SUPPLEMENTARY DATA

Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding

Tsu-Pei Chiu[1], Satyanarayan Rao[1], Richard S. Mann[2], Barry Honig[2,3], and Remo Rohs[1,*]

[1]Computational Biology and Bioinformatics Program, Departments of Biological Sciences, Chemistry, Physics & Astronomy, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

[2]Departments of Systems Biology and Biochemistry & Molecular Biophysics, Mortimer B. Zuckerman Institute, Columbia University, New York, NY 10032, USA

[3]Howard Hughes Medical Institute, New York, NY 10032, USA

*To whom correspondence should be addressed:

Remo Rohs
Computational Biology and Bioinformatics Program
Department of Biological Sciences
University of Southern California
1050 Childs Way RRI 413H
Los Angeles, CA 90089, USA
Tel: +1-213-740-0552
Fax: +1-213-821-4257
Email: rohs@usc.edu

SUPPLEMENTARY MATERIALS AND METHODS

Non-linear Poisson–Boltzmann equation (NLPB)

The NLPB equation is widely used to calculate electrostatic interactions in ionic solutions (1). The equation reads

$$\nabla \cdot [\varepsilon(\vec{r})\nabla \cdot \phi(\vec{r})] - \varepsilon(\vec{r})\kappa(\vec{r})^2 \sinh[\phi(\vec{r})] + 4\pi\rho^f(\vec{r})/\mathrm{kT} = 0,$$

where $\phi(\vec{r})$ is the electrostatic potential (EP) at any location $\vec{r}$. The unit of $\phi(\vec{r})$ is $\mathrm{kT/e}$, where k is the Boltzmann constant, T is the temperature, and e is the elementary charge. $\varepsilon(\vec{r})$ is the dielectric constant in the solute or solution. The term $\kappa^2 = 1/\lambda^2 = 8\pi\mathrm{q}^2 I/\mathrm{ekT}$, where I is the ionic strength (moles/L) of the bulk solution. ρ^f is the fixed charge density. The variables $\phi, \varepsilon, \kappa$ and ρ are functions of the position vector $\vec{r}$. The linearization $\sinh\phi(\vec{r}) = \phi(\vec{r})$ for small potentials leads to the linear Poisson–Boltzmann (LPB) equation in which contributions from different chemical groups are additive (2).

DelPhi – a finite-difference method for solving the NLPB equation

The DelPhi program (1,3) utilizes a finite-difference method to solve the NLPB equation for biomolecules in aqueous solution. The method maps the charge density and dielectric constant onto a three-dimensional cubic grid and solves the NLPB equation iteratively. The program allows users to specify ionic strength and dielectric constants of the solute and solvent. In addition, users can assign charges to the solute and measure the particular charge effects on the molecule in terms of EP.

Generation of pentamer query table for EP prediction

To compile the pentamer query table, we generated a large training dataset of all-atom Monte Carlo (MC) predictions for 2,297 different DNA fragments ranging 12 to 27 base pairs (bp) in length. We performed NLPB calculations to profile EP on these average structures using the DelPhi program (1). For each bp, we derived the EP at the midpoint of the minor groove and at 26 points that were equally distributed on a sphere with 1 Å radius surrounding this midpoint, with a total of 27 EP values for a sphere (4) (Figure 2C). After filtering out extreme EP values (> 0 or < –20 kT/e), we calculated the mean and standard deviation (SD) for the remaining EP values. We assigned an average value to the sphere if its SD was < 3 kT/e, thereby excluding mean values with large EP fluctuations. As the sphere lies in the approximate center of the minor groove, EP can be defined as a function of sequence, with one value per bp.

After mapping EP as a function of sequence for 2,297 DNA fragments, we applied a pentameric sliding-window approach to each fragment. EP values at the central bp in a pentamer were recorded for the 512 unique pentamers. To remove outlier effects, we kept 80% of the data points and removed the extreme 10% of data points from the head and tail end of the data. Using this filtering approach, we generated a query table of average values for each occurrence of 512 possible pentamers in our dataset. The average occurrence of possible pentamers was 45.2 with a SD of 0.3. This query table was integrated in a sliding-window approach for high-

throughput (HT) minor-groove EP prediction, similarly to our previous approach for DNA shape (5).

High-throughput prediction of EP

We provide a stand-alone web server (DNAphi; http://rohslab.usc.edu/DNAphi) for HT prediction of EP in the minor groove. Input data are nucleotide sequences in FASTA format, which can be pasted into a web form or uploaded as a separate file. The function '**φ Prediction**' allows users to perform predictions and view the results in a graphical representation that can be downloaded as a quantitative data table for further analysis. The function '**φ Learning**' requires the binding strength per given sequence in a user-defined unit as additional input data or response variable. The statistical machine-learning (ML) approach of L2-regularized multiple linear regression (MLR) was applied to this function (6). We also integrated the HT EP prediction function into DNAshapeR (5), our R/Bioconductor package (https://www.bioconductor.org/packages/devel/bioc/html/DNAshapeR.html). This package provides an easy-to-use and easy-to-extend interface that can be readily integrated into other HT genomic analysis platforms (5).

Fis binding site data

We used Fis-DNA binding site data for the eight sequences of F1, F24, F25, F26, F27, F28, F29 and F36 (7,8), which exhibit Fis-binding affinities differing by three orders of magnitude. Seven of these sites only differ in their five central bp (colored red in Supplementary Table S1). F36 has one additional bp that differs in the flank of the five central bp (bold in Supplementary Table S1). Sequences with their logarithms of binding affinity K_d are listed in Supplementary Table S1. Analysis of additional Fis binding sites is shown in Supplementary Figure S6.

HT-SELEX data for 215 mammalian transcription factors (TFs)

We used HT-SELEX data for 215 TFs from 27 protein families originally published by Jolma *et al.* (9) and recently re-sequenced with an on average 10-fold increase in sequencing depth, resulting in more accurate binding models (10). Sequencing data are available at the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under study identifier PRJEB14744. Data were pre-processed as described in Yang *et al.* (11) and are available at http://rohslab.usc.edu/MSB2017/.

SELEX-seq data for *Drosophila* Hox proteins

We used experimental data for 21 Exd-Hox heterodimers derived from binding assays followed by deep sequencing (SELEX-seq) (6). Data included the anterior and posterior Hox proteins, the Scr mutants containing mutated Arg3, His–12 or Arg5 and the Antp mutants in which minor groove-contacting residues from Scr were engineered into the Scr linker, all in complex with the cofactor Exd. All sequences selected in SELEX-seq experiments with a count ≥ 25 were aligned based on the core motif 5'-TGAYNNAY-3', where N can be any nucleotide and Y represents C or T. Raw data can be downloaded from the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE65073 (6). Sequences with multiple occurrences of the core motif were removed from this analysis.

gcPBM data for human basic helix-loop-helix (bHLH) proteins

We used experimental data for three bHLH dimers Mad1/Max ('Mad'), Max/Max ('Max') and c-Myc/Max ('Myc') derived from gcPBM experiments (12). Data contained 36-bp genomic sequences centered at a putative E-box binding site. The number of sequences for Mad was 6,927, for Max was 8,569 and for Myc was 7,535. Data can be downloaded from GEO under accession number GSE59845 (12).

Feature vector encoding

Given a DNA sequence s of length l, the corresponding feature vector can be derived from the following basis functions, in a similar way to those introduced by Zhou *et al.* (12) and Yang *et al.* (11). The basis function for EP features is:

$$\Omega_i^{EP}(s) = \begin{cases} EP_i, \text{if } s^{(i)} \in \{\text{A, C, G, T}\} \\ EP_{avg}, \text{otherwise} \end{cases}, \ i = 3,4, \dots, l-2$$

where EP_{avg} is the average value of EP over all possible pentamers. Basis functions for mononucleotide sequence features are:

$$\Omega_{4\times(i-1)+1}^{seq}(\text{s}) = \begin{cases} 0, \text{if } s^{(i)} \neq \text{A} \\ 1, \text{if } s^{(i)} = \text{A} \end{cases}, i = 1, \cdots, l$$
$$\Omega_{4\times(i-1)+2}^{seq}(s) = \begin{cases} 0, \text{if } s^{(i)} \neq \text{C} \\ 1, \text{if } s^{(i)} = \text{C} \end{cases}, i = 1, \cdots, l$$
$$\Omega_{4\times(i-1)+3}^{seq}(s) = \begin{cases} 0, \text{if } s^{(i)} \neq \text{G} \\ 1, \text{if } s^{(i)} = \text{G} \end{cases}, i = 1, \cdots, l$$
$$\Omega_{4\times i}^{seq}(s) = \begin{cases} 0, \text{if } s^{(i)} \neq \text{T} \\ 1, \text{if } s^{(i)} = \text{T} \end{cases}, i = 1, \cdots, l$$

where i is the position of the given sequence. Basis functions for the DNA shape features minor-groove width (MGW), propeller twist (ProT), Roll and helix twist (HelT) are:

$$\Omega_i^{MGW}(s) = \begin{cases} MGW_i, \text{if } s^{(i)} \in \{\text{A, C, G, T}\} \\ MGW_{avg}, \text{otherwise} \end{cases}, \ i = 3,4, \dots, l-2$$
$$\Omega_i^{ProT}(s) = \begin{cases} ProT_i, \text{if } s^{(i)} \in \{\text{A, C, G, T}\} \\ ProT_{avg}, \text{otherwise} \end{cases}, \ i = 3,4, \dots, l-2$$
$$\Omega_i^{Roll}(s) = \begin{cases} Roll_i, \text{if } s^{(i)} \in \{\text{A, C, G, T}\} \\ Roll_{avg}, \text{otherwise} \end{cases}, \ i = 2,3, \dots, l-1$$
$$\Omega_i^{HelT}(s) = \begin{cases} HelT_i, \text{if } s^{(i)} \in \{\text{A, C, G, T}\} \\ HelT_{avg}, \text{otherwise} \end{cases}, \ i = 2,3, \dots, l-1$$

where MGW_{avg}, $ProT_{avg}$, $Roll_{avg}$ and $HelT_{avg}$ are average values with respect to MGW, ProT, Roll and HelT, respectively, over all possible pentamers.

EP and DNA shape features, denoted Ω_i^{EP}, Ω_i^{MGW}, Ω_i^{ProT}, Ω_i^{Roll} and Ω_i^{HelT}, respectively, were generated and normalized by DNAshapeR (5). Normalization was performed by:

$$\Omega_i^{EP}(\mathrm{s}) = (EP_i - EP_{min})/(EP_{max} - EP_{min})$$

where EP_i is the predicted EP value, EP_{min} is the minimum EP value and EP_{max} is the maximum EP value over all possible pentamers. Similarly, normalization for DNA shape features was performed by using:

$$\Omega_i^{MGW}(s) = (MGW_i - MGW_{min})/(MGW_{max} - MGW_{min})$$
$$\Omega_i^{ProT}(s) = (ProT_i - ProT_{min})/(ProT_{max} - ProT_{min})$$
$$\Omega_i^{Roll}(s) = (Roll_i - Roll_{min})/(Roll_{max} - Roll_{min})$$
$$\Omega_i^{HelT}(s) = (HelT_i - HelT_{min})/(HelT_{max} - HelT_{min})$$

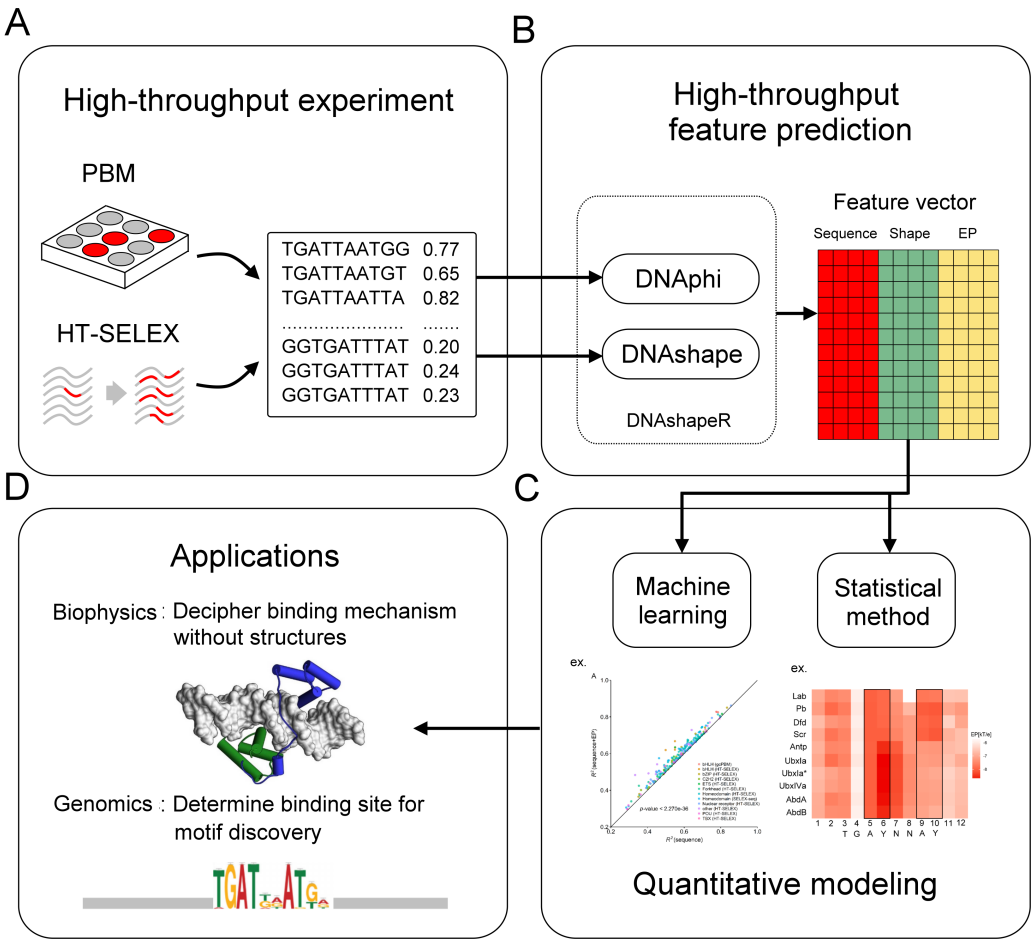
The complete feature vector for the given sequence can be obtained by concatenating the six numeric vectors:

$$\Omega(x) = [\Omega^{EP}(s), \Omega^{seq}(s), \Omega^{MGW}(s), \Omega^{ProT}(s), \Omega^{Roll}(s), \Omega^{HelT}(s)]^{\mathrm{T}}$$

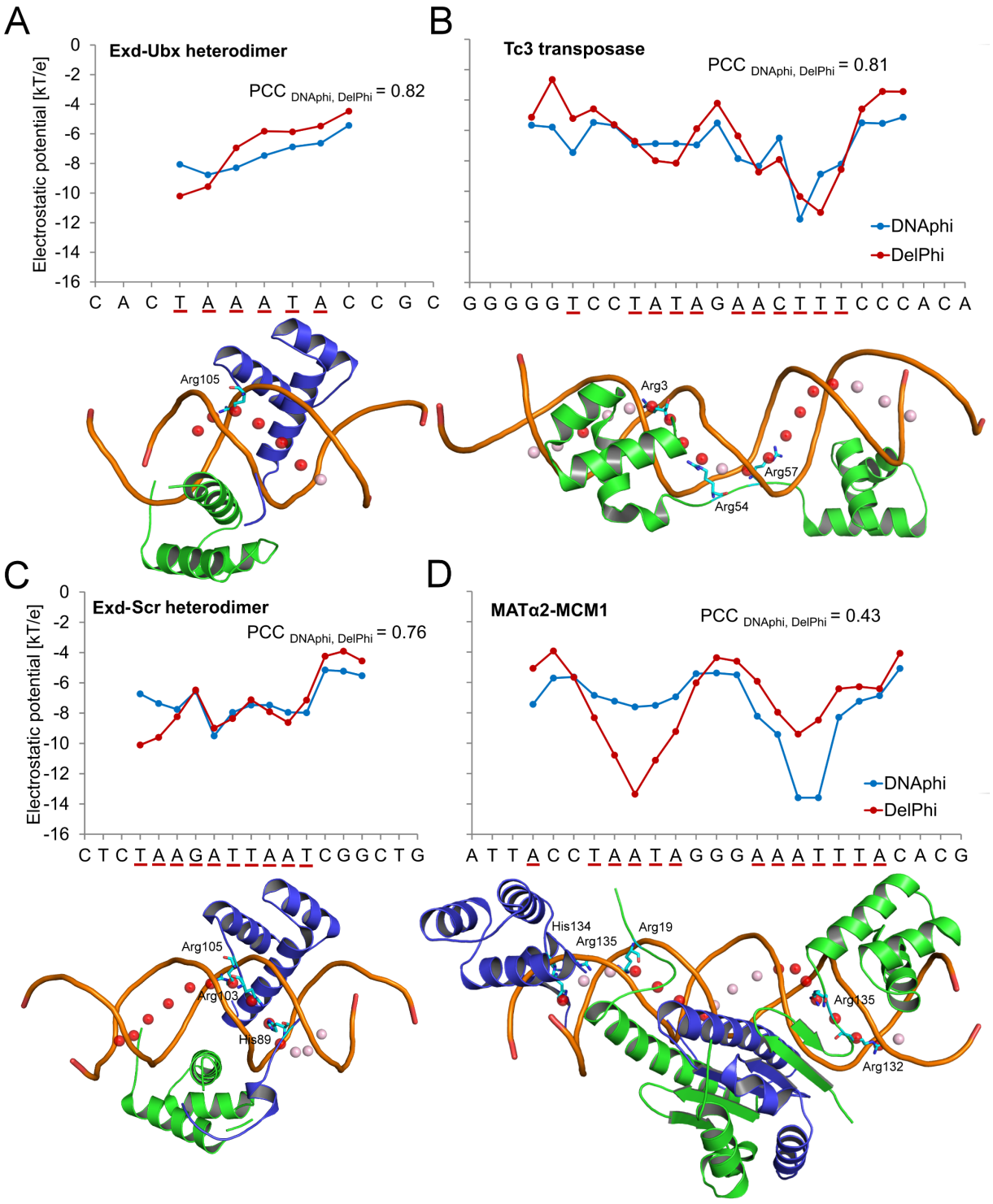
Statistical tests and significance levels

We applied *t*-test hypothesis testing to determine whether there was a significant difference between two experimental groups. For the predictive-power comparison in quantitative modeling (Figure 7, Supplementary Figures S8 and S11), we assumed a performance increase in terms of R^2 for the augmented models as the alternative hypothesis. For the comparison of EP-based models for gcPBM datasets (Supplementary Figure S10), we assumed for the alternative hypothesis that the EP preferences at a particular position were different for two datasets. The calculated *P* value for the hypothesis test represents the probability of mistakenly rejecting the null hypothesis if the null hypothesis is true. The significance level α is the standard value for which a *P* value ≤ α is considered statistically significant. Typical values for α are 0.1 (*), 0.05 (**), 0.01 (***) and 0.001 (****). For example, the *P* value 1.549×10^{-34} (Figure 7B) indicates that the probability of making a mistake by rejecting a true null hypothesis (i.e. probability that there is no improvement achieved by an EP-augmented model) is 1.549×10^{-34}, which falls below the significance level of 0.001 and is therefore considered highly statistically significant.

SUPPLEMENTARY FIGURES

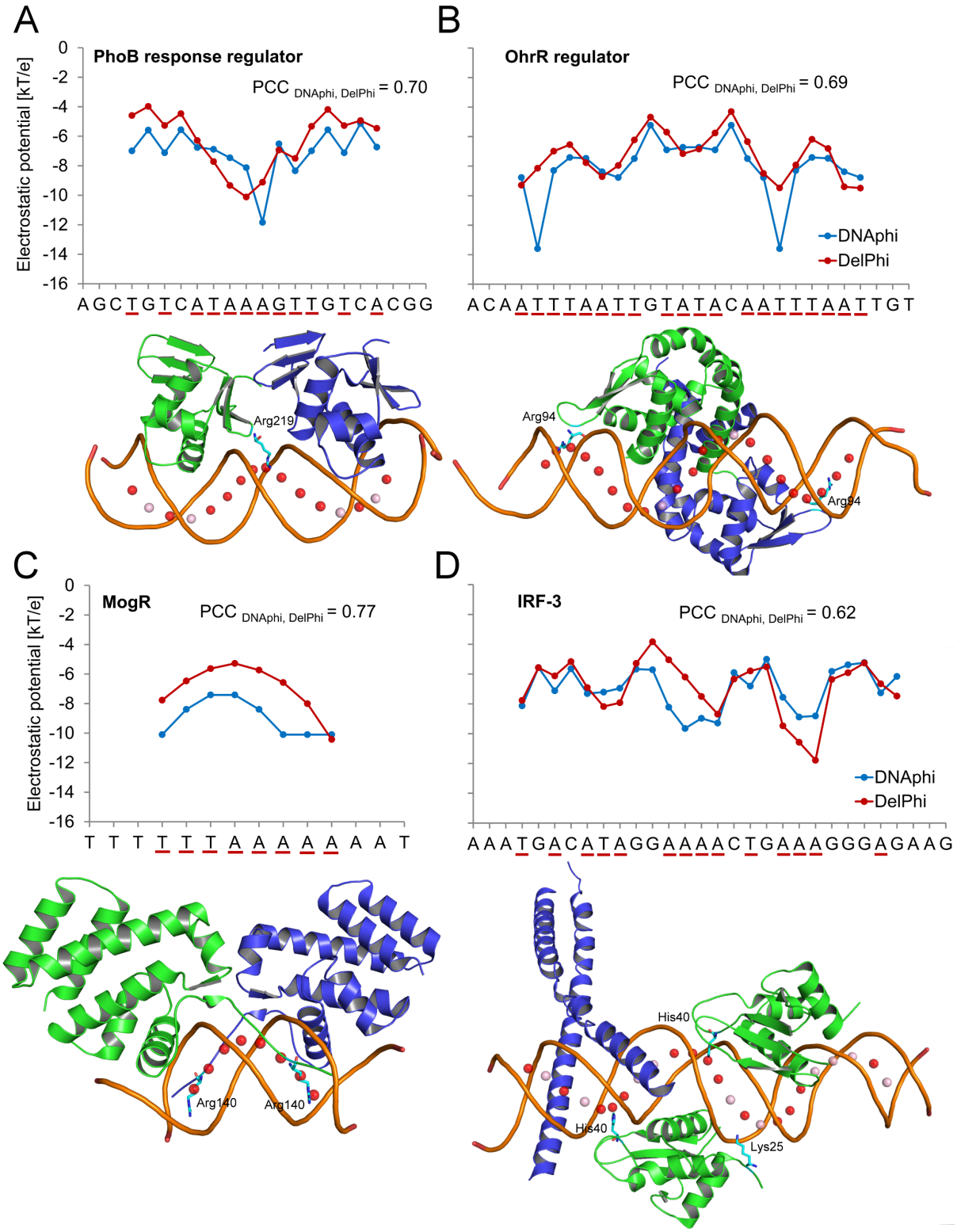


Supplementary Figure S1. Illustration of possible applications of DNAphi. (A) Protein-DNA binding affinities can be measured by several types of HT assays, including PBM (13), HT-SELEX (9,10) and their variants (14,15). Outputs of these experiments are represented as sequences followed by corresponding binding affinities. (B) Output sequences are used as inputs to the HT prediction programs DNAphi (this study) and DNAshape (16) (both methods are included in the DNAshapeR package (5)) for prediction of minor-groove EP and DNA shape features. DNAshapeR can encode EP, sequence and shape features as a concatenated feature vector. (C) Resulting vectors can be used as input of statistical ML methods for further analysis and modeling. (D) Resulting models can be used to infer specific mechanisms of protein-DNA recognition (without requiring 3D structures) or to identify TF binding sites in the genome.



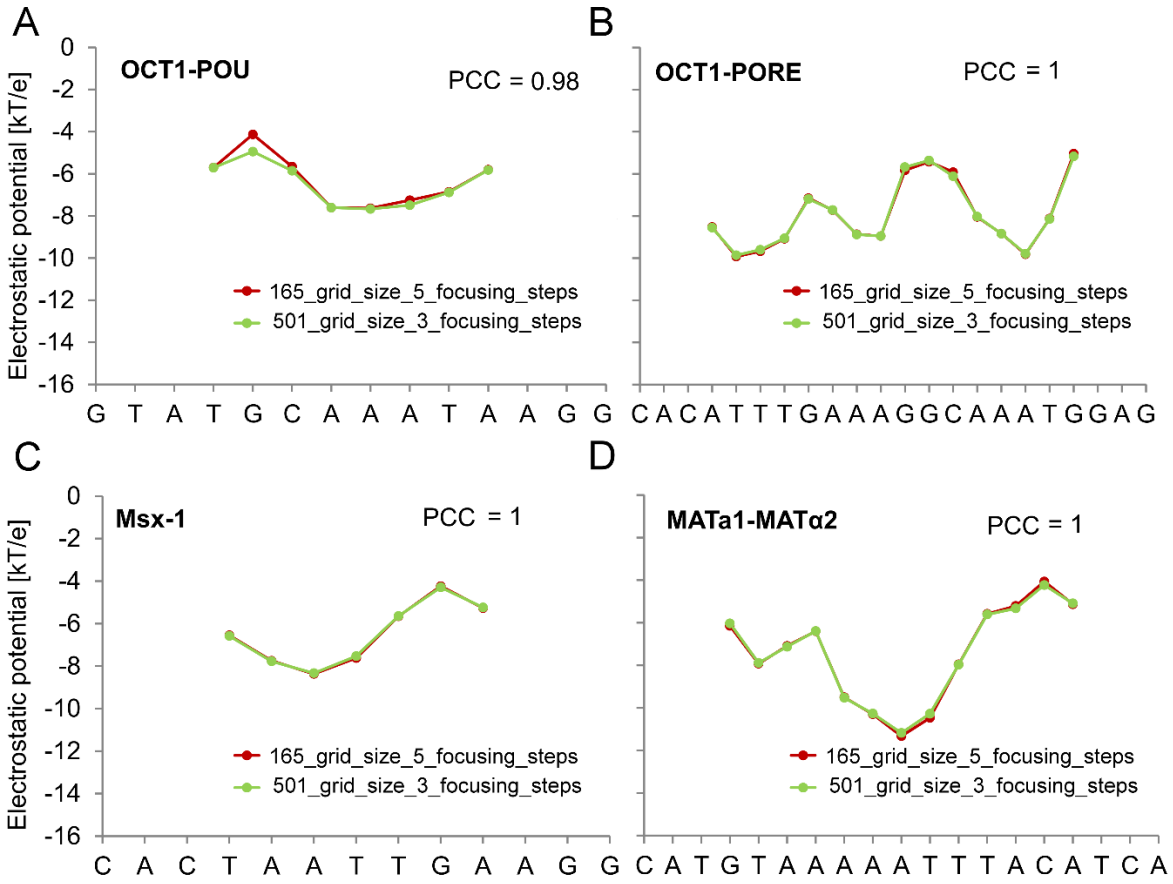
Supplementary Figure S2. Validation of DNAphi predictions at TF-DNA binding sites. Minor-groove EPs of binding sites of (A) Exd-Ubx heterodimer (PDB ID 1B8I) (17), (B) Tc3 transposase (PDB ID 1V78) (18), (C) Exd-Scr heterodimer (PDB ID 2R5Z) (19) and (D) MATα2-MCM1 (PDB ID 1MNM) (20), whose binding interfaces include arginine residues inserted into the minor groove, were predicted by using DNAphi (blue) and

DelPhi (red). Pearson Correlation Coefficients (PCCs) demonstrate the statistical similarity between EP profiles derived by these two approaches. We highlighted the more negative minor-groove EP values (≤ –6.505 kT/e) predicted by DNAphi by underlining respective positions on the *x*-axis. Corresponding spheres defined by DNAphi are represented by spheres in each structure, with red indicating below-average EP values ≤ –6.505 kT/e and pink indicating EP values > –6.505 kT/e. Protein residues that form minor-groove contacts as defined by DNAproDB (21) are shown in each structure.

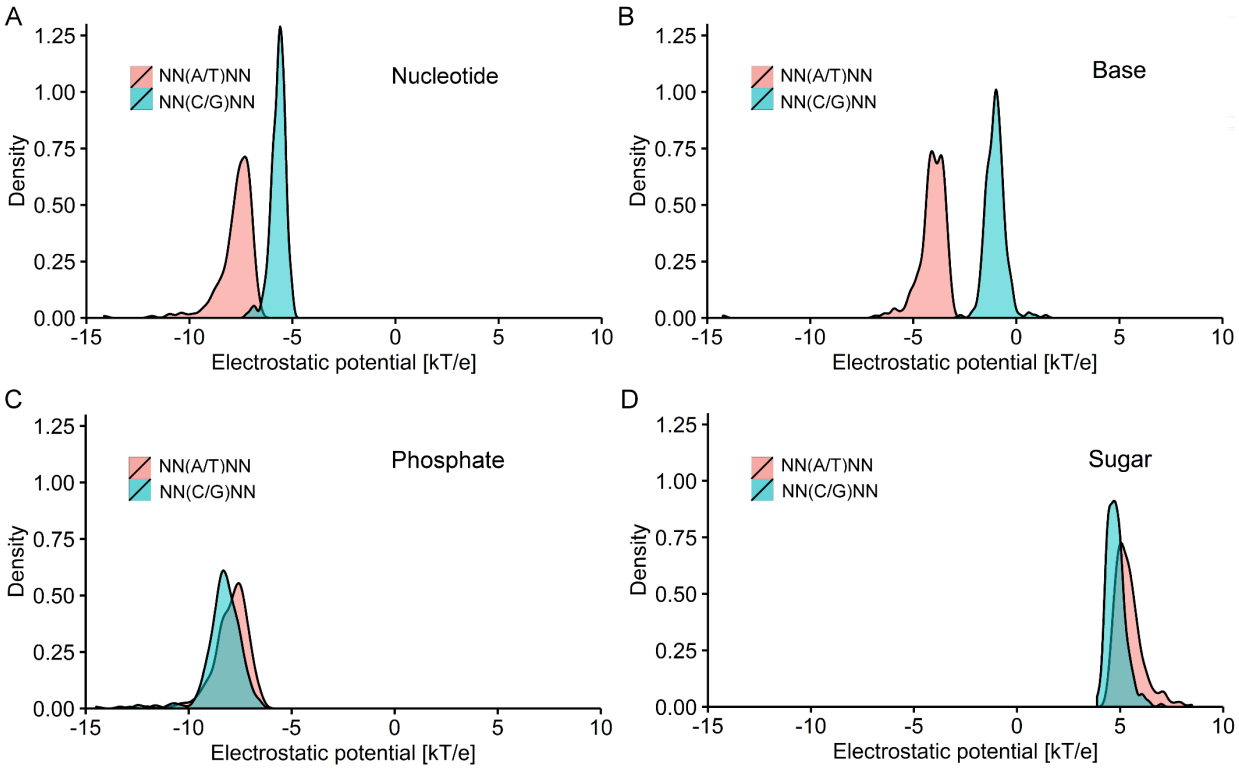


Supplementary Figure S3. Validation of DNAphi predictions using TF-DNA binding sites. Minor-groove EPs of binding sites of (A) PhoB response regulator (PDB ID 1GXP) (17), (B) OhrR regulator (PDB ID 1Z9C) (18), (C) MogR (PDB ID 3FDQ) (19) and (D)

IRF-3 (PDB ID 1T2K) (20), whose binding interfaces include arginine residues inserted into the minor groove, were predicted by using DNAphi (blue) and DelPhi (red). PCCs demonstrate the statistical similarity between EP profiles derived from these two approaches. We highlighted the more negative minor-groove EP values (≤ –6.505 kT/e) predicted by DNAphi by underlining respective positions on the *x*-axis. Corresponding spheres defined by DNAphi are represented by spheres in each structure, with red indicating below-average EP values ≤ –6.505 kT/e and pink indicating EP values > –6.505 kT/e. Protein residues that form minor-groove contacts as defined by DNAproDB (21) are shown in each structure.



Supplementary Figure S4. Comparison of minor-groove EP predictions using DelPhi with different parameter settings. (A-D) EPs for DNA binding sites of four proteins (shown in Figure 3) were predicted by using DelPhi with a grid size of 165 and five focusing steps (red) vs. a grid size of 501 and three focusing steps (green). PCCs demonstrate the statistical similarity between EP profiles derived from DelPhi with varying parameter sets.

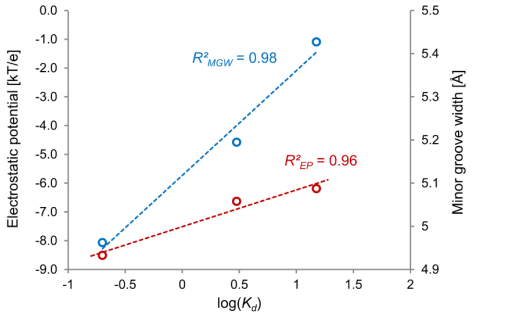


Supplementary Figure S5. EP distributions of 512 unique pentamers based on LPB calculations. (A) Distributions for complete nucleotides based on LPB calculations. Patterns of distributions are similar to those derived with the NLPB (Figure 4C). (B-D) Distributions based on partial charges of the (B) bases, (C) phosphate groups or (D) sugar moieties.

A

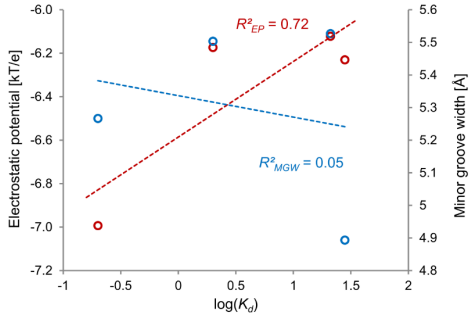
B

Label / Pos	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	K_d	log(K_d)
F1	A	A	A	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	A	T	T	T	0.2	-0.70
F35	A	A	A	T	T	A	G	T	T	T	G	A	A	T	C	T	C	G	A	G	C	T	A	A	T	T	T	15	1.18
F36	A	A	A	T	T	T	G	T	T	T	G	A	A	T	C	T	C	G	A	G	C	A	A	A	T	T	T	3	0.48



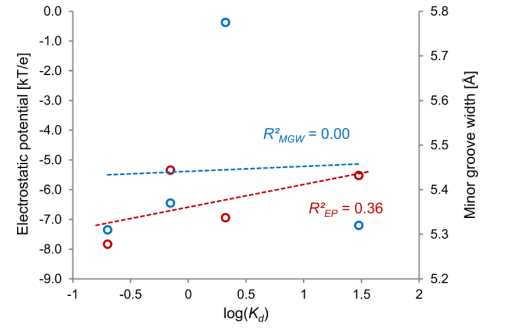
C

Label / Pos	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	K_d	log(K_d)
F1	A	A	A	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	A	T	T	T	0.2	-0.70
F18	A	A	A	T	T	T	G	T	T	G	A	A	T	T	T	T	C	A	G	C	A	A	A	T	T	T	T	2	0.30
F31	A	A	A	T	T	T	G	T	A	G	A	A	T	T	T	T	C	T	G	C	A	A	A	T	T	T	T	21	1.32
F32	A	A	A	T	T	T	G	G	A	G	G	A	A	T	T	T	T	C	T	C	C	A	A	A	T	T	T	28	1.45



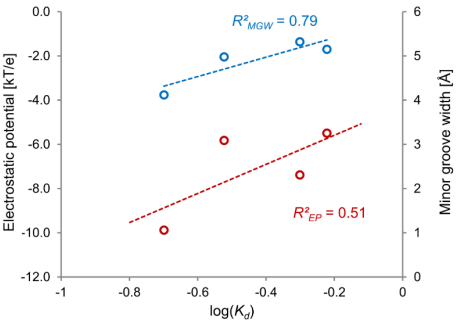
D

Label / Pos	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	K_d	log(K_d)
F1(+-8T)	A	A	A	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	A	T	T	T	0.2	-0.70
F1+-8A	A	A	A	T	T	A	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	T	A	A	T	T	T	2.1	0.32
F1+-8C	A	A	A	T	T	C	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	G	A	A	T	T	T	0.7	-0.15
F1+-8G	A	A	A	T	T	G	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	C	A	A	T	T	T	30	1.48



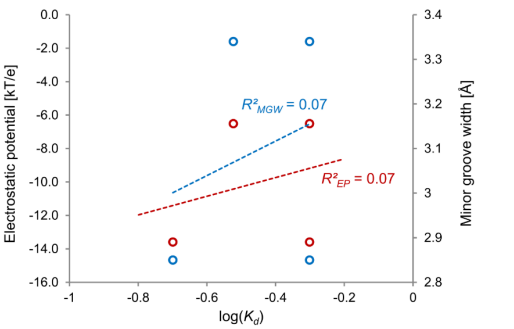
E

Label / Pos	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	K_d	log(K_d)
F1(+-8T)	A	A	A	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	A	T	T	T	0.2	-0.70
F1+-9A	A	A	A	T	A	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	T	A	T	T	T	0.5	-0.30
F1+-9C	A	A	A	T	C	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	G	A	T	T	T	0.3	-0.52
F1+-9G	A	A	A	T	G	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	C	A	T	T	T	0.6	-0.22



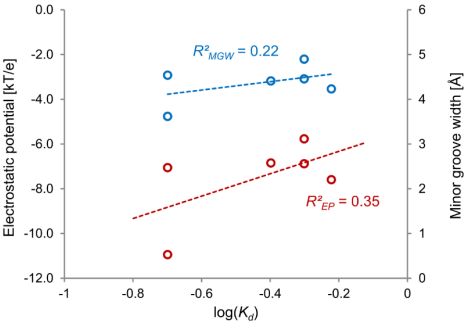
F

Label / Pos	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	K_d	log(K_d)
F1(+-8T)	A	A	A	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	A	T	T	T	0.2	-0.70
F1+-10A	A	A	A	A	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	T	T	T	T	0.5	-0.30
F1+-10C	A	A	A	C	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	G	T	T	T	0.3	-0.52
F1+-10G	A	A	A	G	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	C	T	T	T	0.5	-0.30



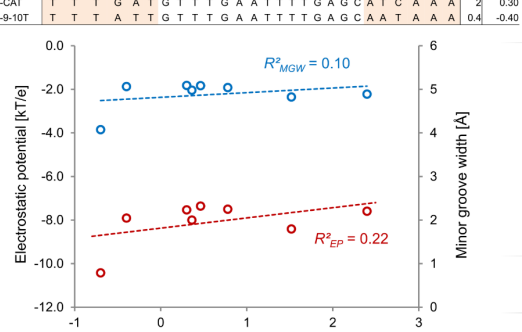
G

Label / Pos	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	K_d	log(K_d)
F1(+-8T)	A	A	A	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	A	T	T	T	0.2	-0.70
F14	G	G	G	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	C	C	C	C	0.5	-0.30
F15	C	C	C	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	G	G	G	G	0.4	-0.40
F16	G	C	G	G	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	C	C	G	C	0.5	-0.30
F33	A	A	A	G	G	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	C	C	T	T	T	0.6	-0.22
F34	A	A	G	C	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	C	G	T	T	0.2	-0.70

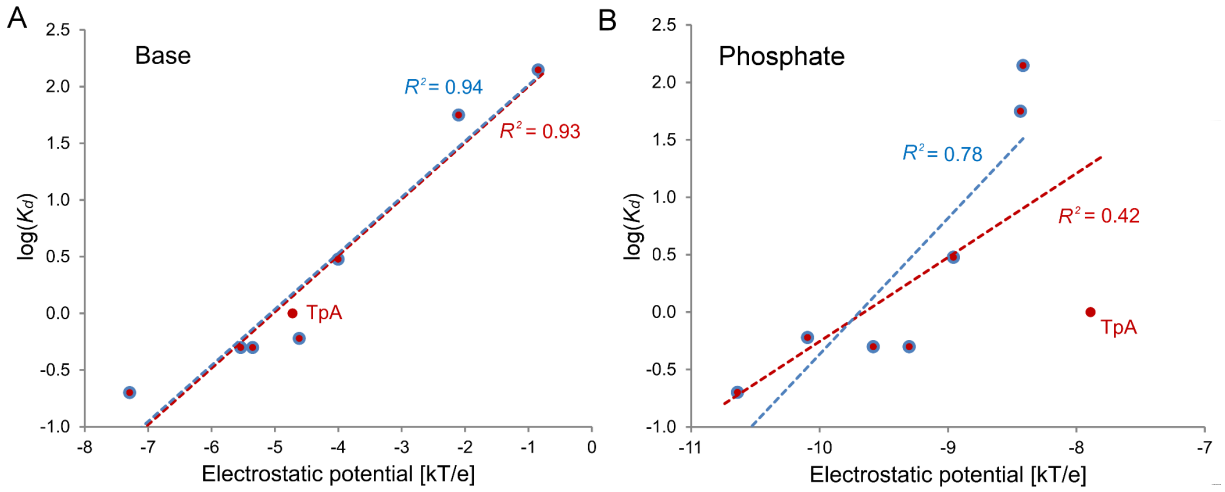


H

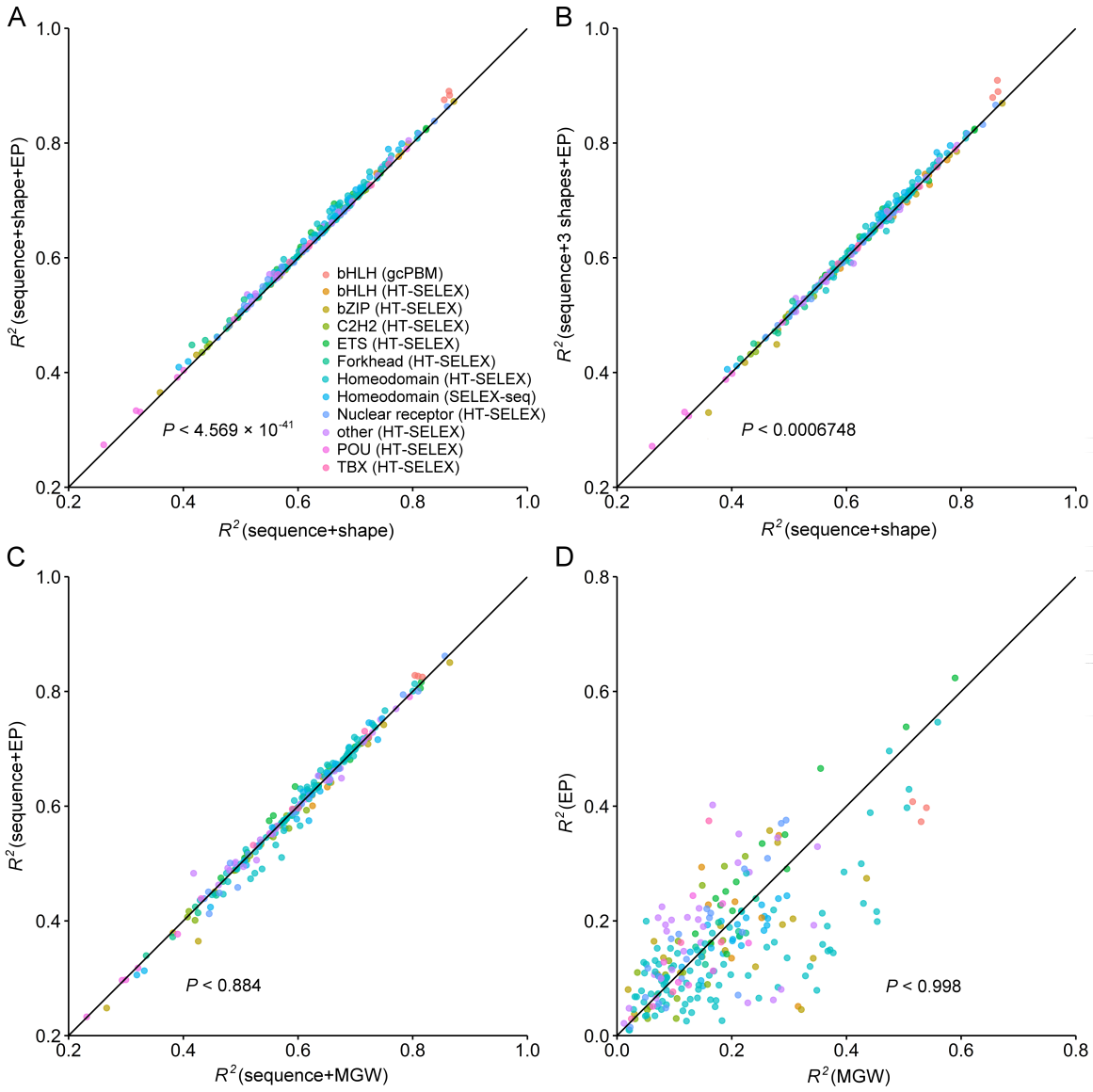
Label / Pos	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	K_d	log(K_d)
F1(+-8T)	A	A	A	T	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	A	T	T	T	0.2	-0.70
INV	T	T	T	A	A	A	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	T	T	T	A	A	A	33	1.52
INV+-8G	T	T	T	A	A	G	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	C	T	T	A	A	A	250	2.40
INV+-8C	T	T	T	A	A	C	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	G	T	T	A	A	A	6	0.78
INV+-8T	T	T	T	A	A	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	T	T	A	A	A	2.3	0.36
INV+-CAT	T	T	T	C	A	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	T	G	A	A	A	2.9	0.46
INV+-CAT	T	T	T	G	A	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	T	C	A	A	A	2	0.30
INV+-9-10T	T	T	T	A	T	T	G	T	T	T	G	A	A	T	T	T	T	G	A	G	C	A	A	T	A	A	A	0.4	-0.40



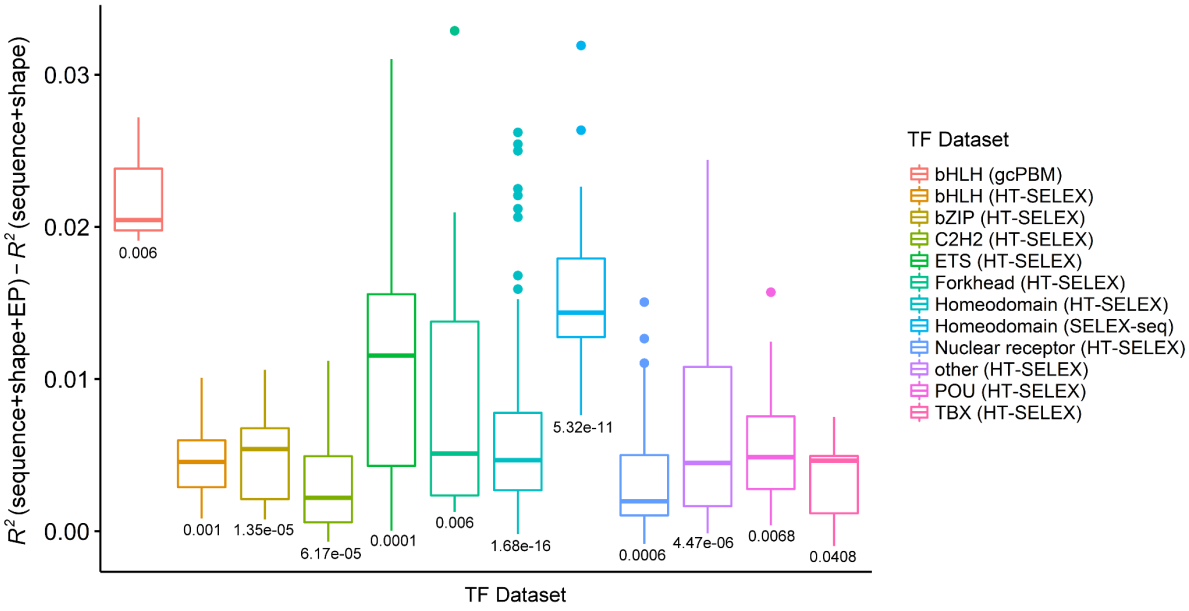
Supplementary Figure S6. HT predictions of MGW and EP for the mutated regions with respect to the high-affinity Fis binding site F1. The reference Fis-F1 co-crystal structure (PDB ID 3IV5) and quantitative Fis binding site logo (7) are shown in (A). The altered nucleotide positions include (B) mutations introducing asymmetry in the structure, (C) mutations introducing a YpR bp step at the ±(4-5) positions (F18) or ±(5-6) positions (F31) or eliminating the YpR bp steps (F32), mutations substituting bp flanking the core at the ±8 positions (D), ±9 positions (E), ±10 positions (F) or further outside the core binding site (G) and (H) mutations inverting AT-rich flanking sequences. We highlighted mutated regions in the respective table (orange shaded columns). HT predictions of MGW and EP within the mutated regions of the Fis binding sites correlate with the logarithm of binding affinity log(K_d). Correlation coefficients R^2 between the logarithm of K_d and EP (blue) or MGW (red) were calculated for all Fis binding sites.



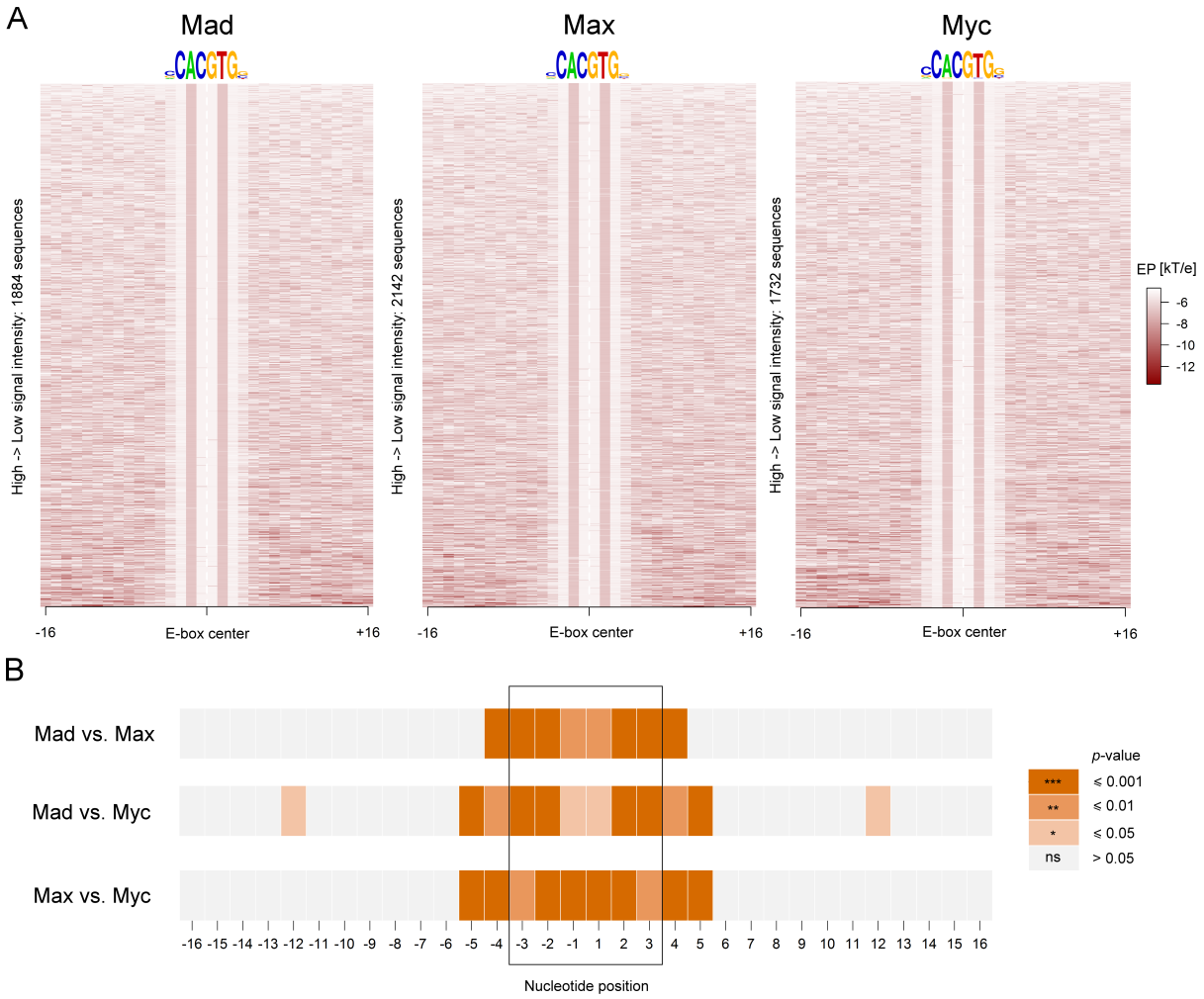
Supplementary Figure S7. HT predictions of EP based on the partial charge of (A) bases and (B) phosphates over the five central bp of eight Fis binding sites correlated with the logarithm of binding affinity K_d. Contribution from the partial base charges has a higher correlation with the logarithm of binding affinity than the contribution from phosphate groups, particularly when including the sequence with a central TpA bp step.



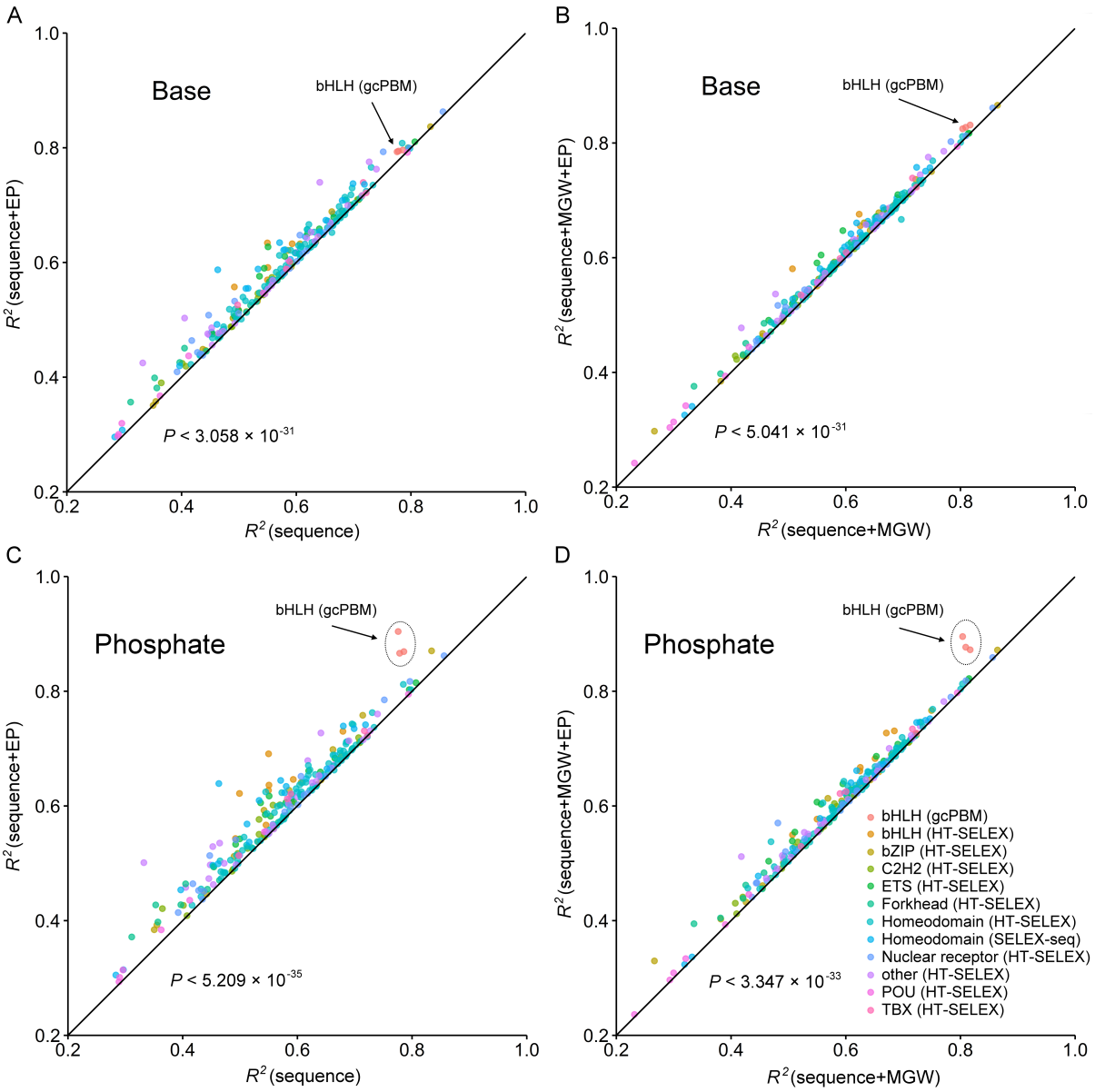
Supplementary Figure S8. Performance comparison of binding specificity predictions for multiple TFs derived from HT-SELEX, SELEX-seq and gcPBM HT binding assays. (A) Comparison of sequence+shape+EP and sequence+shape models. (B) Comparison of sequence+3shapes+EP models with three shape features (HelT, ProT and Roll) and sequence+shape models with four shape features (HelT, ProT, Roll and MGW). (C) Comparison of sequence+EP and sequence+MGW models. (D) Comparison of EP and MGW models. The *P* values were calculated by the *t*-test hypothesis testing method with performance increase in terms of R^2 as the alternative hypothesis.



Supplementary Figure S9. Box plot analysis for quantifying performance gain of EP-augmented models across TF families. Sequence+shape+EP models were compared with sequence+shape models for 239 TFs. Most of the sequence+shape+EP models outperform sequence+shape models (225 of 239 tested proteins). The *P* values were calculated by using the *t*-test hypothesis testing method with performance increase in terms of R^2 as the alternative hypothesis. The *P* value for each TF dataset (TF family/HT binding experiment) is shown below the corresponding box plot.



Supplementary Figure S10. Minor-groove EP preferences of human bHLH TFs. (A) Heat maps illustrate minor-groove EP preferences of the Mad1/Max heterodimer ('Mad'), Max/Max homodimer ('Max') and c-Myc/Max heterodimer ('Myc'). Sequence data were derived from gcPBM experiments (22) using the top-25% signal intensities. (B) Nucleotide positions with significant EP differences based on a *t*-test are indicated for comparisons of Mad vs. Max, Mad vs. Myc, and Max vs. Myc (nucleotide positions with different minor-groove EP distributions are shown in different colors for *P* value). EP features were symmetrized based on the palindromic E-box, which is located at the central positions –3 to +3 marked by a black frame.



Supplementary Figure S11. Performance comparison of binding specificity predictions for TFs derived from HT-SELEX, SELEX-seq and gcPBM binding assays based on LPB calculations (for bases and phosphate groups). (A+C) Sequence+EP models outperform sequence-only models and contribute to the prediction accuracy of DNA binding specificities based on L2-regularized MLR and 10-fold cross-validation. (B+D) Sequence+shape+EP models outperform sequence+MGW models. The *P* values were calculated by the *t*-test hypothesis testing method with performance increase in terms of R^2 as the alternative hypothesis.

SUPPLEMENTARY TABLE

Supplementary Table S1. Eight sequences and logarithms of binding affinities of Fis-DNA sites. Co-crystal structures were solved for six of these binding sites.

Label	PDB ID	Sequence	log(K_d)	Reference
F1	3IV5	AAATTTGTTTGAATTTTGAGCAAATTT	-0.69897	(7,8)
F24	3JRB	AAATTTGTTTGTTTTTTGAGCAAATTT	-0.30103	(7,8)
F25	3JRD	AAATTTGTTTGTTAAATGAGCAAATTT	0	(7,8)
F26	3JRE	AAATTTGTTTGAAAAATGAGCAAATTT	-0.30103	(7,8)
F27	3JRF	AAATTTGTTTGAACTTTGAGCAAATTT	-0.22185	(7,8)
F28		AAATTTGTTTGAGCGTTGAGCAAATTT	1.748188	(7,8)
F29	3JRC	AAATTTGTTTGGGCGCTGAGCAAATTT	2.146128	(7,8)
F36		AAATTTGTTTGAATCT**C**GAGCAAATTT	0.477121	(7)

AUTHOR CONTRIBUTIONS

T.P.C. performed MC simulations and Poisson–Boltzmann calculations, designed and validated the DNAphi method, carried out the statistical ML work, and updated the DNAshapeR/Bioconductor package to include EP. S.R. contributed to the statistical ML work. R.S.M. and B.H. contributed to conception of the project, analysis of the data and writing of the manuscript. T.P.C. and R.R. wrote the manuscript. R.R. conceived and supervised the project.

SUPPLEMENTARY REFERENCES

1. Honig, B. and Nicholls, A. (1995) Classical electrostatics in biology and chemistry. *Science*, **268**, 1144-1149.
2. Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248-1253.
3. Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A. and Honig, B. (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.*, **23**, 128-137.
4. Harris, R.C., Mackoy, T., Dantas Machado, A.C., Xu, D., Rohs, R. and Fenley, M.O. (2012) Opposites attract: shape and electrostatic complementarity in protein-DNA complexes. Chapter 3, vol. 2. In: T. Schlick (ed.) *Innovations in Biomolecular Modeling and Simulations*, 53-80. The Royal Society of Chemistry. Cambridge, UK.
5. Chiu, T.P., Comoglio, F., Zhou, T., Yang, L., Paro, R. and Rohs, R. (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, **32**, 1211-1213.
6. Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R. and Mann, R.S. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell*, **161**, 307-318.
7. Hancock, S.P., Stella, S., Cascio, D. and Johnson, R.C. (2016) DNA Sequence Determinants Controlling Affinity, Stability and Shape of DNA Complexes Bound by the Nucleoid Protein Fis. *PLOS ONE*, **11**, e0150189.
8. Stella, S., Cascio, D. and Johnson, R.C. (2010) The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes Dev.* **24**, 814-826.
9. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861-873.
10. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327-339.
11. Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R. and Rohs, R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.

12. Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R. and Rohs, R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. USA*, **112**, 4654-4659.
13. Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep Iii, P.W. and Bulyk, M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429-1435.
14. Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R. and Bulyk, M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093-1104.
15. Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270-1282.
16. Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R. and Rohs, R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56-W62.
17. Passner, J.M., Ryoo, H.D., Shen, L., Mann, R.S. and Aggarwal, A.K. (1999) Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature*, **397**, 714-719.
18. Watkins, S., van Pouderoyen, G. and Sixma, T.K. (2004) Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. *Nucleic Acids Res.*, **32**, 4306-4312.
19. Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B. and Mann, R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530-543.
20. Tan, S. and Richmond, T.J. (1998) Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature*, **391**, 660-666.
21. Sagendorf, J.M., Berman, H.M. and Rohs, R. (2017) DNAproDB: an interactive tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res.*, **45**, W89-W97.
22. Mordelet, F., Horton, J., Hartemink, A.J., Engelhardt, B.E. and Gordân, R. (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics*, **29**, i117-i125.