Supporting Information

Zhou et al. 10.1073/pnas.1422023112

SI Methods

Sequence and Shape Feature Vectors. We used 1-mer, 2-mer, and 3-mer feature vectors of the form

$$\overrightarrow{\text{1-mer}} = (\text{A}_1,\ \text{C}_1,\ \text{G}_1,\ \text{T}_1, \ldots,\ \text{A}_L,\ \text{C}_L,\ \text{G}_L,\ \text{T}_L)$$

$$\overrightarrow{\text{2-mer}} = (\text{AA}_1,\ \text{AC}_1,\ \text{AG}_1,\ \text{AT}_1, \ldots,\ \text{TG}_{L-1},\ \text{TT}_{L-1})$$

$$\overrightarrow{\text{3-mer}} = (\text{AAA}_1,\ \text{AAC}_1,\ \text{AAG}_1,\ \text{AAT}_1, \ldots,\ \text{TTG}_{L-2},\ \text{TTT}_{L-2})$$

with lengths $4L$, $16(L-1)$, and $64(L-2)$, respectively, where N_i, NN_i, and NNN_i have a value of 1 if the particular k-mer occurs at position i, and a value of 0 otherwise.

The prediction of each shape feature constituted a numeric vector,

$$\overrightarrow{\text{MGW}} = (\text{MGW}_3,\ \text{MGW}_4, \ldots,\ \text{MGW}_{L-3},\ \text{MGW}_{L-2})$$

$$\overrightarrow{\text{ProT}} = (\text{ProT}_3,\ \text{ProT}_4, \ldots,\ \text{ProT}_{L-3},\ \text{ProT}_{L-2})$$

$$\overrightarrow{\text{Roll}} = (\text{Roll}_2,\ \text{Roll}_3, \ldots,\ \text{Roll}_{L-3},\ \text{Roll}_{L-2})$$

$$\overrightarrow{\text{HelT}} = (\text{HelT}_2,\ \text{HelT}_3, \ldots,\ \text{HelT}_{L-3},\ \text{HelT}_{L-2})$$

where MGW_i and ProT_i represent MGW and ProT, respectively, at nucleotide position i, and Roll_i and HelT_i represent Roll and HelT, respectively, of the dinucleotide between positions i and $(i+1)$. Shape values at two positions of both the 5′ and 3′ ends were unavailable because a pentamer model was used to derive the structural features (1).

Each shape feature vector was further expanded to include second-order shape features by adding products of the same shape parameter at two adjacent positions. For example, the resulting $\overrightarrow{\text{MGW}_{\text{1st+2nd}}}$ vector was of the form

$$\overrightarrow{\text{MGW}_{\text{1st+2nd}}} = (\text{MGW}_3,\ \text{MGW}_4, \ldots,\ \text{MGW}_{L-3},\ \text{MGW}_{L-2},\ \text{MGW}_3 * \text{MGW}_4, \ldots,\ \text{MGW}_{L-3} * \text{MGW}_{L-2}).$$

The complete shape feature vector used in this study was obtained by concatenating the four numeric vectors $\overrightarrow{\text{MGW}_{\text{1st+2nd}}}$, $\overrightarrow{\text{ProT}_{\text{1st+2nd}}}$, $\overrightarrow{\text{Roll}_{\text{1st+2nd}}}$, and $\overrightarrow{\text{HelT}_{\text{1st+2nd}}}$.

We provide an easy-to-use tool (with source code and documentation) for computing DNA sequence and shape feature vectors at rohslab.cmb.usc.edu/PNAS2015/.

Implementation of Support Vector Regression. We used the ε-SVR algorithm (2) implemented in the LIBSVM toolkit (3) to train linear regression models for predicting the natural logarithm of the PBM signal intensities (response variable) based on the encoded sequence and shape features. The ε-SVR contains two user-defined hyperparameters: C, the penalty factor used for regularization; and ε, the parameter in the loss function (i.e., maximum distance between the predicted and actual values for which no penalty is incurred). An internal 10-fold cross-validation was used at each step to identify the hyperparameters C and ε that yielded the best performance (i.e., lowest mean-squared error). The hyperparameter space was manually specified as

$$\varepsilon \in \{0.001,\ 0.01,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.7,\ 0.9,\ 1.0\}$$

$$C \in \{0.001,\ 0.01,\ 0.05,\ 0.1,\ 0.5,\ 1,\ 5,\ 10,\ 50\}.$$

The response variables were the natural logarithms of the fluorescence signal intensities. The performance of the different models was evaluated and compared based on the squared Pearson correlation coefficient R^2 between predicted and observed values of the response variable.

gcPBM Data. The gcPBM data for His-tagged human TF dimers, Mad1 (Mxd1)–Max, Max–Max, and c-Myc–Max (Mad, Max, and Myc, respectively), were generated essentially as previously described (4). Briefly, a 4 × 180k microarray design (AMADID 041707; Agilent) was used, which contained 36-bp genomic sequences selected from the ChIP-seq peaks of Mad, Max, or Myc in the HeLa S3 or K562 cell line [encyclopedia of DNA elements (ENCODE)]. Each 36-mer was centered at a putative TF binding site for Mad, Max, or Myc (see ref. 4 for further details on the microarray design). The 24-mer GTCTTGATTC GCTTGA-CGCT GCTG (representing the reverse complement of a primer to be used in the double-stranding step of the PBM assay) was appended to each 36-bp genomic region.

Following the PBM protocol (5), a primer extension step was performed to obtain double-stranded DNA oligonucleotides on the microarray. Each microarray chamber was incubated with a 2% milk blocking solution for 1 h, followed by incubations with the protein-binding mixture for 1 h and with Alexa488-conjugated anti-His antibody (1:20 dilution, Qiagen) for 1 h (5). The array was gently washed and then scanned with a GenePix 4400A scanner (Molecular Devices) at 2.5-μm resolution. Data were normalized with standard analysis scripts (5, 6). Although this study used a previously described microarray design (4), the experimental protocol was slightly modified. Specifically, the milk concentration in the protein-binding mixture was increased from 2% to 4%, to reduce the background signal and to permit higher-quality data to be obtained (Fig. S1).

After the gcPBM data for Mad, Max, and Myc were obtained, each dataset was filtered to remove sequences that contained more than one putative TF binding site. For each probe on the array, flanking regions of the central TF binding site were scanned with the uPBM 8-mer data for Mad, Max, or Myc (7). Any probe for which the flanking regions contained at least one 8-mer with an enrichment score (E-score) ≥ 0.3 was removed. The E-score is a modified form of the Wilcoxon–Mann Whitney statistic. The E-score ranges from –0.5 (least-favored sequence) to +0.5 (most-favored sequence). As reported previously for uPBM assays (5), a false discovery rate of 0.01 typically corresponds to an E-score cutoff of ∼0.32–0.36. Thus, by selecting only probes for which all 8-mers in the flanking regions had an E-score < 0.3, we ensured that each probe contained only one Mad/Max/Myc binding site.

To ensure that each probe was centered at the Mad/Max/Myc binding site, the 8-mer with highest E-score had to be located in the center of the 36-mer, with the next-highest 8-mer adjacent to it. After these filtering criteria were applied, 6,927 probes for Mad, 8,569 probes for Max, and 7,535 probes for Myc were obtained. Raw and processed gcPBM data were submitted to the Gene Expression Omnibus (GEO) under accession number GSE59845.

uPBM Data. This study used the uPBM data (5, 6) for 66 mouse TFs reported by ref. 8 and used in the DREAM5 challenge. Briefly, uPBMs contain artificial DNA sequences designed using a de Bruijn sequence of order 10 over the {A, C, G, T} alphabet, which ensures that all 10-bp DNA sequences are represented on the array. Unlike gcPBM probes, there is no guarantee that a TF binding site will occur in the center of the uPBM probe. In this context, a bias may occur because the location of the TF binding site within the probe affects the TF binding signal as measured by PBM. Specifically, binding sites located close to the free DNA end of a probe generally result in higher PBM signals than binding sites located close to the glass slide (6). Computational methods for training TF binding models from uPBM data either try to learn the positional bias from the data or they use median 8-mer intensities and 8-mer E-scores (8) to average out the positional bias.

Because the goal of this study was to evaluate DNA shape-augmented models of DNA binding specificities compared with traditional sequence-based models, and not to train models that can predict uPBM data, the positional bias was not explicitly modeled. The DREAM5 uPBM data were processed with the intent of minimizing the effect of positional bias and making the data suitable for position-based regression models, according to the following six steps.

i) For each of the 66 mouse TFs, we obtained the normalized uPBM signal intensities for all probes on the array, the 8-mer E-scores derived from the uPBM data, and the best PWM for that TF (as reported by Weirauch et al. after analyzing PWMs obtained with 26 different algorithms) (8).

ii) For each of the 66 TFs, we scanned each uPBM probe to identify the best PWM match on either the forward or reverse strand, accounting for all of the putative sites in the 35-bp variable probe region. The best PWM matches were used to align the uPBM probes to each other.

iii) For each TF, the selected probes were those for which the best PWM match fell at least L positions from the left end of the probe (corresponding to the free DNA end) and at least R positions from the right end of the scanned probe region. Restricting the location of the TF binding site by the L and R parameters minimized the effect of the positional bias on the uPBM signal intensities used in the analyses. As shown in our previous work (9), flanking DNA sequences outside the PWM match can significantly affect TF binding and contribute to the PBM signal. For this reason, flanking regions outside the PWM match were included, as long as the PWM match fell within the limits defined by the L and R parameters.

Several L/R pairs (L2R12, L2R15, L2R18, L5R5, L5R10, L5R12, L5R15, L5R18, L8R12, and L8R15) were tested. On average, L5R10 resulted in the most accurate models and, therefore, was chosen. However, using different L/R pairs did not markedly change the results of the comparisons between DNA shape-augmented models and traditional sequence-based models of DNA binding specificities.

iv) The preceding step identified the best PWM match within each probe, within the limits defined by the L and R parameters, regardless of the PWM score. In this step, any probes for which the best PWM match did not correspond to a putative TF binding site (defined as a site containing at least two consecutive 8-mers with uPBM E-score > 0.3) were filtered out. This criterion is not a stringent cutoff for defining TF binding sites (6, 9), but it ensures that most of the selected probes do contain a specific TF binding site.

v) Although not common, some DNA probes on uPBMs can contain two or more TF binding sites. In such cases, it is not clear how much each binding site contributes to the PBM signal. To remove such probes from consideration, the probe region outside the central 12 bp (corresponding to the best PWM match) was scanned. Probes that contained a second potential binding site were filtered out.

vi) Finally, for each TF, the selected probe sequences were trimmed to a length of T = Length(PWM) + $2L$ bp, to ensure that the same amount of flanking sequence was used for each putative TF binding site.

For one of the 66 TFs in the DREAM5 study (8), Nhlh2, none of the DNA probes passed the filtering criteria. Therefore, this TF was not included in further analyses.

Training and Testing on Different uPBM Array Designs. Most uPBM analyses presented here used cross-validation on the uPBM data obtained from a specific array design. An alternative approach for testing shape-augmented specificity models is to train the models on uPBM data from one array design and use them to predict the binding data obtained from a different array design, similar to the procedure proposed by ref. 8. The latter approach has the advantage that it can test whether the shape features capture array-specific biases and artifacts. However, this approach can only be applied if the data obtained using the two array designs agree well with each other and are of similar quality.

We analyzed the uPBM data for both array designs used by ref. 8 for the 66 mouse TFs. For each TF, we computed the squared Pearson correlation coefficient (R^2) between the 8-mer E-scores derived from the two array designs. Next, we selected the top 10 TFs (Oct1, Pit1, Prdm11, Sox3, Zkscan1, Dmrtc2, Foxo6, Nkx2-9, Pou1f1, and Sdccag8) with the highest correlation between the array designs (i.e., with $R^2 > 0.45$). For each of the 10 selected TFs, we trained specificity models on one array design (as described above, but without performing an embedded cross-validation). We used those models to predict the binding data for the second array design. Data from both array designs were processed in the same way, as described in *uPBM Data* above.

Differences Between gcPBM and uPBM Data. The preprocessing of the uPBM data for the 66 mouse TFs (8) and the experimental protocol for the generation of the gcPBM data for the human bHLH TFs Mad, Max, and Myc yielded three differences. First, sequences selected from the uPBM data were shorter than gcPBM sequences; therefore, fewer positions in the flanking regions were used in the SVR modeling. Second, the gcPBM data did not suffer from positional bias, whereas the selected uPBM probes might still have had some positional bias even after the processing steps. Third, the gcPBM intensity for each probe represented the median over six replicate measurements, whereas the uPBM intensity for each probe corresponded to a single measurement. Thus, it was not surprising that models trained on gcPBM data were more accurate than models trained on uPBM data.

Cross-Platform Testing of Shape-Augmented Models: Training the Models on gcPBM Data and Testing Them on SELEX-seq DNA Sites. To assess how well our PBM-trained models are able to predict TF binding data obtained using other in vitro technologies, we generated SELEX-seq data for one of the TFs in our study, the human protein Max. The SELEX-seq experiment was carried out as described previously (10, 11). Max (66 nM) was incubated with the following randomized oligonucleotide library (at 200 nM): GTTCAGAGTT CTACAGTCCG ACGATCTGG (N_{16}) CCAGAACTCG TATGCCGTCT TCTGCTTG. The following oligonucleotide was used to track the mobility of the Max–DNA complex: ATACATAAGA TCGCATTATG TGGCTTATCA AACCACGTGG TTTATCAAAA TAATAAGTGA TCTGTCATTG ATC. Sequencing was performed with an Illumina HiSeq 2500.

After two rounds of SELEX, we calculated the relative affinities of all 12-mers as described previously (10, 11). Briefly, a fifth-order Markov model was constructed by using Round 0 sequences to predict the number of 12-mer sequences in the initial library (10, 11). Then, the relative affinity of each 12-mer was generated by calculating the square root of the enrichment ratio (counts in Round 2/expected counts in Round 0 from the Markov model).

We used our regression-based models trained on gcPBM data to predict Max binding as measured by SELEX-seq. This task is not trivial because the binding measurements obtained from the two technologies are very different: The gcPBM measures TF binding to 36-bp genomic regions centered at putative TF binding sites, whereas SELEX-seq measures average TF binding to short sequences (typically 6–12 bp), which might contain a TF binding site at any position. Thus, to train our models on gcPBM data and test them on SELEX-seq data, both data types were preprocessed as described below.

First, we trimmed the 36-bp genomic sequences in the Max gcPBM data to the central 10 bp. For each 10-mer obtained at this step, we calculated its average gcPBM signal (which was used as the response variable in the SVR analysis) as the average of the PBM log intensity for all 36-bp probes centered at that 10-mer. To ensure that the average 10-mer binding signal was not biased toward specific flanking regions, we did not consider 10-mers that were present in fewer than ten 36-bp probes. Then, we trained the sequence- and shape-based regression models on the set of selected 10-mers and their corresponding average gcPBM log intensities.

We processed the SELEX-seq data using a method similar to the one described above for uPBM data, with the goal of obtaining SELEX-seq 12-mers that contain only one putative binding site, located in the center of the 12-mer. The putative binding site in each SELEX 12-mer was determined by scanning the sequence with an 8-bp Max PWM and selecting the best PWM match. We discarded the 12-mer sequences for which the putative binding site had a uPBM E-score < 0.3 because these sequences were unlikely to be bound specifically by the TF of interest. After this processing step, 12-mers centered at putative Max binding sites were immediately included in our analyses.

To include as many SELEX-seq sequences as possible, we also selected 12-mers in which the putative binding site was shifted from the center of the 12-mer by 1 bp. Finally, to ensure that all sequences used in our analyses were the same size, we trimmed the 12-mer SELEX sequences to 10-mers by keeping only 1 bp on each side of the putative binding site. The relative affinities of the resulting 10-mers were calculated as the average relative affinities of all 12-mers that were trimmed to a particular 10-mer. The SELEX-seq data for Max were submitted to the Gene Expression Omnibus (GEO) under accession number GSE60200.

To determine the significance of the differences in Spearman's rank correlation coefficients (SRCC) between 1mer+shape and sequence-based models when trained on gcPBM data and tested on SELEX-seq data, we used bootstrapping to generate distributions of such differences under the null hypothesis that two models perform equally well. For each sequence-based model, we performed the following steps. We generated 10,000 bootstrap training samples from the gcPBM data. Next, we trained the sequence-based model on each bootstrap sample, tested it on the SELEX-seq data, and computed the corresponding SRCC. Next, we computed the differences in SRCC values between the original sequence-based model and the models trained on the bootstrap samples, which are expected to perform equally well. Thus, we obtained the null distribution of SRCC differences, which we used to compute empirical *P* values. For each sequence-based model we asked: Under the hull hypothesis, what is the probability of obtaining an SRCC difference at least as extreme as the difference between that sequence-based model and the 1mer+shape model? The computed empirical *P* values were: $P <$ 0.0001 for the 1mer model, $P = 0.0015$ for the 1mer+2mer model, and $P = 0.0058$ for the 1mer+2mer+3mer model.

Number of Parameters in *k*-mer-Based Regression Models. For each nucleotide position in the TF binding sites, our *k*-mer models use 4 features to encode 1-mer identity, 16 features to encode 2-mer identity, 64 features to encode 3-mer identity, and 8 features to encode DNA shape (i.e., 4 first-order and 4 second-order shape features). Therefore, the 1mer+shape, 1mer+2mer, and 1mer+2mer+3mer models used a total of 12, 20, and 84 features per nucleotide position, respectively (Figs. 4*A* and 5*A*). Simple motif models in which the parameters represent probabilities of having specific nucleotides at specific positions in the TF binding site (e.g., position-specific frequency matrices) have only three independent parameters per position. However, in our models, the parameters represent weights related to the contribution of specific nucleotides to the TF binding signal. The sum of the four weights (corresponding to nucleotides A, C, G, and T) can be different at different positions, which is why we use all four 1-mer features at each position in the binding site.

The *k*-mer features used in our regression models are not independent of each other. In theory, the 2-mer features can capture the contributions of 1-mers, and 3-mer features can capture the contributions of 1-mers and 2-mers. However, using just 2-mers or just 3-mers while training *k*-mer regression models is problematic when (some of) the real contributions to the binding affinity are due to 1-mers.

For example, assume that nucleotide A at position i in the binding site contributes x to the TF binding signal. To capture this contribution with a 1mer+2mer model, the regression algorithm only needs to learn one weight (x) for the feature A at position i. However, capturing this contribution using only 2-mer features is not trivial. One possibility would be for the model to learn a weight $x/2$ for the features AA at position $i - 1$, CA at position $i - 1$, GA at position $i - 1$, TA at position $i - 1$, AA at position i, AC at position i, AG at position i, and AT at position i. If the important contributions at positions $i - 1$ and $i + 1$ are also due to 1-mers, then the regression algorithm will have an even harder time finding the correct weights for 2-mer features to capture 1-mer effects. Thus, training models that include both 1-mers and 2-mers (or 1-mers, 2-mers, and 3-mers) is a valid approach, which we have used when reporting the numbers of features in the main text of our manuscript. We also tested SVR models using only 2-mer or 3-mer features; however, they performed slightly worse than models using 1-mers + 2-mers or 1-mers + 2-mers + 3-mers.

Methodology for Calculation of DNA Shape Features. The Monte Carlo sampling that underlies the DNAshape method (1) uses all-atom simulations of DNA fragments starting from canonical B-form conformations. The set of collective and internal variables in the Monte Carlo sampling included 12 degrees of freedom per nucleotide, which comprise three rigid-body translations, three rigid-body rotations, the glycosidic torsion angle, phase and amplitude of the sugar moiety, and two endocyclic torsion and one bond angle in the phosphodiester backbone (12). All remaining endocyclic torsion and bond angles in the backbone were sampled as dependent variables. For the thymine base, the rotation of the thymine methyl group was used as an additional independent variable. The sampling included an analytic chain closure using associated Jacobians (13) and charges and atom sizes as specified in a previously published protocol (12). The system was neutralized with explicit sodium counter ions whose positions were independently sampled. The solvent was described implicitly using a sigmioidal distance-dependent dielectric function (14). The simulation protocol and analysis used for data generation were identical to the one previously described (15).

The DNA shape features were derived with our DNAshape method based on the mining of all-atom Monte Carlo simulations for 2,121 different DNA fragments of 12–27 bp in length (1). Using a sliding pentamer window, shape parameters were calculated at the central nucleotide (MGW and ProT) or two central bp steps (Roll and HelT) using Curves (16). Average shape parameters were then calculated for each of the unique 512 pentamers based on all occurrences of a pentamer in our Monte Carlo-generated dataset.

1. Zhou T, et al. (2013) DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 41(web server issue):W56–W62.
2. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. *Neural Information Processing Systems 9* (MIT Press, Cambridge, MA), pp 155–161.
3. Chang CC, Lin CJ (2011) LIBSVM: A Library for Support Vector Machines. *ACM Trans Intell Syst Technol* 2(3):27.
4. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordân R (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* 29(13):i117–i125.
5. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* 4(3):393–411.
6. Berger MF, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24(11): 1429–1435.
7. Munteanu A, Gordân R (2013) Distinguishing between genomic regions bound by paralogous transcription factors. *Research in Computational Molecular Biology 2013* (Springer, New York), p 145.
8. Weirauch MT, et al.; DREAM5 Consortium (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* 31(2):126–134.
9. Gordân R, et al. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports* 3(4):1093–1104.
10. Slattery M, et al. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147(6):1270–1282.
11. Riley TR, et al. (2014) SELEX-seq: A method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Hox Genes: Methods and Protocols, Methods in Molecular Biology*, eds Graba Y, Rezsohazy R (Springer, New York), Vol 1196, pp 255–278.
12. Rohs R, Sklenar H, Shakked Z (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* 13(10):1499–1509.
13. Sklenar H, Wüstner D, 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(3):309–315.
14. Rohs R, Etchebest C, Lavery R (1999) Unraveling proteins: A molecular mechanics study. *Biophys J* 76(5):2760–2768.
15. Zhang X, et al. (2014) Conformations of p53 response elements in solution deduced using site-directed spin labeling and Monte Carlo sampling. *Nucleic Acids Res* 42(4): 2789–2797.
16. Lavery R, Sklenar H (1989) Defining the structure of irregular nucleic acids: Conventions and principles. *J Biomol Struct Dyn* 6(4):655–667.
17. Brownlie P, et al. (1997) The crystal structure of an intact human Max-DNA complex: New insights into mechanisms of transcriptional control. *Structure* 5(4):509–520.
18. Joshi R, et al. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* 131(3):530–543.
19. Dror I, Zhou T, Mandel-Gutfreund Y, Rohs R (2014) Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic Acids Res* 42(1):430–441.
20. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315(5809):233–237.

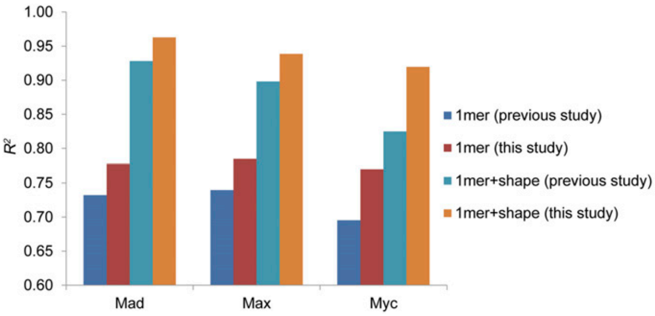

Fig. S1. High quality of experimental gcPBM data generated for this study. Comparison of the performances of the SVR-based 1mer and 1mer+shape models with previously published gcPBM data (4) to their performances with the new gcPBM data generated for this study. Models of the same type always performed better on the new gcPBM data. We note that the Mad TF used in this study is the Mad1 (Mxd1) protein, which is closely related to Mad2 (Mxi1).

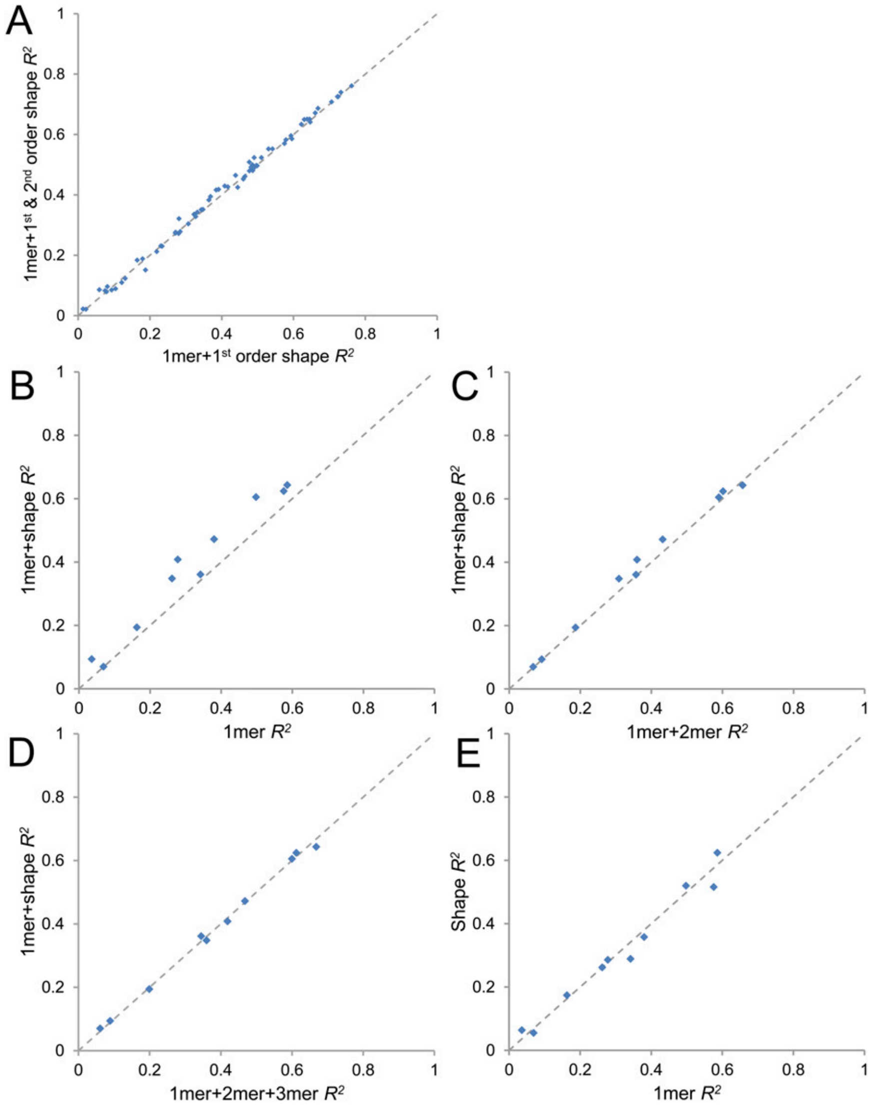

Fig. S2. Performances of various models on the uPBM data for 65 mouse TFs. (*A*) Comparison of model performances on uPBM data of 65 mouse TFs from the DREAM5 dataset (8). Models combining sequence (1mer) with first- and second-order DNA shape features were compared with models combining sequence and only first-order DNA shape features. (*B–D*) Comparison of model performances for shape-augmented models (1mer+shape) with performances for sequence-based (*B*) 1mer, (*C*) 1mer+2mer, and (*D*) 1mer+2mer+3mer models in a cross-array evaluation (i.e., models were trained on one uPBM array design and then tested on a different uPBM array design, as in ref. 8). Results are shown for 10 TFs from the DREAM5 study, chosen based on high agreement between the uPBM data generated using the two array designs (see *Training and Testing on Different uPBM Array Designs* for more details). (*E*) Performance comparison of the shape-only model to the sequence-only (1mer) model in a cross-array evaluation for the 10 TFs shown in *B–D*.

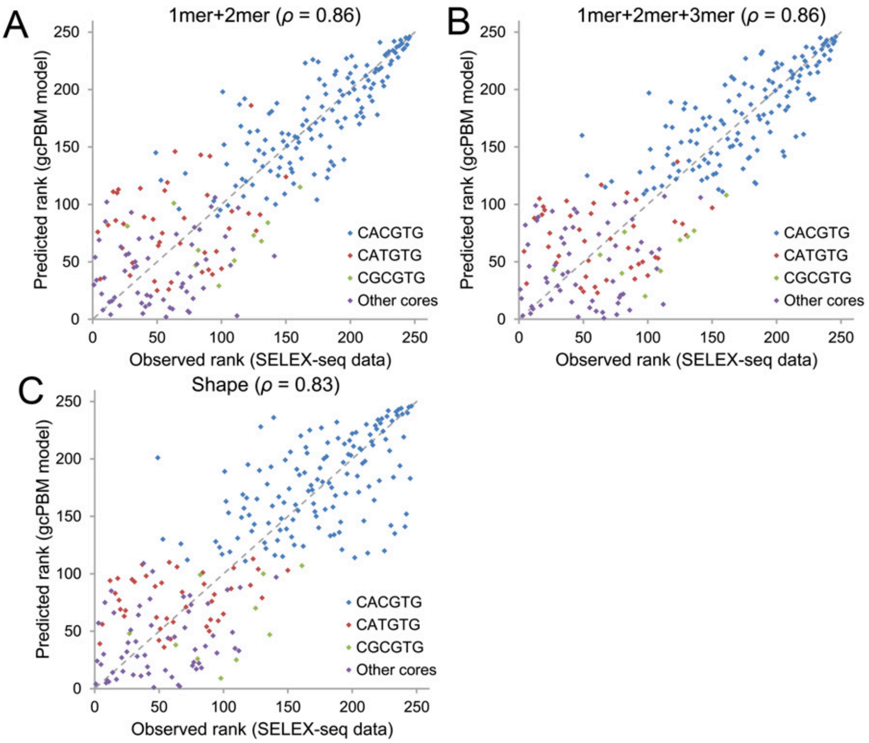

Fig. S3. Performances of binding specificity models across experimental platforms. (*A–C*) Scatter plots of predicted versus observed binding site ranks, illustrating the performances of the (*A*) 1mer+2mer, (*B*) 1mer+2mer+3mer, and (*C*) shape-only models trained on gcPBM data and tested on SELEX-seq data. Here, higher ranks represent higher-affinity binding sites. See Fig. 3*B* and *SI Methods* for the *P* values of the performance comparison between the 1mer+shape model and the sequence-based models.

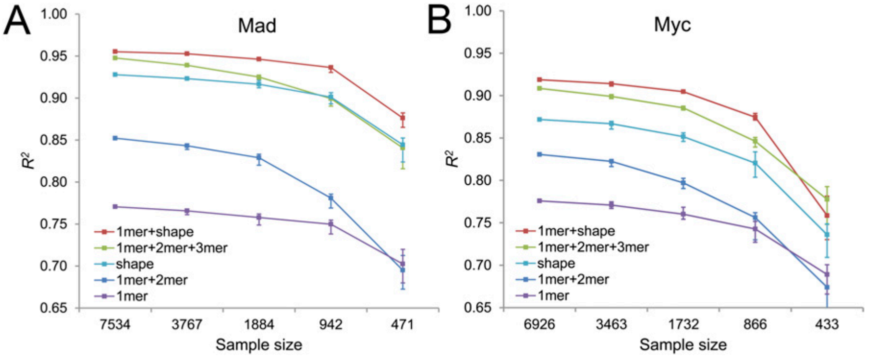

Fig. S4. Performance comparison of various models for human Mad and Myc TFs. Performances of the sequence- and shape-based models for (*A*) Mad and (*B*) Myc binding to DNA as the sample size was decreased.

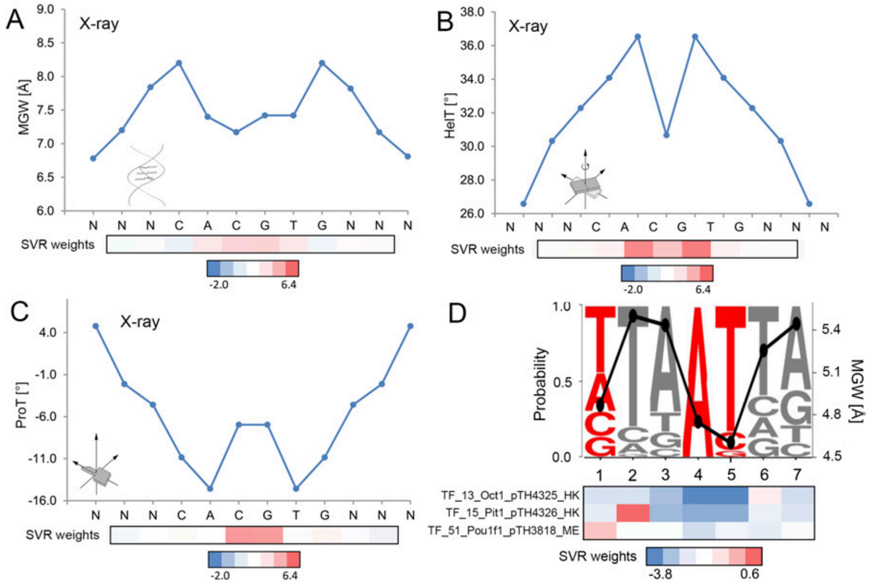

Fig. S5. SVR feature weights reveal structural mechanisms of DNA readout for bHLH and homeodomain TFs. (*A*–*C*) Structural characteristics observed in the cocrystal structure of the ternary Max–Max/DNA complex (17) correspond, to different degrees, with the SVR feature weights derived from the 1mer+first order shape models for (*A*) MGW, (*B*) HelT, and (*C*) ProT based on the gcPBM data for Max. (*D*) SVR feature weights for MGW derived from the 1mer+first order shape model based on uPBM data (8) for the homeodomain TFs Oct1, Pit1, and Pou1 agree with X-ray (18), SELEX-seq (10), and uPBM (19) data. The sequence probability matrix and MGW plot were adapted from our earlier publication (19) and compared with SVR feature weights derived in this study.

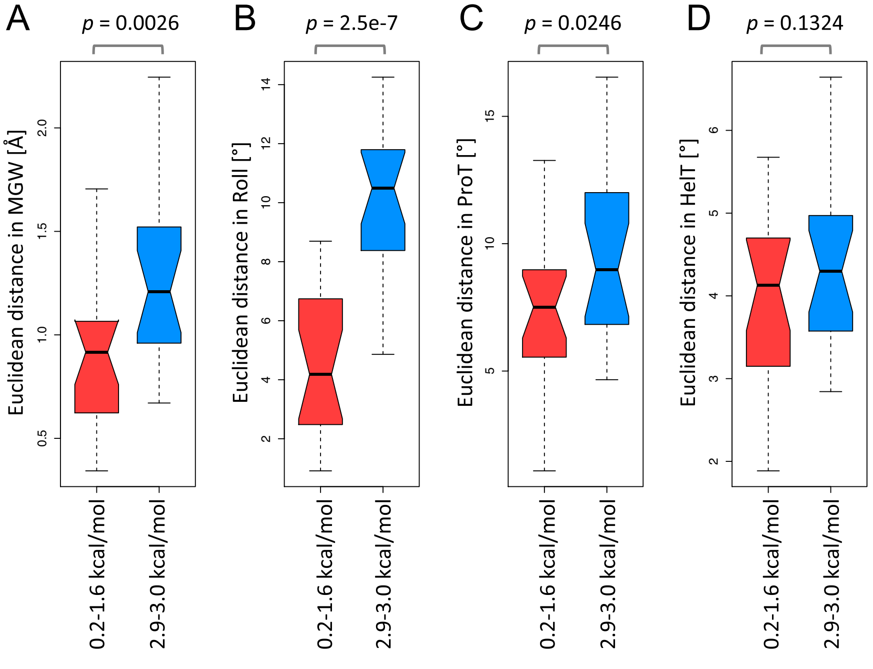

Fig. S6. DNA shape profiles differ significantly between high- and low-affinity binding sites of the human bHLH TF Max. Using mechanically induced trapping of molecular interactions (MITOMI), Maerkl and Quake (20) reported DNA binding affinities for specific and nonspecific target sites of the Max TF, matching the pattern TTGnnnnGTGGGTG. We generated DNA shape profiles for these target sites with our DNAshape method (1) and compared them to the profile of the highest-affinity site, TTGCCACGTGGGTG, based on Euclidean distances and taking all nucleotide positions of the binding site into account. Box plots show the Euclidean distances in (*A*) MGW, (*B*) Roll, (*C*) ProT, and (*D*) HelT between the highest-affinity site and either specific (red) or nonspecific (blue) binding sites. The top 20 and bottom 20 sites, according to binding affinity, were included in the analysis. The observed differences between specific and nonspecific sites are statistically significant for MGW, Roll, and ProT (Mann–Whitney U test *P* values 0.0026, 2.5e-7, and 0.0246, respectively). The range of Gibbs free energy differences ΔΔG (relative to the highest affinity site) is shown on the *x* axis.