relative affinities for oligomers of length *k* were estimated by calculating the oligomer enrichment for R1 counts compared to the expected count as obtained from a Markov model prediction of R0. Fold-enrichment was normalized based on the most enriched oligomer. In the case of ATF4, the most highly enriched 10-bp sequence differed between Lib-U and Lib-M, due to the presence of a repressive effect of methylation at the central CpG dinucleotide. Therefore, libraries were normalized by their shared most enriched oligomer. For the remaining datasets, the most enriched data points were the same, so we simply combined them (8). Data shown in Figure 6 AB

and Supporting Figure S4 are PSAM-filtered Lib-M and Lib-U relative affinities and their respective oligomer sequences (8). Data for ATF4 and C/EBPβ were downloaded from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo) under accession number GSE98652. The pipeline to process EpiSELEX-seq and SMiLE-seq data in DeepRec is available on GitHub (https://github.com/satyanarayan-rao/public_data_processing_for_deeprec).

DeepRec model construction

Data preparation

The refined dataset tables generated by the SELEX R-package (7) are input subjected to DeepRec. The input consists of two variables, *X* and *Y*. Variable *X* encompasses the *k*-mer sequences denoted '*K*mer'. Variable *Y* encompasses the corresponding binding affinities denoted 'SymmetrizedAffinity' that averages binding affinities from forward and reverse complement sequences. Rows with reverse complement sequences were removed, and data were split into training and validation sets with a 90:10 split. Variable *X* was appended with the corresponding reverse complement sequences with a padding of zeros and was encoded by following the encoding schema (see Material and Methods). Notably, the encoding schema can apply to modified DNA and non-Walson-Crick DNA without expanding the data dimension. The multidimensional arrays are stored in HDF5 format.

Feature encoding

Major groove encoding

In the major groove, a G/C base pair has a unique physicochemical signature 'AADN' ('A': H-bond acceptor; 'D': H-bond donor; 'N': nonpolar hydrogen), which differs from the signature of a C/G base pair 'NDAA', the A/T signature 'ADAM', and the T/A signature 'MADA' ('M': thymine methyl group) (Figure 1, S1). We encode the physicochemical signatures as binary vectors A=[0,0,0,1], D=[0,0,1,0], M=[0,1,0,0], and N=[1,0,0,0]. We further encode the major-groove signature of each base pair as a 4×4 binary vector, as follows:

$$\text{A/T} = \begin{bmatrix} A \\ D \\ A \\ M \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \text{T/A} = \begin{bmatrix} M \\ A \\ D \\ A \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\text{G/C} = \begin{bmatrix} A \\ A \\ D \\ N \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad \text{C/G} = \begin{bmatrix} N \\ D \\ A \\ A \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Minor groove encoding

In the minor groove, the A/T base pair shares the same pattern ('ANA') with the T/A base pair. Likewise, the G/C and C/G base pairs share the same pattern ('ADA'). Using the same signature vector for encoding, we can encode the minor-groove signature of each base pair as a 3×4 binary vector:

$$\text{A/T} = \begin{bmatrix} A \\ N \\ A \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \text{T/A} = \begin{bmatrix} A \\ N \\ A \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\text{G/C} = \begin{bmatrix} A \\ D \\ A \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \text{C/G} = \begin{bmatrix} A \\ D \\ A \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The encoding method can be extended to non-Watson–Crick base pairs, including Hoogsteen, synthetic, and mismatched base pairs, 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 for a modified base pair

also implies independence, for instance of a methylated cytosine from cytosine despite the largely overlapping chemical characteristics of C/G and 5mC/G base pairs (13).

DeepRec model design

DeepRec consists of a major-groove physicochemical module, a minor-groove physicochemical module with two convolutional layers, a joint module that integrates signals from the two physicochemical modules, and an output layer that computes the relative TF-binding affinities for each probe (Figure 2). Each module stacks two convolutional layers with 2×2 filters followed by filters of 4×6 for the major groove and 3×6 for the minor groove. The smaller filters mimic the behavior of a residue that recognizes the physicochemical signatures of bases. The larger filters extract patterns in a wider scope, mimicking the behavior of a protein-binding domain.

In the DeepRec model, the output $f(s)$ for sequence s is computed by a feed-forward expression beginning with convolution and ending in a fully connected neural network. The parameters of the two layers' filters $M_{major_1}, M_{major_2}, M_{minor_1}, \text{and } M_{minor_2}$, the thresholds of activation functions $b1_{\text{major}}, b2_{\text{major}}, b1_{\text{minor}} \text{ and } b2_{\text{minor}}$, and the parameters of joint layer and fully connected neural networks $W_1 \text{ and } W_2$ are trainable:

$$f(s) = net_{W_1}(join_{W_2}(f_{major}(s), f_{minor}(s)))$$
$$f_{major}(s) = pool(acti_{b2_{major}}(conv_{major_2}(pool(acti_{b1_{major}}(conv_{major_1}(enco_{major}(s))))))$$
$$f_{minor}(s) = pool(acti_{b2_{minor}}(conv_{minor_2}(pool(acti_{b1_{minor}}(conv_{minor_1}(enco_{minor}(s))))))$$

Major groove physicochemical module

The major-groove module comprises CNNs with one convolutional and one pooling layer, and a fully connected hidden layer. Each layer of CNNs executes a linear transformation of the output from the previous layer by multiplying a weight matrix, followed by a nonlinear transformation. The weight matrix is learned during training to minimize predictive errors. The module takes major-groove physicochemical signatures of user-defined length (*l*) as input, which is represented as a binary matrix in dimensions of $4 \times 4l$.

The convolution layer convolves the 2D physicochemical profile (a)2D image) at every position (*i*, *j*) by using filter W^k:

$$conv(X)_{ijk} = ReLu\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{d=0}^{D-1} W_{mnd}^{k} X_{i+m,j+n,d}\right)$$

where *X* is the input, *i* and *j* are the indices of the output position (*i* is the index of the physicochemical signature position; *j* is the index of the base-pair position), and *k* is the index of the filters. Each convolution filter W^k is an $M \times N \times D$ weight matrix, where *M* is the filter height, *N* is the filter width, and *D* is the number of input channels (here, *D* = 4, i.e., H-bond acceptor, H-bond donor, methyl group and nonpolar hydrogen). *ReLU* represents the rectified nonlinear function:

$$ReLU(x) = \begin{cases} x, if\ x \geq 0 \\ 0, if\ x < 0 \end{cases}$$

The pooling layer computes the maximum value in a window of the convolution layer outputs for each filter. To decrease the dimension of the input and the number of model parameters, non-overlapping pooling with window size $M \times N$ is applied. The pooling operation is defined as:

$$pool(X)_{ijk} = max(\{X_{A,B,k}\})$$

where *X* is the input, *i* and *j* are indices for the output position, *k* is the index of filters, $A = \{iM, iM+1, \cdots, (iM+M-1)\}$, and $B = \{jN, jN+1, \cdots, (jN+N-1)\}$, where *M* and *N* are the height and width of the pooling window size, respectively.

Minor groove physicochemical module

The minor-groove module is nearly identical to the major-groove module. The only difference is that the minor-groove module takes as input the minor-groove physicochemical signatures, represented as a binary matrix with dimensions of $3 \times 4l$.

Joint module

To model interactions between extracted physicochemical signatures from the major- and minor-groove modules, the joint module has one hidden layer with the *ReLU* activation function, which is connected to all neurons of the last layers of the major- and minor-groove modules. One output neuron with the sigmoid activation function is connected to outputs from the hidden layer, which predicts the relative TF binding affinity:

$$Sigmoid(x) = \frac{1}{1+e^{-x}}$$

Model training with DeepRec

Model parameters are initialized randomly (14) and learned on the training dataset by minimizing the Mean Squared Error (MSE) loss function and regularization terms for controlling overfitting:

$$Loss(w) = MSE_w(\hat{y}, y) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2$$
$$MSE_w(\hat{y}, y) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \widehat{y_n})^2$$

The loss function is optimized by mini-batch stochastic gradient descent with a batch size of 128 and a global learning rate of 0.001. The learning rate was adapted from Adam (15). DeepRec was implemented in Python 2.7.18 using Tensorflow 2.0+ and Keras 2.3.1. We initially trained and tested the DeepRec model using a single NVIDIA tesla K80 GPU.

Model tuning

To select an ideal model architecture automatically, DeepRec performs hyperparameter optimization using an exhaustive grid search through a specified subset of the hyperparameter space. The search is guided by a performance metric, i.e. *r*-squared, measured by three-fold cross-validation on the training set. The grid search outputs the settings that achieve the highest *r*-squared. In this study, we randomly chose 100 hyperparameters from the hyperparameter space summarized in Table S1.

Ensemble modeling

Neural network models have high variance by nature due in part to nonlinearity. To reduce the variance of predictions and prevent generalization errors, DeepRec trains multiple models instead of a single model and combines predictions from these models. DeepRec selects the model with the best performance obtained from model tuning and trains 100 models with different random seeds. The resulting models were sorted by their validation *r*-squared, and the best 0.5 quantile was selected for later analysis.

Model interpretation

To interpret the predictive models, DeepRec uses a perturbation-based forward propagation method, which perturbs the input and probes its possible effects on the prediction of the network. By nullifying each signature of the input physicochemical signatures with [0.25, 0.25, 0.25, 0.25], DeepRec investigates the relative binding affinity changes in the output. The difference in terms

of relative binding affinity will be converted to energy. To assess a physicochemical signature of a TF that contributes to binding free energy ($\Delta\Delta G/RT$), we compared the difference between the binding free energies with and without such the physicochemical signature and calculated $\Delta\Delta\Delta G/RT$ for the physicochemical signature of the top binding site, where

$$\frac{\Delta\Delta\Delta G}{RT} = \frac{\Delta\Delta G}{RT}_{reference} - \frac{\Delta\Delta G}{RT}_{reference\ with\ nullified\ signature}$$

We calculate this for each selected model and average the difference to plot the height of each physicochemical signature. The letter (A, D, M, N) represent H-bond acceptor, donor, methyl-group, and nonpolar hydrogen, respectively. The height represents $-\Delta\Delta\Delta G/RT$. We also calculated the standard error to the mean (SEM), which is represented as a vertical bar.

DeepRec sequence model

We developed a DeepRec sequence model and compared it to the original DeepRec version (which encodes DNA as a set of physicochemical signatures). The architecture of the DeepRec sequence model is similar to DeepBind (16). The model considers DNA as a 1D string and encodes input DNA sequence as one-hot binary vectors, where A=[1,0,0,0], C=[0,1,0,0], G=[0,0,1,0], and T=[0,0,0,1]. The DeepRec sequence model consists of a single convolutional layer with 1×6 filters, followed by a pooling layer and a target layer to predict relative binding affinities.

The DeepRec sequence model uses the same perturbation-based forward propagation method as the original version of DeepRec to generate sequence logos and interpret the model. The DeepRec sequence model calculates the binding free energy ($\Delta\Delta G/RT$) changes in output by nullifying each base pair of the input sequence with the vector [0.25, 0.25, 0.25, 0.25] at a time. To assess a base pair's contribution to binding free energy ($\Delta\Delta G/RT$), we compared the difference between the binding free energies with and without that base pair and calculated $\Delta\Delta\Delta G/RT$ for the base pair in reference to the top binding site, where

$$\frac{\Delta\Delta\Delta G}{RT} = \frac{\Delta\Delta G}{RT}_{reference} - \frac{\Delta\Delta G}{RT}_{reference\ with\ nullified\ base\ pair}$$

Performance comparison

DeepRec was compared with DeepBind (16) and FeatureREDUCE (17, 18), the top two highest-scoring methods compared in (16). DeepBind is a method based on deep convolutional neural networks, and FeatureREDUCE is a method that combines position weight matrix (PWM) and *k*-mer models. For the DeepBind method, we selected the best parameter setting from 100 calibrations and ran 10 replicates for each comparison. For FeatureREDUCE, we used the default setting and did 10 random training test partitions for each comparison. To compare how a *k* value affects model performance, we tested the *k*-mer value in FeatureREDUCE and the filter size in DeepBind from 4 to 10. The adapted codes for DeepBind and FeatureREDUCE are available on GitHub: https://github.com/TsuPeiChiu/DeepBind and https://github.com/satyanarayan-rao/FeatureREDUCE.

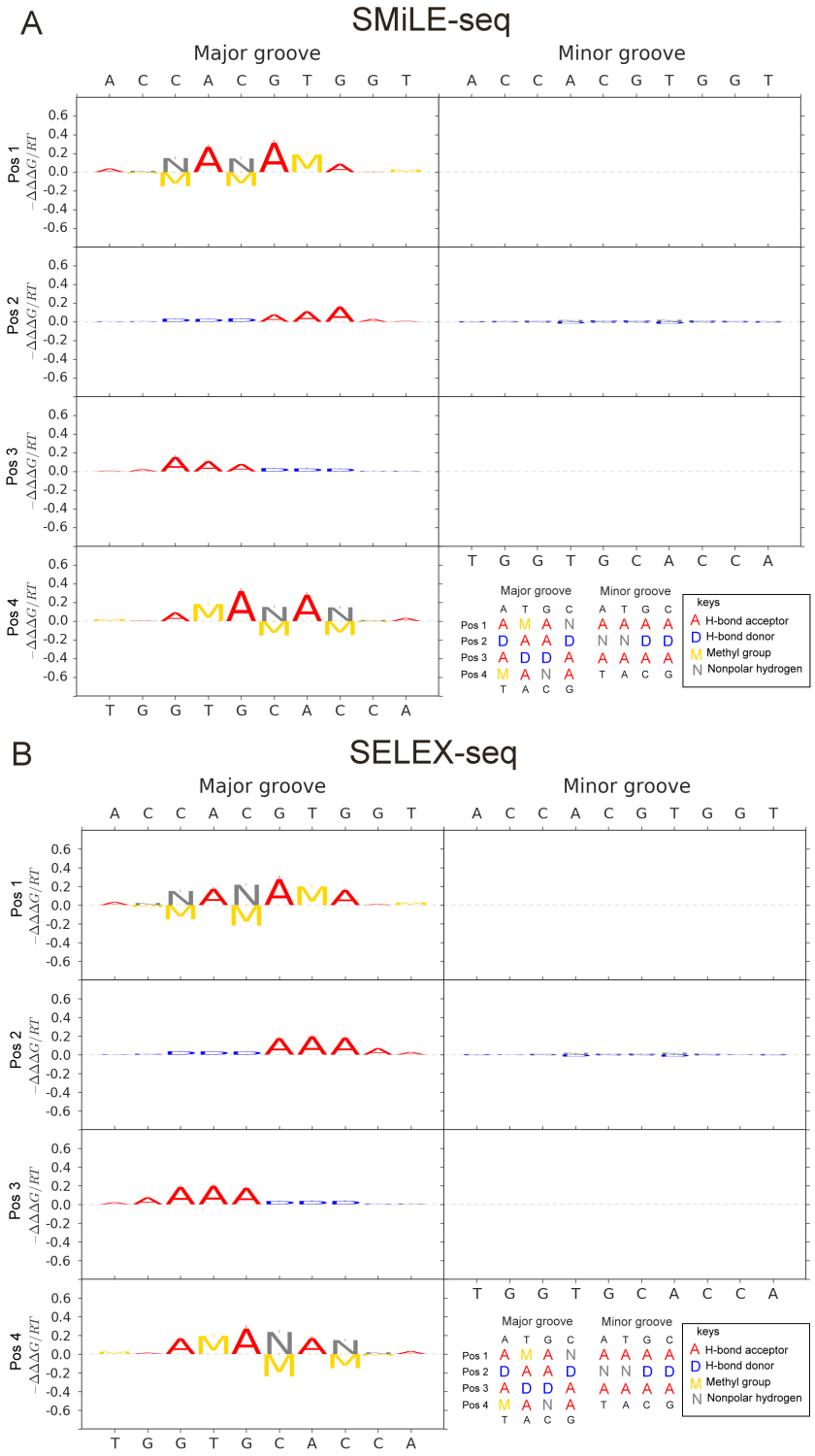
Detailed author contributions

T.P.C. and R.R. designed this study. T.P.C. independently led and executed all aspects of the study. The original concept and idea of the physicochemical signature encoding of DNA was first developed in T.P.C.'s dissertation and made public through the University of Southern California (https://uosc.primo.exlibrisgroup.com/permalink/01USC_INST/273cgt/cdi_proquest_journals_2178722224). S.R. participated in the initial phase of the development of DeepRec and continued to contribute to testing of the method. T.P.C. and R.R. directed the project and wrote the paper.

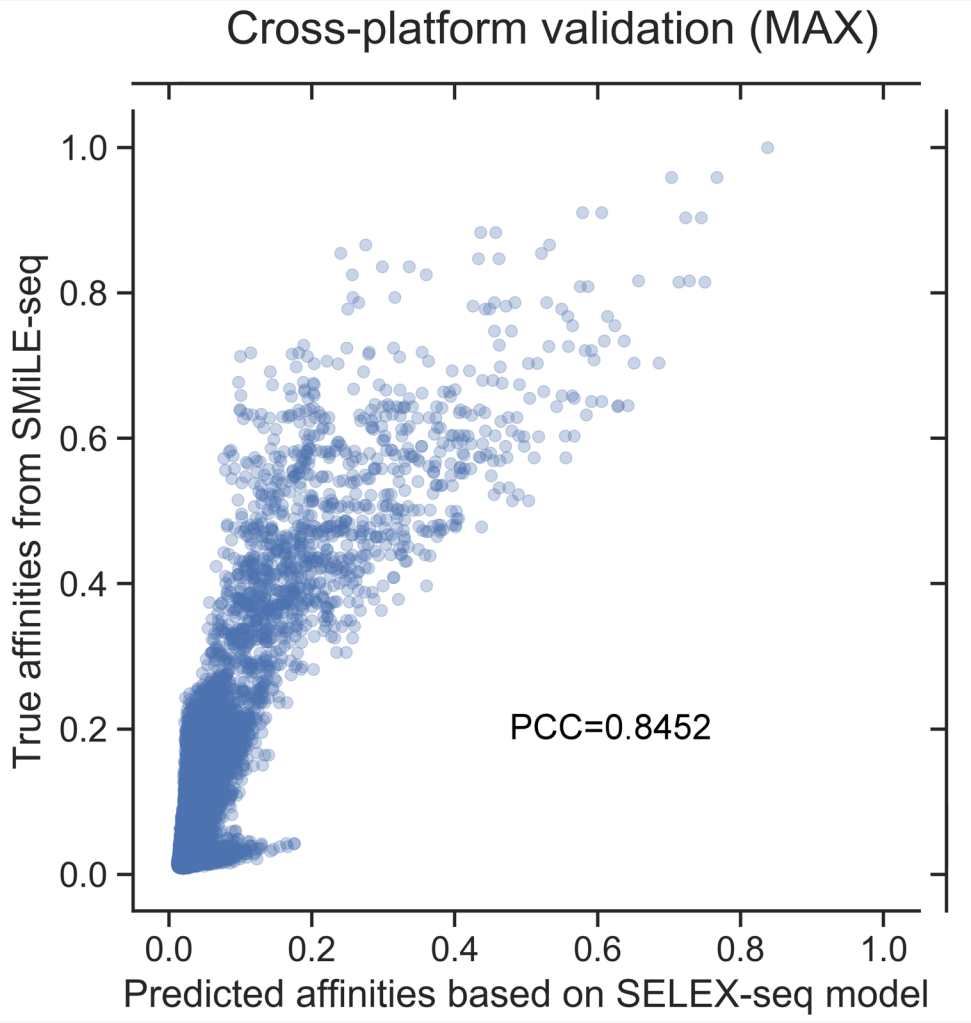
Supporting References

1. T. Zhou*, et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.* **112**(15), 4654-4659 (2015).
2. C. Rastogi*, et al.* Accurate and sensitive quantification of protein-DNA binding affinity. *Proc. Natl. Acad. Sci. U.S.A.* **115**(16), E3692-E3701 (2018).
3. A. Jolma*, et al.* DNA-binding specificities of human transcription factors. *Cell* **152**(1-2), 327-339 (2013).
4. L. Yang*, et al.* Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.* **13**(2), 910 (2017).
5. A. Isakova*, et al.* SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods* **14**(3), 316-322 (2017).
6. M. Djordjevic & A.M. Sengupta Quantitative modeling and data analysis of SELEX experiments. *Phys. Biol.* **3**(1), 13-28 (2005).
7. T.R. Riley*, et al.* SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol. Biol.* **1196**, 255-278 (2014).
8. J.F. Kribelbauer*, et al.* Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *Cell Rep.* **19**(11), 2383-2395 (2017).
9. 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**(11), eaao1799 (2017).
10. Y. Yin*, et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**(6337), (2017).
11. D. Tillo*, et al.* The Epstein-Barr Virus B-ZIP Protein Zta Recognizes Specific DNA Sequences Containing 5-Methylcytosine and 5-Hydroxymethylcytosine. *Biochemistry* **56**(47), 6200-6210 (2017).
12. M. Slattery*, et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**(6), 1270-1282 (2011).
13. A.C. Dantas Machado*, et al.* Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief Funct. Genomics* **14**(1), 61-73 (2015).
14. X. Glorot & Y. Bengio Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, eds Yee Whye T & Mike T (PMLR, Proceedings of Machine Learning Research), pp 249--256 (2010).
15. D.P. Kingma & J. Ba (2014) Adam: A Method for Stochastic Optimization. *ArXiv e-prints.*
16. 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**(8), 831-838 (2015).
17. M.T. Weirauch*, et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**(2), 126-134 (2013).
18. T.R. Riley, A. Lazarovici, R.S. Mann, & H.J. Bussemaker Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *Elife* **4**, (2015).

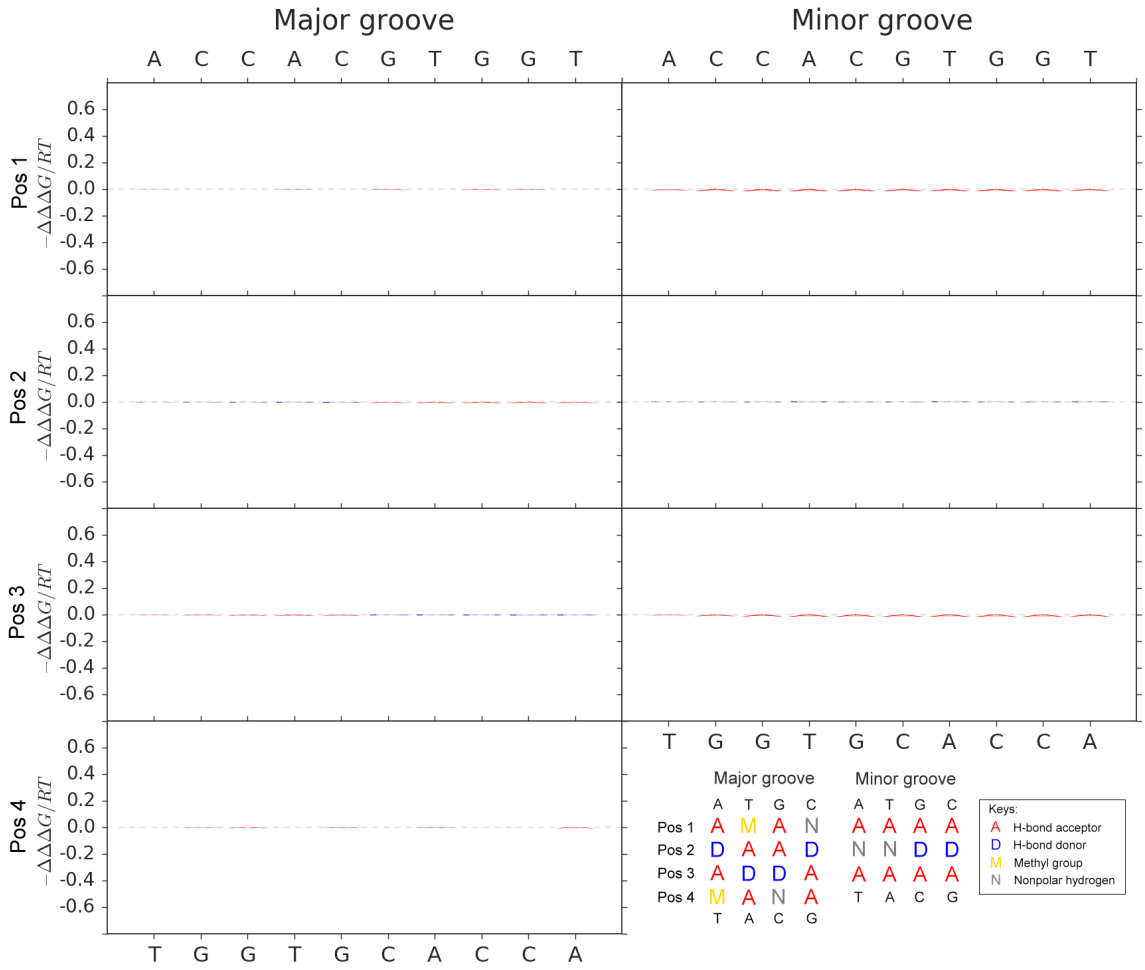
Supporting Figures



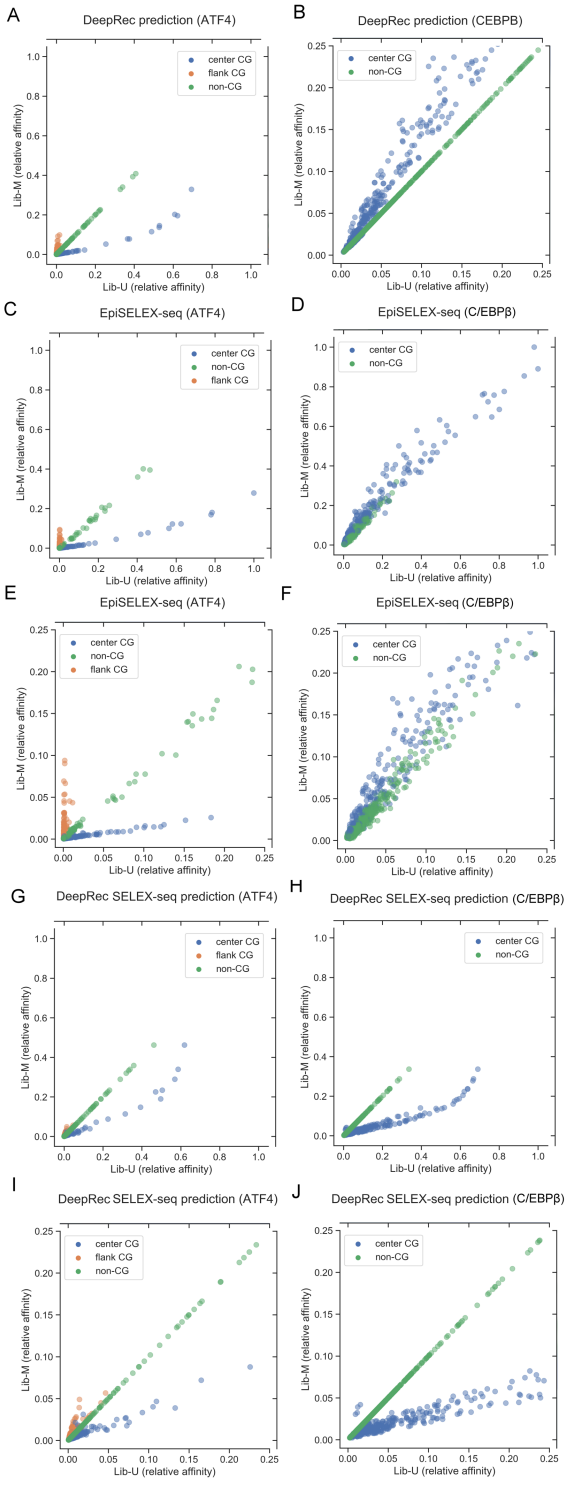
Supporting Figure S1. Comparison of physicochemical energy logos of MAX based on the (A) SMiLE-seq and (B) SELEX-seq data.



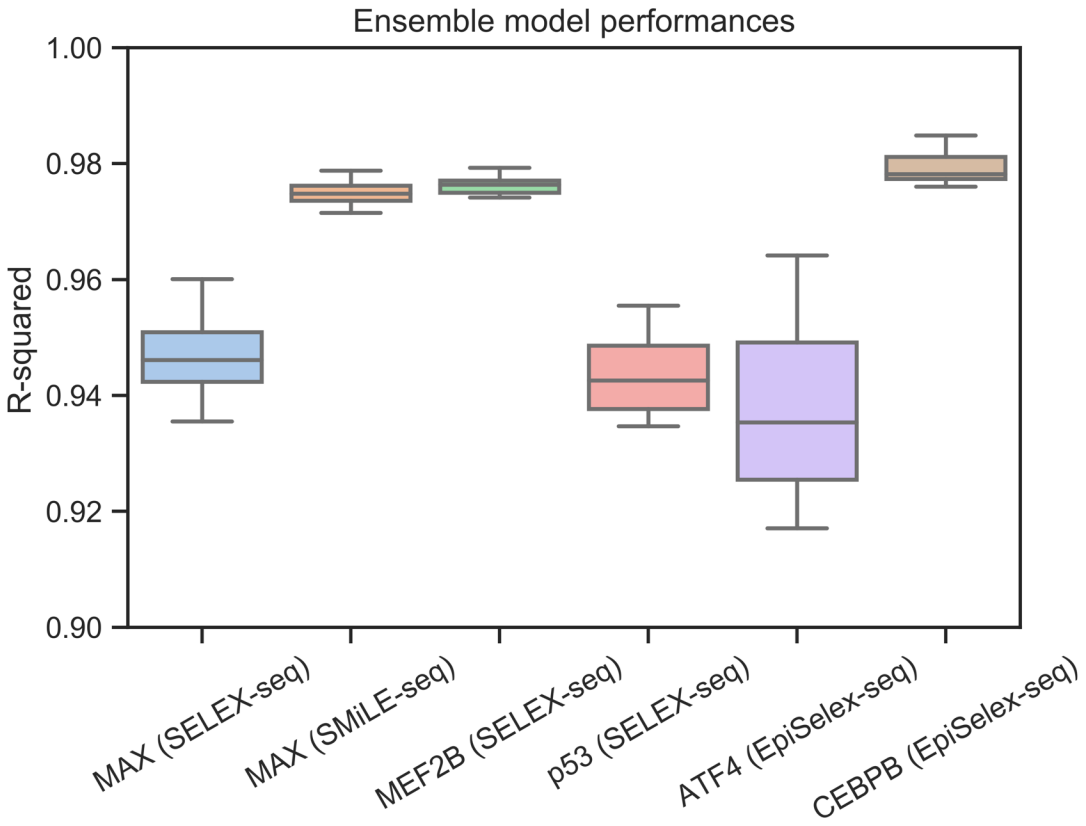
Supporting Figure S2. Comparison of relative binding affinities measured by SMiLE-seq and predicted by the model trained by the SELEX-seq data.



Supporting Figure S3. Selected DNA physicochemical energy logos for control, where the relative binding affinities of the training data were shuffled.



Supporting Figure S4. Comparison between relative binding affinities for a methylated sequence library (Lib-M) versus an unmethylated sequence library (Lib-U). (A, B) Binding affinities were predicted by DeepRec on a combined dataset of Lib-M and Lib-U for ATF4 (A) and C/EBPβ (B). (C-F) Binding affinities were obtained from EpiSELEX-seq for ATF (C, E) and C/EBPβ (D, F) with different affinity ranges. (G-J) Binding affinities were predicted by DeepRec on an unmethylated dataset derived from SELEX-seq for ATF4 (G, I) and C/EBPβ (H, J) with different affinity ranges.



Supporting Figure S5. DeepRec models are robust. For a given TF dataset (X-axis), 50 DeepRec models were trained with different random seeds. Model performance measured as *R*-squared shows that the performance is consistent.

Supporting Table S1. Hyperparameters for DeepRec models.

Hyperparameters	**Values**
learning rate	[0.01, 0.001, 0.0001]
number of epoch	[100]
batch size	[128, 256]
filter length 1	[2]
filter length 2	[6]
number of filter 1	[8, 16]
number of filter 2	[8, 16]
alpha 1	[0.001, 0.0001, 0.00001]
l1 ratio 1	[0, 0.1, 0.2, 0.4, 0.8, 1]
number of node in joint layer	[16, 32, 64]
alpha in joint layer	[0.001, 0.0001, 0.00001]
l1 ratio in joint layer	[0, 0.1, 0.2, 0.4, 0.8, 1]
drop out in joint layer	[0, 0.01, 0.02]