Methods

Promoter library construction

For pTpA and Abf1TATA libraries, a single-stranded oligonucleotide pool was ordered from IDT containing the random 80 bp oligonucleotide flanked by arms complementary to the promoter scaffold for use with Gibson assembly. These oligonucleotides were double stranded with a complementary primer sequence and Phusion polymerase master mix (NEB), gel purified and cloned into the dual reporter vector, ensuring a complexity of at least 10^8 for each library.

The promoter scaffold sequences were:

For pTpA:

(poly-T; distal)

GCTAGCAGGAATGATGCAAAAGGTTCCCGATTCGAACTGCATTTTTTTCACATC

(poly-A; proximal)

GGTTACGGCTGTTTCTTAATTAAAAAAAGATAGAAAACATTAGGAGTGTAACACAAGACTTTCGGATCCTGAGCAGGCAAGATAAACGA (up to the theoretical TSS).

For Abf1TATA:

(Abf1 site; distal)

GCTAGCTGATTATGGTAACTCTATCGGACTTGAGGGATCACATTTCACGCAGTATAGTTC

(TATA-box; proximal)

GGTTTATTGTTTATAAAAATTAGTTTAAACTGTTGTATATTTTTTCATCTAACGGAACAATAGTAGGTTACGCTAGTTTGGATCCTGAGCAGGCAAGATAAACGA. In both cases, 80 Ns were inserted in between distal and proximal regions.

For the scaffold library (sequences in **Table S1**), the library was cloned in two stages. In the first, the promoter scaffolds (synthesized by microarray synthesis) were amplified and cloned using Gibson Assembly. The resulting library had a common restriction site into which the N80 was cloned by ligation.

Reporter assay

Libraries were transformed into yeast (strain Y8205 (70)) using the lithium acetate method (71), starting with 1L of yeast harvested at an OD of 0.3-0.4, ensuring at least 10^8 cells were transformed (with the exception of the high-quality pTpA library, where a dilution series was performed to achieve the desired lower complexity). The yeast were then grown in SD-Ura for two days, diluting the media by 1:4 three times during this period. Media was then either changed to YPD, growing for at least 5 generations prior to cell sorting, or to YPGly and YPGal, with culture grown for at least 8 generations (due to the different carbon source). In the final 10 hours of growth prior to cell sorting, all cultures were allowed to grow continuously in log phase, never achieving an OD above 0.6, by diluting in fresh media. All cultures were grown in a shaker incubator, at 30°C and approximately 250 RPM.

Prior to sorting, yeast were spun down, washed once in ice-cold PBS, and then suspended in ice-cold PBS and kept on ice until cell sorting. Cells were sorted by $\log_2$(RFP/YFP) signal (using mCherry and GFP absorption/emission) on a Beckman-Coulter MoFlo Astrios, using 18 uniform bins, done in three batches of six bins each, with the exception of the scaffold library, which was sorted into non-uniform bins to account for the higher variance at low expression levels and the larger dynamic range of the library. The FACS configuration varied between experiments (*e.g.*, different laser intensities), resulting in different baseline expression values. Post sort, cells were spun down and resuspended in SC-Ura (supplemented with 1% Gal for Gal sort), grown for 2-3 days, shaking at 30°C. The plasmids were then isolated, the promoter region amplified, Nextera adaptors and multiplexing indices added, and the resulting libraries sequenced with 76 bp, paired-end reads, using 150 cycle kits on an Illumina NextSeq sequencer, achieving complete coverage of the promoter, including overlap in the center. For the scaffold library, the libraries were instead sequenced with a 300 cycle kit using a 190 bp read 1 and 112 bp read 2.

MNase-Seq experiment

Aliquots of the pTpA library, expected to correspond to ~100,000 (sample A) or ~200,000 (sample B) viable cells were each cultured in duplicate (Rep 1 and 2) in YPD for ~16 hours to an OD of ~0.4-1.0. For each sample, 0.5 mL of culture was pelleted and frozen to prepare input genomic DNA, and 3 mL of culture was crosslinked with 1% formaldehyde, washed twice with 1mL H_2O supplemented with a protease inhibitor cocktail, and the pellet frozen for MNase treatment. These pellets were next spheroplasted using zymolyase, and spheroplasts were lysed in NP buffer (10 mM Tris pH 7.4, 50 mM NaCl, 5 mM $MgCl_2$, 1 mM $CaCl_2$, and 0.075% NP-40, freshly supplemented with 1 mM β-mercaptoethanol, 500 μM spermidine, and EDTA-free protease inhibitor cocktail) at a concentration of $2*10^6$ cells/ μl of NP buffer. 0.125 units of Worthington MNase were added per 10μl of lysed spheroplasts and MNase digestion was preformed at 37°C for 20 minutes. MNase digestion was stopped by addition of equal volume of 2X MNase Stop Buffer (220 mM NaCl, 0.2% SDS, 0.2% sodium deoxycholate, 10 mM EDTA, 2% Triton X-100, EDTA-free protease inhibitor cocktail). MNased chromatin samples were treated with RNase A and proteinase K, reverse cross linked, separated on a 4% agarose gel and mononucleosome bands were isolated. Genomic DNA was prepared using the Masterpure Yeast Genomic DNA Preparation Kit (Epicenter). For both MNase and genomic DNA, the variable region of the promoter library was amplified, and adaptors added for sequencing using an Illumina NextSeq with 76 bp single-end reads.

Theoretical TFBS abundance

We estimated the abundance of TFBSs in random DNA by analyzing the information contents (ICs) of known motifs associated with yeast TFs (25). The IC of a motif (IC_{motif}) is proportional to the frequency (f_{motif}) with which that motif is expected to be found on either strand of random DNA with the following relationship, where IC_{motif} is expressed in bits:

$$f_{motif} = 2^{-(IC_{motif}-1)}$$

The number of instances present in a library of a given TFBS motif, assuming that binding sites are independent, is the number of positions in the library that could potentially contain a complete binding site multiplied by the expected frequency of the TFBS motif. For a library with a complexity of 10^7, comprised of 80 bp sequences, the number of possible TFBSs is $(80 - length_{motif} + 1) * 10^7$.

For **Figure 1B**, we used the average motif length as the $length_{motif}$ for all motifs so that the x axis could include frequency and the expected number of binding sites. For this analysis, motifs for zinc cluster monomers were excluded, since these are abundant in the database (25) and are likely to represent only a half TFBS. Several TFBS motifs that are long but generally have low IC content, were also excluded since they are unlikely to represent true TF specificities (**Table S2**).

Promoter sequence consolidation

The paired end reads representing both sides of the promoter sequence were aligned using the overlapping sequence in the middle, constrained to have 40 (+/-15) bp of overlap for pTpA and Abf1TATA libraries and 16 (+/-10) bp for the scaffold library, and discarding any reads that failed to align well within these constraints. Note that only ~0.3μg of N80 DNA were received from IDT, and only ~10^8 of these were successfully cloned; these are only a vanishingly small portion of the possible 4^{80} sequences in N80 (which would weigh ~10^{26} kg even with just one copy of each possible molecule). Thus, any very similar sequences we observe represent the same source promoter with high probability, with minor differences likely corresponding to PCR or sequencing errors. Consequently, promoters of pTpA and Abf1TATA libraries were aligned to themselves using Bowtie2 (version 2.2.1) (72) to identify clusters of related sequences, merging these clusters and taking the sequence with the most reads as the "true" promoter sequence for each cluster. For the scaffold library, promoters were first clustered into those sharing a common scaffold, using Bowtie2 to align to the known scaffold sequences (using the following parameters: -L 18 -p 4 -f --no-sq --no-head --np 0 --n-ceil C,100). Promoters were then sub-clustered within each scaffold using the sequences of the random 80-mers using CD-HIT (version 4.6.5, using the following parameters: -g 1 -p 1 -r 0 -c 0.96 -uS 0.05 -uL 0.05 -mismatch -1) (73), yielding a single consensus sequence for each promoter.

Estimating the proportion of active random promoters

We estimated the proportion of random promoters (those with both random 80-mers and scaffolds designed to mimic random DNA) that were expressed at detectable levels using the empirical log(YFP/RFP) distributions of regrown, previously-sorted, cells (**Figure 2B**). We considered any bin above the lowest expression bin to be "expressed", but since some cells might end up in this lowest expression bin upon re-sorting, we attempted to estimate the number of cells that would remain expressed upon resorting. AUROC statistics were calculated to estimate how well the cells sorted into each bin can be distinguished from those sorted into the not-expressed bin. Here, each AUROC is equivalent to the probability that a cell sorted into the corresponding expressing bin is expressed higher than a randomly selected cell from the not-expressed bin. Thus, cell proportions in expressing bins were weighted by the corresponding AUROC for that bin to get an estimate of the number of expressing random promoters, 83%.

Linear transcription model

TF motifs (**Table S2**) were taken from the YeTFaSCo database (25) and supplemented with the poly-A motif (AAAAA), which we initialized to 100% A at all five positions. Motifs were trimmed to fill 30 bp 1-d convolutional filters, centering the motif if it was less than 30 bp, and, where motifs were longer than 30 bp, trimming off the least informative bases until it was 30 bp.

To identify dissociation constants, K_d, for each TFBS motif and each potential binding site instance, motif filters were applied to DNA sequences and their reverse complements by scanning them with the TFBS motif position weight matrix. Binding to each site in the DNA was determined by the GOMER method using a fixed $[TF_x]$ that corresponds to the minimum K_d possible with the motif (and therefore a perfect match corresponds to 50% occupancy) (26). The expected binding (sum of all binding to all binding sites), assuming Michaelis-Menten equilibrium binding occupancies for all possible binding sites (location *l*, strand *s*) for TF *x* in promoter *p*, where K_ds for each binding site are calculated from the position weight matrix:

$$Binding_{x,p} = \sum^{\substack{s \in Strand \\ l \in Location}} \frac{1}{1 + \frac{[TF_x]}{K_{d\{x,p,l,s\}}}}$$

Correlations between predicted occupancy for each individual TF and expression level were done using these values ($Binding_{x,p}$). We optimized a single weight for each TF ($Activity_x$), representing the ability of that TF to activate or repress transcription.

$$EL_p = c + \sum^{x \in TF} Binding_{x,p} Activity_x$$

This model was implemented in Tensorflow, as described for the other models below, but without a regularization term.

Billboard model of transcription

The billboard model includes parameters for TF concentration ($[TF_x]$), TF activity (Act_x), TF potentiation (Pot_x), and TF activity limits (AL_x). The concentration parameter is unlikely to be comparable to measured cellular TF concentration, since its magnitude also depends on TF affinity and PWM scale, and possibly other factors that affect TF binding. Motifs were trimmed, as before, but filling 25 bp 1-d convolutional filters. As described above, we use these filters, the DNA sequence, and the (now learned) TF concentration parameter to gain an initial estimate for TF binding, here called Raw Binding ($RB_{x,p}$).

We calculate $Open_p$, which corresponds to a probability the promoter *p* will be accessible to binding, as a logistic function on the sum of each TF's (*x*) predicted Raw Binding, weighted by Pot_x, their learned ability to potentiate the binding of other factors:

$$Open_p = \frac{1}{1 + e^{-\sum^{x \in TF} RB_{x,p} Pot_x + c}}$$

Because our promoters are small, we can reasonably assume that a TF that opens chromatin would open it for the entire 80-bp variable region: if the promoter is open, all TFs can bind unimpeded; if the promoter is closed, no TFs can bind. Thus, we re-weight the Raw Binding scores with $Open_p$ to get $Binding_{x,p}$, the amount of binding of each TF *x* to each promoter *p*, as:

$$Binding_{x,p} = RB_{x,p} Open_p$$

Finally, the predicted expression level (EL_p) is the sum of binding values for each TF *x*, weighted by their learned effect on expression (Act_x):

$$EL_p = \sum^{x \in TF} Act_x Binding_{x,p} + c$$

Here, the measured and predicted expression levels are in log space, corresponding to the log-space bins of YFP/RFP. One possible interpretation of the formulation above is that TF activities are proportional to how much the TF affects the zero-order rate constants for different steps of mRNA production, which would be multiplicative in linear space or additive (as above) in log space.

When activity limits for TFs (*AL$_x$*) were included as a learned parameter, the expression level was instead calculated as follows, putting an upper limit on TF activity:

$$EL_p = c + \sum^{x \in TF} \begin{cases} min(Act_x Binding_{x,p}, AL_x), & \text{if } Act_x \geq 0 \\ max(Act_x Binding_{x,p}, AL_x), & \text{otherwise} \end{cases}$$

Position-specific activity model

Position-specific activity models were built as an extension of the billboard model that included binding limits. Here, each potential TFBS position was allowed its own (learned) activity parameter. Position-specific occupancy was estimated similarly to before, but accounting for the strand (*s*) and binding location (*l*) of each TF (*x*) to each promoter (*p*):

$$Binding_{x,p,l,s} = RB_{x,p,l,s} Open_p$$

The transcriptional effect of each TF on each promoter (*Effect$_{x,p}$*) was estimated using the position-specific activity parameters (*Act$_{x,l,s}$*), which were implemented as a local scale of the overall TF activity (*Act$_x$*):

$$Effect_{x,p} = \sum_{\substack{s \in Strand \\ l \in Location}} Act_{x,l,s} Binding_{x,p,l,s} + c$$

We then re-implement the binding limits as follows:

$$EL_p = c + \sum^{x \in TF} \begin{cases} min(Effect_{x,p}, AL_x), & \text{if } Effect_{x,p} \geq 0 \\ max(Effect_{x,p}, AL_x), & \text{otherwise} \end{cases}$$

Model learning

Parameters were learned iteratively, first learning TF activity and potentiation, then TF concentration, then allowing the motifs themselves to be changed, then including a parameter that limited the maximum binding of each TF, and finally learning position-specific activity parameters, each time, learning the new parameters and updating those previously included with a single pass through the data.

Transcriptional models were implemented in Tensorflow (74), minimizing the mean squared error between predicted and measured expression level using the AdamOptimizer and learning in batches of 1,024 promoters. In all cases (except the linear model above), potentiation and activity parameters were regularized with an $L1$ penalty (0.00001), motifs were regularized with an $L2$ penalty (0.000001), and position-specific activity biases (when present) were regularized with an L2 penalty (0.00001) on the difference between adjacent (by location *l*) activity biases. Learning rate was set to 0.04 for the epoch learning activity and potentiation parameters, 0.01 when also learning concentration, and 0.001 when also learning motifs, activity limits, and position-specific activities. All analyses used the models that did not include activity limits, with the exception of the comparisons to Miller et al. (24) data, and the position-specific activity model.

Applying models to other promoter libraries

We applied the pTpA+glucose and Abf1TATA+glucose models to predict expression of our scaffold library. Because each model was trained on much shorter sequences than those on which it was tested, we applied the model to each promoter in five overlapping windows of 110 bp (pTpA) and 115 bp (Abf1TATA), ending at -183, -136, -99 (which included the random 80-mer), -43, and -13 relative to the theoretical TSS, yielding five expression predictions for the different regions. The five predictions tiling the region were combined into a single expression prediction using a linear model that predicted measured expression level using predicted expression and accessibility for each of the five windows, trained on a random 20% of the scaffold library data (approximately 2 million sequences). This was then used to predict expression of the remaining 80% of the scaffold library. For comparisons within promoter scaffolds or only including the random 80-mer, the predicted expression levels for the bin ending at -99 (relative to the TSS) were used directly.

Applying models to native sequences

Since the models above were designed to operate on relatively short sequences (~110 bp), scanning the yeast genome (R64) was done in tiling windows of 110 bp each, spaced at 1 bp intervals, yielding expression and accessibility predictions for nearly all bases in the genome.

To compare to chromatin organization in core promoters, the accessibility predictions were averaged across all yeast promoter sequences to yield a metagene plot, as was done for DNase (43) and nucleosome occupancy (42) data.

To compare the models' predictions to RNA synthesis rates, the model's predicted expression levels for sequences from -450 to -75 relative to the TSS were averaged; to avoid overfitting, this range was optimized on unrelated RNA-seq data (30). We then compared this predicted average expression to the inferred RNA synthesis rate for each gene (24).

Comparing refined and original motifs

The original and model-refined motifs were evaluated for their ability to predict independent ChIP binding and TF mutant gene expression data. The GOMER method (26) was used to get a predicted binding occupancy of each sequence for the original and model-refined motifs. For ChIP data (44), ChIP-chip probes were scanned with the motifs, and their ability to predict ChIP binding for the corresponding TF was evaluated. For TF perturbation experiments (25, 45) promoter sequences were scanned with motifs, and their ability to predict expression changes when the cognate TF is perturbed (mutated, over-expressed, or deleted) was evaluated. In both cases, there were often multiple experiments for the same TF. We repeatedly sampled the data from each experiment (50% of the data sampled randomly 100 times, without replacement), and with each sample calculated the Pearson correlation coefficient between motif-predicted binding and biological measurement (gene expression, ChIP intensity) for both model-refined and original motifs. If the model-refined motif had a Pearson r^2 greater than the original in at least 95% of samples, we considered the experiment to be predicted better by the refined motif. Conversely, if the original motif was better in at least 95% of samples, the experiment was considered to be predicted worse by the refined motif. A model-refined motif was considered to be better than the original if at least one experiment was predicted better and no experiment was predicted worse, while it was considered worse if at least one experiment was predicted worse and no experiment was predicted better. In all other cases, the motifs were considered equal.

Motifs that were regularized out of the model (*i.e.* became neutral PWMs) were not considered in this analysis.

Classifying TFs into activators and repressors by GO annotation

GO terms for yeast genes were downloaded from SGD (75) on Jan. 14, 2017. TFs annotated with a term containing any of "positive regulation of transcription", "transcriptional activator", "activating transcription factor binding", or "positive regulation of RNA polymerase II" were labeled as activators. TFs annotated with "negative regulation of transcription", "transcriptional repressor", "repressing transcription factor binding", or "negative regulation of RNA polymerase II" were labeled as repressors. Any annotated as both or neither were ignored for the purposes of testing for enrichment.

Identifying TFs that act non-linearly

To identify cases where TF activity was not captured accurately by the model, we first examined the relationship between expression level and TF binding directly, but found this to be misleading in many cases, because many TFs have related motifs, leading to seeming non-linearities merely due to multiple TFs acting on related TFBSs (*e.g.* **Figure S2C**). As an alternative, we identified lingering relationships between predicted TF binding and residual expression level (actual minus predicted expression; **Figure 4A**), since the residual expression level is calculated after accounting for the activity of other TFs. Here, the model-learned PWMs and concentration parameters were used to identify promoters containing each TFBS (predicted occupancy 5% or above, relaxing this to 1% if there were fewer than 10^6 such promoters, and subsampling to approximately 10^6 promoters if there were more than 10^7). For each TF, lines of best fit were learned between predicted occupancy of the TFBS and the promoter's residual expression level after the model's fit, and the slopes of these lines calculated at 300 points along the curve, each spanning 1/300th of the data points. The maximum absolute value of the slope of each curve was used to rank TFs by their lingering non-linear relationships (**Figure 4B**).

MNase-Seq analysis

Sequencing reads were mapped to all known promoters in any pTpA library using Bowtie2 (72). Only promoters with at least 20 reads in the input DNA and 1 read in the MNase data were kept for subsequent analysis. Input and MNase counts were scaled within each sample to yield counts per million (CPM) per promoter and the log ratio of MNase to input was compared between replicates and to the model's predicted occupancy, corresponding to log(1-predicted accessibility). To combine MNase replicates, the log ratio of MNase to input was averaged for promoters present in both samples – those in only one sample were ignored. Similarly, pairwise correlations between samples in **Figure 3C** reflect only the promoters common to both samples, and all promoters within the sample when comparing to the model's predictions.

Zinc cluster monomer analysis

The zinc cluster monomeric model was created as above, training on the pTpA+glucose data, but only TFBS motifs representing the canonical zinc cluster monomeric motif (CGG) and one base pair variants (CGG-variants) were provided as motif features and the only parameters learned were motif activities and potentiations, whereas the motifs themselves were held static. In the model, expected binding is proportional to the number of exact motif matches, and so is equivalent to counting the number of the corresponding CGG-variant in the sequence.

Potentiation and activity parameters for each CGG-variant were then compared to those learned for predicting protein binding microarray (PBM) data (47). To this end, we learned a linear model relating the number ($N_{m,p}$) of each CGG-variant (m) within each PBM probe (p) to PBM binding signal for the probe (S_p) for each zinc cluster TFs in the UniPROBE database (47), learning a binding coefficient for each CGG-variant (B_m).

$$S_p = \sum^{m} N_{m,p} B_m + c$$

The degree of binding that could be captured by CGG-variants was then estimated by calculating Pearson's correlation coefficient r between measured PBM binding and predicted binding by these linear models. The Pearson correlation coefficients between PBM-learned binding weights for each TF and the billboard model's CGG-variant coefficients were calculated to see which TF's binding profile was most similar to the learned potentiation/activity weights.

Other models comprised only of simple (1- to 3-mers) motifs were created similarly to the CGG-variant model in order to determine the predictive power of lower-order features (*e.g.*, 1- and 2-mers, reflecting GC content and dinucleotide frequencies, respectively) and how much performance could be gained by including additional simple features in the CGG-variant model (*e.g.*, 1- and 2-mers, and other 3-mers). We also aimed to estimate the degree to which the lower-order models were simply capturing the activity of CGG-variants (*e.g.*, since %G+C and occurrence of CGG are correlated, %G+C is predictive of CGG content and therefore expression). Thus, linear models that take as features the abundance of mono- and di-nucleotide features within each promoter and predict the CGG-variant model's expression level predictions were created, training on the first 8,000 high-quality pTpA+Glu promoters. These were then applied to the last ~2,000 high-quality pTpA+Glu promoters, and a Pearson r^2 (red bars) was calculated for the correlation between this model's predictions and measured expression level. All data in **Figure S5A** are using these same ~2,000 test promoters. We conclude that CGG and related motifs are likely to be the true active motifs because models using only 1- and 2- mers can predict expression about as well as they are able to capture the features of the CGG-variant model, adding 1-, 2-, or 3-mers to the CGG-variant model adds little predictive value, and a model including only the two most impactful CGG variants (CGG and CGC) can explain nearly 53% of expression (**Figure S5B**).

Position and orientation-specific TF activities

In order to identify the approximate fraction of TFs displaying a 10.5 bp helical activity bias, the position-specific activities across the variable promoter region were compared to a 10.5 bp sine wave. First, we regressed out the overall positional activity bias using loess regression (span=0.5; **Figure S6G** – green curves). These long-range trends were subtracted from the data, leaving only the short-range trends (**Figure S6G** – blue curves), which were then compared to a 10.5 bp sine wave for 100 possible alignments of the sine wave, taking the largest magnitude correlation for each TF and strand, and calculating Spearman's correlation coefficient, ρ. As background, the same procedure was performed after first shuffling the position-specific activity biases for 100 permutations of the data per TF. A P-value and AUROC were calculated describing the difference between the randomized and actual data for each model using Wilcoxon's rank sum test.

Data and software availability

Data are available at NCBI's GEO: GSE104903, GSE104878. Open source code for our transcriptional models is available on: https://github.com/Carldeboer/CisRegModels.

Supplemental Figures

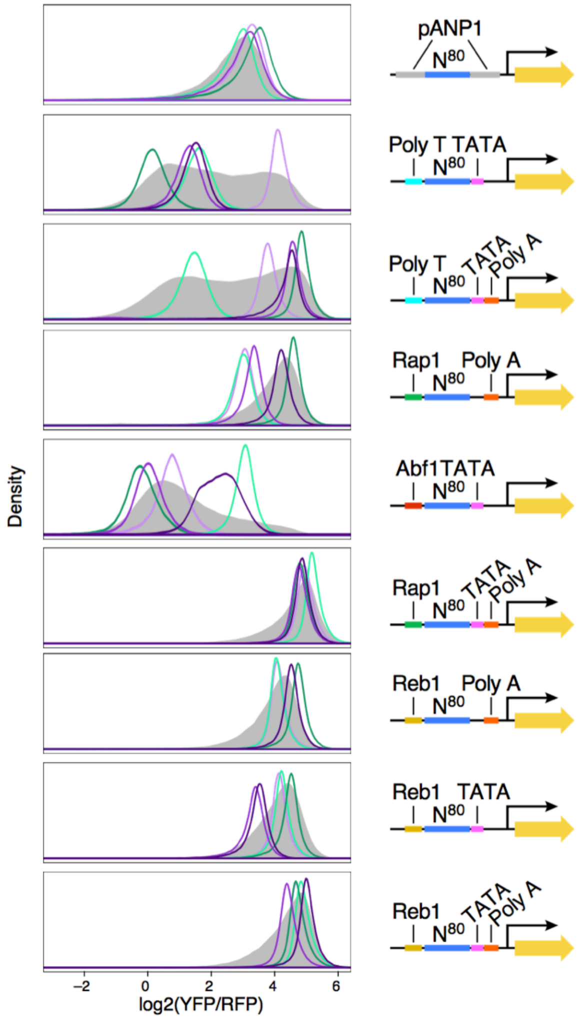

Figure S1. Random DNA yields diverse expression levels in all promoter scaffolds tested. For each promoter scaffold (right), shown are the distributions of expression levels ($\log_2$(YFP/RFP), *x* axis) measured by flow cytometry for the entire library (gray filled curves) and for a few selected clones, each from a different single promoter from the library (colored line curves).

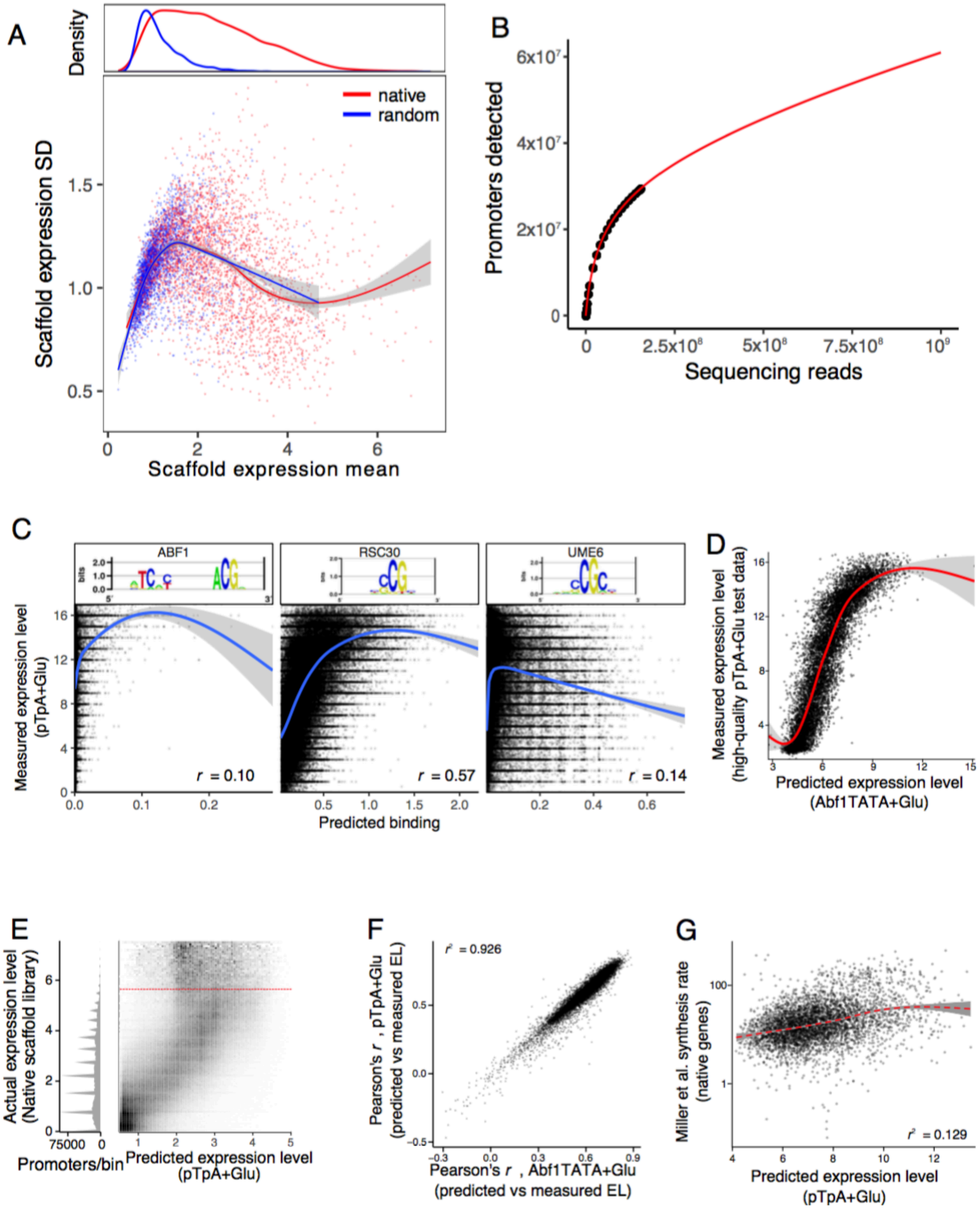

Figure S2: Predictions of billboard model on other test data sets. (**A**) Impact of promoter scaffolds. Relationship between mean (*x* axis) and standard deviation (SD, *y* axis) of expression for random sequences embedded within different promoter scaffolds based on either native promoters (red) or random sequences (blue). Each point is the mean and SD calculated for a single scaffold from the measured expression of all promoters with that scaffold, only considering those appearing in one bin. (**B**) Saturation analysis. Shown are the numbers of distinct promoters detected when subsampling the pTpA+glucose sequencing data (black points), after combining reads from all expression bins. Red curve: promoters projected to be detected with additional sequencing (76). (**C**) Relationship between predicted binding of individual TFs and expression level. Measured expression level (pTpA+Glu data; *y* axis) *vs*. predicted binding (*x* axis) for Abf1 (left), Rsc30 (middle), and Ume6 (right). Top: Motifs. Blue lines: GAM lines of best fit. Gray shaded areas: 95% confidence intervals. (**D**) Predictions of the Abf1TATA+glucose trained model on the high-quality pTpA+glucose test data. Shown are the measured expression levels in the high-quality pTpA+glucose test data (*y* axis) *vs*. the corresponding predictions for these sequences by the billboard model trained on the Abf1TATA+glucose data (*x* axis). Red: GAM fit; Grey shaded area: 95% confidence interval. (**E**) Partial prediction of hybrid native-random promoters for pTpA model. As in **Figure 2F**. (**F**) Models' performance across scaffolds. Pearson *r*'s for how well Abf1TATA (*x* axis) and pTpA (*y* axis) models predict measured expression level for each native promoter scaffold in the scaffold library (points). (**G**) Inferred mRNA synthesis rates of yeast promoters (from (24), *y* axis) *vs*. their predicted expression by the pTpA model. Red line: GAM line of best fit (**Methods**).

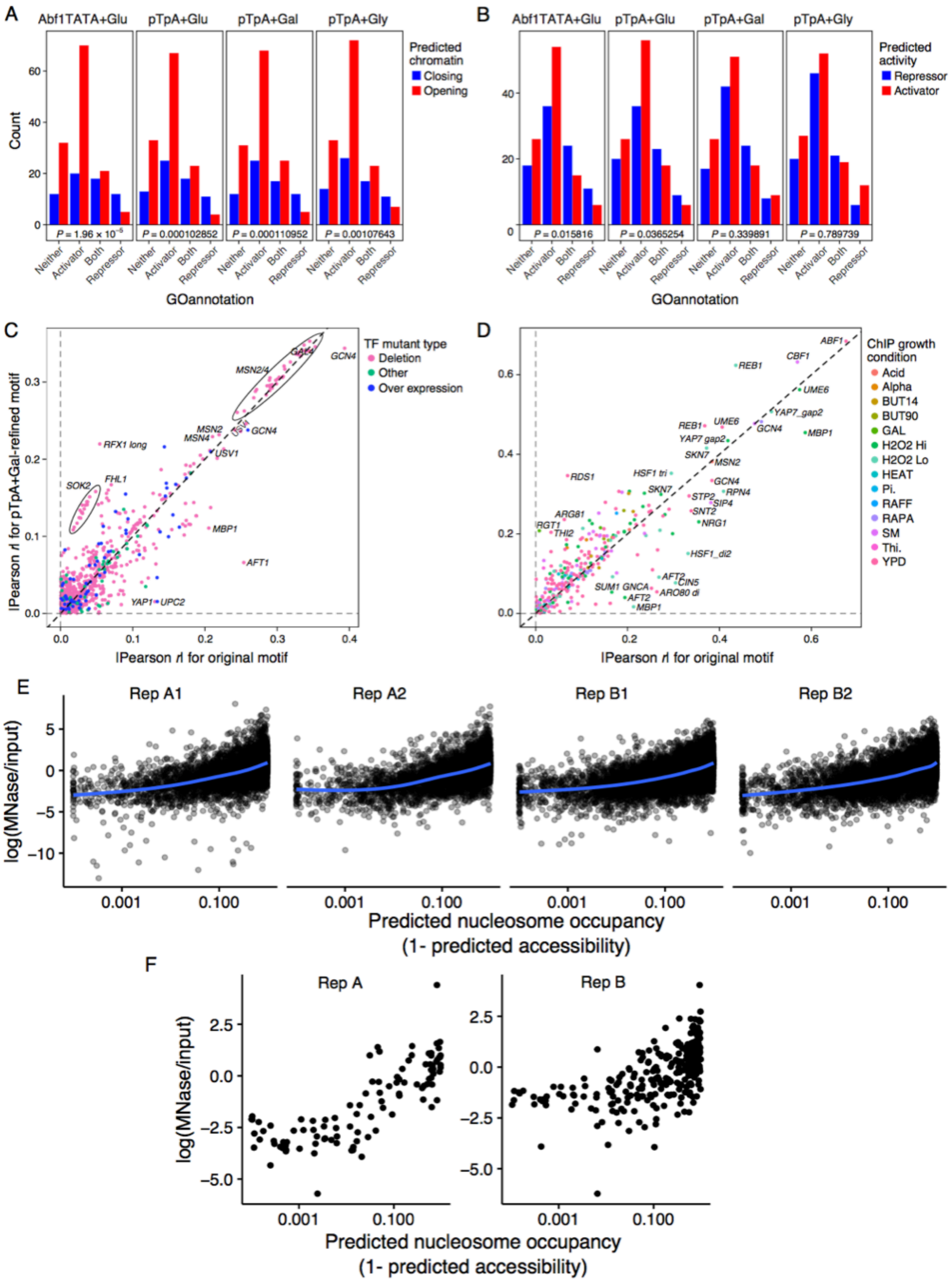

Figure S3. The billboard models identify biochemical activities of TFs. (**A,B**) TF classification into activators and repressors. Shown are the number of TFs classified as activators, repressors, neither, or both in the yeast Gene Ontology (GO, **Methods**) (bars) and whether they are predicted as (**A**) closing (blue) or opening (red) chromatin; or (**B**) repressor (blue) or activator (red), by each model (label on top). Hypergeometric P-values for overlaps between predicted activator/repressor (or chromatin opener/closer), compared with activator/repressor GO annotations are as shown ("neither" and "both" categories are ignored). (**C,D**) Model-refined motifs perform better in predicting TF binding and knockout effects in independent experiment. Shown are the absolute values of the Pearson correlation coefficient (|*r*|) when using either the original motifs (*x* axis) or the pTpA+Gal model-refined motifs (*y* axis) to predict whether (**C**) the gene's expression will change in the corresponding TF mutant (compared to wild type; (25, 45)) based on predicted binding to the promoter, or (**D**) a ChIP probe will be bound by the TF in a ChIP assay (44) based on predicted binding to ChIP probe. (Here, data were not subsampled). Overall, model-refined motifs perform better (points above diagonal), but some perform worse. Reduced performance can be due to condition specific regulators that are minimally active in our tested growth conditions (*e.g.*, Gcn4), redundancy between motifs (*e.g.*, Hsf1 has mono-, di-, and trimeric motifs), and overfitting of the original motif to the test data (*e.g.*, ChIP-derived motifs tested on ChIP data). (**E**,**F**) Prediction of nucleosome occupancy. (**E**) Model predicted (*x* axis) *vs*. measured (MNase-Seq, *y* axis) nucleosome occupancy. Four MNase biological replicates are shown (**Methods**). (**F**) As in E, with replicates averaged, and only promoters present in both replicates shown.

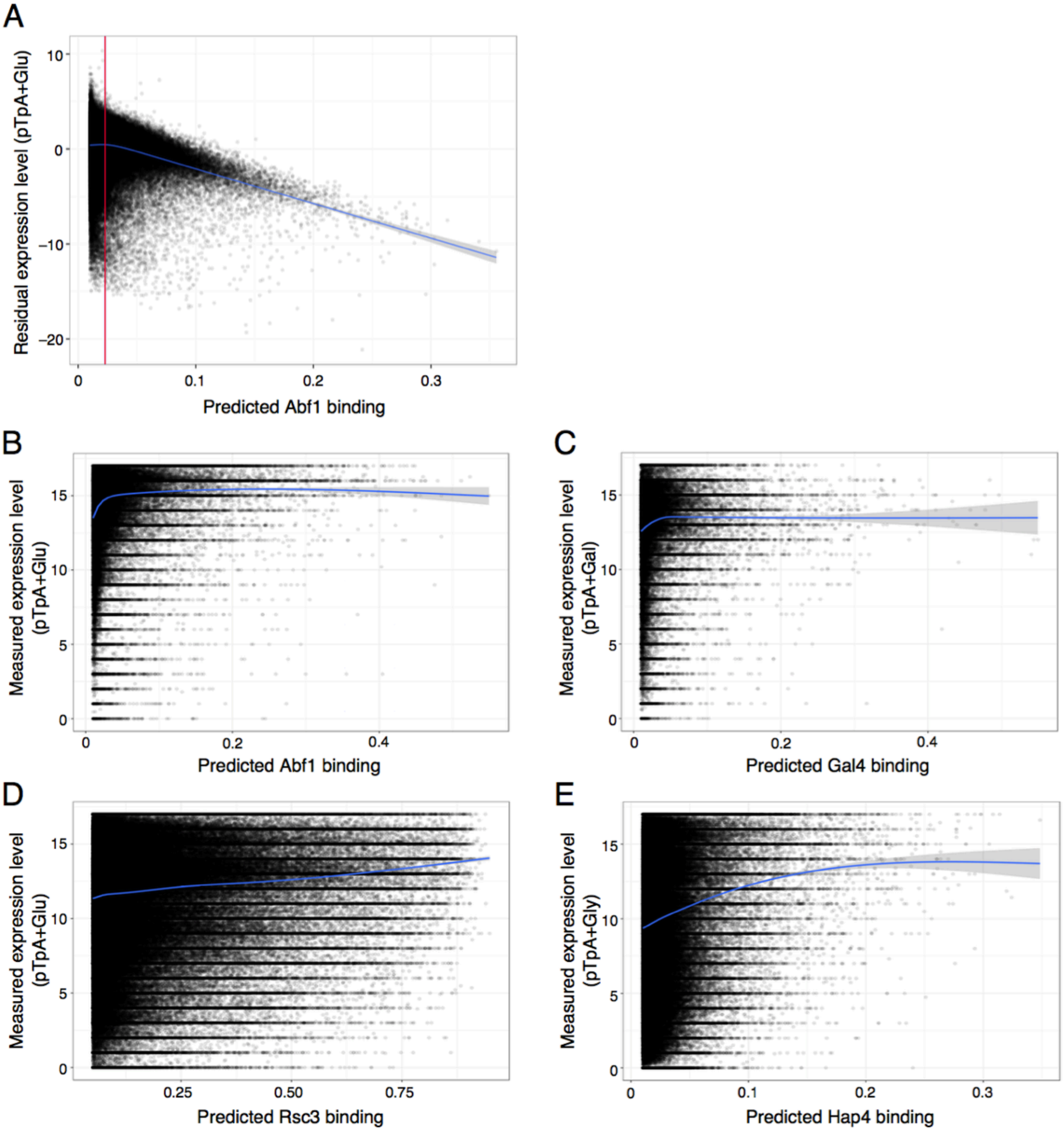

Figure S4: The GRFs and Gal4 have saturating activity. (**A**) Lingering relationship for Abf1. Relationship between predicted Abf1 binding (*x* axis) and residual expression level (*y* axis). Blue line: GAM line of best fit. Vertical red line: estimated saturation point. (**B**-**E**) Relationship between measured expression level (*y* axis) and predicted binding strength (*x* axis) for Abf1 (**B**, in pTpA+glucose), Gal4 (**C**, in pTpA+galactose), Rsc3 (**D**, in pTpA+glucose), and Hap4 (**E**, in pTpA+glycerol).

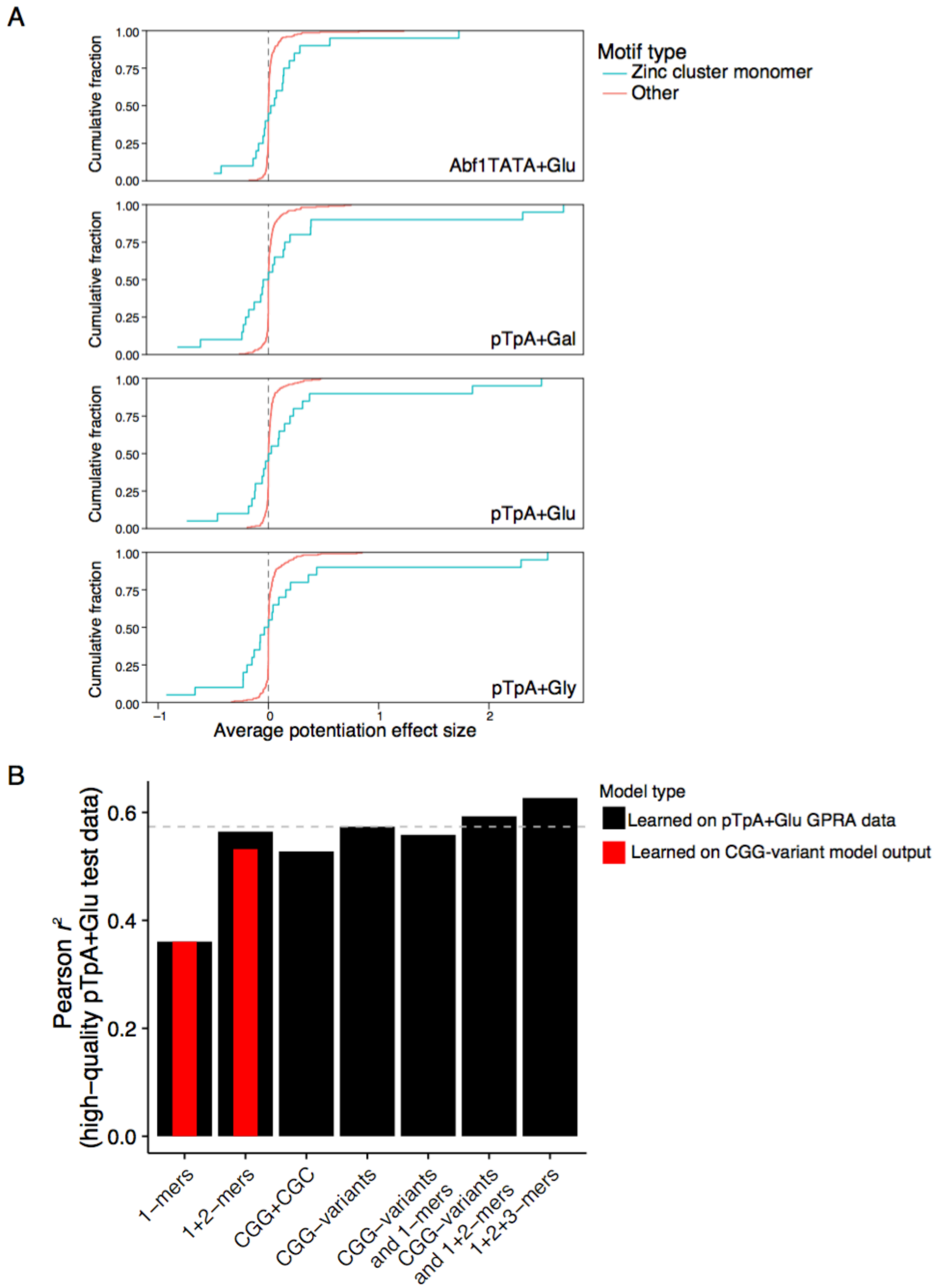

Figure S5. Zinc cluster monomeric TFBSs have large potentiation effect sizes. (**A**) Shown are the cumulative distribution functions of the average potentiation effect sizes (*x* axis) for zinc cluster monomeric TFBS variants (blue), and all other TFBS motifs (pink), in each of the four learned models. (**B**) CGG-variants best explain CGG-variant model performance. Ability of models containing only simple sequence features (up to 3-mers) (bars, *x* axis) to predict high-quality pTpA+glucose test data (Pearson r^2, *y* axis). Models were trained to predict either pTpA+glucose GPRA expression data directly (black bars), or the CGG-variant model's expression output (red bars). The latter asks how well the included features are able to (indirectly) capture CGG-variants, and so how much of their performance can be attributed to CGG-variant activity. Gray dashed line: CGG-variant model performance. Marginal gain in performance of CGG-variant model supplemented with 2- and 3-mers could result from other important motifs being partly captured (*e.g.*, poly-A).

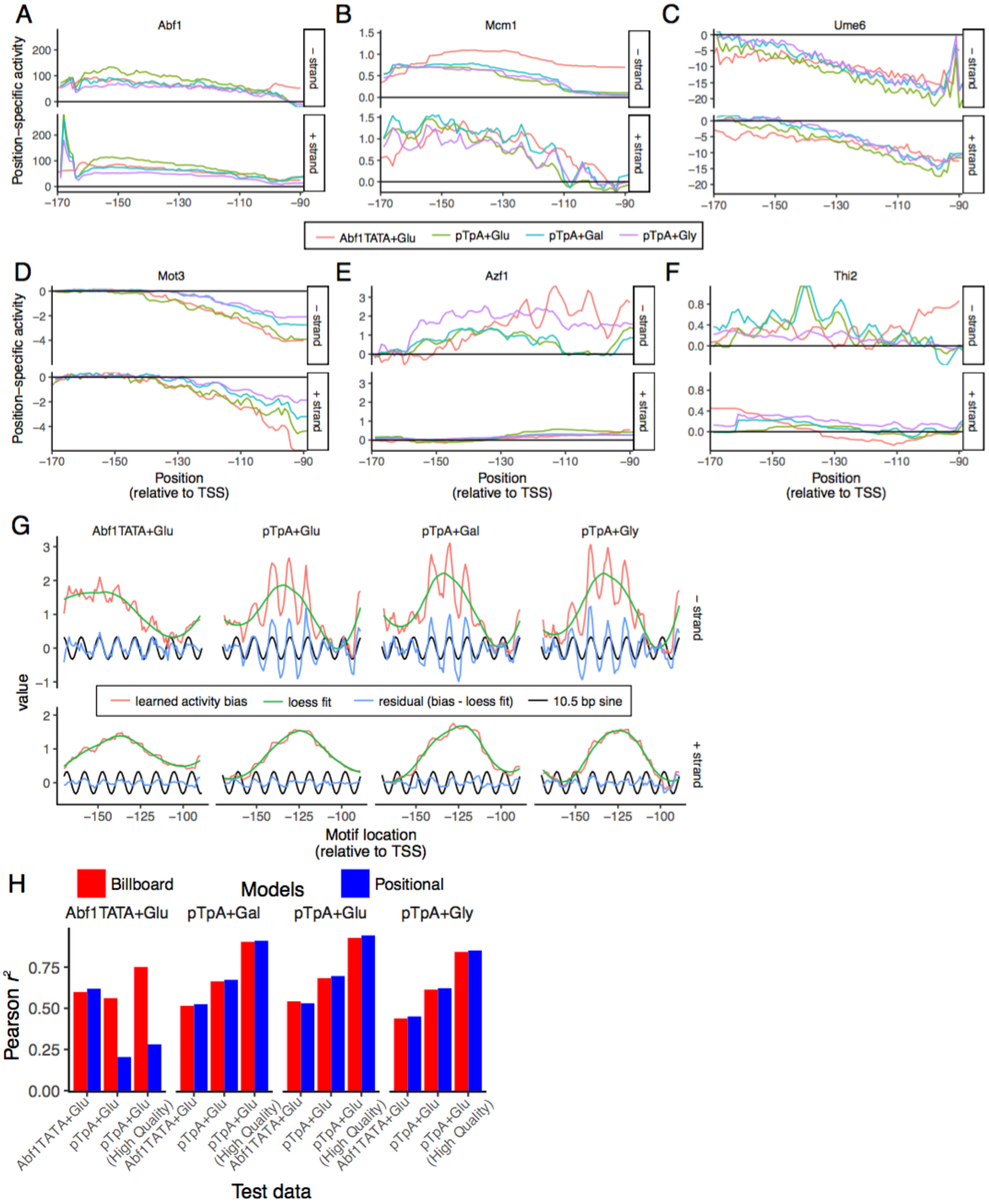

Figure S6. Positional preferences of TFs are prevalent and context-dependent. (**A-F**) Position and strand preferences. Learned activity parameter values (*y* axis) for motifs in each position (*x* axis) and strand orientation (upper and lower panels) for each model (colors), for (**A**) Abf1, (**B**) Mcm1, (**C**) Ume6, (**D**) Mot3, (**E**) Azf1, and (**F**) Thi2. (**G**) Capturing helically biased positional preferences. Plot shows, for each location within the promoter (*x* axis), the learned activity bias parameters (red curve; as in **Figure 6B**) for the poly-A motif, long-range trend captured by a loess fit (green), and short-range residual activity bias after subtracting loess fit (blue) with reference 10.5 bp sine waves (black) for the minus strand (top) and plus strands (bottom) for the four different models (columns). (**H**) Modeling positional preferences increases predictive accuracy within the same scaffold but can drastically decrease it between scaffolds. For each training data set (four sub-panels) for both model types (colors), the Pearson r^2 (*y* axis) capturing performance on each test dataset (*x* axis).

Table S1: Promoter scaffolds included in the scaffold library. Sequences include 80 Ns in place of the random 80-mers and begin 13 bp upstream of the theoretical TSS.

Table S2: Motifs used in this study. Motif IDs are from the YeTFaSCo database (25). Motifs excluded from the motif frequency analysis (**Figure 1B**) are indicated.