Supplementary Data for "DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding."

Wenxiu Ma[1], Lin Yang[2], Remo Rohs[2], and William Stafford Noble[3]

[1]Department of Statistics, University of California Riverside, Riverside, CA 92521, USA
[2]Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA
[3]Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA.

Supplementary Notes

1 Computation of the DNA shape features

Recently, the Rohs lab has developed DNAshape, a computational approach to predicting the three-dimensional double-helix DNA structure features from one-dimensional nucleotide sequences (Zhou *et al.*, 2013; Chiu *et al.*, 2016). The DNAshape method uses a five-base sliding window method where the double helix DNA structural features unique to each of the 512 distinct pentamers are defined by minor groove width (MGW), Roll, propeller twist (ProT), and helix twist (HelT). MGW and ProT represent base-pair parameters whereas Roll and HelT represent base pair-step parameters. The values for four DNA shape feature as function of its pentamer sequence were derived from Monte Carlo simulations. The DNAshape method is computationally efficient and its predictions are compatible with DNA structures solved by X-ray crystallography and NMR spectroscopy, hydroxyl radical cleavage data, statistical analysis and cross-validation, and molecular dynamics simulations.

The DNAshape prediction method is publicly available at the DNAshape web server (http://rohslab.cmb.usc.edu/DNAshape) and also in the DNAshapeR Bioconductor package (https://bioconductor.org/packages/release/bioc/html/DNAshapeR.html).

2 Comparison between compositional kernel and positional kernel

2.1 Pre-processing details

We summarized the difference of the pre-processing details of the DREAM5 universal PBM (uPBM) data between the Zhou *et al.* (2015) study and our study, in Supplementary Table 2.

Starting from the raw PBM data, there are $40,330 - 40,526$ probes (each 35 bp long) on each array. After the initial normalization step, in our study we used the entire set of full-length probes as our training and testing data. In the Zhou *et al.* (2015) study, after all the pro-processing, filtering, trimming steps, there are $27 - 2,700$ probes left for each array, where the probes are of length $17 - 20$ bp.

2.2 Dimensionality of kernel space

We summarized the kernel space dimensionality of our compositional k-spectrum kernel and k-spectrum+shape kernel with the positional k-mer and k-mer+shape kernels proposed by Zhou *et al.* (2015), in Supplementary Table 3. In Zhou *et al.* (2015), the $4L - 14$ shape features were further expanded to include second-order shape feature by adding products of the same shape parameter at two adjacent positions. To facilitate a direct comparison with the k-spectrum+shape kernel, we didn't include the second-order shape features. In the following discussion, the k-mer+shape kernel only contains the first-order shape features.

As we can see, the positional kernel space depends on the input nucleotide length L. For a typical PBM experiment, the raw probe length is about $25 - 35$ bp and the preprocessed and trimmed probe length is about $17 - 20$ bp. Thus, the positional k-mer and k-mer+shape kernels have much higher dimension.

2.3 Performance evaluation on the pre-processed DREAM5 data

We compared our compositional k-spectrum+shape kernel with the positional k-mer+shape kernel (Zhou *et al.*, 2015) on the pre-processed DREAM5 data. Since the pre-processed probes were well aligned at known motif sites, the positional kernel has better performance than our compositional kernel on these pre-processed data (Supplementary Figure 4).

2.4 Performance evaluation on the raw DREAM5 data

We also compared our compositional k-spectrum+shape kernel with the positional k-mer+shape kernel (Zhou *et al.*, 2015) on the raw DREAM5 data. As shown in Supplementary Figure 5, our compositional kernel has outperformed the positional kernel on the raw, unaligned probe data.

Supplementary References

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. *Nature Biotechnology*, **24**(11), 1429–1435.

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**(8), 1211–1213.

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 Reports*, **3**(4), 1093–1104.

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., DREAM5 Consortium (including W. S. Noble), Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolovitzky, G., and Hughes, T. R. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, **31**(2), 126–134.

Zhou, T., Yang, L., Lu, Y., Dror, I., Machado, A. C. D., 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 Research*, **41**(W1), W56–W62.

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. *Proceedings of the National Academy of Sciences*, **112**(15), 4654–4659.

Supplementary Tables

	k-spectrum n	k-spectrum+shape $(3+4k)n$	di-mismatch n	di-mismatch+shape $5n$
$k=1$	2	14		
$k=2$	10	110		
$k=3$	32	480	32	160
$k=4$	136	2,584	136	680
$k=5$	512	11,776	512	2,560
$k=6$	2,080	56,160	2,080	10,400
$k=7$	8,192	253,952	8,192	40,960
$k=8$	32,896	1,151,360	32,896	164,480

Supplementary Table 1: Number of features in different kernel space. Here n is the number of unique k-mers. Note that for the di-mismatch kernel and di-mismatch+shape kernel, the dimensionality of the kernel space do not depend on the di-mismatch parameter m.

Preprocessing steps	**Zhou *et al.* (2015)**	**Our study**
0. For each of the 66 mouse TFs, we took the median signal intensity values of the replicates of each probe on the PBM and then take log2 as the normalized uPBM signal.	✓	✓
1. 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 (Weirauch *et al.*, 2013)	✓	
2. 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.	✓	
3. 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 previous work (Gordân *et al.*, 2013), 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.	✓	
4. 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 (Berger *et al.*, 2006; Gordân *et al.*, 2013), but it ensures that most of the selected probes do contain a specific TF binding site.	✓	
5. 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.	✓	
6. 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.	✓	

Supplementary Table 2: Preprocessing details for DREAM5 dataset. The preprocessing steps were copied from Zhou *et al.* (2015) with minor formatting changes.

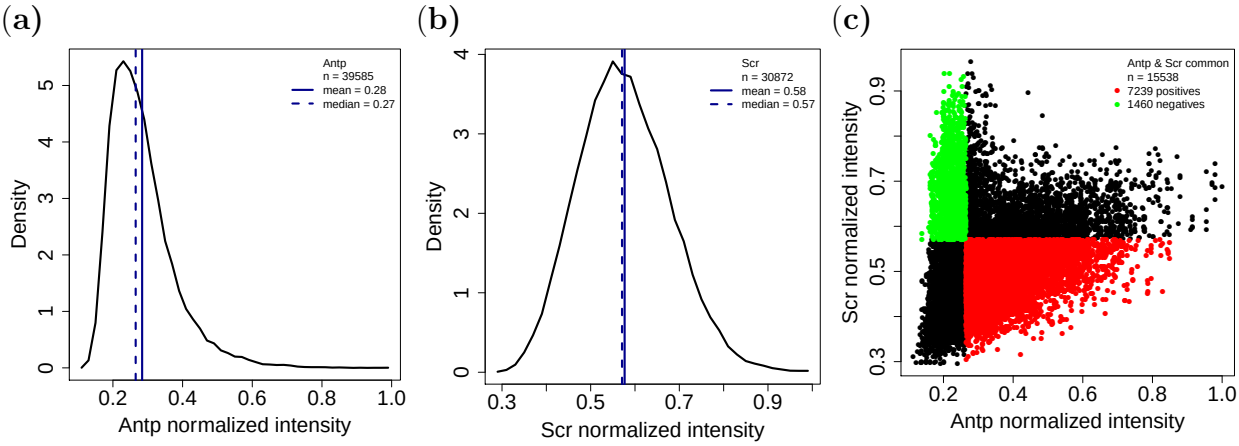
	compositional k-spectrum n	**compositional k-spectrum+shape** $(3+4k)n$	**positional k-mer** $4^k(L-k+1)$	**positional k-mer+shape** $4^k(L-k+1)+(4L-14)$
$k=1$	2	14	$4L$	$8L-14$
$k=2$	10	110	$16L-16$	$20L-30$
$k=3$	32	480	$64L-128$	$68L-142$

Supplementary Table 3: Number of features in different kernel space. Here n is the number of unique k-mers. L is the length of the input nucleotide sequences. The positional k-mer+shape kernel only include first-order shape features.

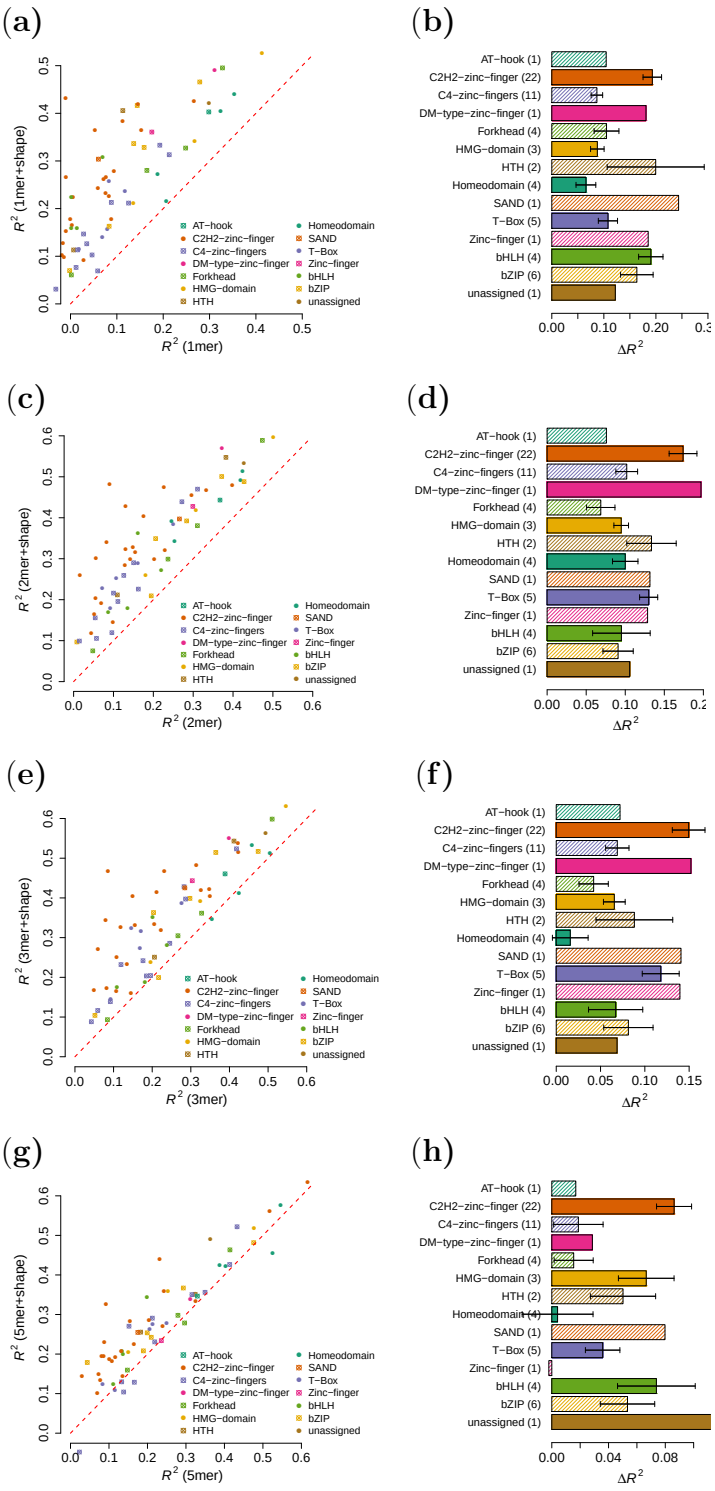
	k-spectrum	k-spectrum+shape	di-mismatch m=1	di-mismatch+shape m=1	di-mismatch m=2	di-mismatch+shape m=2	di-mismatch m=3	di-mismatch+shape m=3
k=1	0.6256 (+/-0.0058)	**0.8899** (+/-0.0029)						
k=2	0.6602 (+/-0.0090)	**0.9789** (+/-0.0026)						
k=3	0.9562 (+/-0.0055)	**0.9843** (+/-0.0028)	0.6915 (+/-0.0068)	**0.9579** (+/-0.0045)				
k=4	0.9767 (+/-0.0027)	**0.9847** (+/-0.0023)	0.9772 (+/-0.0022)	**0.9813** (+/-0.0013)	0.9572 (+/-0.0030)	**0.9796** (+/-0.0014)		
k=5	0.9805 (+/-0.0027)	**0.9855** (+/-0.0015)	**0.9815** (+/-0.0025)	0.9740 (+/-0.0035)	**0.9822** (+/-0.0022)	0.9808 (+/-0.0021)	0.9798 (+/-0.0025)	**0.9838** (+/-0.0022)
k=6	0.9717 (+/-0.0024)	**0.9778** (+/-0.0068)	0.9752 (+/-0.0043)	**0.9770** (+/-0.0018)	0.9854 (+/-0.0030)	**0.9869** (+/-0.0019)	0.9874 (+/-0.0022)	**0.9885** (+/-0.0018)
k=7	0.9766 (+/-0.0023)	**0.9796** (+/-0.0024)	0.9797 (+/-0.0025)	**0.9797** (+/-0.0016)	**0.9876** (+/-0.0019)	0.9832 (+/-0.0018)	0.9860 (+/-0.0014)	**0.9867** (+/-0.0025)
k=8	0.9751 (+/-0.0025)	**0.9831** (+/-0.0019)	**0.9761** (+/-0.0033)	0.9739 (+/-0.0039)	**0.9865** (+/-0.0020)	0.9812 (+/-0.0027)	**0.9884** (+/-0.0018)	0.9777 (+/-0.0028)

Supplementary Table 4: AUROC performance of distinguishing Exd-Antp and Exd-Scr binding sites in SELEX-seq data. Numbers in the parentheses represent the variance of AUROC scores among the five outer cross-validation folds. For each pairwise comparison between k-spectrum and k-spectrum+shape models and between di-mismatch and di-mismatch+shape models, higher AUROC scores are marked in bold font. The highest AUROC score among all models is colored in blue.

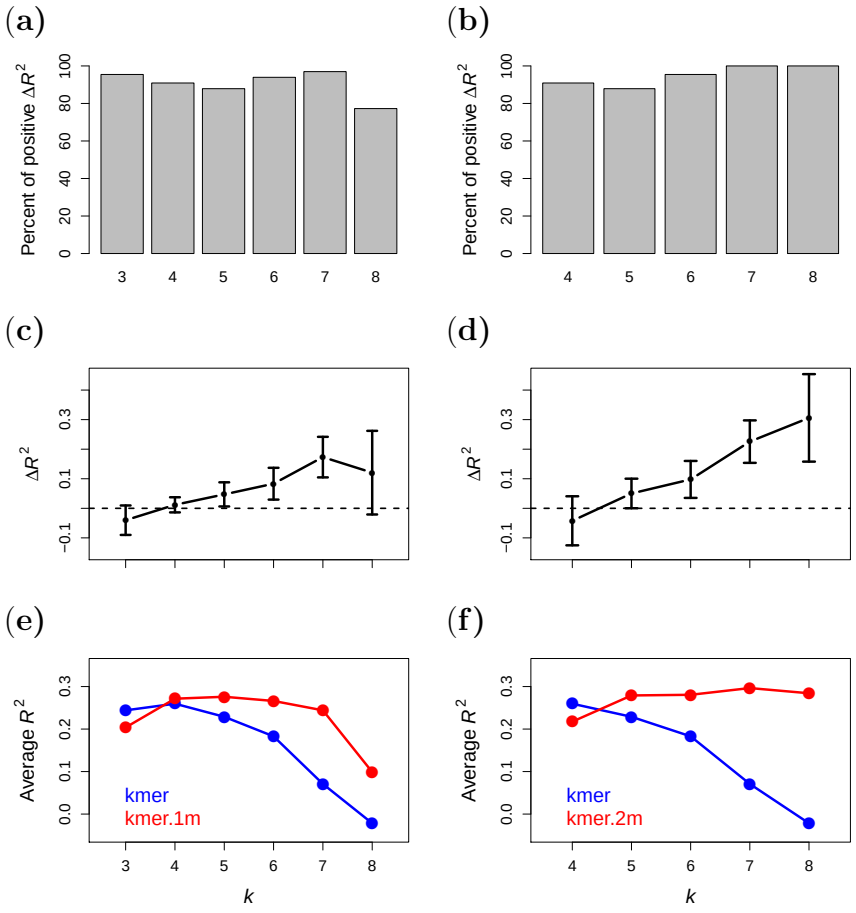
Supplementary Figures



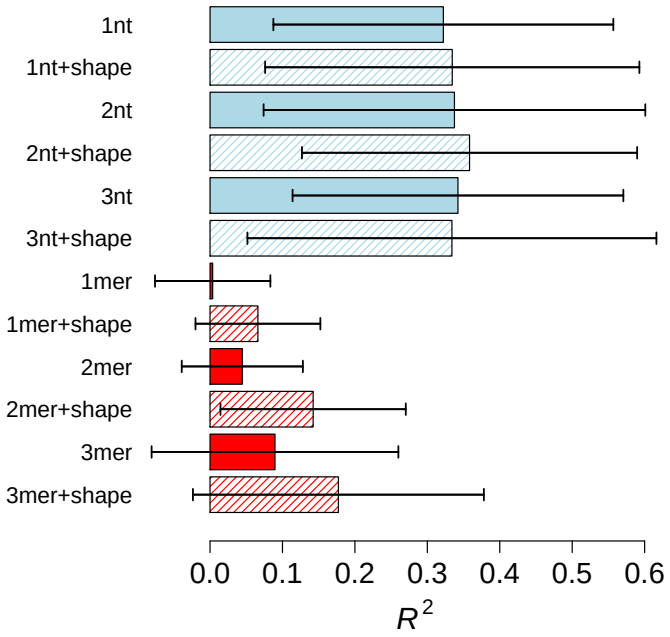
Supplementary Figure 1: Exd-Hox SELEX-seq data. **(a)** Histogram of normalized binding affinity values for the Exd-Antp SELEX-seq data. **(b)** Histogram of normalized binding affinity values for the Exd-Scr SELEX-seq data. **(c)** Scatter plot of the normalized Antp and Scr binding affinities for the common sequences. Each dot represents one DNA sequence. Red dots are sequences with positive labels, which have relatively higher binding affinity for Antp. Green dots are sequences the negative labels, which have relatively higher binding affinity for Scr.



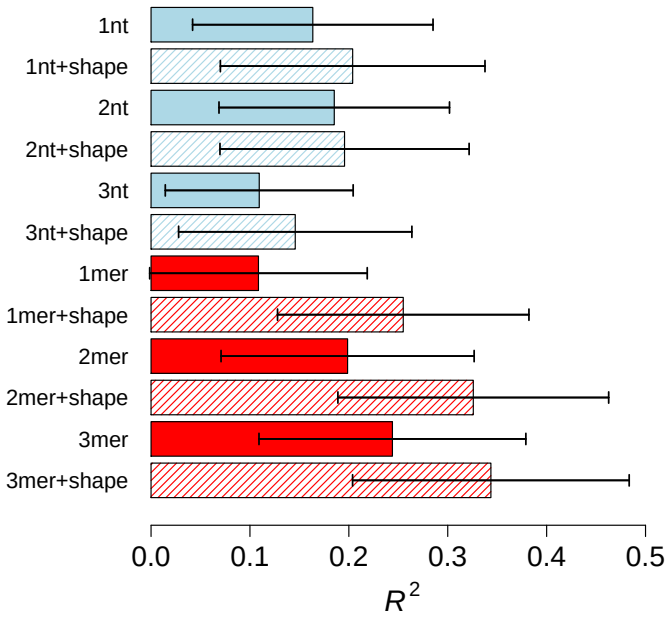
Supplementary Figure 2: R^2 **performance for** k**-spectrum model verse** k**-spectrum+shape model on uPBM data set,** $k = 1, 2, 3, 5$**. (a)(c)(e)(g)** Scatter plots of the R^2 performance values between the two models. Each dot represents one TF, colored corresponding to its protein family. **(b)(d)(f)(h)** Barplots of R^2 improvements for various protein families. Numbers in the parentheses are the number of DREAM5 TFs in each TF family. The x-axis shows the differences of R^2 values between the two models. The length of the bars represent the mean of R^2 differences and the error bars mark the standard error of the mean.



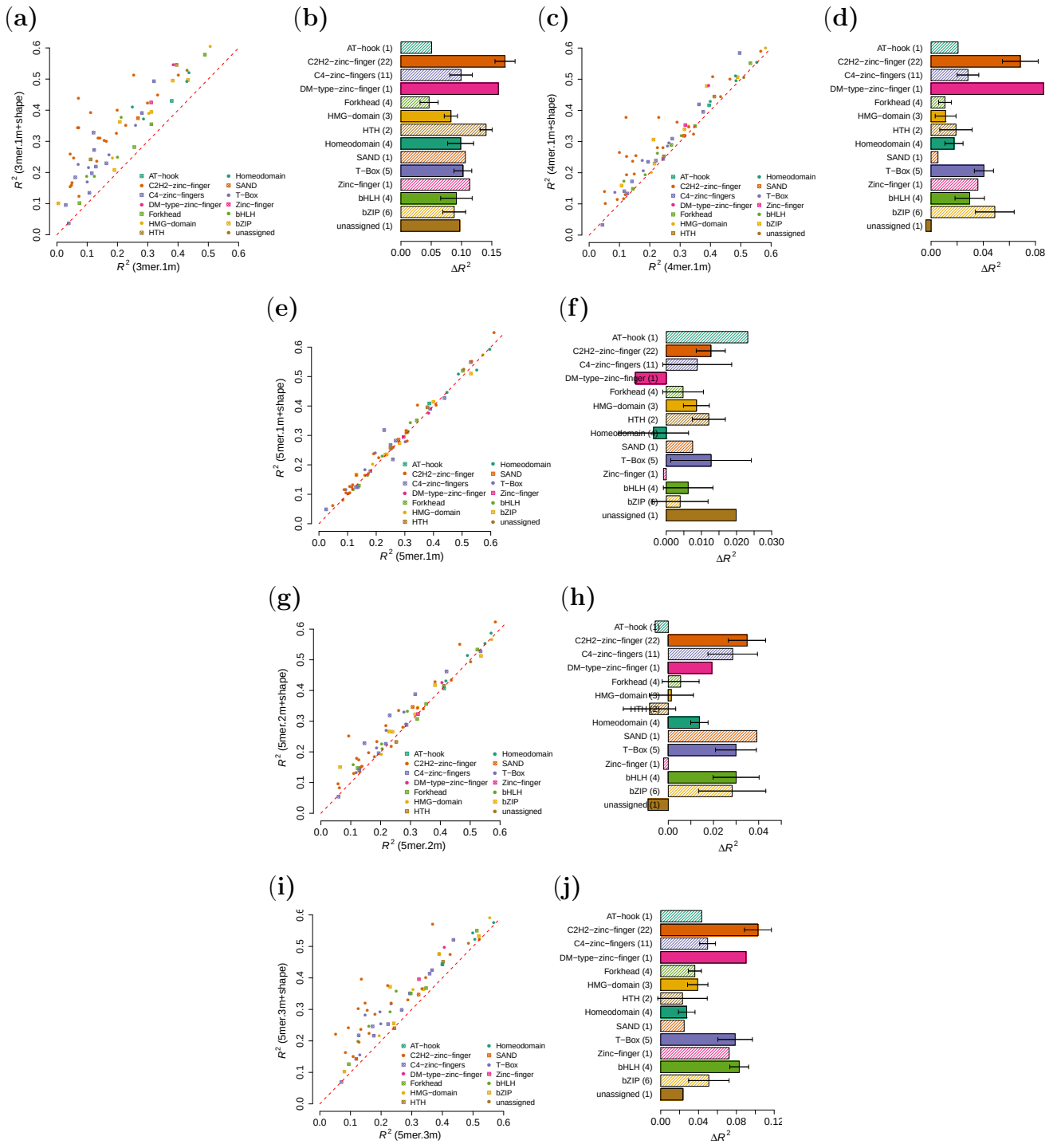
Supplementary Figure 3: Comparison between k-spectrum and di-mismatch models. **(a)** Percent of DREAM5 TF datasets that have higher R^2 values using the di-mismatch model than using the k-spectrum model, for $k = 3, \cdots, 8$ and $m = 1$. **(c)** Differences of R^2 values between the two models, for $k = 3, \cdots, 8$ and $m = 1$. **(e)** R^2 performance scores of various k-specturm models (blue) and di-mismatch models (red), for $k = 3, \cdots, 8$ and $m = 1$. **(b)(d)(f)** Corresponding plots for di-mismatch paramters $k = 4, \cdots, 8$ and $m = 2$.



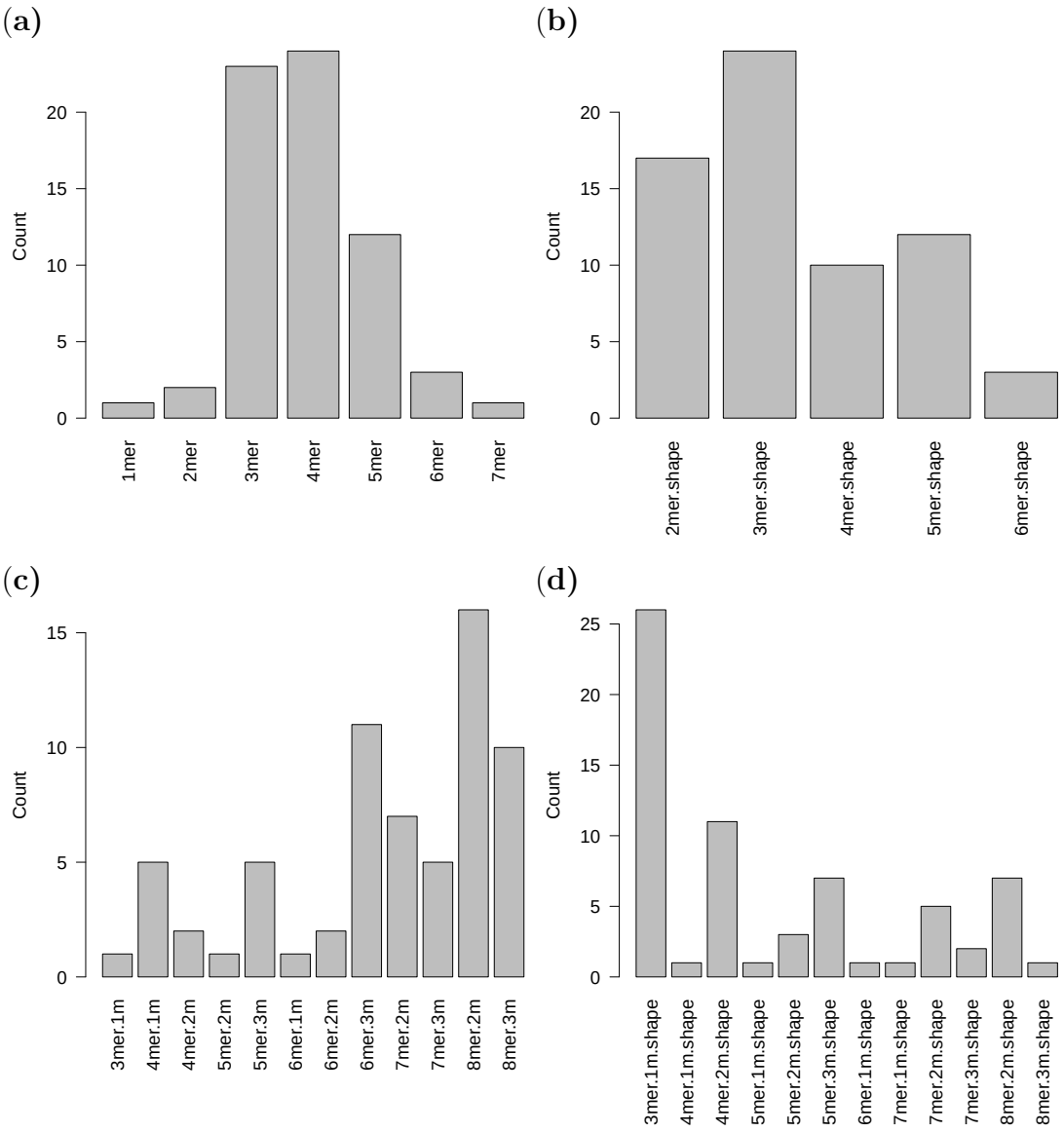
Supplementary Figure 4: Comparison between positional kenels and compositional kernels on the pre-processed DREAM5 uPBM data. The k-nt and k-nt+shape models represent the positional kernel approaches as described in Zhou *et al.* (2015); whereas k-mer and k-mer+shape models represent our compositional k-spectrum and k-spectrum+shape kernels. The probe preprocessing steps were described in Supplementary Note 2.1 and the pre-processed probe data was obtained from Zhou *et al.* (2015). The bar heights indicate the average R^2 values among 65 mouse TF uPBM dataset and the error bars mark the standard error.



Supplementary Figure 5: Comparison between positional kernels and compositional kernels on the raw DREAM5 data. The k-nt and k-nt+shape models represent the positional kernel approaches as described in Zhou *et al.* (2015); whereas k-mer and k-mer+shape models represent our compositional k-spectrum and k-spectrum+shape kernels. The bar heights indicate the average R^2 values among all 66 mouse TF uPBM dataset and the error bars mark the standard error.



Supplementary Figure 6: R^2 performance for di-mismatch model verse di-mismatch+shape model on uPBM data set, $k = 3, 4, 5$, $m = 1, \cdots, k-2$. (a)(c)(e)(g)(i) Scatter plots of the R^2 performance values between the two models. Each dot represents one TF, colored corresponding to its protein family. **(b)(d)(f)(h)(j)** Barplots of R^2 improvements for various protein families. Numbers in the parentheses are the number of DREAM5 TFs in each TF family. The x-axis shows the differences of R^2 values between the two models. The length of the bars represent the mean of R^2 differences and the error bars mark the standard error of the mean.



Supplementary Figure 7: Distribution of best k (and m) parameters in each kernel for the mouse DREAM5 dataset. **(a)** Histogram of best k-spectrum kernels. **(b)** Histogram of best k-spectrum+shape kernels. **(c)** Histogram of best di-mismatch kernels. **(d)** Histogram of best di-mismatch+shape kernels.