SUPPLEMENTARY DATA

Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding

Jinsen Li [1], Jared M. Sagendorf [1], Tsu-Pei Chiu[1], Marco Pasi[2], Alberto Perez[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]Centre for Biomolecular Sciences and School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, UK

[3]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, 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

Preparation of the Universal Protein Binding Microarray (uPBM) Dataset

An analysis of uPBM data (1, 2) was compared with results from other experimental binding assays. The uPBM dataset used in this study contained data for 66 mouse transcription factors (TFs) used in the DREAM5 challenge (3). Briefly, uPBM data comprise DNA sequences designed to ensure that all possible 10-bp sequences are represented on the array. The pre-processing of uPBM datasets from the DREAM5 challenge (3) was previously described (4). These data contain sequences of shorter length than sequences from gcPBM (5), SELEX-seq (6), or HT-SELEX (7) datasets.

Monte Carlo (MC) Simulations

All-atom MC simulations of 2,121 DNA fragments of 12–27 bp in length were analyzed to derive DNA shape features. These previously reported MC simulations (8) were performed over 2 million MC cycles with standard B-DNA used as starting configuration (9). MC sampling was based on collective and internal variables combined with analytic chain closure using associated Jacobians (10). Energy calculations employed the AMBER force field (11) and implicit solvent combined with explicit sodium counter ions (12). Average structures were generated by using a previously described simulation protocol (9) and analyzed with CURVES (13). Resulting three-dimensional MC predictions for 2,121 DNA sequences were mined in terms of the 512 unique pentamers (data used in our DNAshape method (8) and labeled as “DNAshape”) or 136 unique tetramers (resulting dataset labeled as “MC4”).

Molecular Dynamics (MD) Simulations

MD trajectories of the µABC dataset of 39 double-stranded B-DNA oligomers (14) were analyzed to obtain information on the microsecond-scale average tetranucleotide-dependent structure of B-DNA (15). In particular, each oligomer in the µABC dataset was 18-bp long, constructed by repeating a particular 4-bp sequence *ABCD* three and a half times: 5′-GC-*CD-ABCD-ABCD-ABCD*-GC-3’. Sequences were designed such that the resulting set of 18-mers covered each of the 136 distinct tetranucleotide sequences (resulting dataset labeled as “MD4”) with a minimum of three occurrences. One-microsecond MD simulations on each of the oligomers in the µABC dataset were performed in explicit solvent (16) with a physiological concentration of K^+Cl^- (17) by using the parmbsc0 modifications to the AMBER parm99 force field. CURVES+ (18) was used to obtain the ensemble averages and standard deviations (SDs) of helical parameters from the centermost occurrence of each tetranucleotide (14, 15). The first 10% of the trajectories were excluded from the analysis. Further details on the MD simulation protocol and analysis were previously described (15).

Feature Encoding and Normalization

For a single sequence, we first encoded its sequence features as 1-mers, 2-mers, or 3-mers, using the methods noted in the main manuscript, and encoded shape features for that sequence. We used the DNAshape method (8) within the DNAshapeR package (19) to generate these features. For a single shape feature category, such as MGW, we predicted values for a given sequence as a vector of $\{MGW_3, MGW_4, \dots, MGW_{n-2}\}$ (the first two positions are not predictable because the method is pentamer-based). To be better compatible with regression models, predicted shape features were normalized by subtracting a global minimum value acquired from the respective query table (see Materials and Methods) and dividing by an SD value. This SD value was calculated for each shape feature and based on all values for this shape feature category from a given dataset. The SD value used in normalization was computed separately for each individual dataset. In addition, if a position weight matrix (PWM) was known for a particular TF and it was palindromic, we averaged each feature vector, including the sequence feature and shape features complemented by the reverse complement.

Once each feature was encoded, normalized, and complemented by its reverse palindrome (if applicable), we concatenated all features in a final feature vector:

$$\{\overrightarrow{S_1},\ldots,\overrightarrow{S_n}; MGW_3,\ldots,MGW_{n-2}; HelT_2,\ldots,HelT_{n-1};\ldots\}$$

where $\overrightarrow{S_i}$ denotes the *k*-mer sequence feature at the i-th position, which should have a length of four for 1mer, 16 for 2mer, and 64 for 3mer features. Other features, such as MGW and HelT, denote shape features; and n denotes the sequence length.

Mean Squared Error (MSE) Calculation

MSE is closely related to R^2 and can be computed as follows:

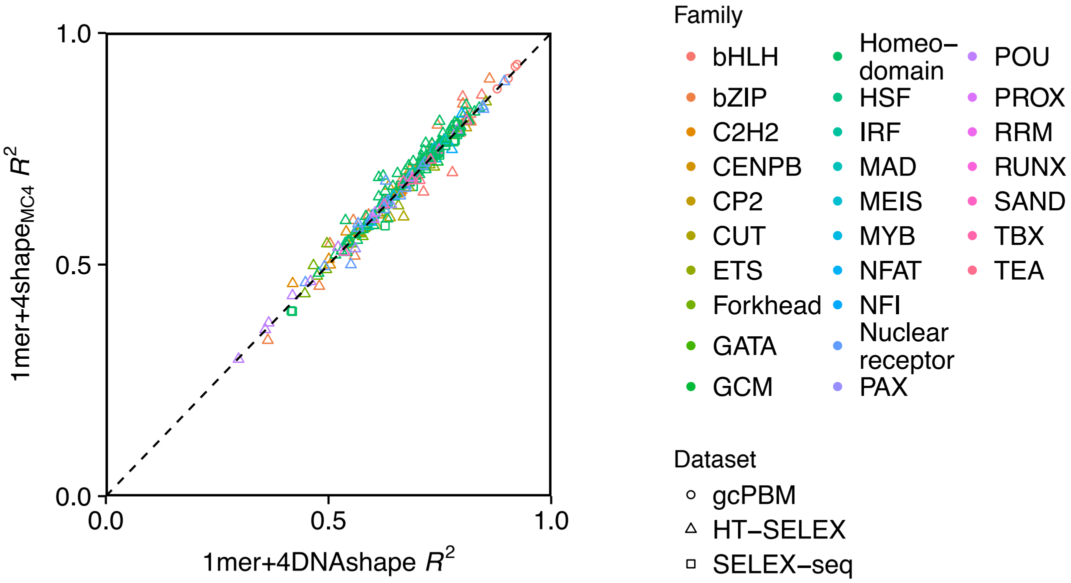
$$\mathrm{MSE} = \frac{\sum_i (y_i - \hat{y}_\mathrm{i})^2}{n-m}$$

where y_i and $\hat{y}_i$ represent the observed and predicted binding affinity scores, respectively; n represents the sample size (e.g., total number of sequences); and m represents the length of the feature vector, including the intercept. Therefore, when R^2 is higher, MSE can also be higher because the length of the feature vector affects it.

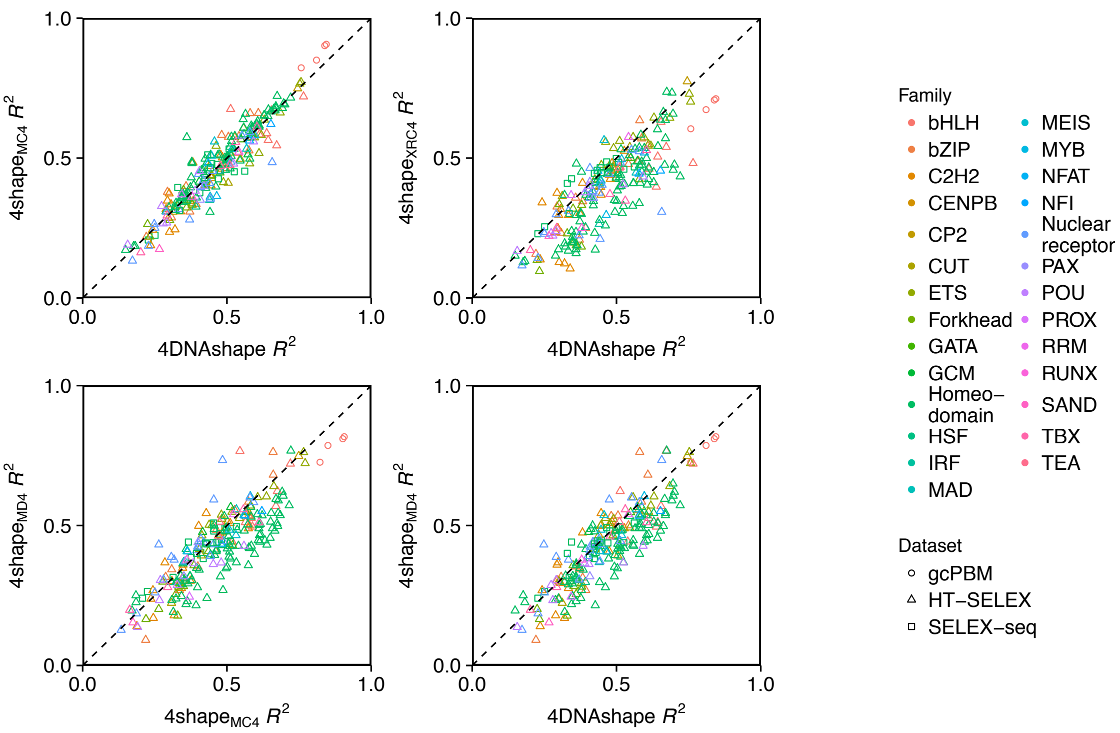
Limitations of DynaSeq-derived Shape Features

While our work was in revision, a different group published alternative DNA shape features derived only from MD simulations (20). These “DynaSeq” shape features have several differences compared to the MD features derived from the µABC dataset (see *MD Simulations* section). Firstly, the DynaSeq shape features were derived from a sequence design in which each of the 136 unique tetramers occurred only once and were always flanked 5’ and 3’ by GCGC tetrads (21). The hydration patterns of GC-rich sequences have been shown to differ from those of other sequence environments; thus, the use of a single sequence environment might not fully sample the conformational space of the central tetramer (22). Secondly, DynaSeq features were based on MD simulations that were an order of magnitude shorter than the µABC data used here. Thirdly, DynaSeq used the Bsc0 force field (23), despite availability of the Bsc1 force field that could improve the conformational sampling of DNA and the accuracy of MD simulations (24). Moreover, DynaSeq-derived shape features were not validated through direct comparison with experimental data. Instead, they were validated indirectly through the ability to identify a sequence within a random set of other sequences based on structural information. However, this validation included many DNA structures affected by drug binding and even bp mismatches (20).

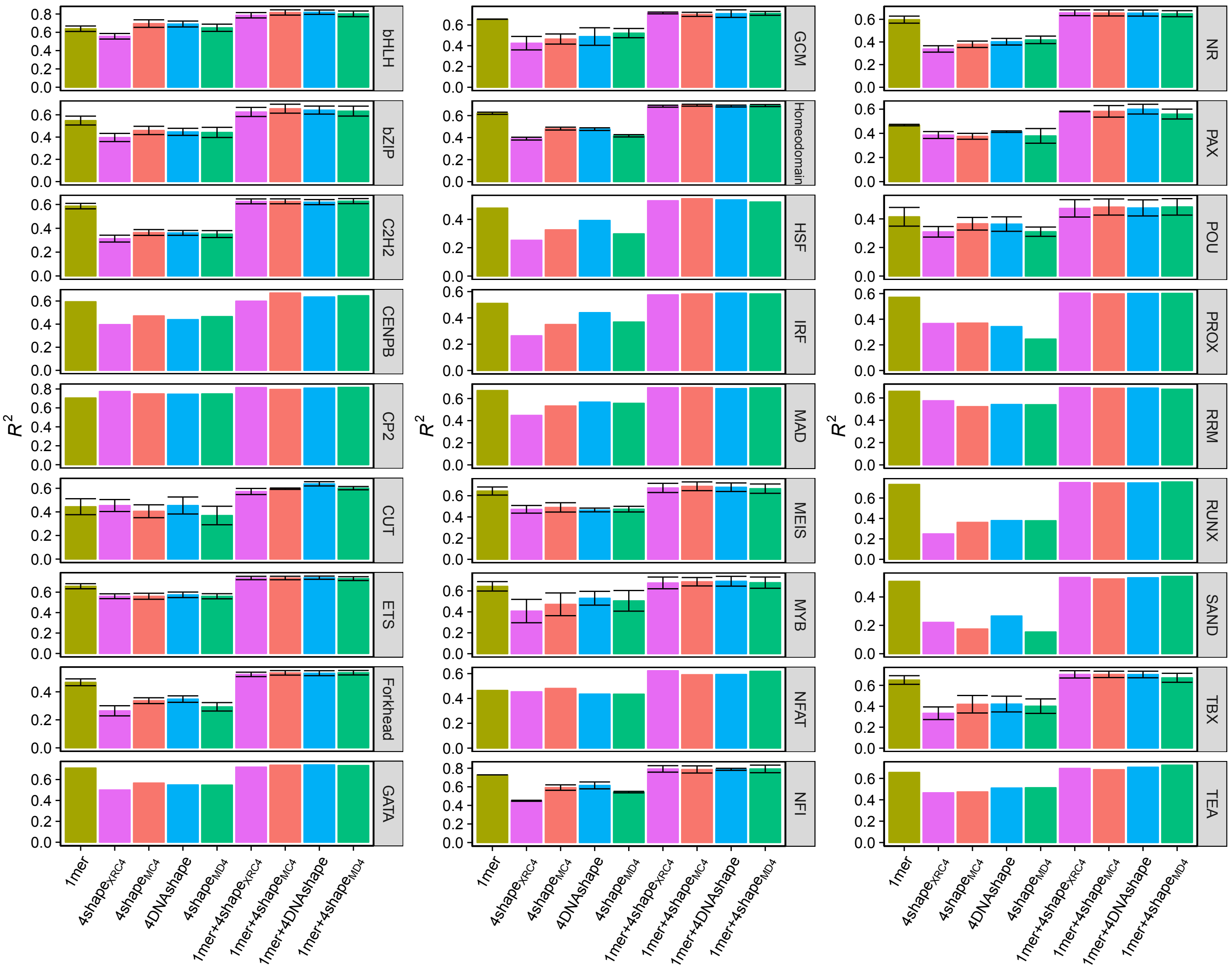
SUPPLEMENTARY FIGURES



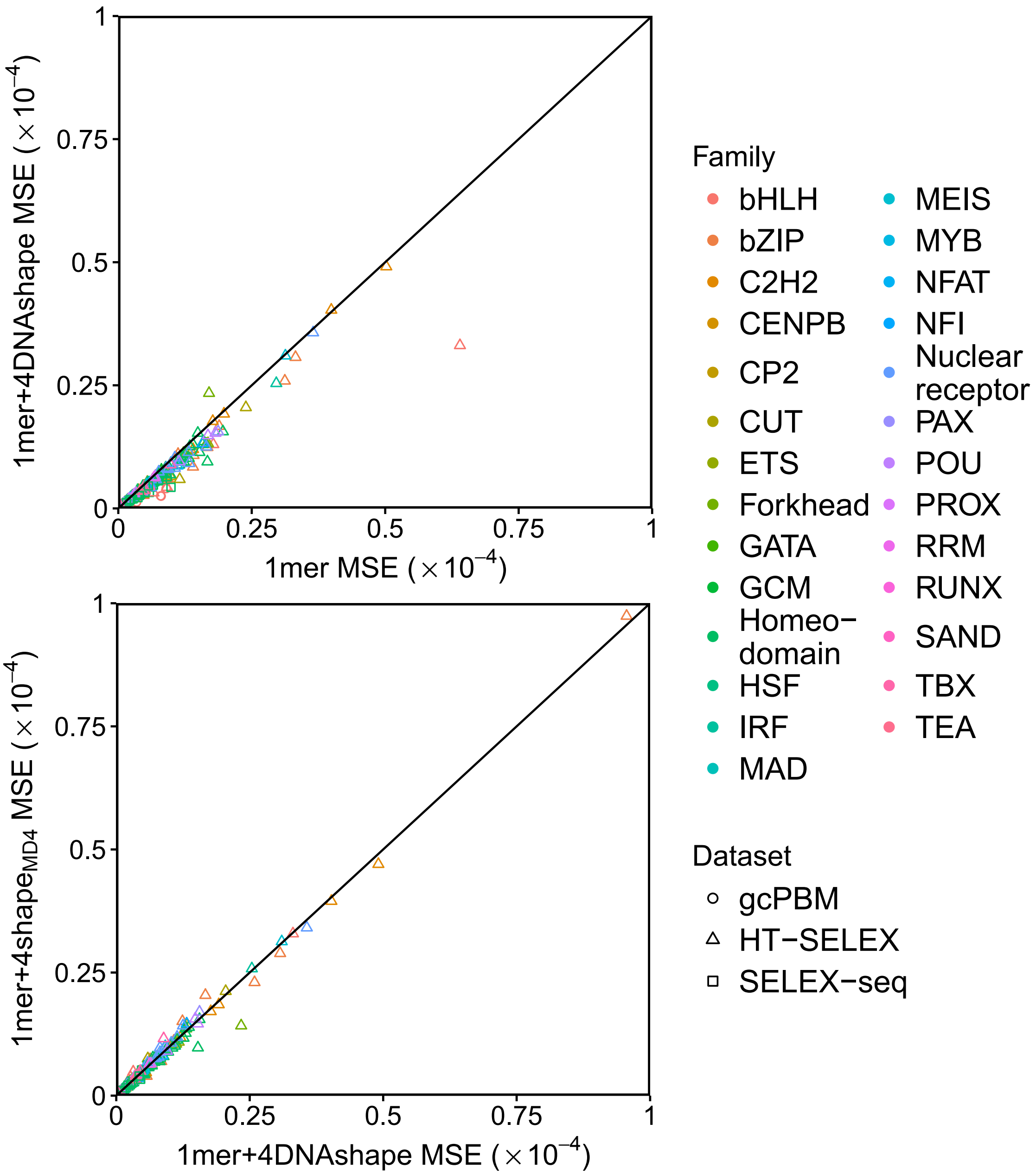
Supplementary Figure S1. Performance comparison using the four shape features MGW, ProT, HelT, and Roll from MC-derived pentamers combined with 1mer sequence (1mer+4DNAshape model) and MC-derived tetramer query tables combined with 1mer sequence (1mer+4shape$_{MC4}$ model) for multiple datasets and TF families.



Supplementary Figure S2. Performance comparison between shape-only models (MC-derived tetramer-based 4shape$_{MC4}$, MC-derived pentamer-based 4shape$_{DNAshape}$, and MD-derived tetramer-based 4shape$_{MD4}$ models) for multiple datasets and TF families. We note that 1mer sequence features were not included in these models.



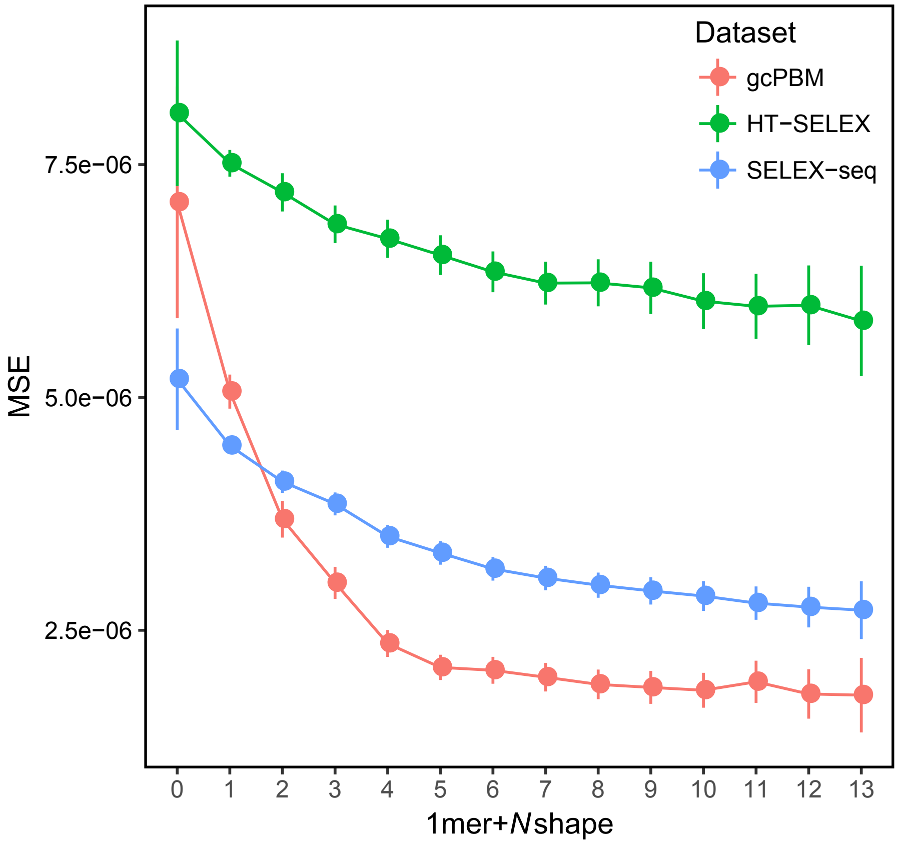
Supplementary Figure S3. Summary of the weighted-average performances for different models. Shape features used here were MGW, ProT, HelT, and Roll (representing the four features derived from MC, MD, or XRC data by mining the 136 unique tetramers, or the pentamer-based DNAshape (8) method). Error bars were calculated based on the standard error of the mean. Model performances were divided into groups based on TF families used in this study. Identical color was applied for models using the same source of shape features. Among labels for TF families, NR represents the nuclear receptor family.



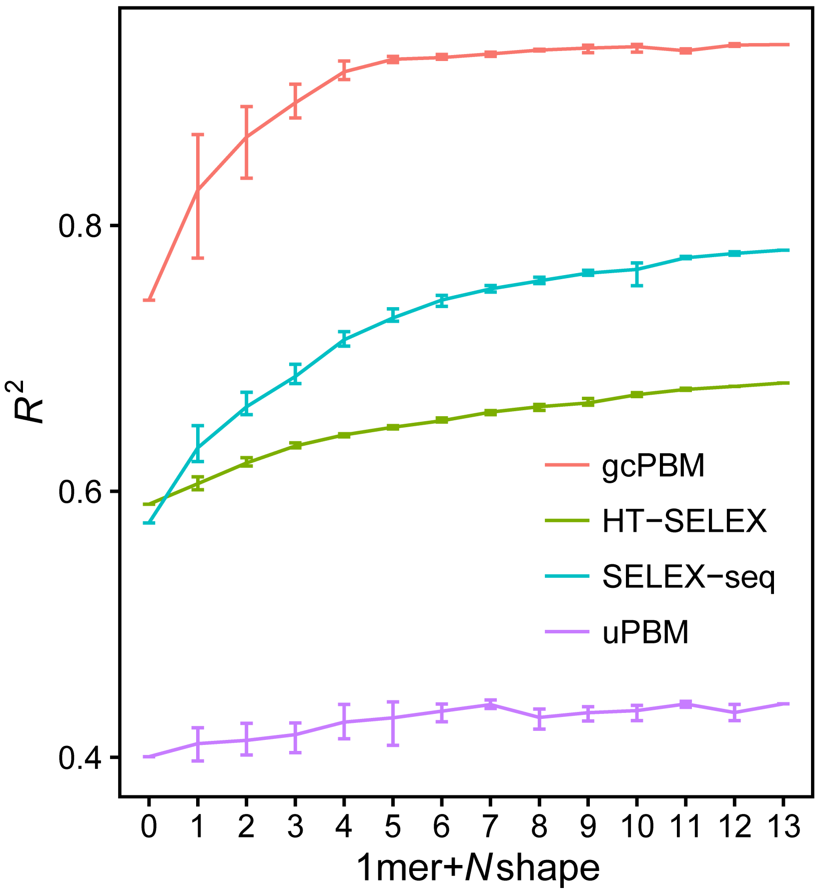
Supplementary Figure S4. Mean squared error (MSE) for different TF datasets for different models. A few datasets may have larger MSE values that are outside the plot.

A) Performance of 1mer models compared to models augmented by four MC-derived and pentamer-based shape features (1mer+4DNAshape models).

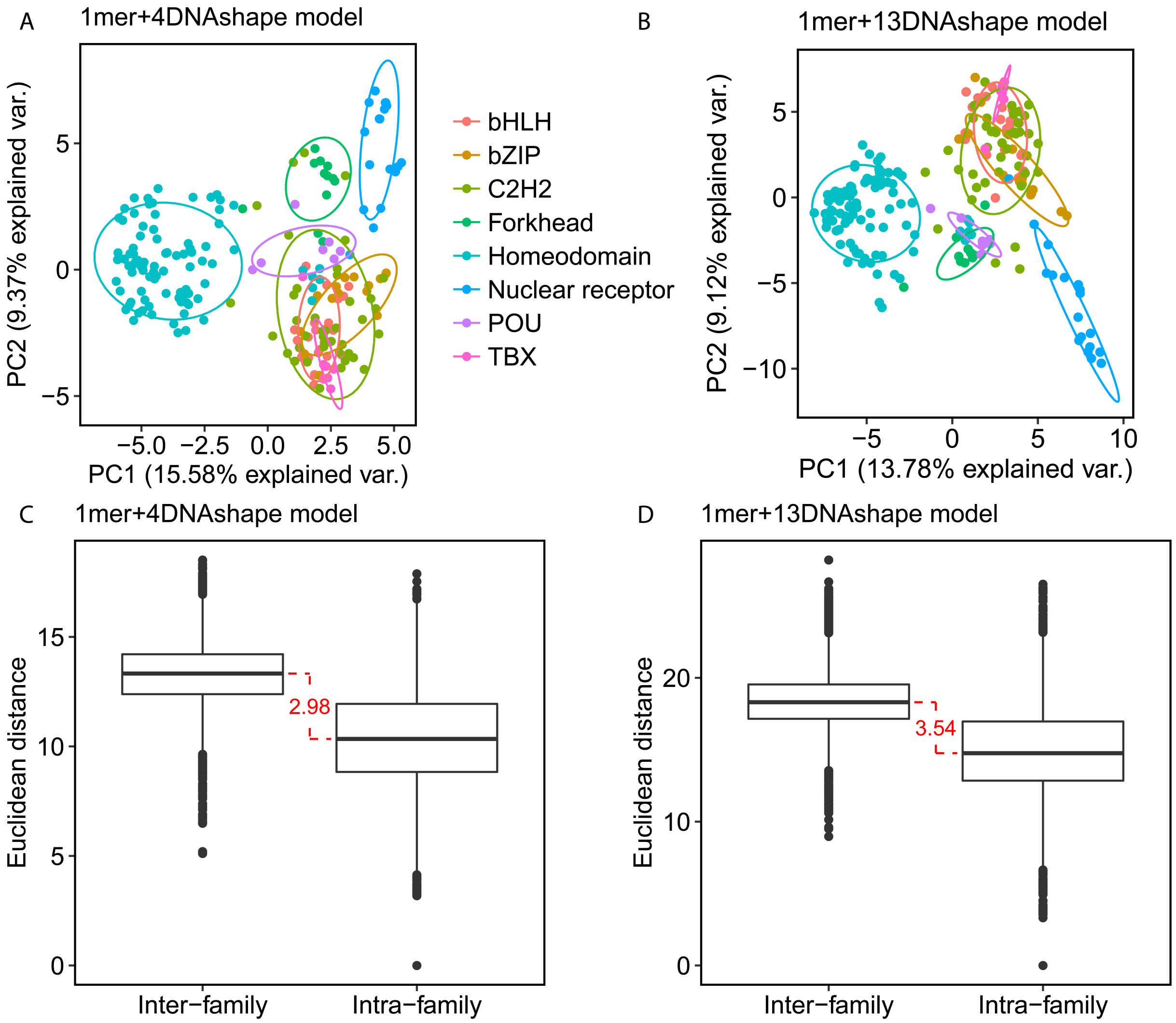
B) Performance of 1mer+4DNAshape models compared to MD-derived and tetramer-based 1mer+4shape$_{MD4}$ models.



Supplementary Figure S5. MSE decreases as *N* increases in 1mer+*N*shape models. MC-derived and pentamer-based DNAshape features (8) were used in this analysis. Each dot represents the mean among all datasets in each category, whereas the error bar represents the standard error of the datasets in each experimental category.

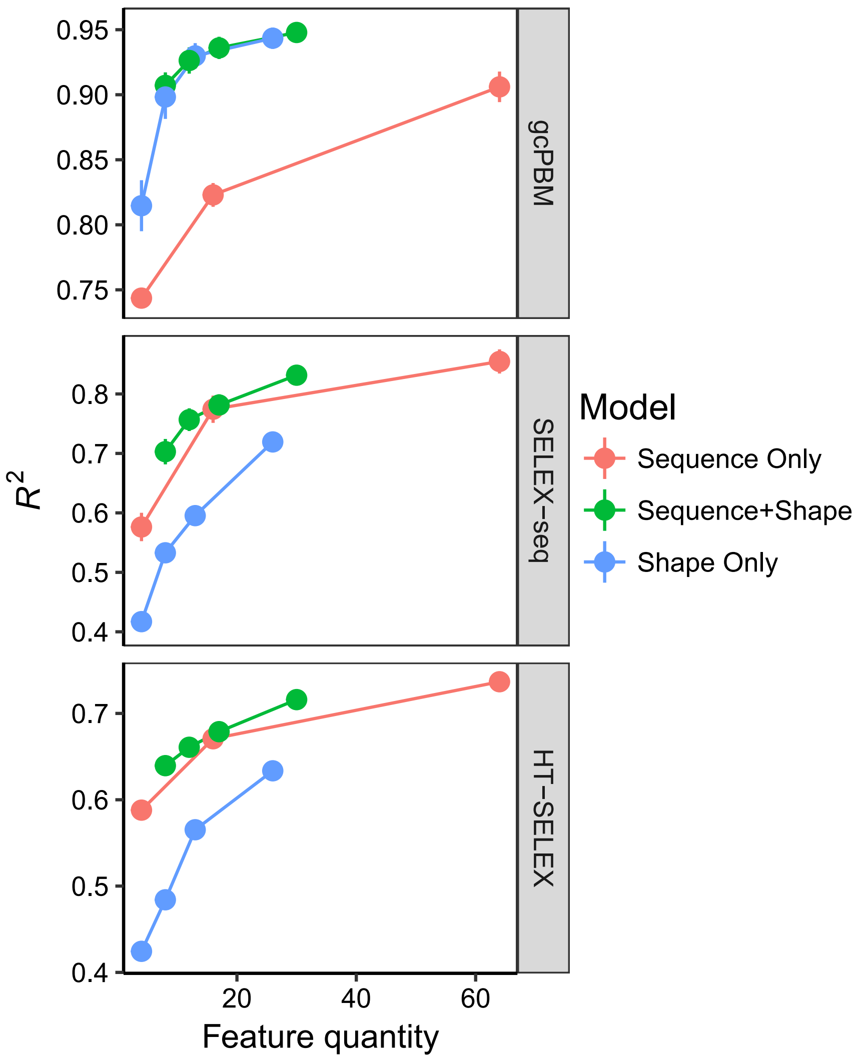


Supplementary Figure S6. Plot equivalent to Figure 5B, with the addition of the DREAM5 uPBM dataset (3), generated after repeatedly adding shape features into the previous model. Shape features used in this plot were only derived from MC data based on pentamers, denoted “DNAshape” (8). In each round, the most-informative feature was added based on model performance by using the DNAshape query table. Performance measures were calculated based on weighted mean R^2 among all datasets in each experimental category. Error bars indicate maximum and minimum performances when *N* shape features were added.

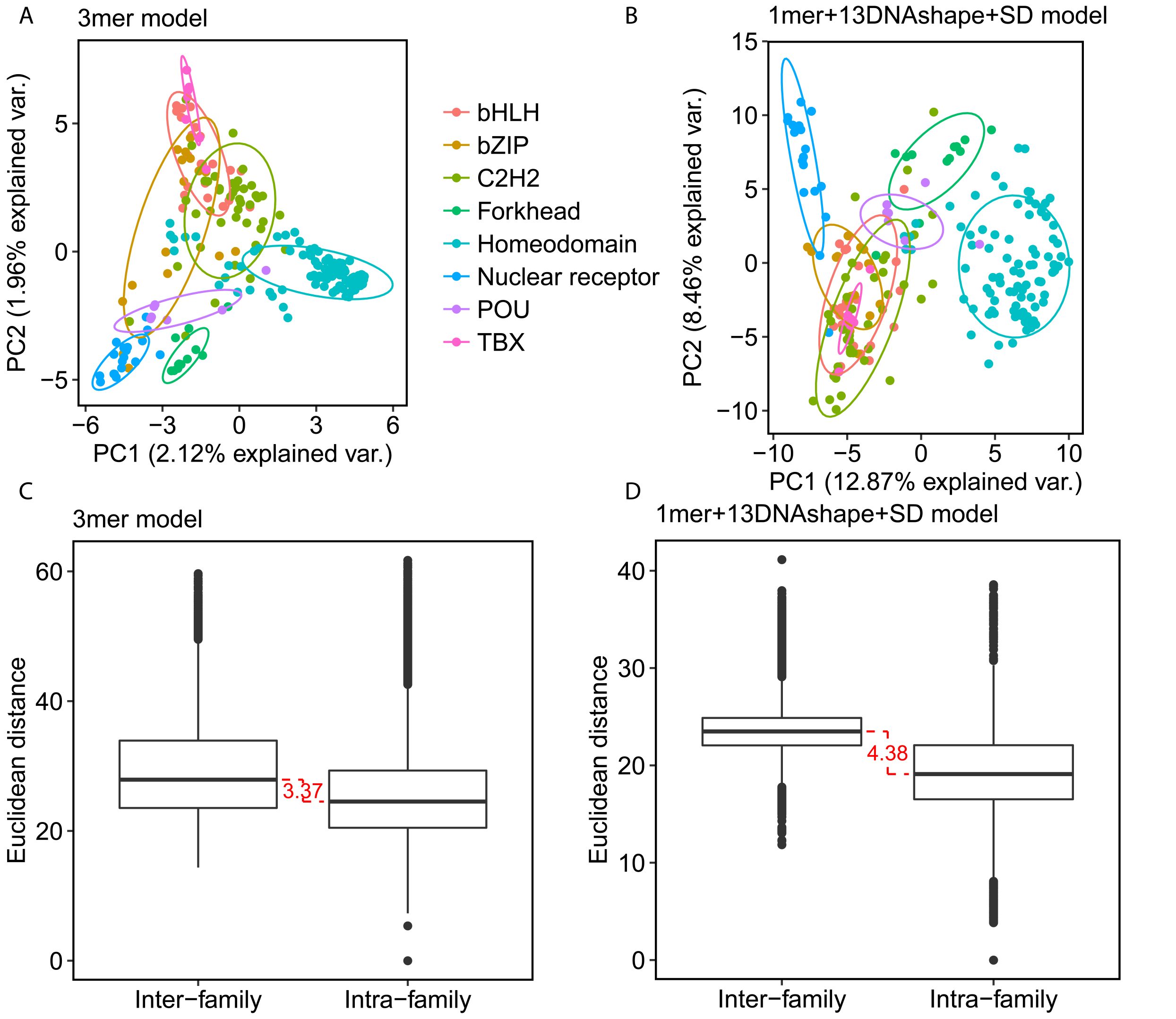


Supplementary Figure S7. Principal component analysis (PCA) reveals different DNA-binding specificities within and between TF families (intra- vs. inter-family).

A) PCA results using 1mer+4shape features. Each dot represents a TF dataset. Dots of identical color belong to the same TF family. An ellipse was drawn for each TF family representing the 0.68 probability contour of the respective fitted two-variate normal distribution.
B) PCA results using 1mer+13shape features (i.e., nine shape features added with respect to (A)).
C) Boxplots of inter- and intra-family TF distances derived from (A).
D) Boxplots of inter- and intra-family TF distances derived from (B).

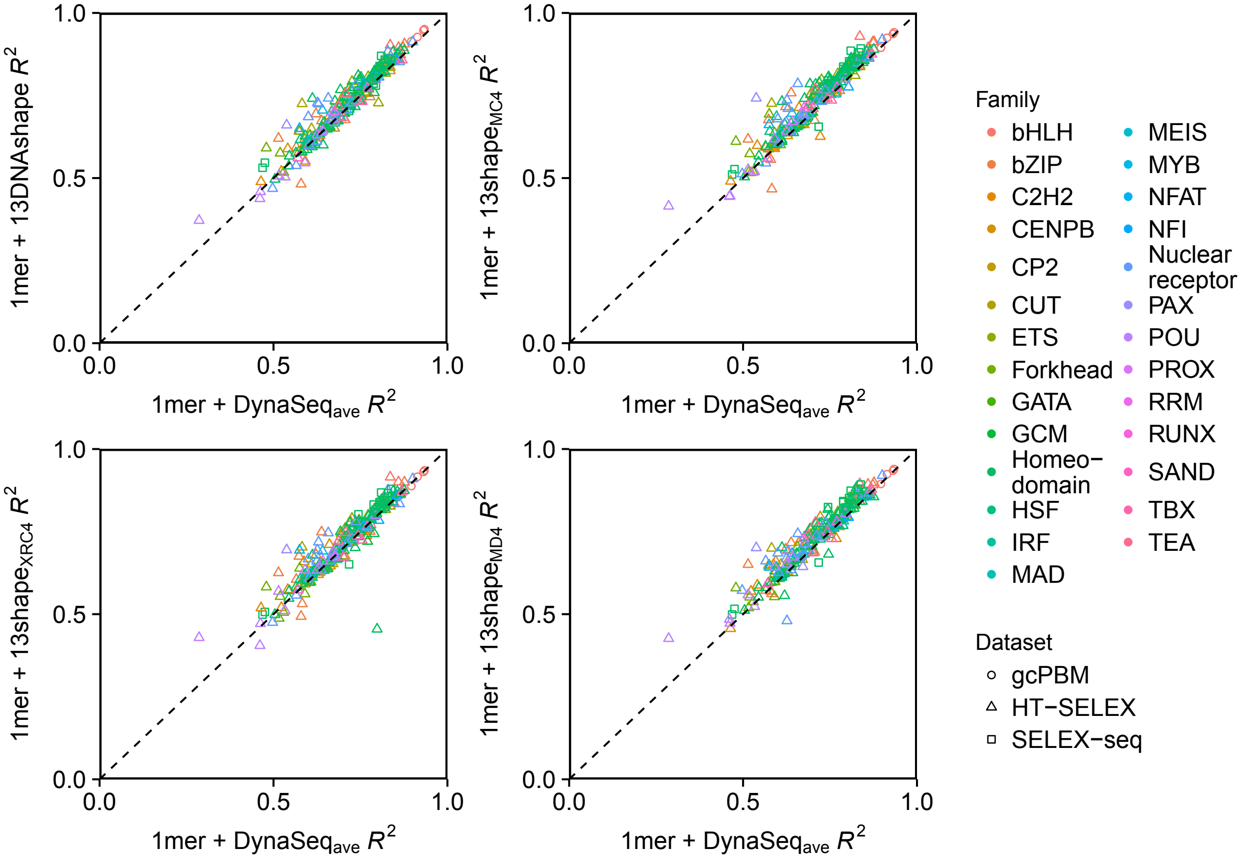


Supplementary Figure S8. Comparison between feature quantity and R^2 performance in three different model settings. Points shown in “Sequence+Shape” group are 1mer+4shape, 1mer+4shape+SD, 1mer+13shape, and 1mer+13shape+SD models. SD represents the standard deviations of shape features and, thus, conformational flexibility. Points shown in “Shape Only” group are 4shape, 4shape+SD, 13shape, and 13shape+SD models. Points shown in “Sequence Only” models are 1mer, 2mer, and 3mer models. Models in the “Sequence+Shape” group performed usually better in achieving higher R^2 scores, particularly when using fewer features.

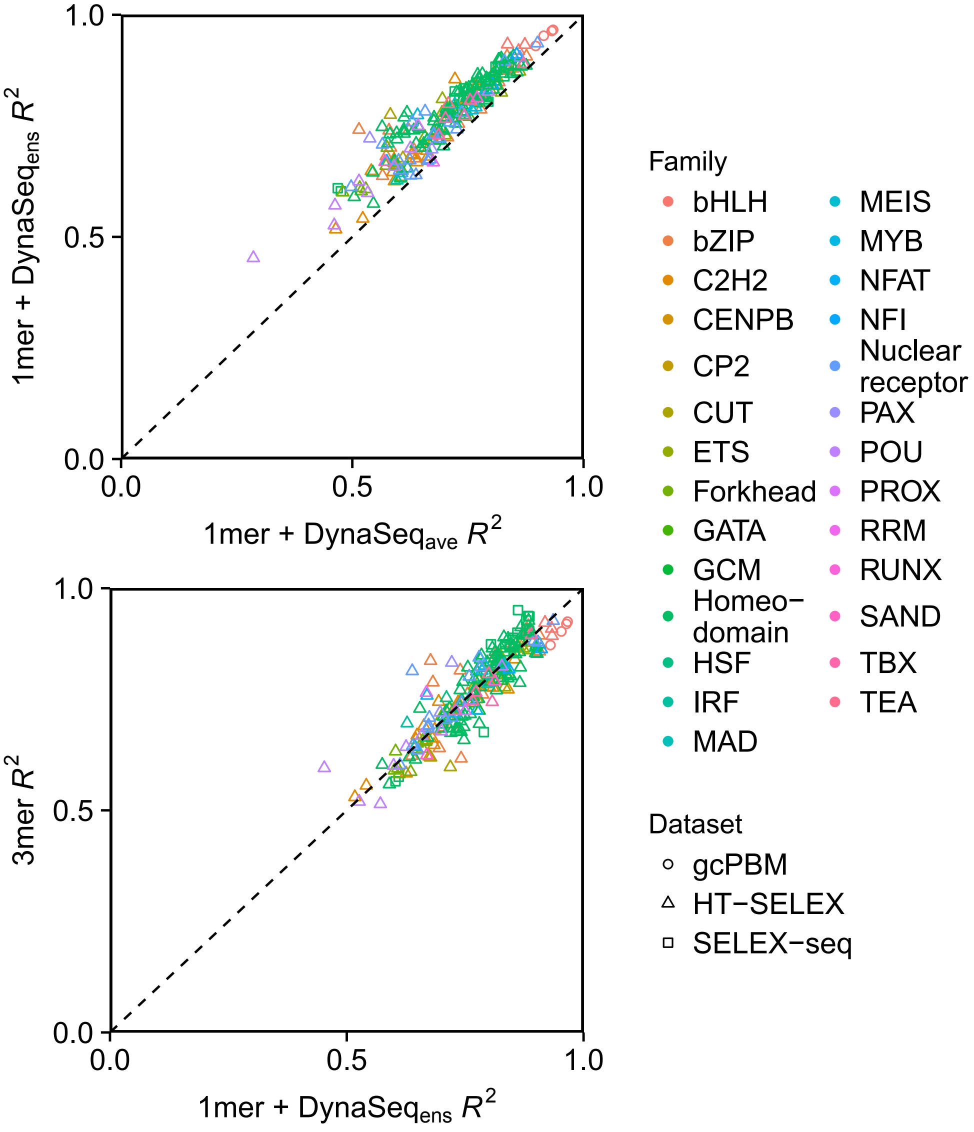


Supplementary Figure S9. Principal component analysis (PCA) reveals different DNA-binding specificities within and between TF families (intra- vs. inter-family).

A) PCA results using 3mer features. Each dot represents a TF. Dots of identical color belong to the same TF family. An ellipse was drawn for each TF family representing the 0.68 probability contour of the respective fitted two-variate normal distribution.
B) PCA results using 1mer+13shape+SD features (30 features per nucleotide position) compared to 3mer features (64 features per nucleotide position) used in (A).
C) Boxplots of inter- and intra-family TF distances derived from (A).
D) Boxplots of inter- and intra-family TF distances derived from (B).



Supplementary Figure S10. R^2 performance comparing 1mer models augmented by DynaSeq$_{ave}$ (averaged version) shape features (20) and 1mer+shape models introduced in this work. DynaSeq$_{ave}$ represents 13 shape features corresponding in number to our 13 shape features. All four models used in this research (1mer+13DNAshape, 1mer+13shape$_{MC4}$, 1mer+13shape$_{MD4}$, 1mer+13shape$_{XRC4}$) achieved higher performance in predicting TF-DNA binding specificity, indicating the limitations of the DynaSeq approach (see Supplementary Materials and Methods).



Supplementary Figure S11. R^2 performance comparing between 1mer models augmented by shape features derived from DynaSeq$_{ave}$ (averaged version) and DynaSeq$_{ens}$ (ensemble version) (20) and 3mer-based models. DynaSeq$_{ave}$ represents 13 shape features comparable to the 13 shape features introduced in this work. DynaSeq$_{ens}$ represents 5-bin ensembles of 13 shape features, which sum to a total of 65 shape features (69 when 1mer features are included in 1mer+DynaSeq$_{ens}$ models). Therefore, DynaSeq$_{ens}$ is superior to DynaSeq$_{ave}$ in this comparison. However, the 1mer+DynaSeq$_{ens}$ model with 69 features does not outperform comparable 3mer models with 64 features.

SUPPLEMENTARY TABLES

Supplementary Table S1. Spearman's rank correlation coefficients between features in different query tables and features derived from 590 protein-bound XRC structures. Coefficients were calculated based on a concatenated vector that was generated from all 590 DNA conformations (see Supplementary Table S2 for PDB IDs). This validation is equivalent to the validation used for validating the DNAshape method (8).

	DNAshape (Pentamer)	MC (Tetramer)
HelT	0.62	0.62
ProT	0.68	0.68
Opening	0.57	0.55
Stretch	0.55	0.60
Shift	0.64	0.64
Buckle	0.59	0.63
Rise	0.72	0.72
Stagger	0.69	0.69
Slide	0.62	0.62
MGW	0.70	0.64
Tilt	0.56	0.60
Roll	0.70	0.69
Shear	0.60	0.60

Supplementary Table S2. XRC dataset comprised of 590 X-ray co-crystal structures used in calculating Supplementary Table S1.

Dataset	**PDB IDs**
XRC bound (Used for generating XRC query table)	1P3O, 1S97, 1C0W, 2H1O, 2ER8, 2C6Y, 1N48, 1R0A, 2NVX, 2ADY, 1EFA, 1AKH, 2R5Y, 2E2I, 1JJ6, 1SRS, 1RZR, 1GU5, 1QN4, 2IS4, 1NJX, 2J6U, 1IGN, 2GLI, 1OWG, 1IAW, 2IIE, 1OH6, 2OST, 1S32, 1ZX4, 2PI5, 1NWQ, 1PUE, 1KB2, 1H88, 2AU0, 1QLN, 1K78, 1MM8, 1P3M, 1NVP, 1OZJ, 1BP7, 1JKQ, 1DDN, 1MUH, 2AGQ, 2PPB, 1DSZ, 2H27, 2IEF, 1IC8, 1RH6, 2AOQ, 1S10, 1LBG, 1F4K, 1OH7, 2EUV, 1P34, 2D5V, 1XO0, 1R8D, 1LWT, 1T8I, 1CQT, 1D2I, 2O8E, 2DY4, 2D45, 1R4R, 1ID3, 2E2J, 1NG9, 1TW8, 2I9T, 4KTQ, 1LE8, 1P3B, 1QNC, 2HMI, 1GJI, 1ZME, 2C2R, 2ERG, 1CDW, 1SKN, 2OWO, 1HDD, 2C28, 1MUS, 2O93, 1TL8, 1KBU, 2I13, 2BNZ, 3PVI, 2JEJ, 1JKR, 1MA7, 1IG9, 2I9K, 1NJY, 1K79, 1KX5, 1AOI, 1P3G, 1LMB, 1EVW, 1IU3, 1F66, 2AS5, 1N6J, 1Z19, 1J1V, 2EWJ, 2HOT, 1MOW, 1C9B, 2AOR, 1Z63, 1FJL, 1TKD, 2ATL, 1PP8, 1HJC, 1S0N, 1OH8, 2P0J, 2A3V, 1TTU, 1HW2, 1AU7, 2BQ3, 1CZ0, 1HCR, 1L3U, 1R0N, 1Q9Y, 1G2F, 1G9Y, 2O5I, 2HVR, 2IRF, 2EUZ, 2KTQ, 1LLM, 1QN8, 2CAX, 1U8B, 1NKB, 2EVG, 1SKM, 2UVR, 2J6S, 1NZB, 1U3E, 1EJ9, 1DUX, 1N6Q, 2AQ4, 2HOS, 1T8E, 2GM4, 1RPE, 1A74, 1IPP, 1SAX, 2HOF, 1P7D, 2I3Q, 1DH3, 2UVV, 1APL, 1R0O, 1PVP, 1P3F, 1P3K, 1J59, 1W7A, 1EOO, 1MUR, 1P8K, 1FJX, 1CEZ, 1BC7, 2OG0, 2CGP, 1YTF, 1ODH, 1A0A, 1P4E, 1WBD, 1KSY, 1GA5, 4CRX, 1ZRE, 1KSX, 1SC7, 2JEF, 1ZLK, 2AJQ, 1DFM, 1HLZ, 1NK0, 2HDD, 2IIF, 2RVE, 1EQZ, 1BL0, 1ECR, 2EZV, 1K7A, 1PVI, 1PYI, 1U0D, 1LPQ, 1H0M, 1R7M, 2F8X, 2EVH, 2ERE, 1TF6, 1JJ4, 1VRR, 1YO5, 1TK8, 1CF7, 2ASJ, 1M5R, 2ASD, 2HHQ, 1ZG1, 1I3J, 1Q0T, 1P7H, 1VKX, 1B72, 3CRX, 1IXY, 1QNE, 1HF0, 2OH2, 2BR0, 1XBR, 1NKC, 2CV5, 2OR1, 1SXQ, 1F0O, 1QN9, 1P3L, 1KB6, 2P6R, 2IVH, 1A6Y, 1FLO, 9ANT, 8MHT, 1TQE, 2A66, 1ZRF, 1W0U, 2BNW, 1R4O, 2F5P, 1LLI, 1NK5, 2PI4, 1VOL, 2GIE, 1JEY, 1HWT, 1LE9, 2F8N, 1PER, 2OAA, 2NVZ, 1XNS, 1TK0, 2AYB, 1BPZ, 1NGM, 2EVI, 1W0T, 1M1A, 1NNE, 2FO1, 1CIT, 2HVS, 2H8R, 2BSQ, 1XHV, 2HOI, 1P47, 2I3P, 1UA1, 2NLL, 1H9T, 3MHT, 1H89, 2HHV, 1U8R, 2B9S, 1XHU, 2HR1, 1RUN, 1FOK, 1QSS, 1Z1G, 1K6O, 1Q3V, 1ZS4, 1G2D, 1D66, 2H1K, 1IO4, 1WBB, 2EVF, 1L5U, 1N5Y, 1RYS, 1H8A, 1QN6, 1FOS, 1ZRC, 1ZYQ, 2ETW, 1A36, 1YSA, 1REP, 1U35, 1HLV, 1F5T, 1T3N, 2GIG, 1BDT, 2C7A, 1NKP, 2R5Z, 1R9T, 1TGH, 2HAN, 1K61, 1ZNS, 1D5Y, 1FW6, 1L3L, 1RZ9, 1M5X, 1G3X, 2CRX, 1O3T, 2HT0, 2O61, 2ACJ, 1GLU, 1L3V, 1IJW, 1QP9, 1NLW, 1TX3, 2BQU, 1WB9, 1EGW, 1M19, 1G9Z, 1QNA, 1ZRD, 1AWC, 1Q3U, 1GTW, 1AM9, 1JKO, 2JEG, 1NH2, 2EVJ, 1MNN, 3CRO, 1OUZ, 1EYU, 2EUX, 1PVR, 2IS6, 1U78, 2EX5, 2BQR, 2HHS, 2IT0, 1LE5, 1A3Q, 1QN3, 1WTE, 1S9F, 1BY4, 1M6X, 1PUF, 1OH5, 1PVQ, 2UVU, 1D3U, 1E3M, 1R49, 1J5O, 2DTU, 1LQ1, 2F5N, 2A07, 1IHF, 2IMW, 1TRR, 2DPD, 1RM1, 2NTC, 1LAT, 2EUW, 4BDP, 1DU0, 1R71, 2P5O, 2GIH, 1OWR, 1T05, 1ZR4, 2Q2T, 1BF5, 1F2I, 1GU4, 1QN7, 1LRR, 1CGP, 1CRX, 2FIO, 1Q9X, 1YF3, 2NRA, 1HLO, 1LB2, 1NK8, 1MJQ, 1N3E, 1Z1B, 2C9L, 2FLD, 1GD2, 1M18, 3HDD, 1KC6, 1PP7, 2HZV, 2IVK, 1IG7, 1QNB, 1DRG, 1RIO, 1NJW, 1JKP, 1MNM, 1CYQ, 1GT0, 1JX4, 1A02, 2GEQ, 5CRX, 1TSR, 2Q2U, 2H7H, 1ZBB, 1MJ2, 2UVW, 1JJ8, 1MEY, 1MHD, 1JNM, 1KB4, 1CKT, 1U0C, 1N3F, 1B3T, 1RAM, 1YFH, 2ASL, 2C2E, 1N56, 1EWQ, 1LWS, 1ZR2, 2C2D, 1EA4, 1OCT, 2DRP, 1O3R, 1BDV, 1IMH, 1KX3, 1S0O, 1S0M, 1GXP, 1GDT, 2ISZ, 2NTZ, 2J6T, 2O6G, 1P3P, 1P3A, 1O3Q, 1OUQ, 1TRO, 1IF1, 2BGW, 2ODI, 2JEI, 1PAR, 1RR8, 2HAP, 1TUP, 2IS2, 2AGO, 1B8I, 2C22, 2Q10, 1UBD, 1YFI, 2GII, 1PT3, 2NP2, 1RTD, 2E2H, 1QN5, 1DC1, 2ATA, 1PZU, 1NK9, 1MJO, 1P3I, 1JFI, 1T2K, 1E3O, 1T2T, 1W36, 1R4I, 2F5O, 1JE8, 1K4T, 1YTB, 2FJ7, 2NZD, 1JT0, 1RYR, 1KX4, 2O5J, 1O3S, 1HBX, 1ZLA, 2GIJ, 1T9J, 1A73, 1S9K, 1RUO, 1F44, 1QTM, 1YNW, 2HHT, 1MDM, 1HCQ, 2AYG, 2AC0, 1TC3, 1OWF, 1YRN, 1ZG5, 1PDN, 1H6F, 2AHI, 1MDY, 1Z9C, 1M0E, 1HJB, 2IS1, 1K82, 1SA3, 6PAX, 1YFJ, 1QSY, 2AGP, 1JWL, 1T9I, 1AN4, 3KTQ, 1RRJ

AUTHOR CONTRIBUTIONS

J.L. and R.R. conceived and designed the project. J.L. analyzed and validated MC data and performed statistical machine-learning work. J.M.S. analyzed XRC data derived from the PDB. T.P.C. updated the DNAshapeR/Bioconductor package with the help of J.L. to include additional DNA shape features. M.P. and A.P. analyzed MD simulations and generated the MD query table. J.L. and R.R. wrote the manuscript with contributions from all other authors. R.R. supervised the project.

SUPPLEMENTARY REFERENCES

1. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
2. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.
3. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S., *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
4. 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. U.S.A.*, **112**, 4654–4659.
5. 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.
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. 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.
8. 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**, W56–62.
9. Rohs,R., Sklenar,H. and Shakked,Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure*, **13**, 1499–1509.
10. Sklenar,H., Wüstner,D. and Rohs,R. (2006) Using internal and collective variables in Monte Carlo simulations of nucleic acid structures: chain breakage/closure algorithm and associated Jacobians. *J. Comput. Chem.*, **27**, 309–315.
11. Cornell,W.D., Cieplak,P., Bayly,C.I. and Gould,I.R. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules J. Am. Chem. Soc., **117**, 5179− 5197.
12. Rohs,R., Etchebest,C. and Lavery,R. (1999) Unraveling proteins: a molecular

mechanics study. *Biophysical J.*, **76**, 2760–2768.

13. Lavery,R. and Sklenar,H. (1989) Defining the Structure of Irregular Nucleic Acids: Conventions and Principles. *J. Biomol. Struct. Dyn.*, **6**, 655–667.
14. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299–313.
15. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
16. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
17. Dang,L.X. (1995) Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J. Am. Chem.Soc.*, **117**, 6954–6960.
18. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
19. 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.
20. Andrabi,M., Hutchins,A.P., Miranda-Saavedra,D., Kono,H., Nussinov,R., Mizuguchi,K. and Ahmad,S. (2017) Predicting conformational ensembles and genome-wide transcription factor binding sites from DNA sequences. *Sci. Rep.*, **7**, 4701.
21. Fujii,S., Kono,H., Takenaka,S., Go,N. and Sarai,A. (2007) Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.*, **35**, 6063–6074.
22. Pérez,A., Lankas,F., Luque,F.J. and Orozco,M. (2008) Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.*, **36**, 2379–2394.
23. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophysical J.*, **92**, 3817–3829.
24. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.