Supplemental Material for:

Designing realistic regulatory DNA with autoregressive language models

Avantika Lal[1,*], David Garfield[2], Tommaso Biancalani[1] and Gokcen Eraslan[1,*]

[1]Biology Research | AI Development, gRED Computational Sciences, Genentech, South San Francisco, CA 94080, USA
[2]OMNI Bioinformatics and Department of Regenerative Medicine, Genentech, South San Francisco, CA 94080, USA

Table of Contents

Supplemental Figures

Supplemental Code

Supplemental Code S1: Source code for regLM.

Supplemental References

Supplemental Methods

Regression models

Regression models were trained to take as input a one-hot encoded CRE sequence and predict as output its activity. All regression models were based on the Enformer architecture (Avsec et al. 2021) and were built using the enformer-pytorch package (https://github.com/lucidrains/enformer-pytorch). All regression models were trained for 10 epochs on 1 NVIDIA A100 GPU using the Adam optimizer. Performance was measured as the Pearson's correlation between measured and predicted CRE activity on the test set.

Yeast promoters

All regression models were Enformer-based models with 3 convolutional blocks followed by 1 transformer encoder layer. The first convolutional block has 384 channels. Each model has a single linear output layer that predicts promoter activity in one of the two media. Models were trained with learning rate 5×10^{-4} and batch size 2048. Validation set loss was measured after each epoch and the model with lowest validation loss was saved.

Human enhancers

For human data, we downloaded the pre-trained Enformer model and reduced its size by dropping the last 8 transformer encoder layers (leaving 7 convolutional blocks and 3 transformer encoder layers). For each of our regression models, we added a single linear output layer that predicts the total measured expression for the input sequence in a specific cell type.

For the regLM-matched regression models, we fine-tuned the model on the same sequences as regLM. These models were fine-tuned with learning rate 10^{-4}, batch size 1024, and MSE loss. During training, examples with each label were sampled from the training set with a weight inversely proportional to the frequency of the label, allowing the model to focus on cell type-specific enhancers that were extremely rare. Validation set loss was measured after each epoch and the model with lowest validation loss was saved.

Additional Models for Human CREs

Lentiviral MPRA model

Training data were obtained from (Agarwal et al. 2023), specifically the three cell line experiment detailed in supplemental tables 6 (200bp sequences) and 7 (log-transformed MPRA

values). The dataset was restricted to sequences from autosomes and not overlapping ENCODE blacklist regions. Sequences from Chromosome 10 were used for validation and Chromosome 11 for testing, and sequences from the remaining autosomes were used for training.

We took the pre-trained Enformer model, dropped all but the first transformer layer, and fine-tuned the model on this dataset using a batch size of 512, learning rate of 10^{-4}, and MSE loss for 10 epochs with the Adam optimizer.

The Pearson's correlation coefficient between the model predictions and the true values, in the test set, was: 0.84 for HepG2, 0.84 for K562, and 0.85 for WTC11.

ATAC-seq model

We downloaded public ATAC-seq data in the form of processed BAM files from the ENCODE Project for the following cell lines: K562 (ENCFF534DCE), HepG2 (ENCFF624SON), GM12878 (ENCSR095QNB), IMR90 (ENCFF715NAV), WTC11 (ENCFF240QKT), and SK-N-SH (ENCFF270AGJ). As standard, SK-N-SH cells were treated with trans-retinoic acid prior to sequencing to induce neural-like differentiation. In addition, we downloaded raw fastq files for Jurkat cells from the SRA archive (SRX7785407). These raw reads were then aligned to the hg38 genome following the ENCODEv4 standards.

Peaks for each cell line were called using MACS3 (Zhang et al. 2008) callpeaks with the additional parameters "--nomodel --shift -100 --extsize 200". We then created a unified peak set as described in (Corces et al. 2018) with an SPM = 2 and an extension of 250bp. This resulted in a uniform peak set consisting of 366,776 500 bp regions to which we added an additional 15 percent of 500 bp regions overlapping no known peak to serve as no-signal background regions during modeling. These peak regions were binarized per cell line by overlapping them with cell line specific MACS3 peak calls using an SPM value of 5. Peaks were then resized to 200bp around the center. Peaks overlapping ENCODE blacklist regions were dropped. Of the remaining peaks, peaks from Chromosome 10 were used for validation, peaks from Chromosome 11 were used for testing, and peaks from all remaining autosomes were used for training.

We took the pre-trained Enformer model, dropped all but the first transformer layer, and added a head (linear layer) with a sigmoid activation function to predict the probability of the input sequence being a peak in each cell line. We fine-tuned the model on this dataset using a batch size of 512, learning rate of 10^{-4}, and binary cross-entropy loss for 10 epochs with the Adam optimizer. Validation set loss was measured after each epoch and the model with lowest validation loss was saved.

The average precision of the trained model on the test set, per cell line, was: 0.73 in GM12878, 0.70 in HepG2, 0.73 in IMR90, 0.73 in Jurkat cells, 0.75 in K562, 0.70 in SK-N-SH, and 0.77 in WTC11.

At inference, predicted probabilities were thresholded with a cutoff of 0.5 to generate final predictions.

CATLAS scATAC-seq model

We downloaded binarized single-cell pseudobulk chromatin accessibility matrices from the CATLAS project (Zhang et al. 2021). Peaks were resized to 200 bp around the center. Cell types with accessibility in less than 3% of peaks were discarded, which reduced the number of cell types from 222 to 204. Of the remaining peaks, peaks from Chromosome 7 were used for validation, peaks from Chromosome 13 were used for testing, and peaks from all remaining autosomes were used for training.

We took the pre-trained Enformer model, dropped all but the first transformer layer, and added a head (linear layer) with a sigmoid activation function to predict the probability of the input sequence being a peak in each cell type. We fine-tuned the model on this dataset using a batch size of 512, learning rate of 10^{-4}, and binary cross-entropy loss for 10 epochs with the Adam optimizer. Validation set loss was measured after each epoch and the model with lowest validation loss was saved.

The average precision of the trained model on the test set was 0.53 across all 204 cell types. For the cell types shown in Fig. 3H, it was: 0.61 in Hepatocytes, 0.56 in Fetal Hepatoblast, 0.55 in Fetal Erythroblast 1, 0.51 in Fetal Erythroblast 2, 0.58 in Fetal Erythroblast 3, and 0.55 in Fetal Erythroblast 4.

At inference, predicted probabilities were thresholded with a cutoff of 0.5 to generate final predictions.

Full-stack ChromHMM model

Annotations for the hg38 genome generated using the full-stack ChromHMM model were obtained from (Vu and Ernst 2022). These annotations were extended to 1024bp from the center and restricted to the autosomes. The fine scale categories were collapsed by stripping off the prefix and suffix values to generate 16 broad categories of annotations (Acet (acetylations), BivProm (bivalent promoter), DNase, EnhA (Enhancers), EnhWk (Weak enhancers), GapArtf (Assembly gaps and artifacts), HET (heterochromatin), PromF (Flanking promoter), ReprPC (Polycomb repressed), Quies (Quiescent), TSS (Transcription start site), Tx (Transcription), TxWk (Weak transcription), TxEnh (Transcribed Enhancer), TxEx (Exon & Transcription), and znf (ZNF genes)). The resulting element set was downsampled to have a maximum of 250,000 instances of any given category. The dataset was also restricted to autosomes with Chromosome 7 used for validation and Chromosome 13 for testing, excluding regions overlapping ENCODE blacklist regions.

We took the pre-trained Enformer model, dropped all but the first transformer layer, and added a linear layer with a Softmax activation function to predict class probabilities. We fine-tuned the model to perform multiclass classification on this dataset. For fine-tuning, we used a batch size of 512, learning rate of 10^{-4}, reverse complement augmentation at randomly selected examples, and cross-entropy loss for 14 epochs with the Adam optimizer. Validation set loss was measured after each epoch and the model with lowest validation loss was saved.

The average precision of the model for each class, in the test set, was: 0.81 for DNase, 0.81 for TSS, 0.67 for GapArtf, 0.64 for PromF, 0.57 for HET, 0.47 for Quies, 0.41 for BivProm, 0.4 for EnhA, 0.40 for EnhWk, 0.32 for Tx, 0.32 for TxEx, 0.31 for Acet, 0.29 for TxEnh, 0.27 for TxWk, 0.20 for ReprPC, and 0.14 for znf.

At inference, each sequence was predicted to belong to the class with the highest predicted probability.

Generating synthetic yeast promoters with benchmark methods

In order to benchmark regLM against existing commonly used approaches, we ran five other methods to generate synthetic yeast promoters: Directed Evolution, Ledidi, AdaLead, FastSeqProp and Simulated Annealing. These are all model-guided methods that iteratively make edits to a starting sequence to maximize a defined objective function using a trained predictive model (the 'oracle').

We randomly chose sequences that had been measured to have low activity in all conditions (label 00) as the starting sequences. To ensure a fair comparison to regLM-generated sequences, the regression models trained on the same data as regLM were used as oracles. All approaches were each run multiple times with a different starting sequence each time, to generate diverse synthetic CREs. We used the CODA software package (Gosai et al. 2023 Aug 9) to run AdaLead, FastSeqProp and Simulated Annealing.

For yeast promoters, we aimed to generate promoters with high activity in both media. The objective function for all methods was the mean predicted activity in the two media. All methods were run 200 times, each time with a different initial sequence, resulting in a diverse set of synthetic promoters from each method. Parameters were tuned to achieve synthetic sequences with similar predicted activity to those generated by regLM.

The following parameters were used:
Directed Evolution: 10 iterations
Ledidi: max_iter=800, l=20, lr=4×10^{-3}
AdaLead: model_queries_per_batch=75
FastSeqProp: n_steps=5, learning_rate=0.1
Simulated Annealing: n_steps=220, n_proposals=5

For each method, each regLM-generated strong promoter was matched to the method-generated sequences that were closest to it in predicted activity (measured by the mean squared error across both conditions), resulting in a matched set of 200 putative strong promoters designed by each method. Thus, since the various groups of synthetic elements have highly similar predicted activity, we can compare their sequence content to assess which approach gives rise to more biologically realistic sequences while reaching the same objective.

Supplemental Notes

Supplemental Note S1: Regulatory programs used by synthetic yeast promoters generated with different methods

We clustered yeast promoters into groups containing different combinations of motifs and partitioned the synthetic promoters generated by each method into clusters. 23% of the strong promoters in the test set mostly fell into cluster 0, which is characterized by ABF1, SKO1, YAP6, and CIN5 motifs, followed by 18% in cluster 1 (RSC3, RSC30, DAL82, SUT1, and TEA1 motifs), 17% in cluster 2 (ECM22, HAL9, ERT1, CAT8, and ASG1 motifs) and 15% in cluster 3 (CHA4, PDR3, PDR1, IME1, and RDS1 motifs). regLM promoters partitioned similarly to test set promoters, with no significant differences. However, the most notable feature in all other methods was a strong enrichment for clusters 6 (CUP2, EDS1, STB3, SUM1, and SFP1 motifs) and 8 (ARR1, CIN5, FKH1, RLM1, and SPT15 motifs). In addition, promoters generated by FastSeqProp were enriched in clusters 2 (ECM22, HAL9, ERT1, CAT8, and ASG1 motifs) and 7 (characterized by the presence of XBP1 motifs).

Supplemental Note S2: Choice of labels

In the above experiments, we generated labels by dividing yeast promoters into 5 equal bins per medium and dividing human enhancers into 4 unequal bins per cell type. In theory, any number of class labels can be used, and our package allows users to choose the number of labels and to define them in any way. However, the more we subdivide our data, the less information the model will have to learn an accurate distribution of each class. In contrast, having fewer subdivisions may make it more difficult for the model to share information across similar categories.

The resolution of the data should also be kept in mind. In the case of the yeast promoter GPRA assay, promoters were sorted into bins based on their measured activity; however, many promoters were measured only once and so their measured values are not precise. Hence, we chose to use a lower resolution.

To demonstrate that our method can work with different labeling schemes, we have also re-trained the yeast model using 10 label classes instead of 5. The 10-class model performs well at generating sequences with label-consistent expression (Fig. S30).

In the case of human enhancers, the cell type specific enhancers we aimed to design were extremely rare in the dataset. Therefore instead of dividing the dataset into equally sized bins, we assigned sequences with extremely high activity (in the 95th percentile for each cell line) to a separate bin.

Supplemental Tables

	Transcription Factor	**Motif**	**Contribution**	**TOMTOM q-value (test set)**	**TOMTOM q-value (regLM generated promoters)**
1	ABF1	MA0265.3	Positive	4.5×10^{-8}	4.4×10^{-3}
2	REB1	MA0363.3	Positive	1.2×10^{-5}	5.0×10^{-4}
3	RAP1	MA0359.3	Positive	1.9×10^{-4}	3.5×10^{-5}
4	TBF1	MA0403.3	Positive	3.4×10^{-3}	N/A
5	RTG3	MA0376.2	Positive	6.3×10^{-3}	N/A
6	RSC3	MA0374.2	Positive	0.03	0.02
7	SFP1	MA0378.2	Positive	0.04	0.01
8	STB3	MA0390.2	Positive	0.05	0.02
9	UME6	MA0412.3	Negative	1.4×10^{-6}	2.9×10^{-4}
10	RPH1	MA0372.2	Negative	0.014	N/A

Supplemental Table S2. Motifs identified by TF-MoDISco and TOMTOM on test set and regLM generated promoters. N/A indicates that no motif with a significant match was found by TF-MoDISco.

	Test Set	**regLM**	**Evolution**	**Evolution (V)**	**Ledidi**	**Ada-Lead**	**FastSeq -Prop**	**Simulated Annealing**

Differential *k*-mers w.r.t. Test Set	N/A	**0**	86	122	51	33	27	71
Fraction of Nearest Neighbors in Test Set (*k*-mer frequency)	0.88	**0.91**	0.45	0.33	0.53	0.73	0.73	0.87
SVM AUROC vs. Test Set (*k*-mer frequency)	N/A	**0.5**	0.52	0.64	0.57	**0.5**	**0.5**	0.51
Differential motifs w.r.t. Test Set	N/A	**0**	22	42	5	6	1	**0**
Fraction of Nearest Neighbors in Test Set (motif frequency)	0.85	**0.83**	0.67	0.47	0.71	0.75	0.82	0.82
SVM AUROC vs. Test Set (motif frequency)	N/A	**0.5**	0.52	0.53	0.53	**0.5**	**0.5**	**0.5**
Differential motif pairs w.r.t. Test Set	N/A	**1**	285	439	153	125	89	21
Differential motif positioning w.r.t. Test Set	N/A	**0**	1	9	2	3	**0**	2
Pearson's Rho (fraction of pair in same orientation) with test set	N/A	**0.52**	0.51	0.34	0.47	0.45	0.45	0.41
Differential inter-motif distance w.r.t. Test Set	N/A	1897	1853	**1846**	2052	1909	1945	1968
Fraction of Nearest Neighbors in Test Set (model embedding)	0.88	**0.88**	0.80	0.53	0.79	0.81	0.86	0.86

Supplemental Table S5. Additional metrics for all sets of synthetic yeast promoters as well as strong promoters from the test set. The best performances for each method are highlighted in bold. Evolution (V) represents synthetic promoters generated by (Vaishnav et al. 2022).

	Specificity	Transcription Factor	Motif	Contribution	TOMTOM q-value
1	HepG2-specific	HNF1A	MA0046.3	Positive	0.000070
2		HNF1B	MA0153.2	Positive	0.000064
3		HNF4A	MA0114.5	Positive	0.01
4		HNF4G	MA0484.3	Positive	0.01
5		FOXD1	MA0031.2	Positive	0.02
6		FOXA2	MA0047.4	Positive	0.06
7	K562-specific	GATA2	MA0036.4	Positive	0.06
8		KLF4	MA0039.5	Positive	0.05
9	SK-N-SH-specific	AP-1	MA0099.4	Positive	0.07

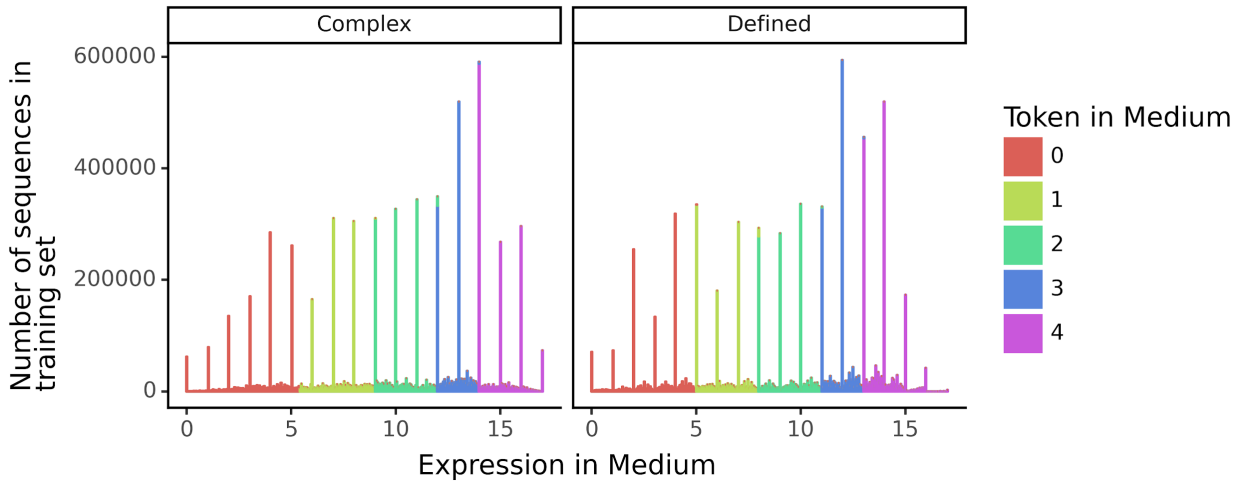
Supplemental Table S9. Motifs identified by TF-MoDISco and TOMTOM on the 100 regLM generated human enhancers specific to each cell type.

Transcription Factor	Motif	Group	Mean no. of sites	Fraction sequences with motif	Log2 Fold Change in abundance	FDR-adjusted p-value
HepG2-specific						
HNF1B	MA0153.2	AdaLead (G)	1.77	0.64	1.96	3.2×10^{-8}
		FastSeqProp (G)	1.76	0.63	1.94	3.1×10^{-7}
		Simulated Annealing (G)	2.13	0.73	2.56	3.0×10^{-13}
		regLM (top 100)	**1.80**	**0.79**	**2.01**	**1.2×10^{-11}**
		regLM (all)	**0.81**	**0.31**	**N/A**	**0**
CEBPA	MA0102.5	AdaLead (G)	1.27	0.62	1.8	2.5×10^{-6}
		FastSeqProp (G)	0.64	0.42	0.29	0.92

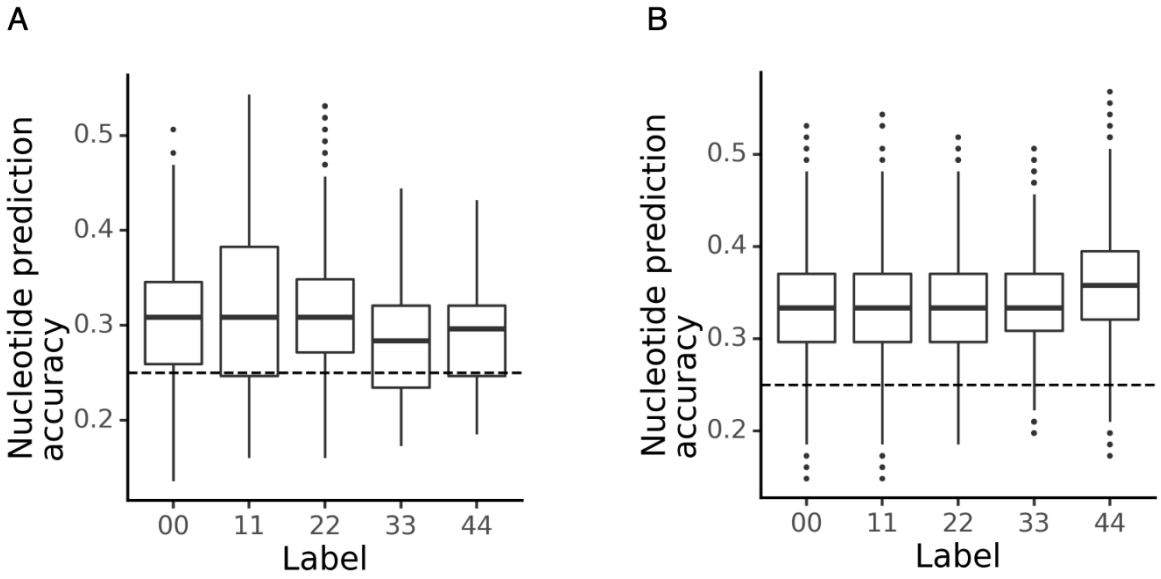
		Simulated Annealing (G)	0.73	0.5	0.55	0.3
		regLM (top 100)	**0.93**	**0.74**	**1.06**	**1.2×10^{-5}**
		regLM (all)	**0.54**	**0.45**	**N/A**	**0**
K562-specific						
GATA1	MA0035.5	AdaLead (G)	1.69	0.89	3.6	2.5×10^{-34}
		FastSeqProp (G)	1.12	0.55	2.5	3.8×10^{-9}
		Simulated Annealing (G)	1.43	0.82	3.1	6.9×10^{-27}
		regLM (top 100)	**0.72**	**0.72**	**1.6**	**3.7×10^{-12}**
		regLM (all)	**0.30**	**0.29**	**N/A**	**0**
GATA1::TAL1	MA0140.3	AdaLead (G)	1.49	0.79	3.8	5.4×10^{-29}
		FastSeqProp (G)	1.95	0.90	4.6	8.1×10^{-43}
		Simulated Annealing (G)	1.62	0.85	4.1	2.7×10^{-35}
		regLM (top 100)	**0.49**	**0.49**	**1.5**	**4.7×10^{-6}**
		regLM (all)	**0.21**	**0.21**	**N/A**	**0**
SNAI3	MA1559.2	AdaLead (G)	1.54	0.77	1.0	1.9×10^{-4}
		FastSeqProp (G)	1.45	0.76	0.8	9.3×10^{-4}
		Simulated Annealing (G)	1.38	0.76	0.7	4.5×10^{-3}
		regLM (top 100)	**0.34**	**0.26**	**-2.1**	**1.1×10^{-8}**
		regLM (all)	**1.02**	**0.56**	**N/A**	**0**
NFKB1	MA0105.4	AdaLead (G)	0.39	0.19	-1.2	0.12
		FastSeqProp (G)	0.31	0.15	-1.5	3.4×10^{-2}
		Simulated Annealing (G)	0.31	0.15	-1.5	3.1×10^{-2}
		regLM (top 100)	**0.09**	**0.07**	**-3.5**	**3.9×10^{-4}**
		regLM (all)	**0.72**	**0.27**	**N/A**	**0**

Supplemental Table S10. Abundance of selected motifs in synthetic cell type-specific human enhancers generated by different methods. Fold changes were calculated for each group using the set of all regLM-generated enhancers (regLM (all)) as the reference group. P-values were calculated using the two-sided Wilcoxon test.

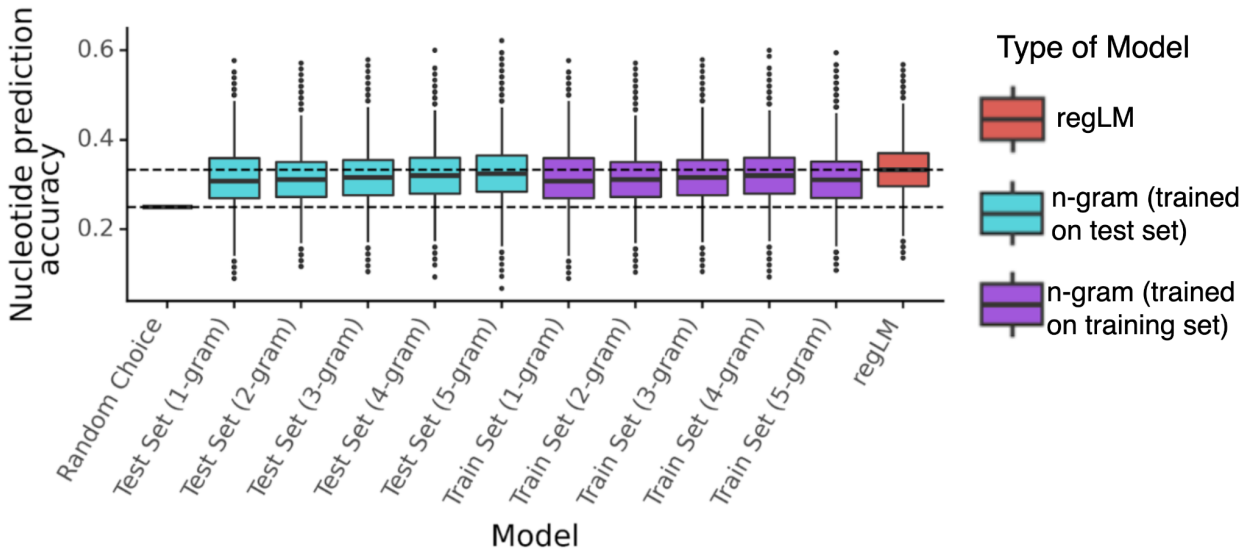
Supplemental Figures



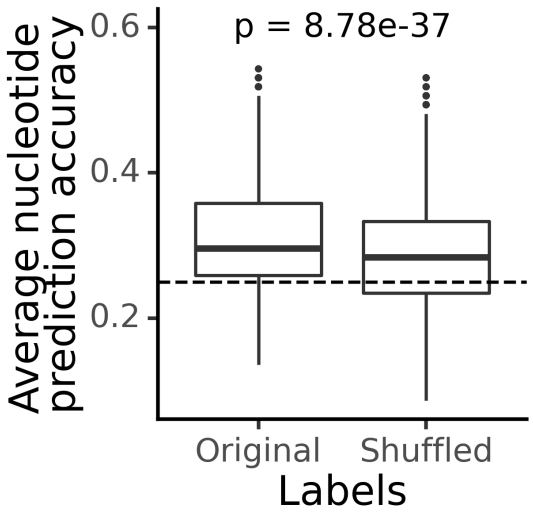
Supplemental Figure S1. Histograms of measured promoter activity in the two media, colored by the assigned token in each medium. Each promoter was assigned a token ranging from 0-4 where 0 corresponds to the lowest quintile of measured activity and 4 corresponds to the highest. This procedure was performed separately for measurements in complex and defined media.



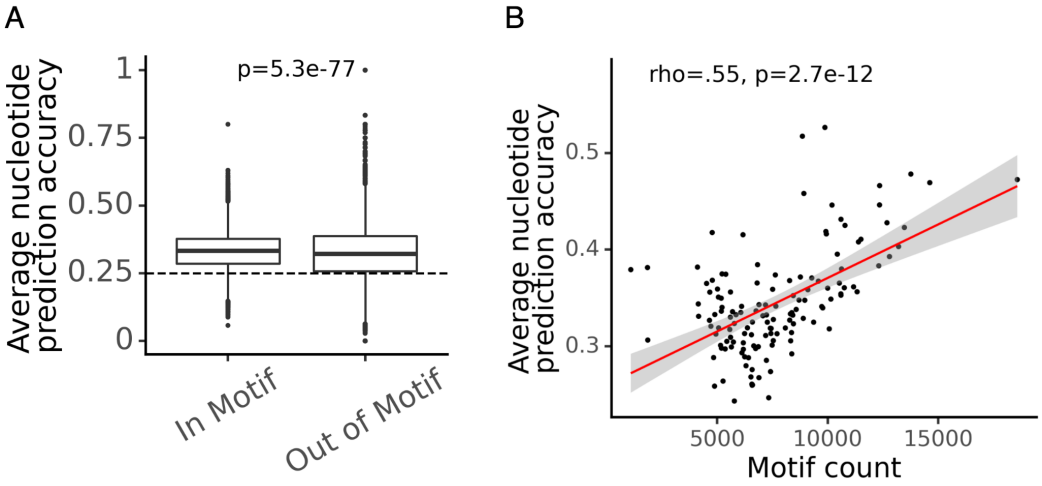
Supplemental Figure S2. A. Boxplots showing the average per-nucleotide prediction accuracy of the yeast regLM model on **A)** 3,922 native yeast promoters and **B)** 50,000 promoters from the test set, separated by the promoter class labels. Only the 5 most common labels (00, 11, 22, 33, 44) are shown. The dashed lines represent the accuracy of 0.25 expected by chance.



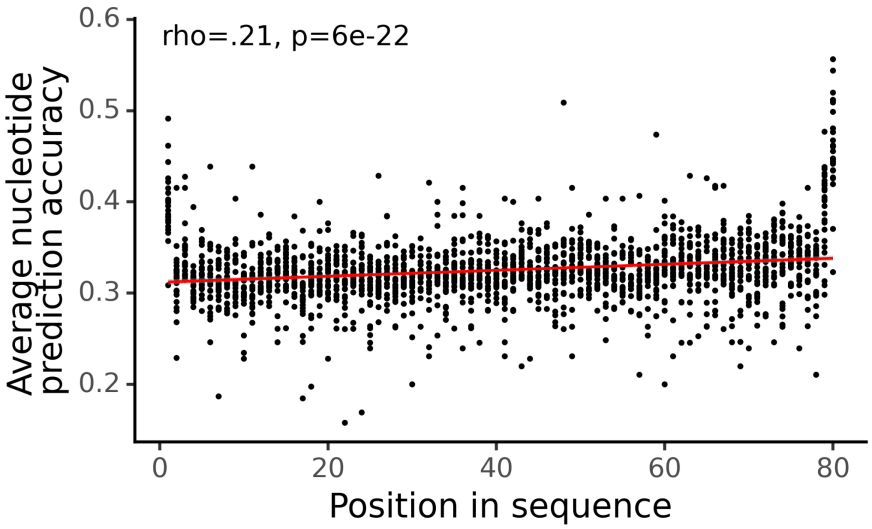
Supplemental Figure S3. Box plots showing the average per-nucleotide prediction accuracy of the yeast regLM model on 50,000 promoters from the test set, compared to baseline models. Dotted lines show the performance of regLM (0.338) and random chance (0.25).



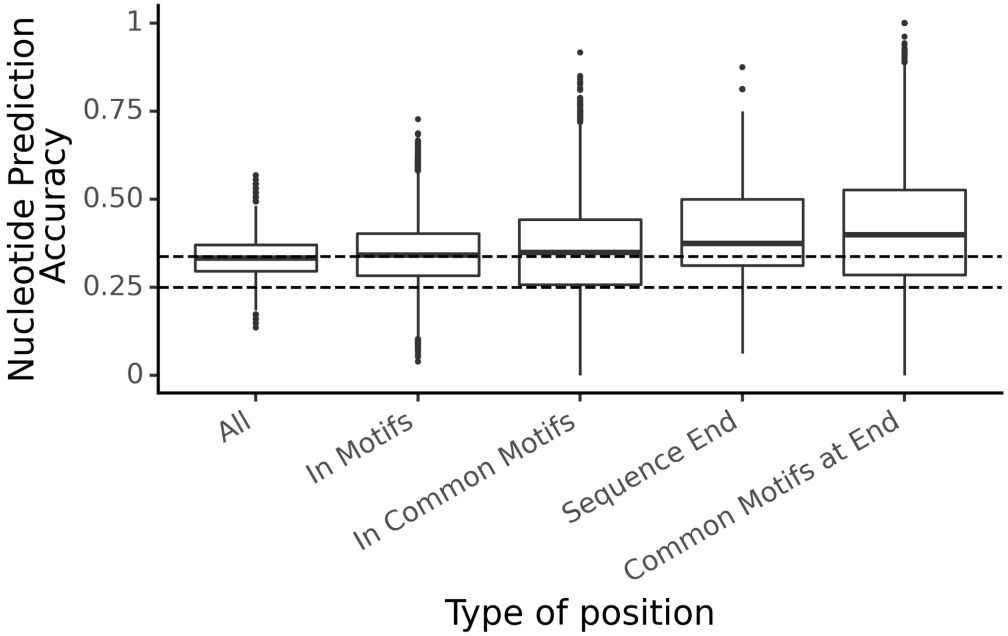
Supplemental Figure S4. Boxplots showing the average per-nucleotide prediction accuracy of the yeast regLM model on 3,922 native yeast promoters, before and after shuffling the labels across sequences. The dashed line represents the accuracy of 0.25 expected by chance.



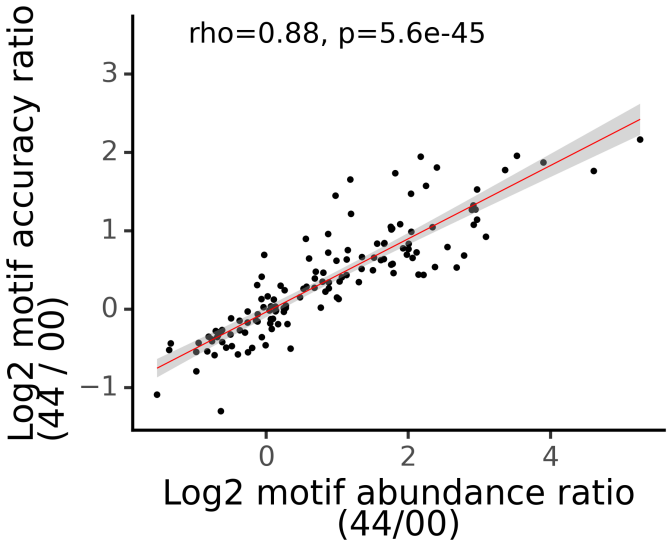
Supplemental Figure S5. A) Average nucleotide prediction accuracy of the regLM model on 50,000 promoters in the test set, for nucleotides within known TF-binding motifs versus those outside. The dashed line represents the accuracy of 0.25 expected by chance. **B)** Scatterplot showing the accuracy of the regLM model on nucleotides within TF binding motifs in the test set. Each point represents a TF binding motif. The x-axis shows the number of occurrences of the motif across the test set. The y-axis shows the average accuracy of the regLM model on all instances of the motif in the test set. The red line shows the linear fit to the data.



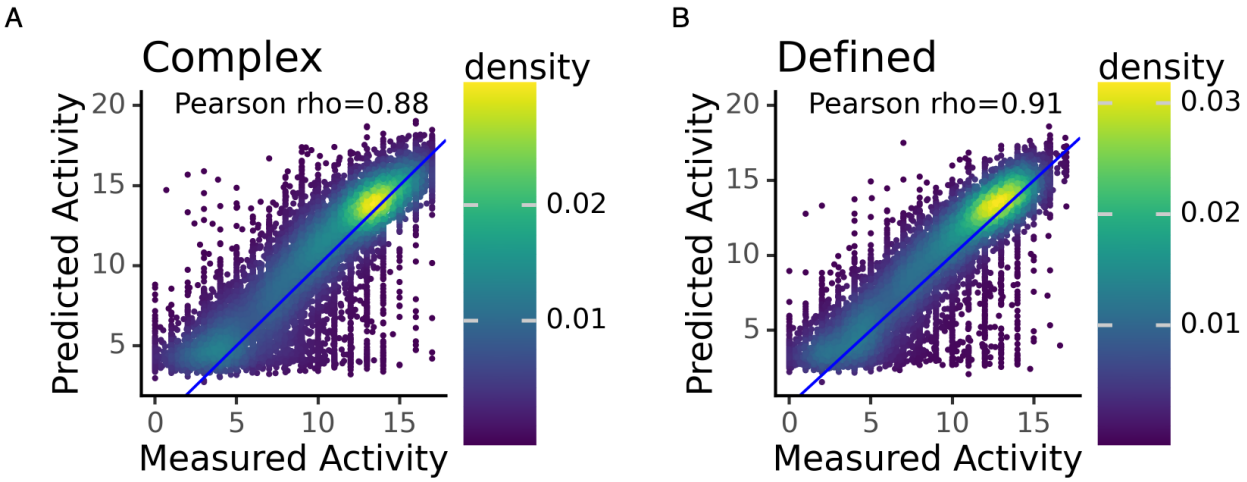
Supplemental Figure S6: Scatter plot showing the accuracy of the regLM model on nucleotides at different positions along the sequence length, for the test set.



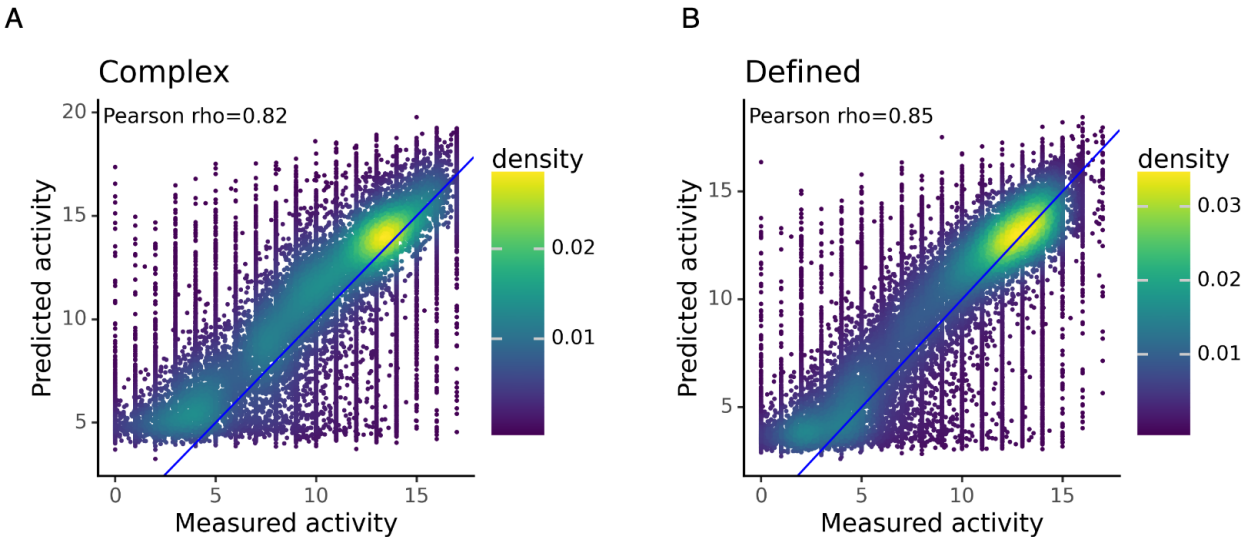
Supplemental Figure S7: Accuracy of the regLM model on 50,000 promoters in the test set, over different categories of nucleotides. “Common Motifs” are nucleotides within the 50 most common motifs in the test set. “Sequence End” comprises the last 15 bases of each 80 bp long sequence. Dotted lines show the mean performance of regLM across all nucleotides (0.338) and random chance (0.25).



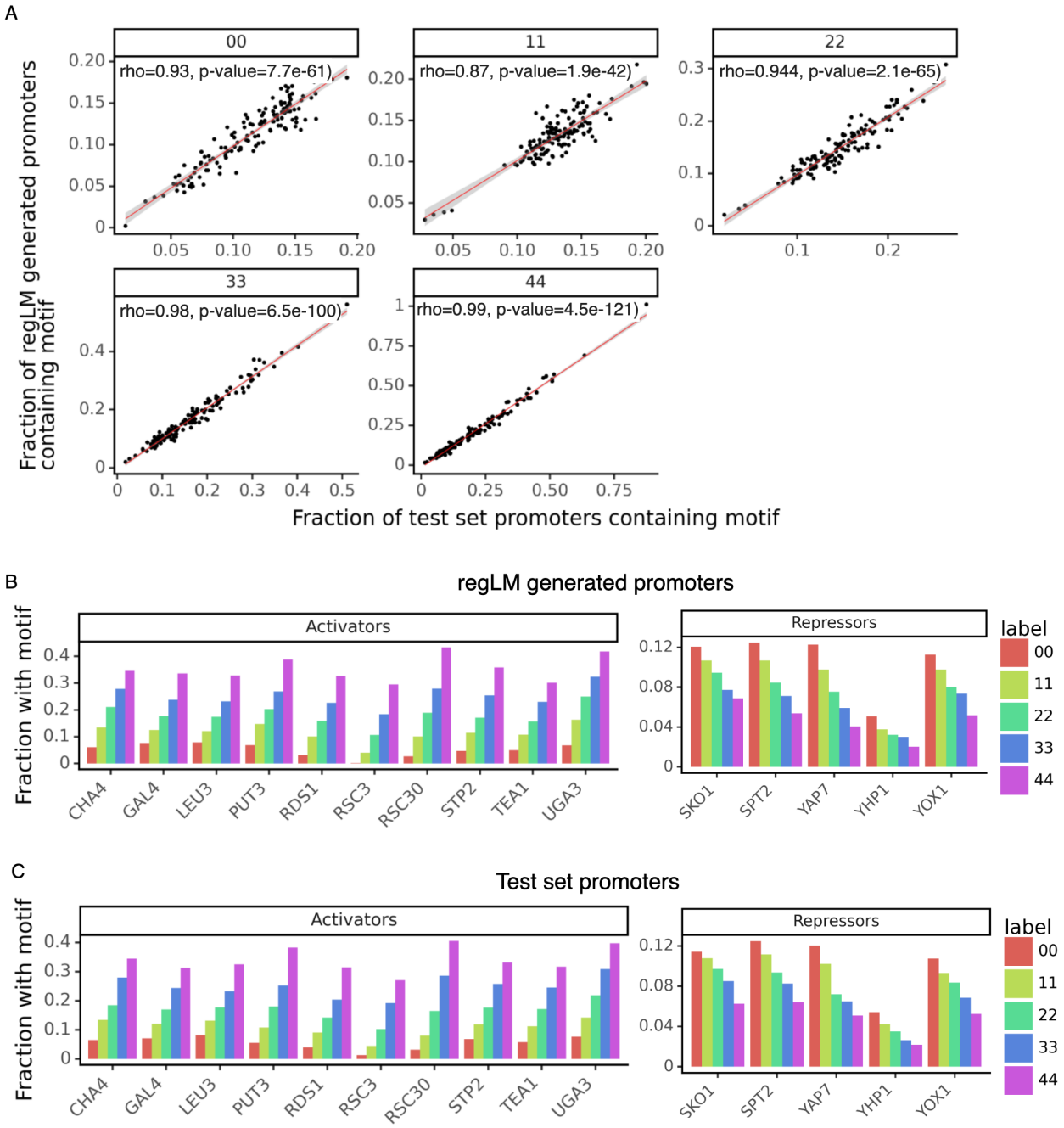
Supplemental Figure S8: Scatterplot showing the log ratio between the abundance of a motif in strong promoters (label 44) vs. weak promoters (label 00) on the x-axis, and the log ratio between the average accuracy of the regLM model on all instances of the motif in strong promoters vs. weak promoters on the y-axis. The red line shows the linear fit to the data.



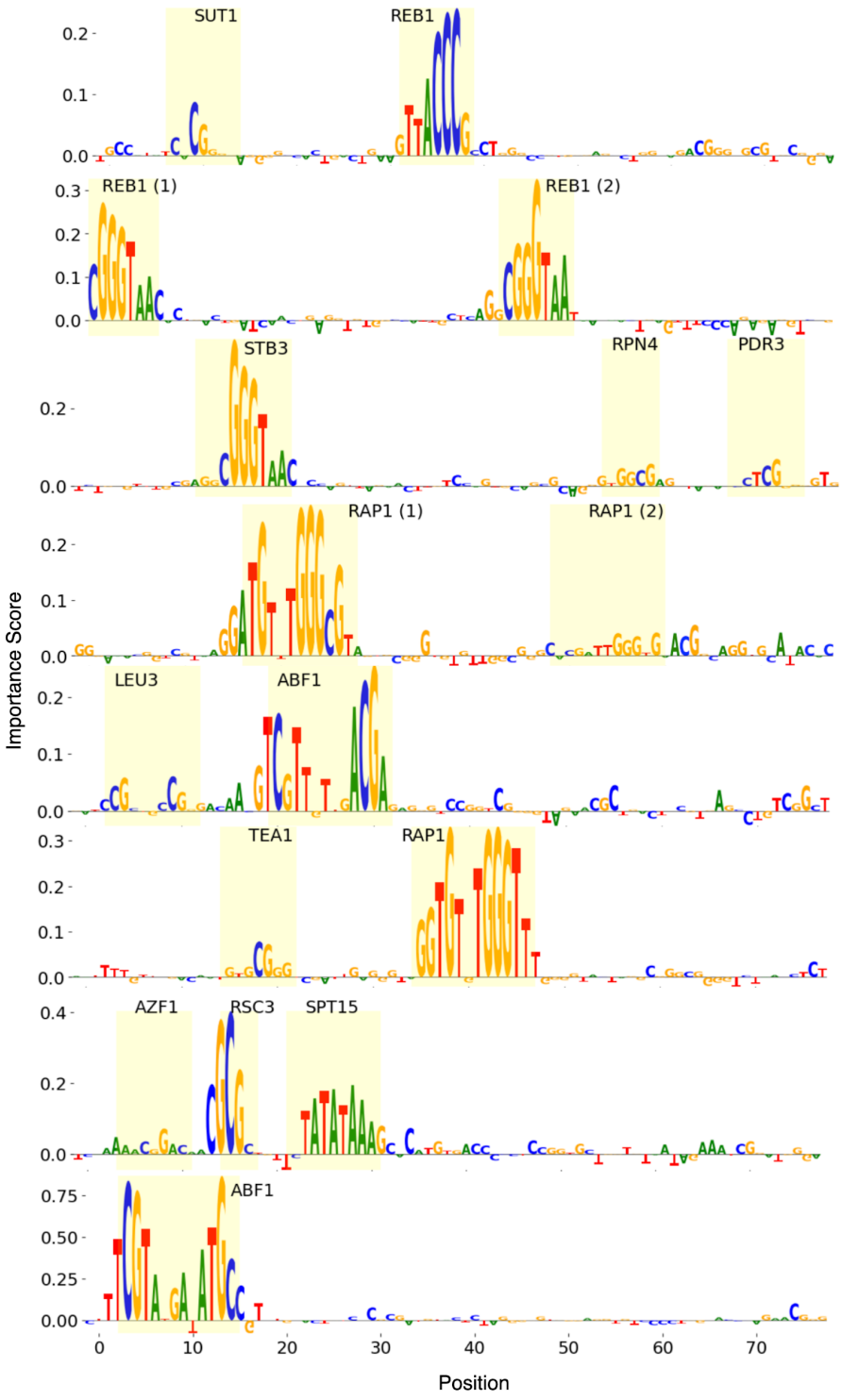
Supplemental Figure S9: Performance of supervised regression models trained to predict promoter activity of yeast promoter sequences, in **A)** complex medium and **B)** defined medium. The models were trained and tested on the same data as the regLM model. Scatterplots show the measured and predicted activity of 50,000 test set promoters.



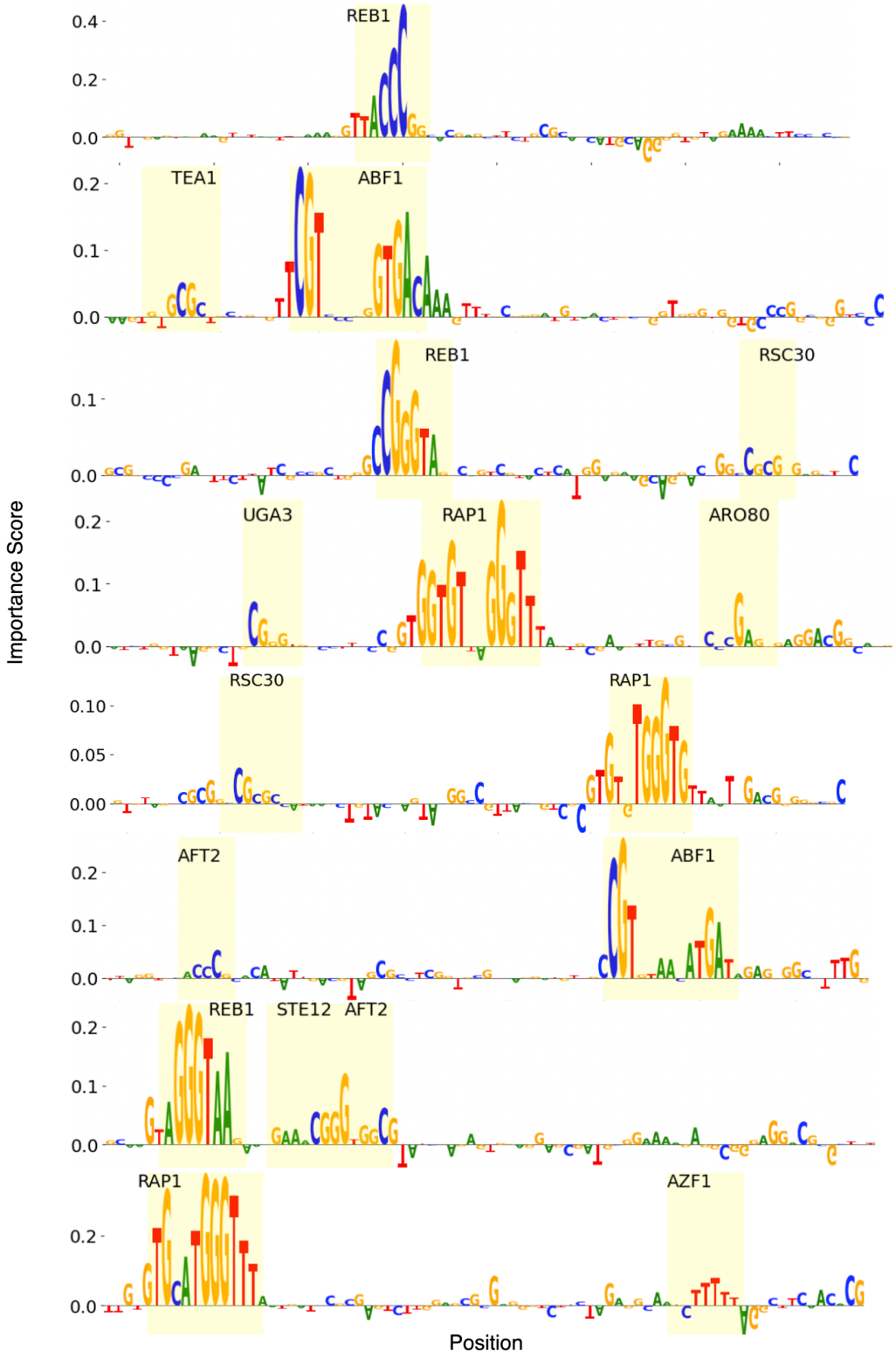
Supplemental Figure S10: Performance of two supervised regression models trained to predict promoter activity of yeast promoter sequences in **A)** complex medium and **B)** defined medium respectively. These models were trained and tested on separate data from the regLM model. Scatterplots show the measured and predicted activity of 50,000 test set promoters each.



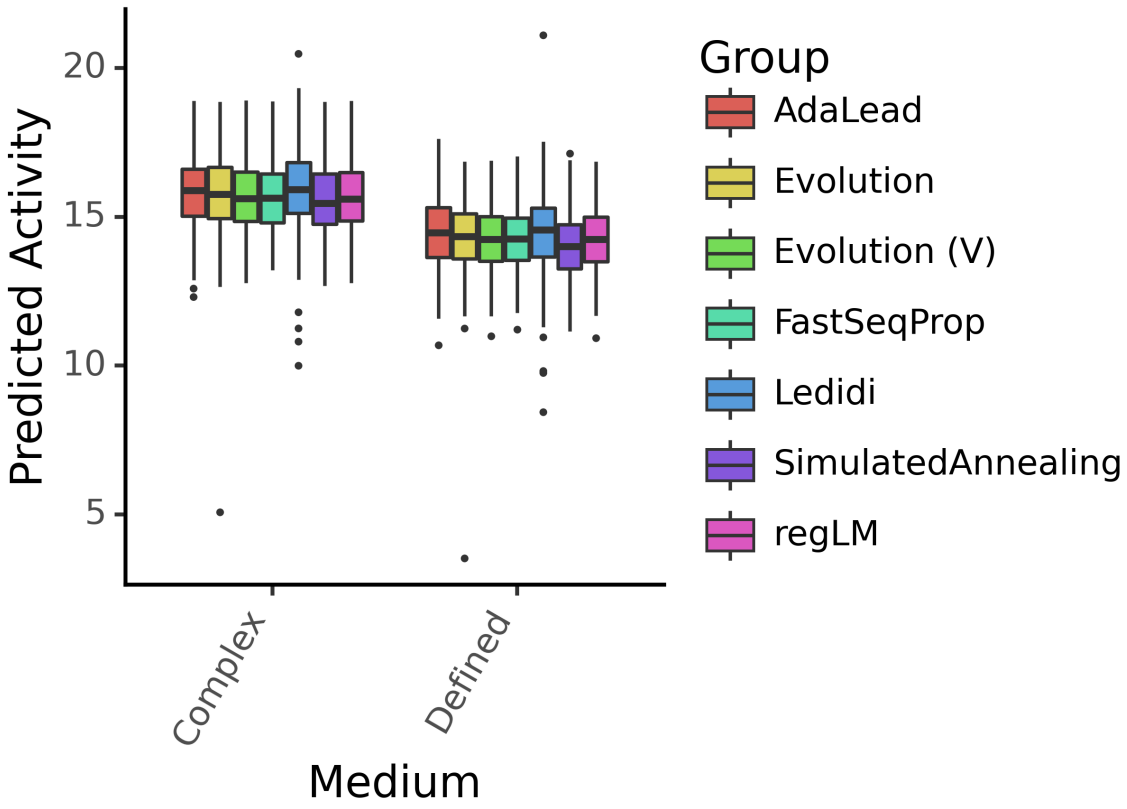
Supplemental Figure S11: A) Scatterplot showing the association between motif abundance in regLM generated promoters versus the test set, for promoters with different labels. Red lines show the linear fit to the data. **B, C)** A closer focus on the motifs that show the strongest differential abundance between strong and weak promoters in the test set, showing the close match between their abundance in the test set and in the generated promoters. Bar plots show the fraction of regLM generated promoters and test set promoters that contain selected activating and repressing TF motifs, separated by label. Only the 5 most common labels (00, 11, 22, 33, 44) are shown.



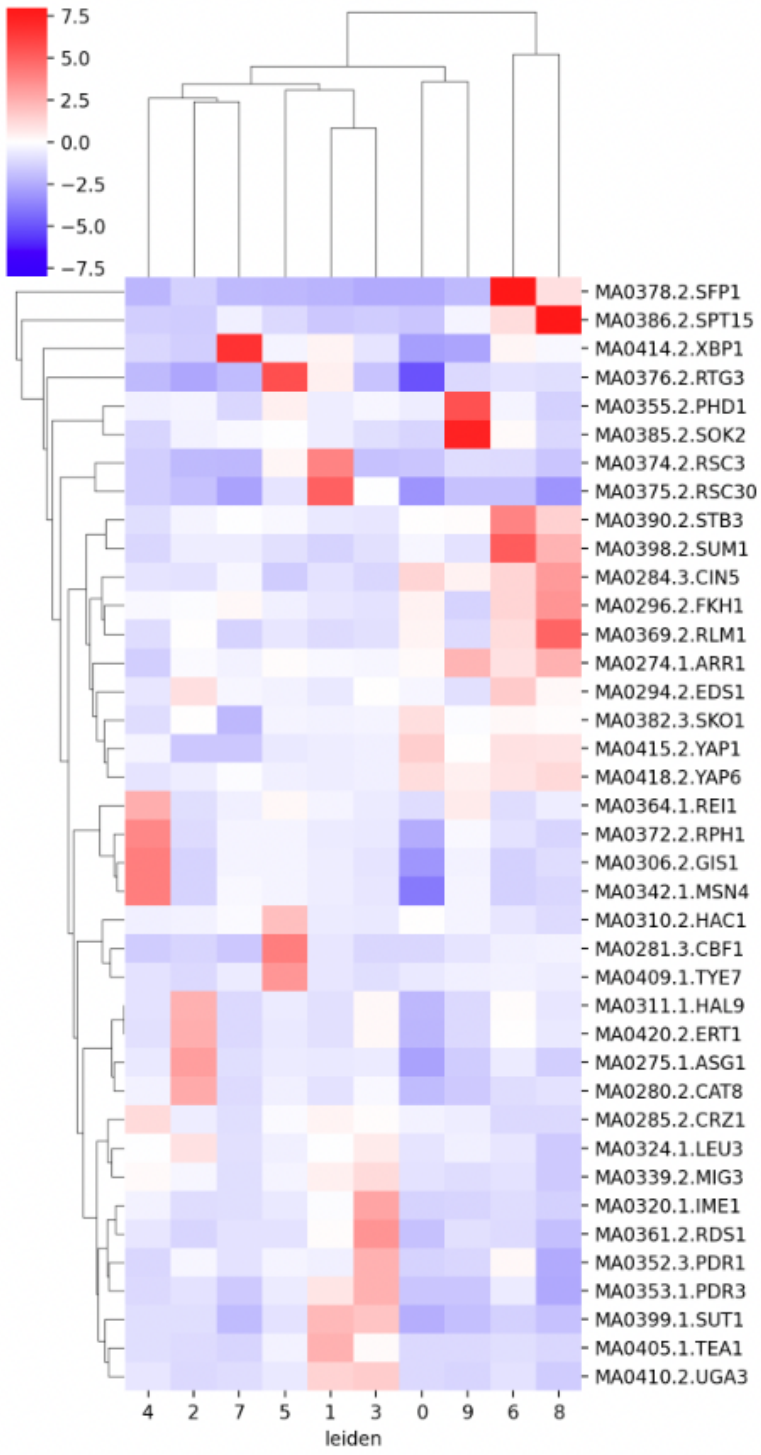
Supplemental Figure S12: Examples of strong promoters in the test set. Height represents the per-nucleotide importance score obtained from the paired regression model using ISM. Motifs with high importance are highlighted.



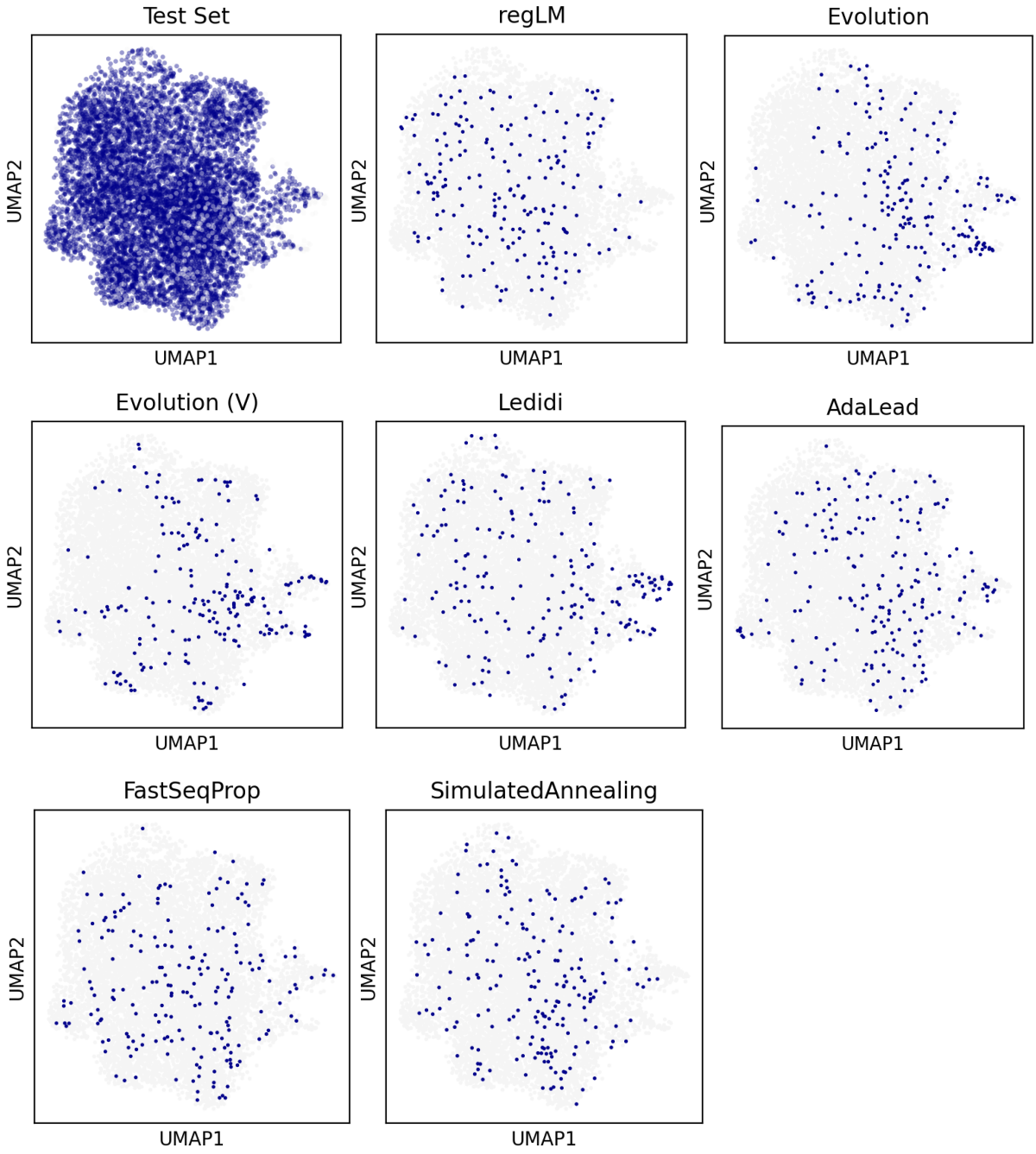
Supplemental Figure S13: Examples of strong promoters generated by regLM. Height represents the per-nucleotide importance score obtained from the paired regression model using ISM. Motifs with high importance are highlighted.



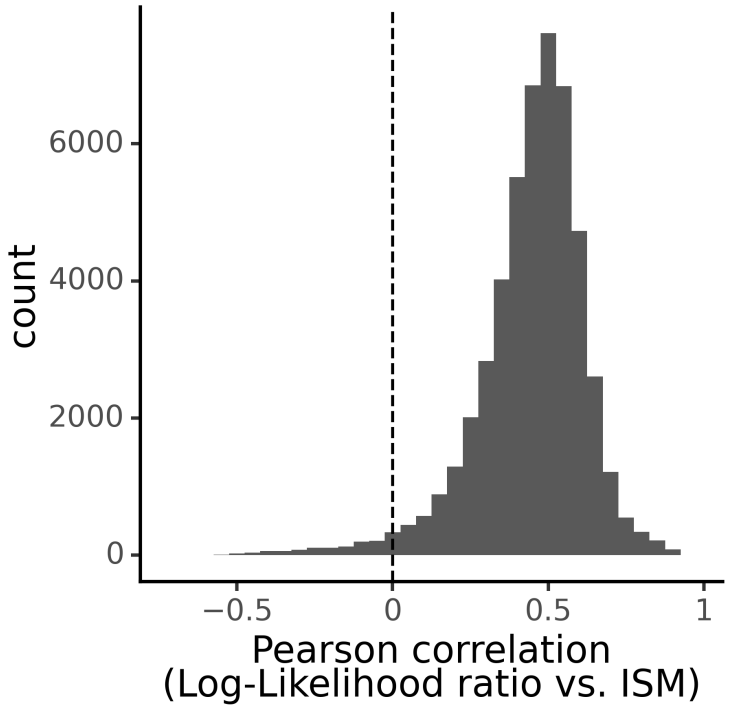
Supplemental Figure S14: Predicted activity of synthetic strong yeast promoters generated by different methods, in complex and defined media. 200 synthetic promoters were generated by each method. Evolution (V) represents synthetic promoters generated by (Vaishnav et al. 2022).



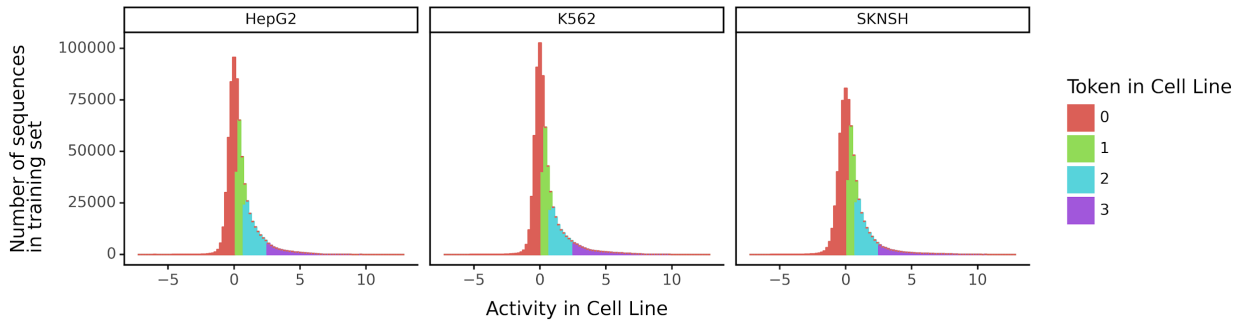
Supplemental Figure S15: Heatmap showing log2 fold changes in motif abundance for the top differentially abundant motifs across clusters of strong promoters. Fold changes were calculated for each cluster relative to the entire dataset.



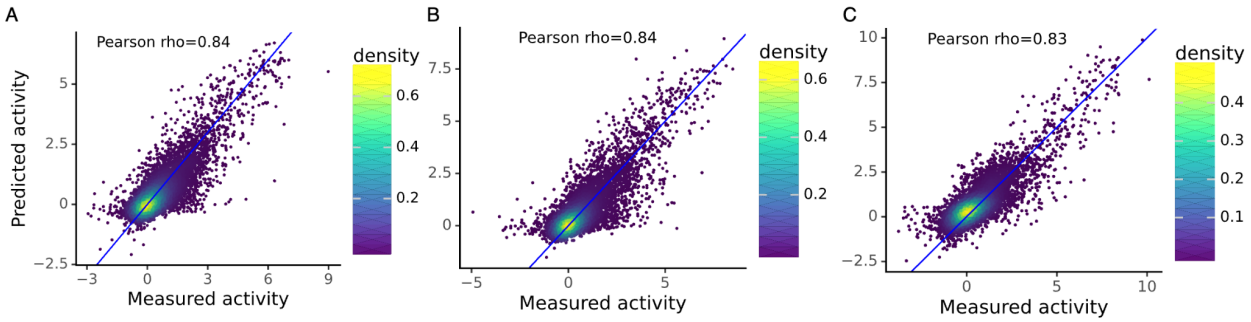
Supplemental Figure S16: UMAP visualization of real (Test Set) and synthetic strong promoters, labeled by the source dataset.



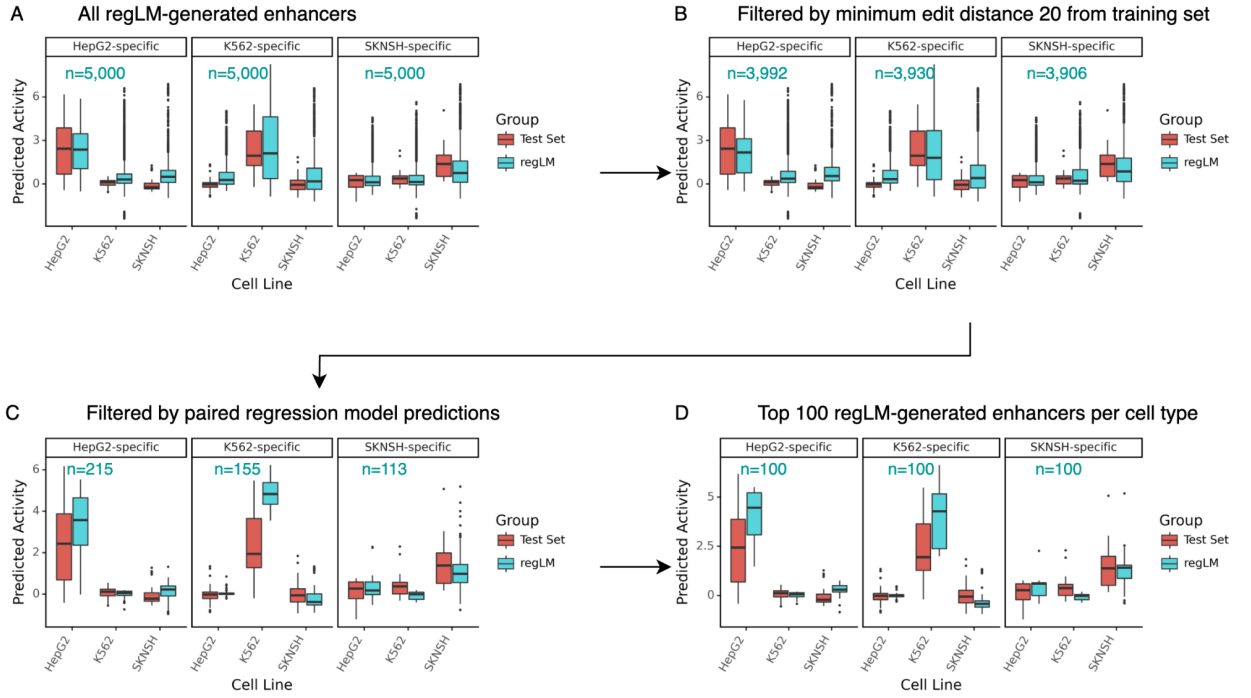
Supplemental Figure S17: Histogram of the Pearson's correlation between per-base ISM scores from the paired regression model, and the per-base log-likelihood ratios from the regLM model (44 vs. 00), across all 50,000 promoters in the test set.



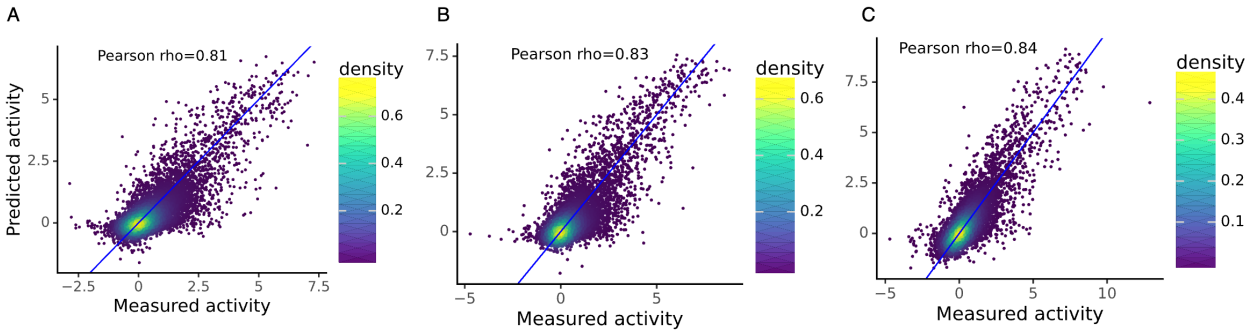
Supplemental Figure S18: Histograms showing the measured enhancer activity in each of 3 cell lines, for cell type-specific enhancers in the training set. Sequences were divided into 4 bins based on their measured activity. Each sequence was assigned a token ranging from 0-3 where 0 corresponds to the lowest bin and 3 corresponds to the highest. This procedure was performed separately for measurements in each cell line. The color corresponds to the assigned token in that cell line.



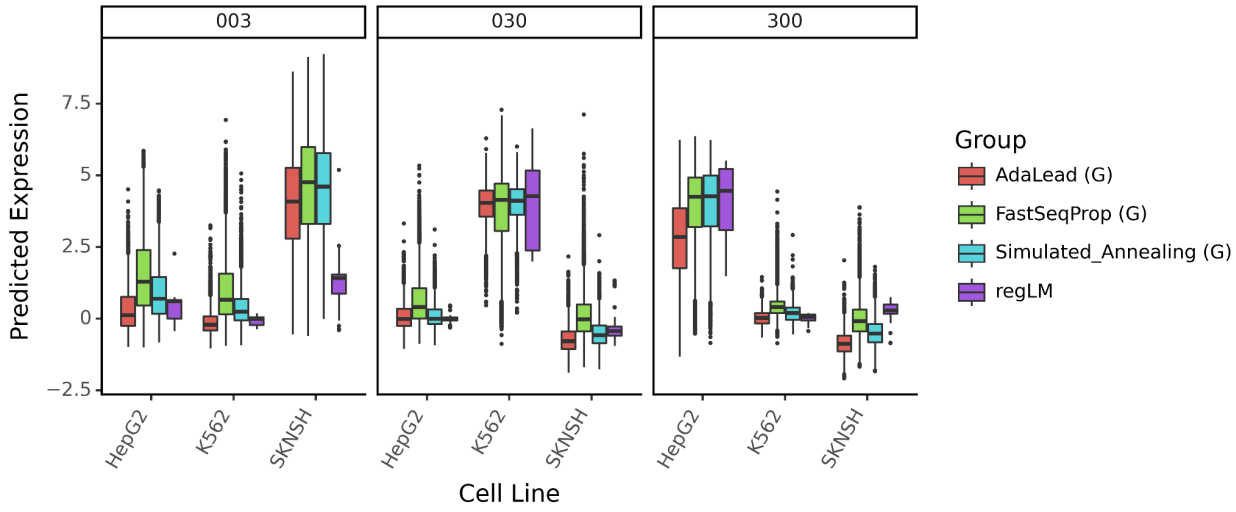
Supplemental Figure S19: Performance of a supervised regression model trained to predict activity of human enhancer sequences in **A)** HepG2 cells **B)** K562 cells **C)** SK-N-SH cells. The models were trained and tested on the same data as the regLM model. Scatter plots show the measured and predicted activity of enhancers in the test set.



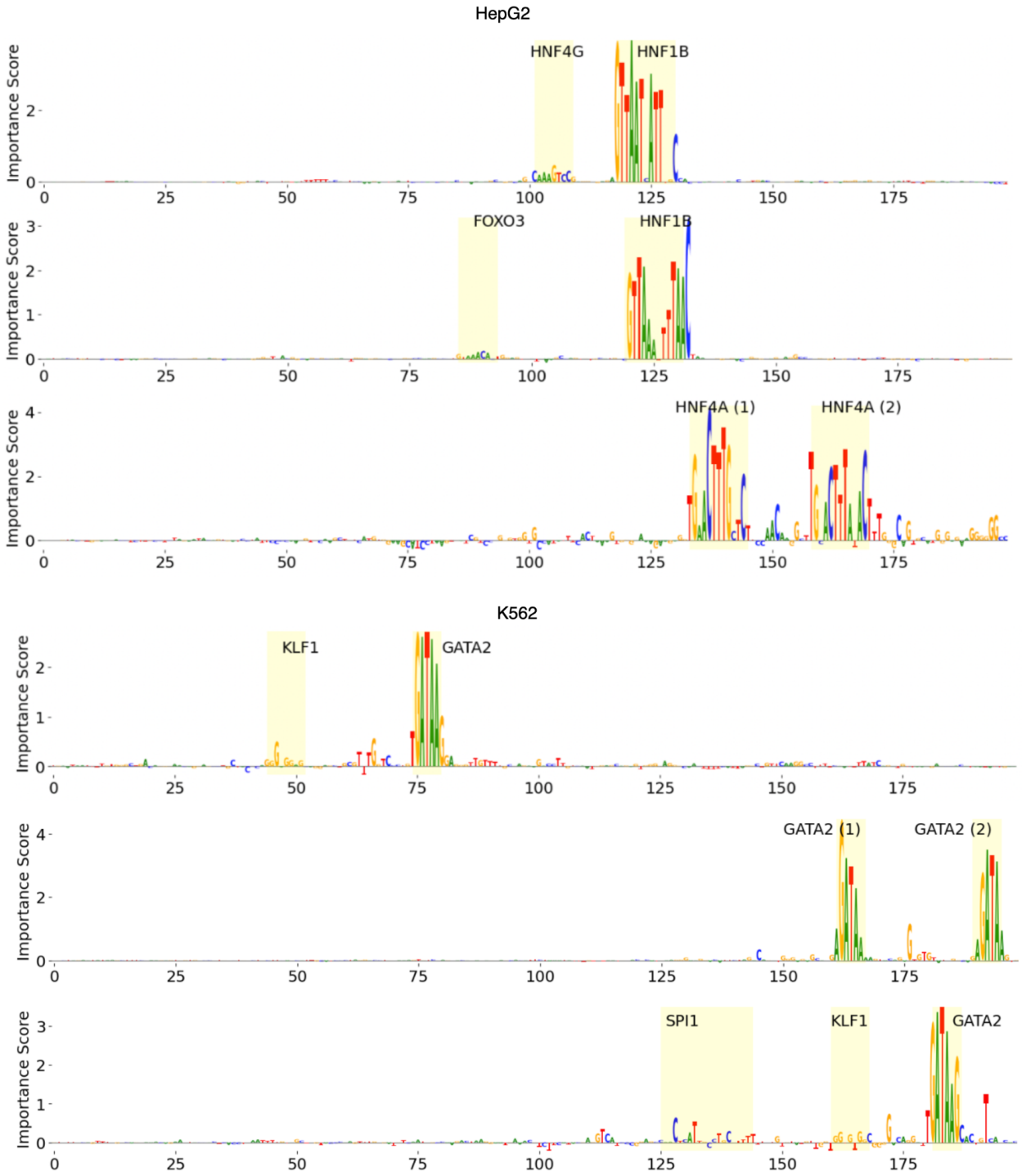
Supplemental Figure S20: Predicted activity of synthetic cell type-specific enhancers generated by regLM. The y-axis represents the activity predicted by independent regression models. Numbers in blue show the number of regLM-generated enhancers remaining after each filtering step. **A)** All 5,000 enhancers generated by regLM for each cell line **B)** Generated enhancers with a minimum edit distance of 20 from the training set **C)** Generated enhancers with a minimum edit distance of 20 from the training set, as well as on-target predictions >= 3.5 and maximum off-target prediction <=0.2, based on the regLM-paired regression models. **D)** The final set of the top 100 regLM-generated enhancers for each cell type, based on cell type specificity predicted by the regLM-paired regression models.



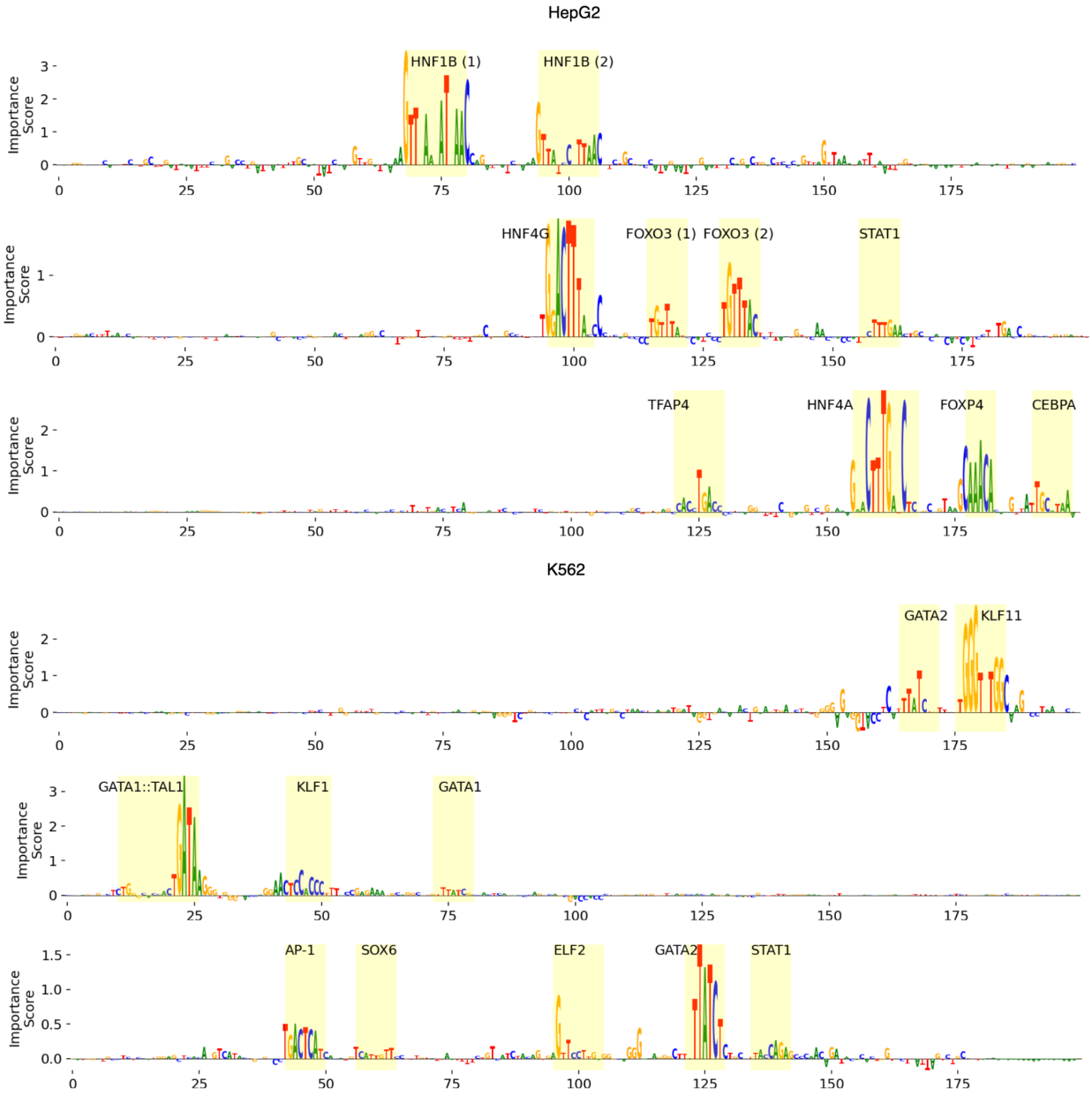
Supplemental Figure S21: Performance of a supervised regression model trained to predict activity of human enhancer sequences in **A)** HepG2 cells **B)** K562 cells **C)** SK-N-SH cells. The models were trained and tested on separate data from the regLM model. Scatter plots show the measured and predicted activity of enhancers in the test set.



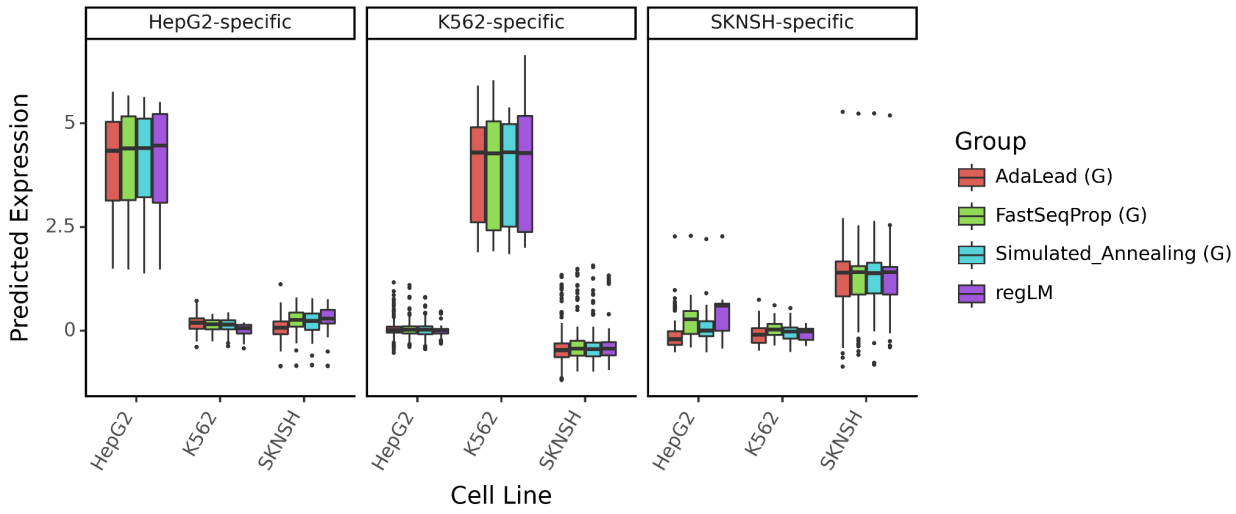
Supplemental Figure S22: Predicted activity of the synthetic cell type-specific enhancers generated by regLM and by Gosai et al. (Gosai et al. 2023), in 3 cell lines. Predictions were generated by regression models trained on separate data from regLM. (G) indicates that the method was performed by Gosai et al.



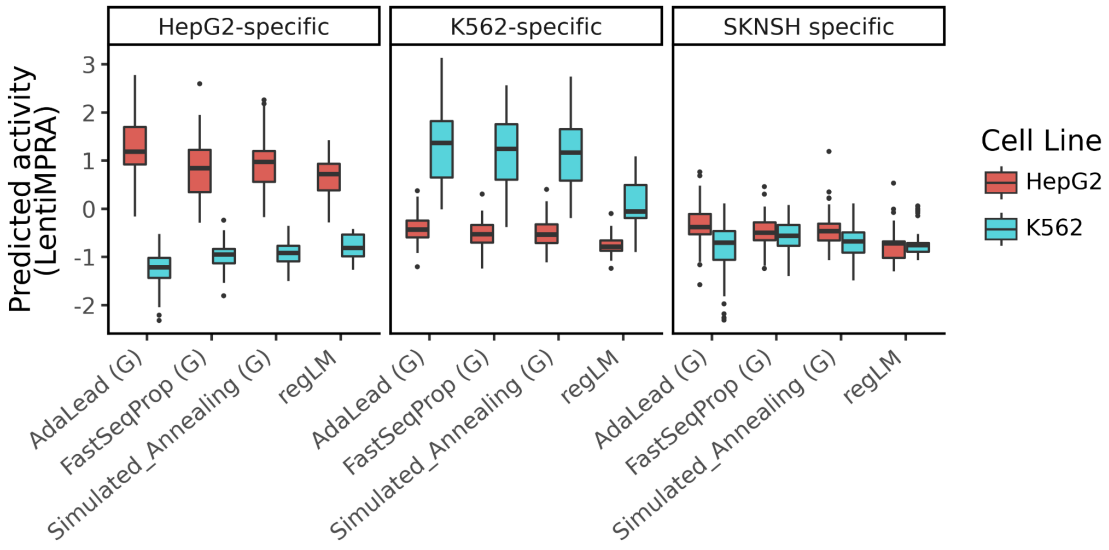
Supplemental Figure S23: ISM-based importance scores for regLM generated K562 and HepG2-specific enhancers, highlighting highly contributing motifs.



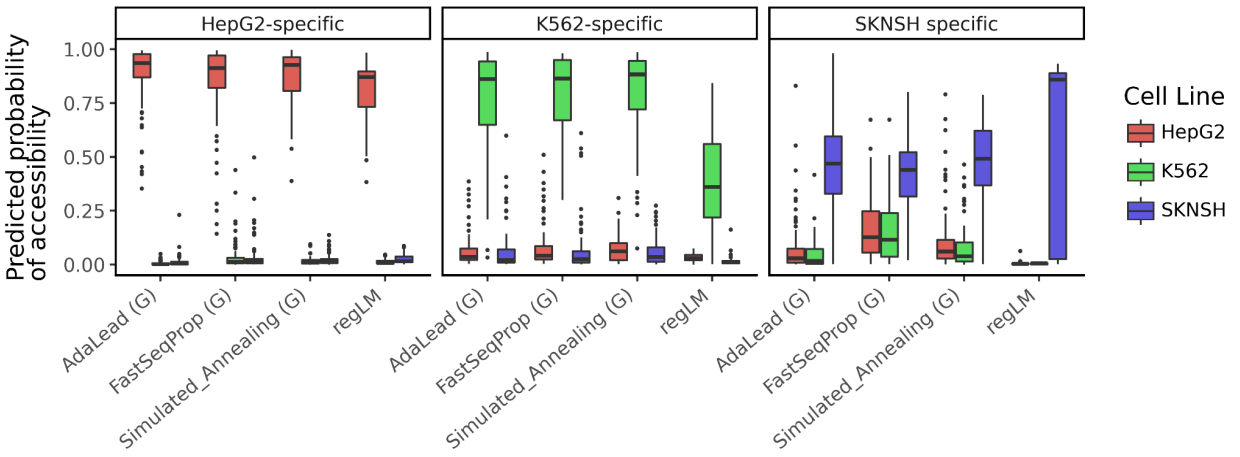
Supplemental Figure S24: ISM-based importance scores for K562 and HepG2-specific enhancers in the test set, highlighting highly contributing motifs.



Supplemental Figure S25: Predicted activity of 100 synthetic cell type-specific enhancers generated by regLM for each cell line, and the 100 Gosai et al. (Gosai et al. 2023) designed elements chosen to have the most similar activity for each cell line. (G) indicates that the method was performed by Gosai et al.

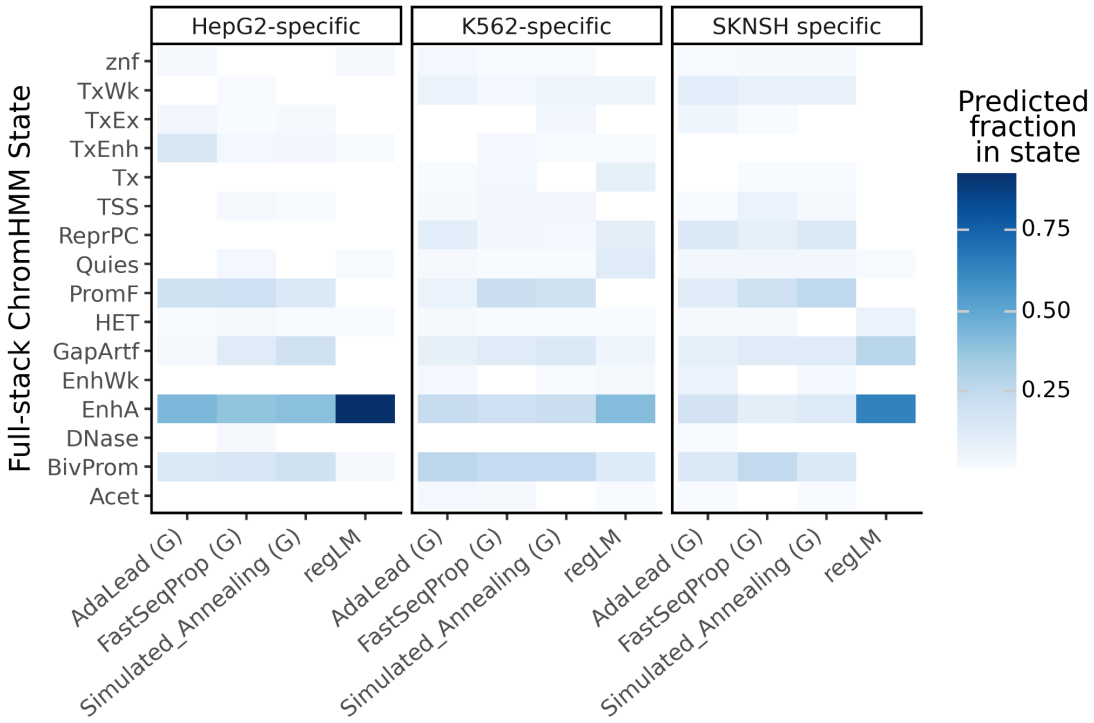


Supplemental Figure S26: Predicted activity of synthetic cell type-specific enhancers generated by different methods, using a model trained on Lentiviral MPRA data. (G) indicates that the method was performed by Gosai et al. (Gosai et al. 2023). The 300 Gosai et al. designed elements chosen based on similar activity to regLM generated enhancers are shown here.

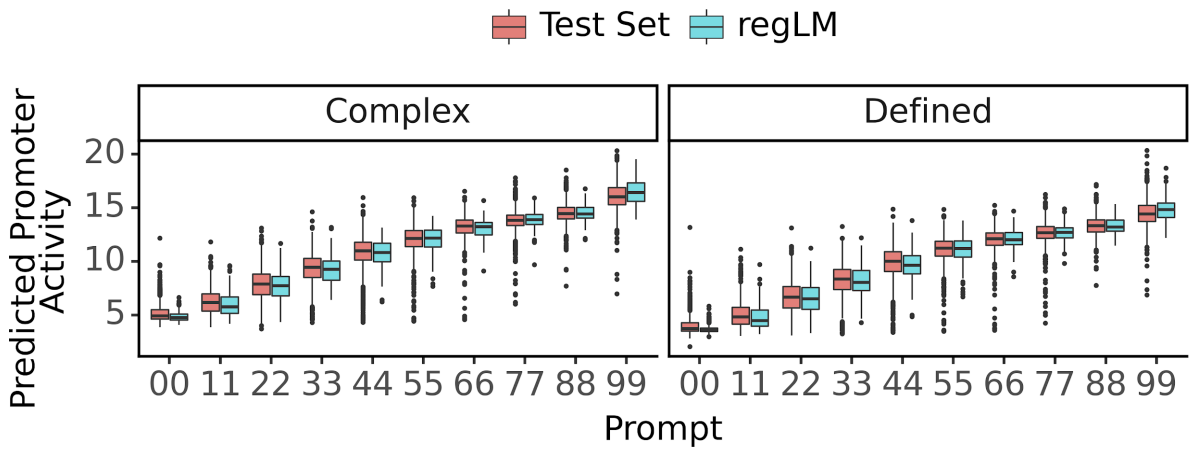


Supplemental Figure S27: Predictions of a binary classification model trained to predict ATAC-seq peaks in three cell lines, on synthetic cell type-specific enhancers generated by different methods. (G) indicates that the method was performed by Gosai et al. (Gosai et al. 2023). The 300 Gosai et al. designed elements chosen based on similar activity to regLM generated enhancers are shown here.

Supplemental Figure S28: Predictions of a binary classification model trained to predict ATAC-seq peaks in 203 cell types, on synthetic cell type-specific enhancers generated by different methods. (G) indicates that the method was performed by Gosai et al. (Gosai et al. 2023). The 300 Gosai et al. designed elements chosen based on similar activity to regLM generated enhancers are shown here.



Supplemental Figure S29: Predictions of a classification model trained to classify genomic DNA into chromatin states defined by the fullstack ChromHMM annotation (Vu and Ernst 2022), on synthetic cell type-specific enhancers generated by different methods. (G) indicates that the method was performed by Gosai et al. (Gosai et al. 2023). The 300 Gosai et al. designed elements chosen based on similar activity to regLM generated enhancers are shown here.



Supplemental Figure S30: The yeast regLM model was re-trained with labels consisting of 10 tokens ranging from 0 (lowest activity) - 9 (highest activity). The trained model was prompted with labels ranging from 00-99 and 100 synthetic promoters were generated from each prompt. The activity of these synthetic promoters was predicted using regression models trained on separate data and compared to the predicted activity of experimentally validated test set promoters with the same labels.

Supplemental References

Agarwal V, Inoue F, Schubach M, Martin BK, Dash PM, Zhang Z, Sohota A, Noble WS, Yardimci GG, Kircher M, et al. 2023. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. bioRxiv doi:10.1101/2023.03.05.531189.

Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 18(10):1196–1203.

Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, et al. 2018. The chromatin accessibility landscape of primary human cancers. *Science* 362(6413).

Gosai SJ, Castro RI, Fuentes N, Butts JC, Kales S, Noche RR, Mouri K, Sabeti PC, Reilly SK, Tewhey R. 2023. Machine-guided design of synthetic cell type-specific cis-regulatory elements. bioRxiv doi:10.1101/2023.08.08.552077.

Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, Regev A. 2022. The evolution, evolvability and engineering of gene regulatory DNA. *Nature* 603(7901):455–463.

Vu H, Ernst J. 2022. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biol* 23(1):9.

Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, et al. 2021. A single-cell atlas of chromatin accessibility in the human genome. *Cell* 184(24):5985–6001.e19.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137.