Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics: Supplementary Figures and Tables

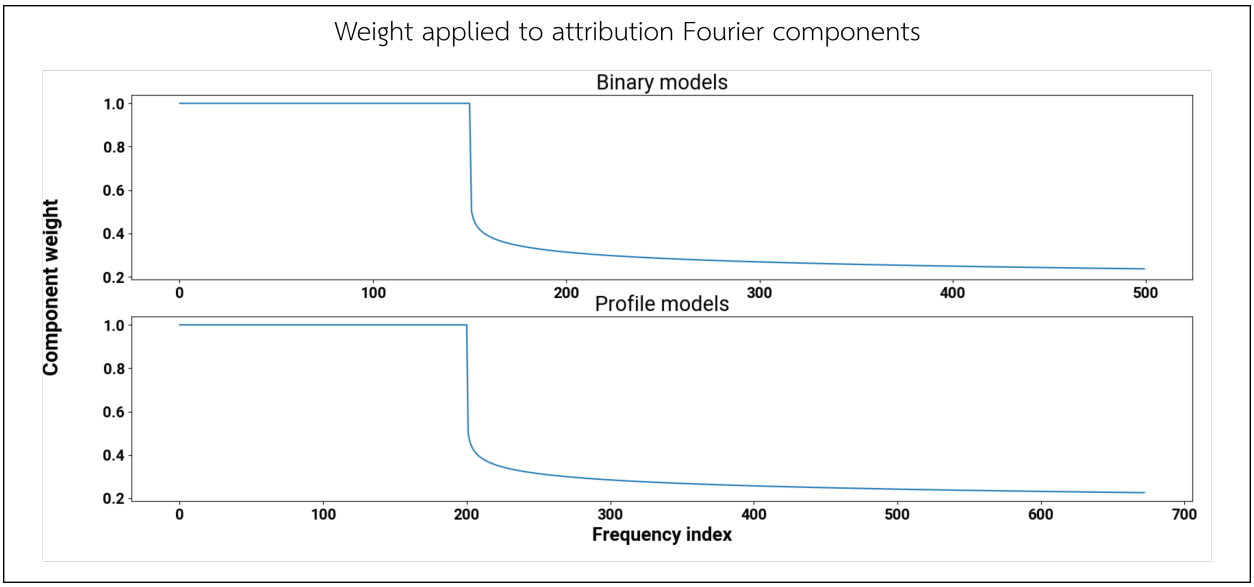

Figure S1: Weight applied to Fourier components in attribution prior loss. To compute the Fourier-based attribution prior loss, the Fourier components corresponding to positive frequencies of the attributions are weighted (these weights w are plotted here). This weighted sum constitutes the score of the attributions, and 1 minus this score becomes the attribution prior loss value. Note that because input sequences to binary and profile models have different lengths, the lengths of the discrete Fourier transform components are also different; in both cases, the frequency threshold T corresponds to a minimum expected motif length of 6–7 bp.

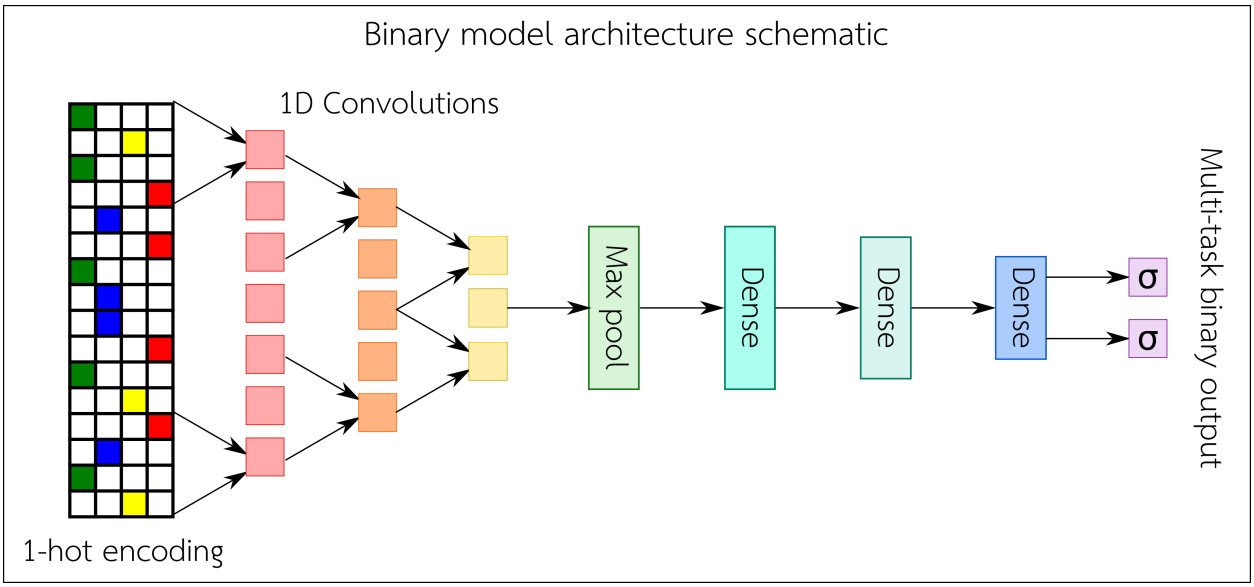

Figure S2: Schematic of binary model architecture. A one-hot encoded sequence is fed into three consecutive convolutional layers. The resulting activations are passed through a max pooling layer, followed by three dense layers, where the final dense layer outputs a sigmoid-transformed binary prediction.

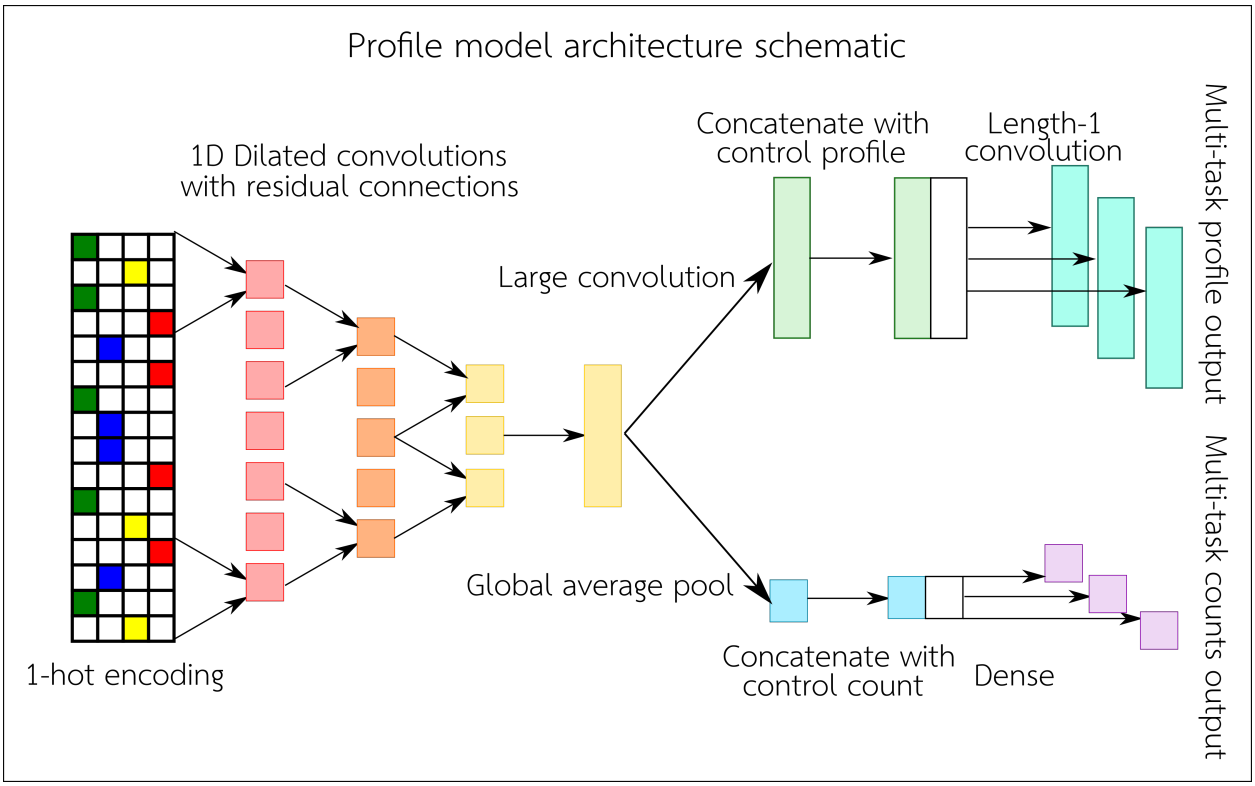

Figure S3: Schematic of profile model architecture, based on the architecture in Avsec et al. [1]. A one-hot encoded sequence is fed into six consecutive dilated convolutional layers with summed residual connections. For each task, the model predicts a profile shape and a read count. The profile shape prediction is obtained by feeding the activations from the dilated convolutions to another convolutional layer with a large kernel size, concatenating the result with a set of control profiles, and performing a length-1 convolution over the concatenation to yield a profile shape prediction. The read count prediction is obtained by feeding the activations from the dilated convolutions through a global average pooling layer, concatenating the result with a set of control read counts, and passing this concatenation through a dense layer to obtain predicted read counts.

Table S1: Predictive performance of single-task K562 binary model versus Basset

	Basset	Single-task model
auROC	0.838	**0.966**
auPRC	0.839	**0.964**

We compare our single-task K562 binary models (trained without any prior—see Supplementary Figure S2) to Basset, a massively multi-tasked binary model trained to predict chromatin accessibility in 164 cell types (including K562) [2]. On our K562 test set, our single-task model achieves better predictive performance than the K562 output head of Basset.

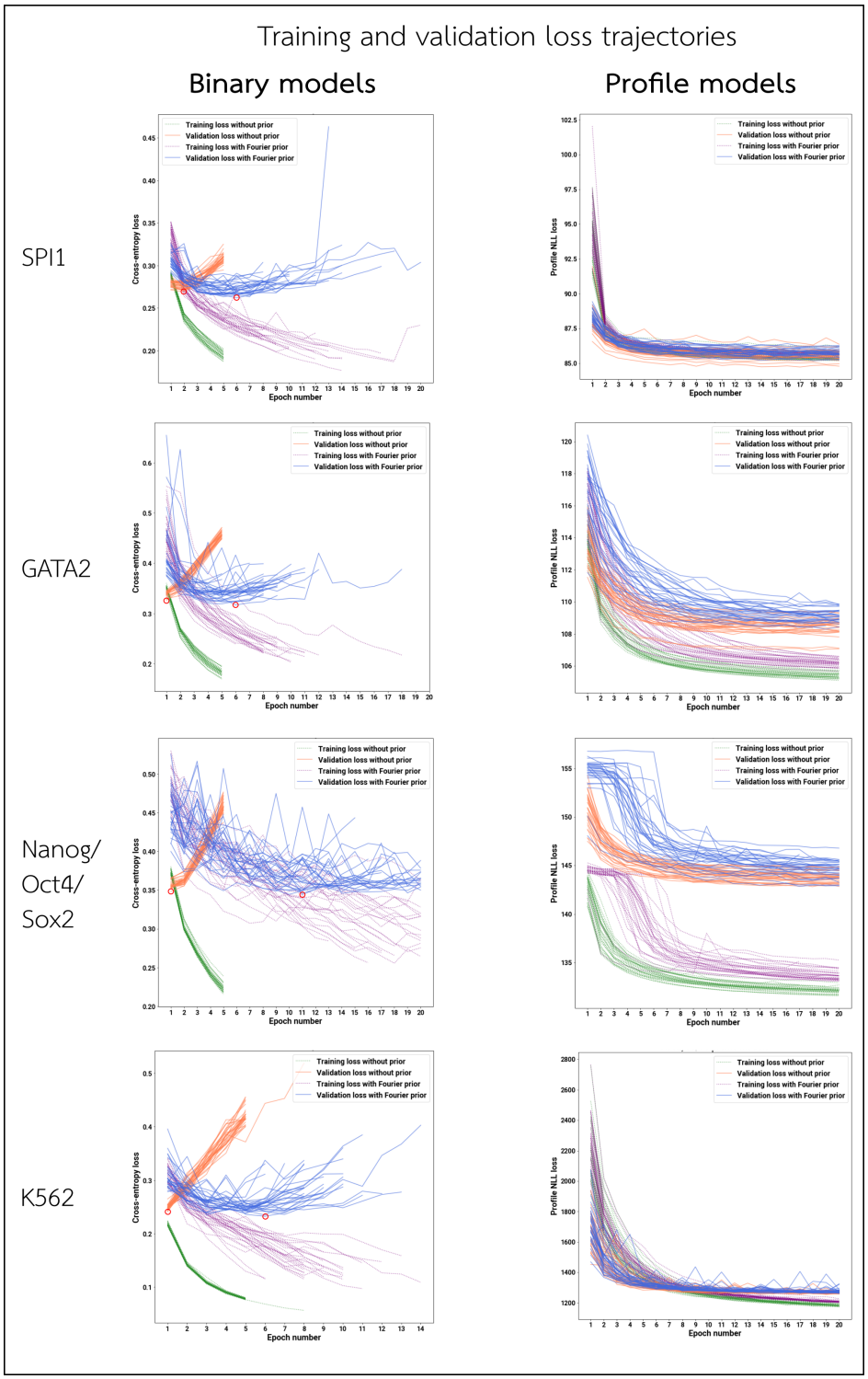

Figure S4: Training and validation correctness loss trajectories. For each architecture and dataset, we show the trajectory of the training and validation correctness losses (i.e. excluding any attribution prior loss) after each epoch of training, over all random initializations. Loss values are plotted beginning after the first epoch of training. In general, binary models (left) overfit very easily, with validation loss visibly growing after the first epoch. Profile models (right), however, are much more resilient to overfitting, as they benefit from extensive data augmentation through random jitters in the input sequences. On binary models (left), red circles indicate the models that achieve the lowest validation loss, trained with and without the Fourier-based prior.

Table S2: Improved interpretability on test-set sequences

		Average sum of high-frequency Fourier components		Average entropy	
		No prior	With prior	No prior	With prior
Binary	SPI1	0.416 ± 0.002	$\mathbf{0.374 \pm 0.002}$	8.35 ± 0.01	$\mathbf{7.89 \pm 0.01}$
	GATA2	0.442 ± 0.002	$\mathbf{0.394 \pm 0.002}$	8.55 ± 0.01	$\mathbf{7.96 \pm 0.02}$
	Nanog/Oct4/Sox2	0.449 ± 0.002	$\mathbf{0.381 \pm 0.002}$	7.98 ± 0.01	$\mathbf{6.76 \pm 0.02}$
	K562	0.494 ± 0.002	$\mathbf{0.465 \pm 0.002}$	8.84 ± 0.01	$\mathbf{8.65 \pm 0.01}$
Profile	SPI1	0.449 ± 0.002	$\mathbf{0.381 \pm 0.002}$	7.98 ± 0.01	$\mathbf{6.76 \pm 0.02}$
	GATA2	0.449 ± 0.002	$\mathbf{0.416 \pm 0.001}$	8.49 ± 0.01	$\mathbf{7.34 \pm 0.02}$
	Nanog/Oct4/Sox2	0.564 ± 0.001	$\mathbf{0.452 \pm 0.001}$	9.01 ± 0.01	$\mathbf{8.01 \pm 0.02}$
	K562	0.567 ± 0.001	$\mathbf{0.452 \pm 0.001}$	9.20 ± 0.01	$\mathbf{8.80 \pm 0.01}$

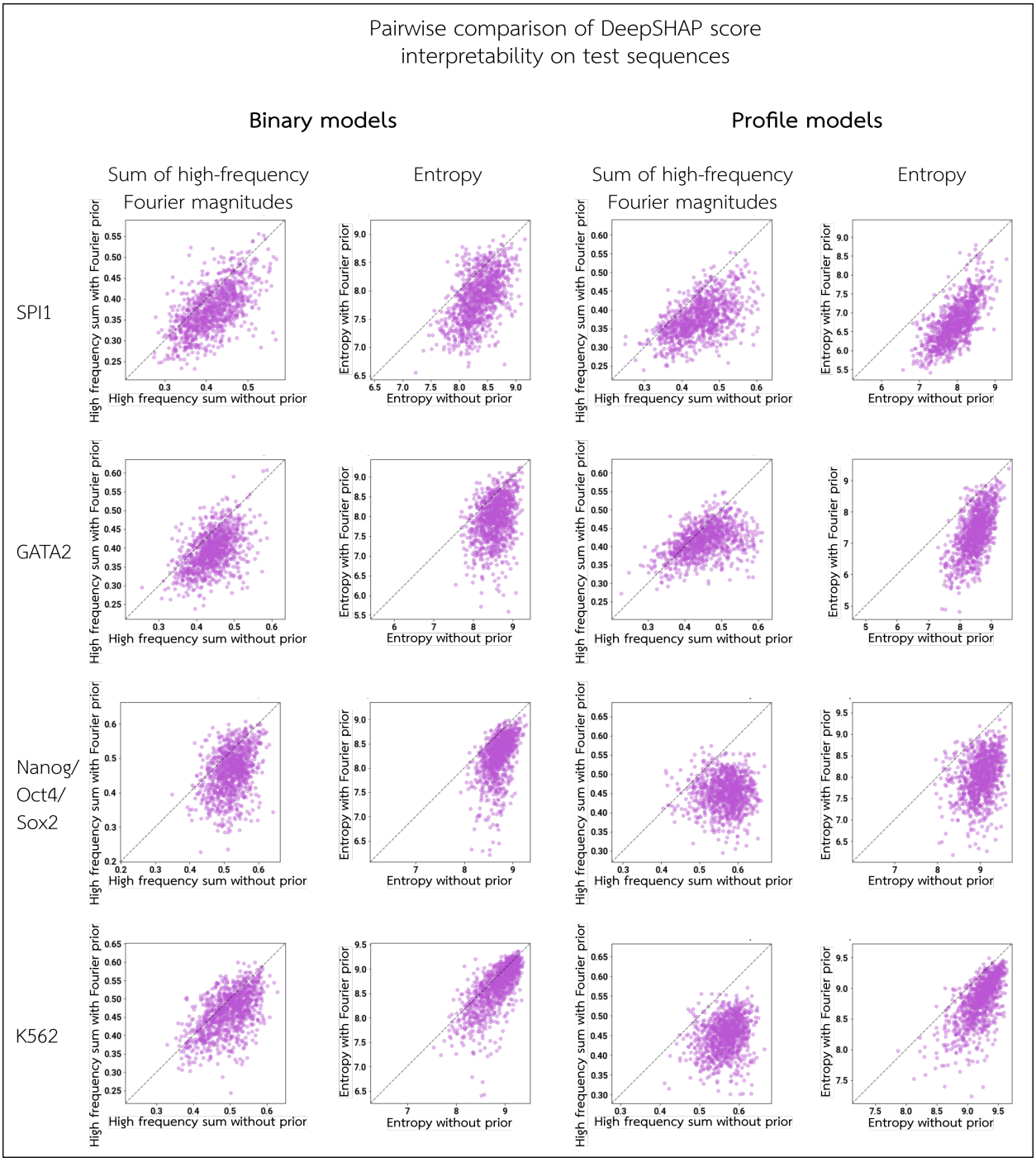

Figure S5: Signal-to-noise ratio of attributions across test-set sequences. For each architecture and dataset, we compute DeepSHAP importance scores for 1000 randomly selected peak sequences from the test set. An improvement in the signal-to-noise ratio of the attributions is quantified as a reduction in the high-frequency Fourier component magnitudes, and as a reduction in Shannon entropy. We compare the sum of normalized high-frequency Fourier components and the Shannon entropy for each sequence, between models trained with versus without the Fourier-based prior.

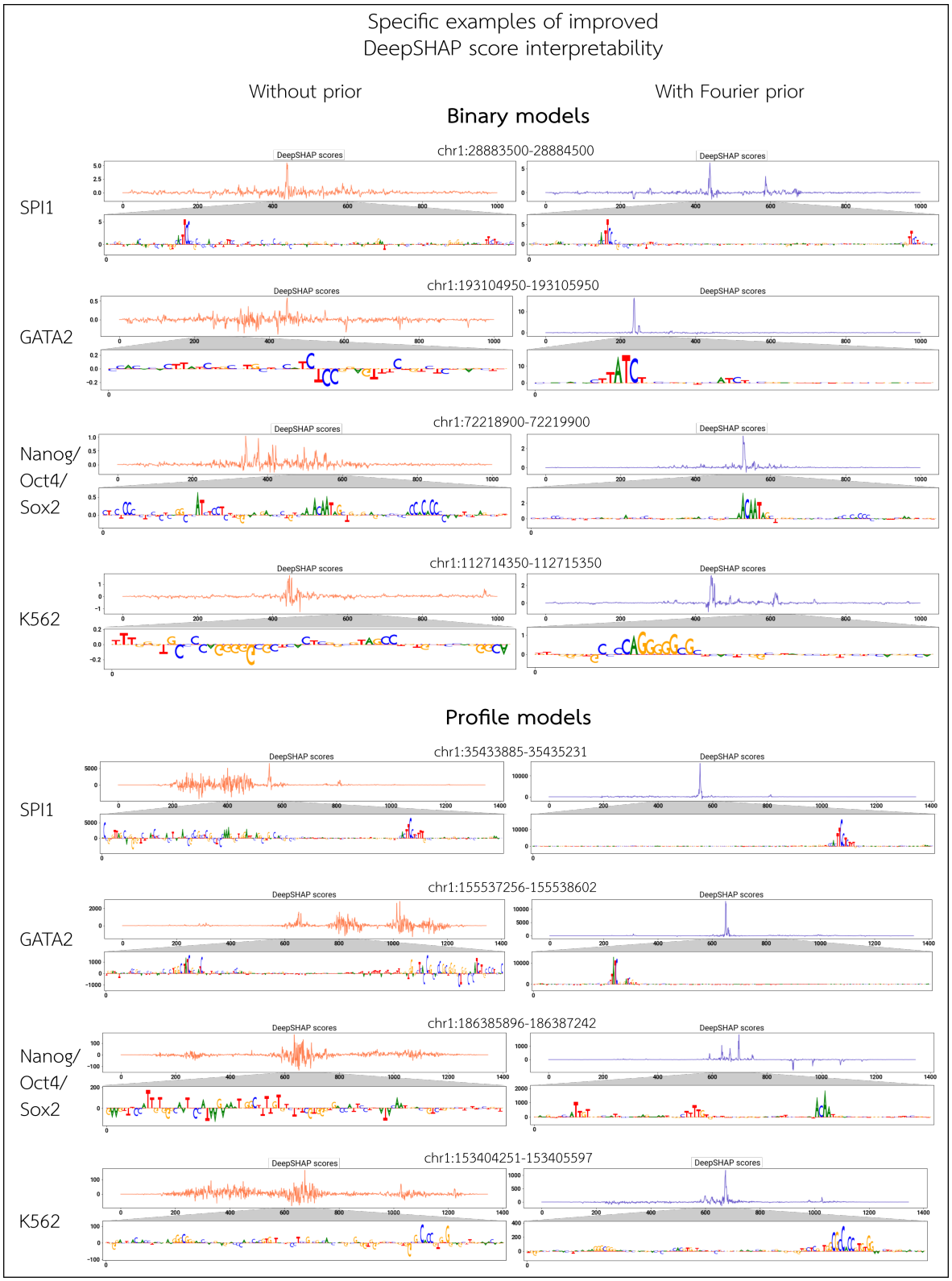

Figure S6: Specific examples of improved interpretability in DeepSHAP scores. For each architecture and dataset, we show the DeepSHAP attribution scores of specific peak sequences. For each selected input sequence, we display the value of the DeepSHAP importance for the bases present along the entire input sequence, as well as the base-pair-level attributions in the summit region.

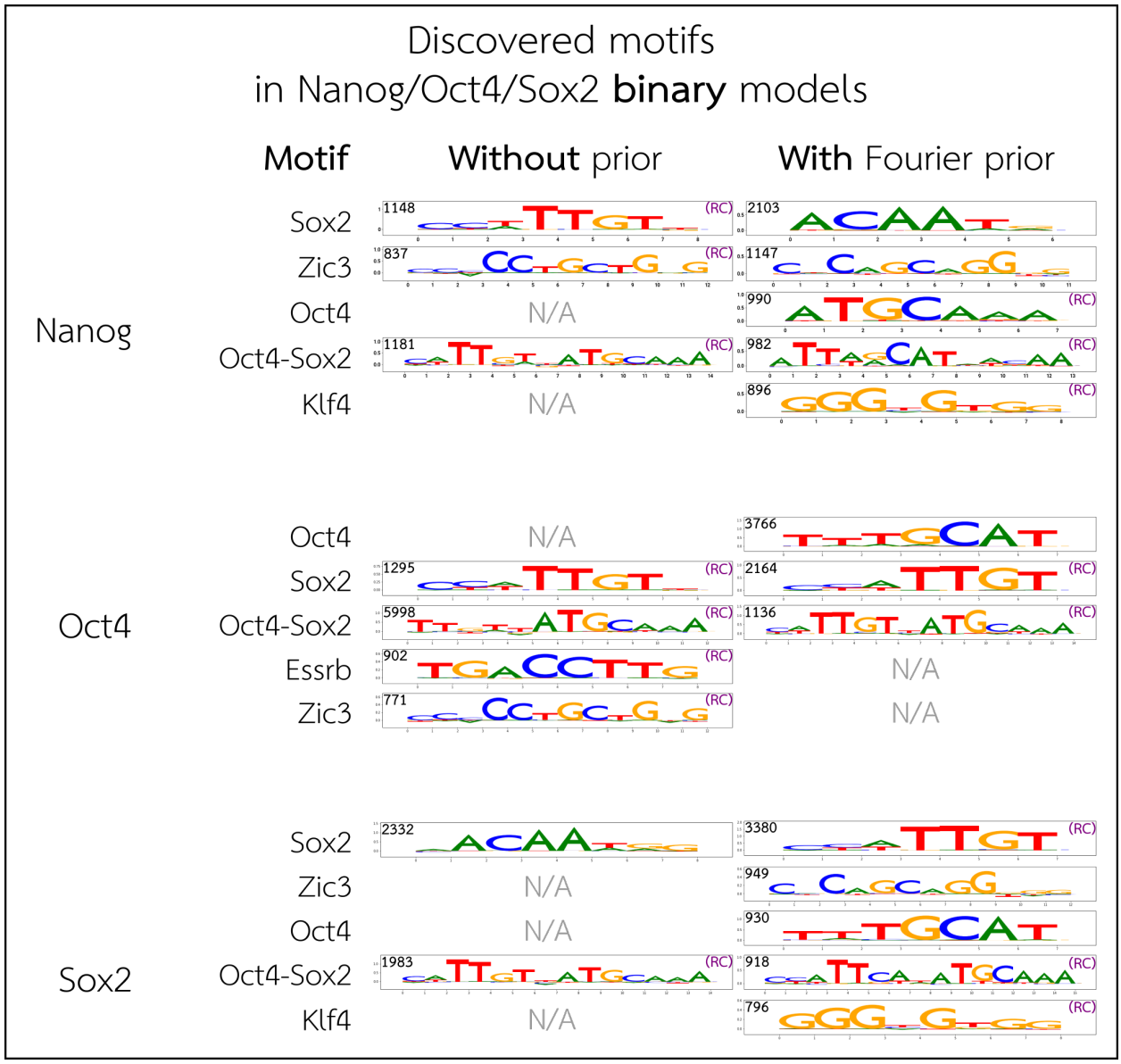

Figure S7: Discovered motifs from Nanog/Oct4/Sox2 binary models. For each TF, we use TF-MoDISco to discover motifs using DeepSHAP attributions of test-set peak sequences. We show the relevant motifs identified by TF-MoDISco which pass our thresholds (Supplementary Methods Sec. 5), and match them to known motifs from Avsec et al. [1]. "(RC)" denotes that the motif shown is reverse-complemented relative to the orientation in Avsec et al. [1]. The number in the top left of each motif indicates the number of seqlets identified by TF-MoDISco that underlie the motif.

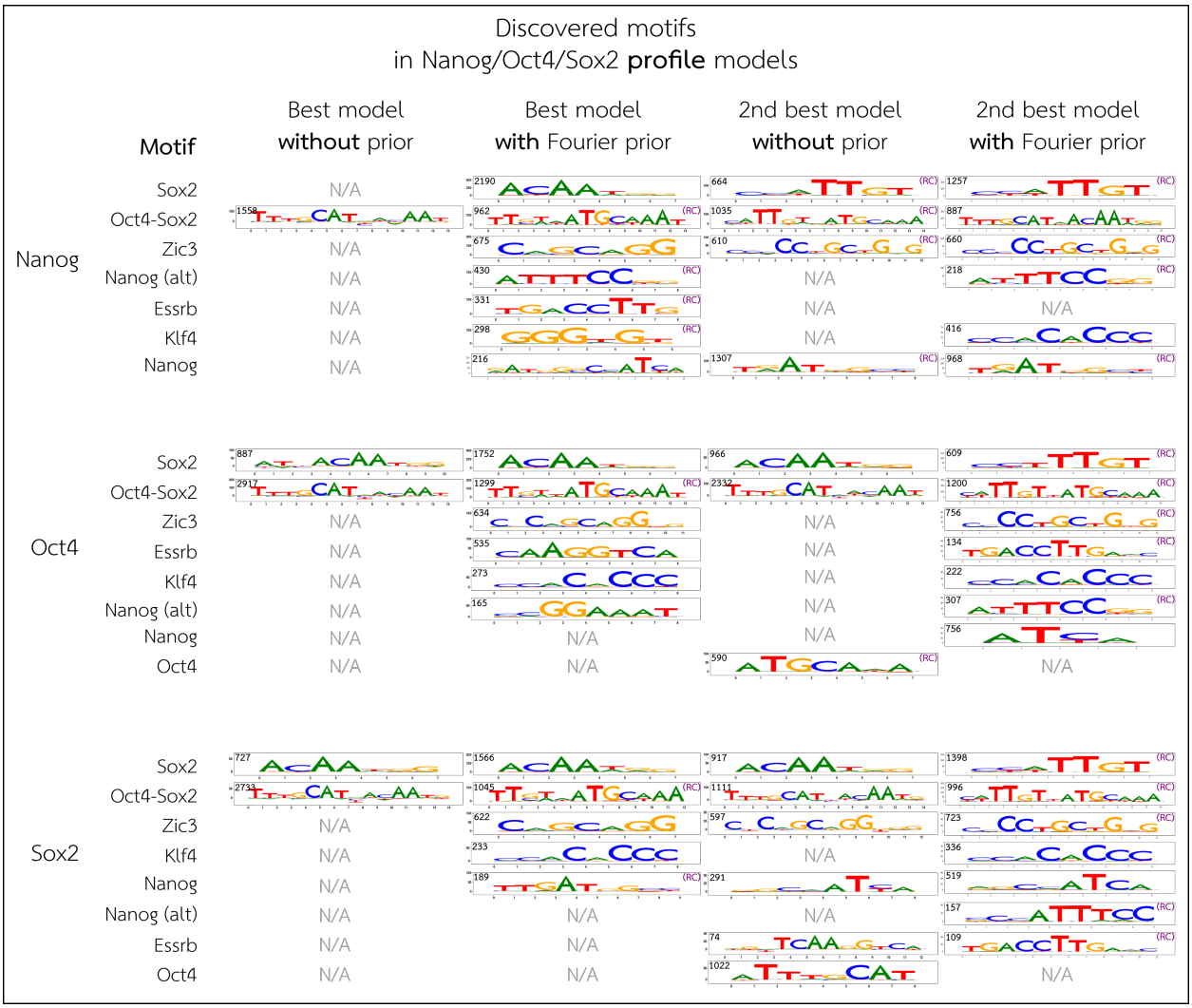

Figure S8: Discovered motifs from Nanog/Oct4/Sox2 profile models. For each TF, we use TF-MoDISco to discover motifs using DeepSHAP attributions of test-set peak sequences. We show the relevant motifs identified by TF-MoDISco which pass our thresholds (Supplementary Methods Sec. 5), and match them to known motifs from Avsec et al. [1]. "(RC)" denotes that the motif shown is reverse-complemented relative to the orientation in Avsec et al. [1]. The number in the top left of each motif indicates the number of seqlets identified by TF-MoDISco that underlie the motif. While we focus our downstream analysis on the best-performing models with and without the Fourier-based prior, the best-performing profile model trained without the prior identified much fewer motifs than the model trained with the prior, so we also show the motifs identified using TF-MoDISco on the second-best-performing models, both with and without the prior. This demonstrates that the profile models without the Fourier-based prior are capable of learning the larger set of motifs expected, but are limited by noisy and irreproducible attributions.

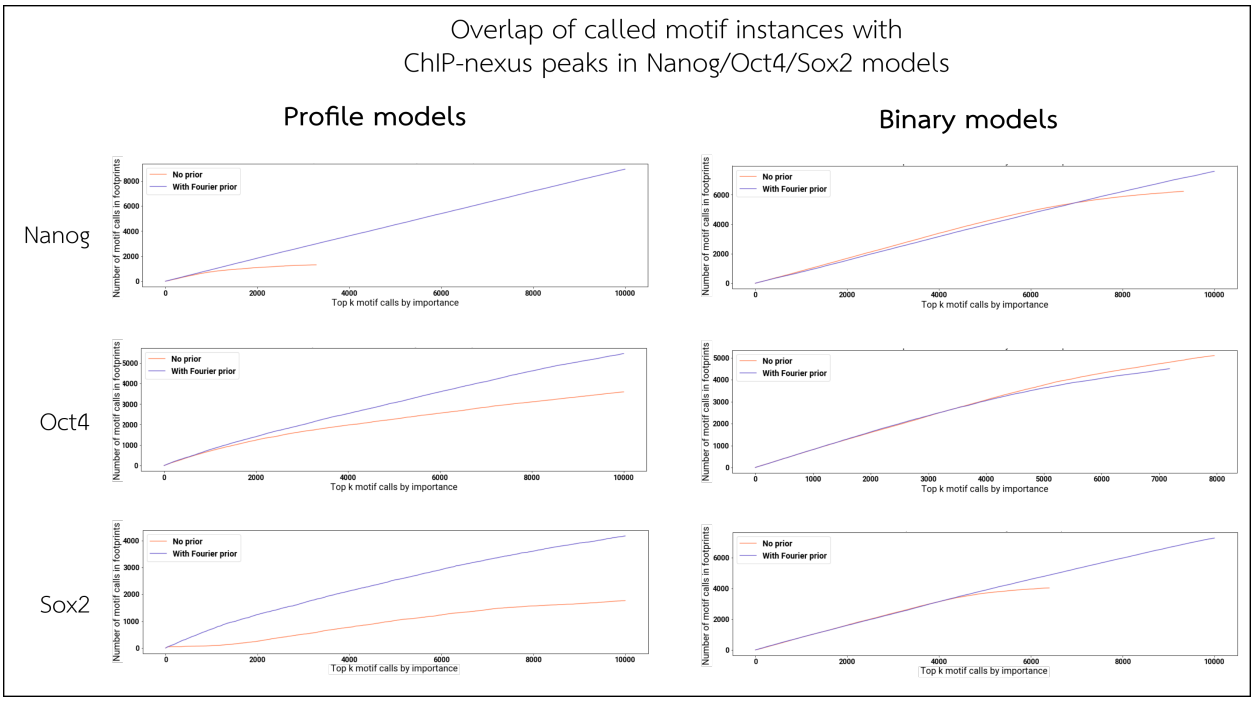

Figure S9: Motif instance call support. For each TF in the Nanog/Oct4/Sox2 models, we perform motif instance calling using the discovered motifs on a sample of 1000 test-set peak sequences. We rank the motif instance calls by total DeepSHAP importance, and compute a cumulative count of how many instances overlap with a ChIP-nexus peak for that TF. Note that the models trained without the prior typically have fewer motif calls in total (due to lower-quality attribution scores and fewer motifs discovered by TF-MoDISco), resulting in the shorter red lines.

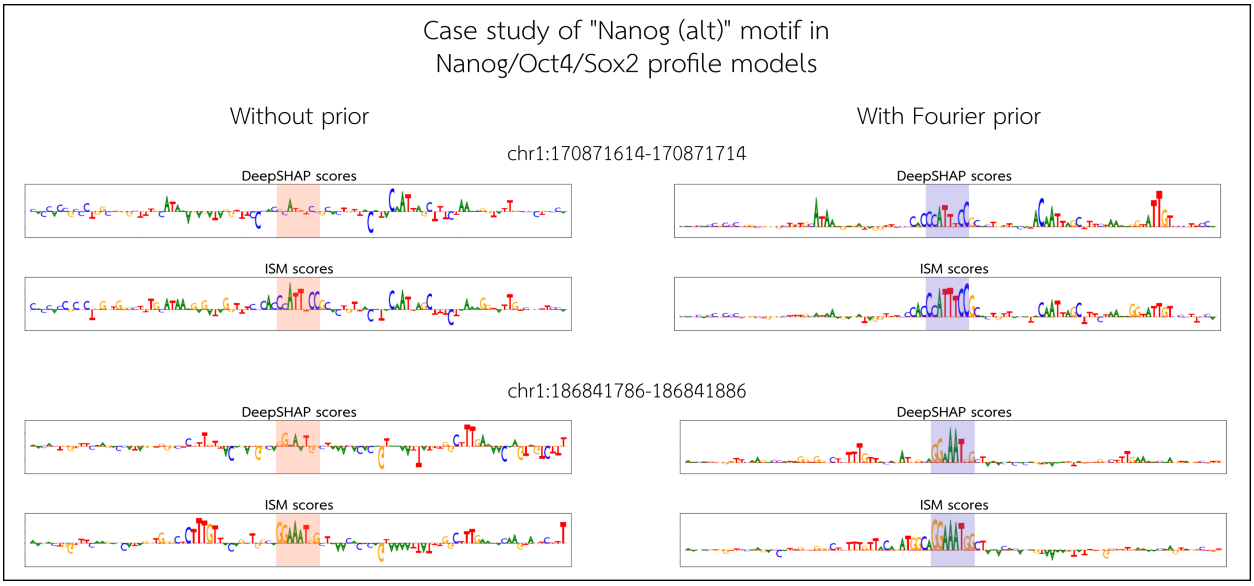

Figure S10: Case study of the "Nanog (alt)" motif in Nanog/Oct4/Sox2 profile models. TF-MoDISco identified the GGAAAT "Nanog (alt)" motif from the profile Nanog/Oct4/Sox2 model trained with the Fourier-based prior, but not from the model trained without the prior, even on the Nanog prediction task specifically (Supplementary Figure S8). Focusing on the Nanog prediction task, we show examples of sequences where TF-MoDISco identified a Nanog (alt) motif from the model trained with the prior, and show the importance scores of the same sequence from the model without the prior. We show both the DeepSHAP importance scores, and the perturbation scores derived from *in silico* mutagenesis (ISM). The highlighted regions indicate the location of the Nanog (alt) motif. Notably, ISM scores from the model trained without the prior are generally noisier compared to the model trained with the prior. More importantly, this demonstrates that the model trained without the prior is learning the Nanog (alt) motif, but it is not visible from DeepSHAP importance scores. The model trained with the Fourier-based prior, however, clearly highlights this motif using both methods of interpretation (i.e. DeepSHAP and ISM). This indicates that the Fourier-based prior allows the model to reveal its learned motifs in a human-interpretable way, especially when it is too computationally expensive to rely on perturbation-based scoring methods like ISM, which take orders of magnitude longer to run than DeepSHAP.

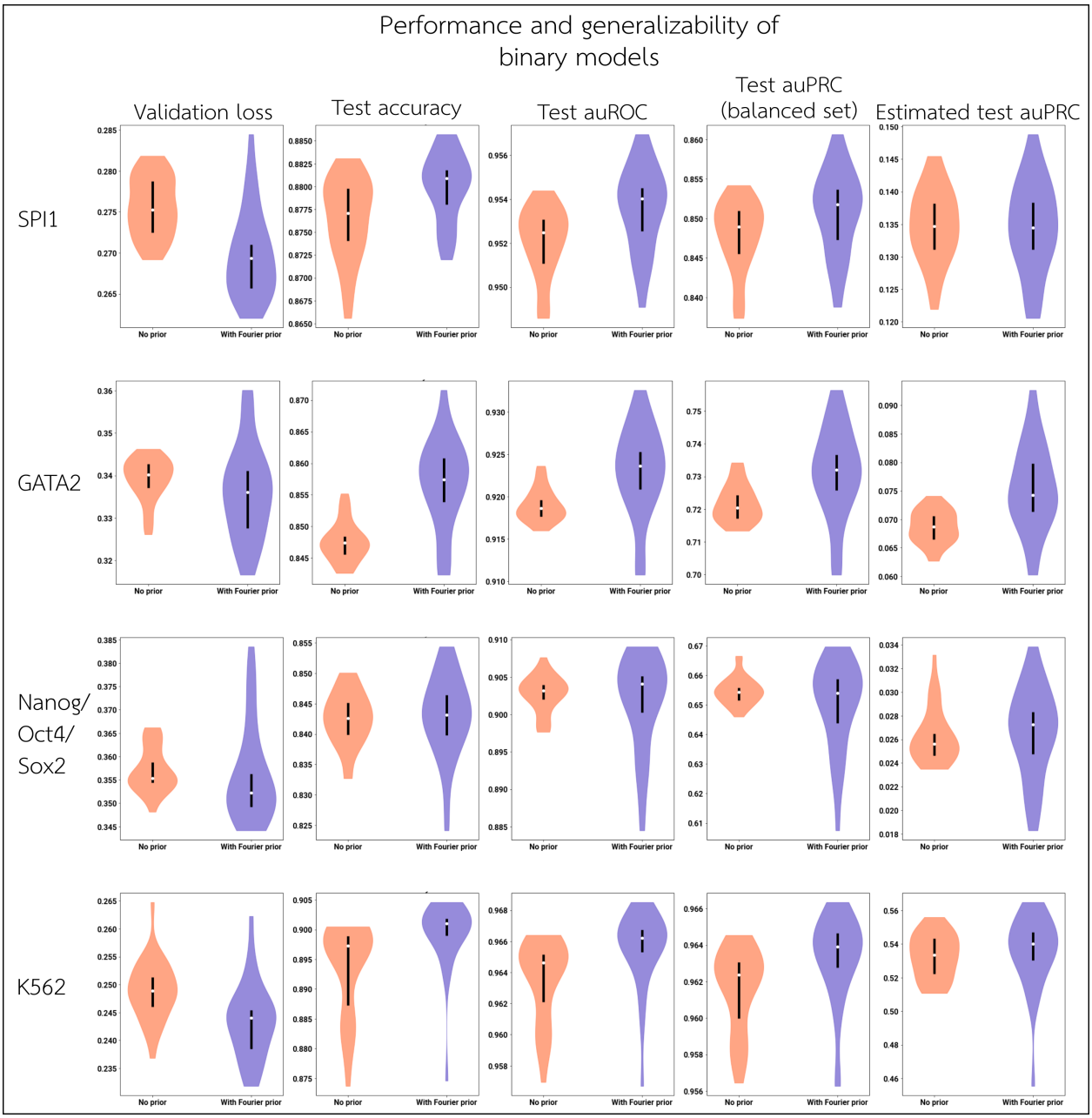

Figure S11: Validation loss and test-set performance of binary models. For each dataset, we consider the validation- and test-set performance of models trained with and without the Fourier-based prior, over 30 random initializations each. Validation loss is computed over all positive examples in the validation set and a equal-sized sample of negative validation examples. Test accuracy, test auROC, and test auPRC are computed over all positive examples in the test set and a equal-sized sample of test negative examples. "Estimated test auPRC" is an estimated measure of auPRC on the full test set without subsampling the negative examples, and is computed by artificially inflating the false positive rate (Supplementary Methods Sec. 2.5).

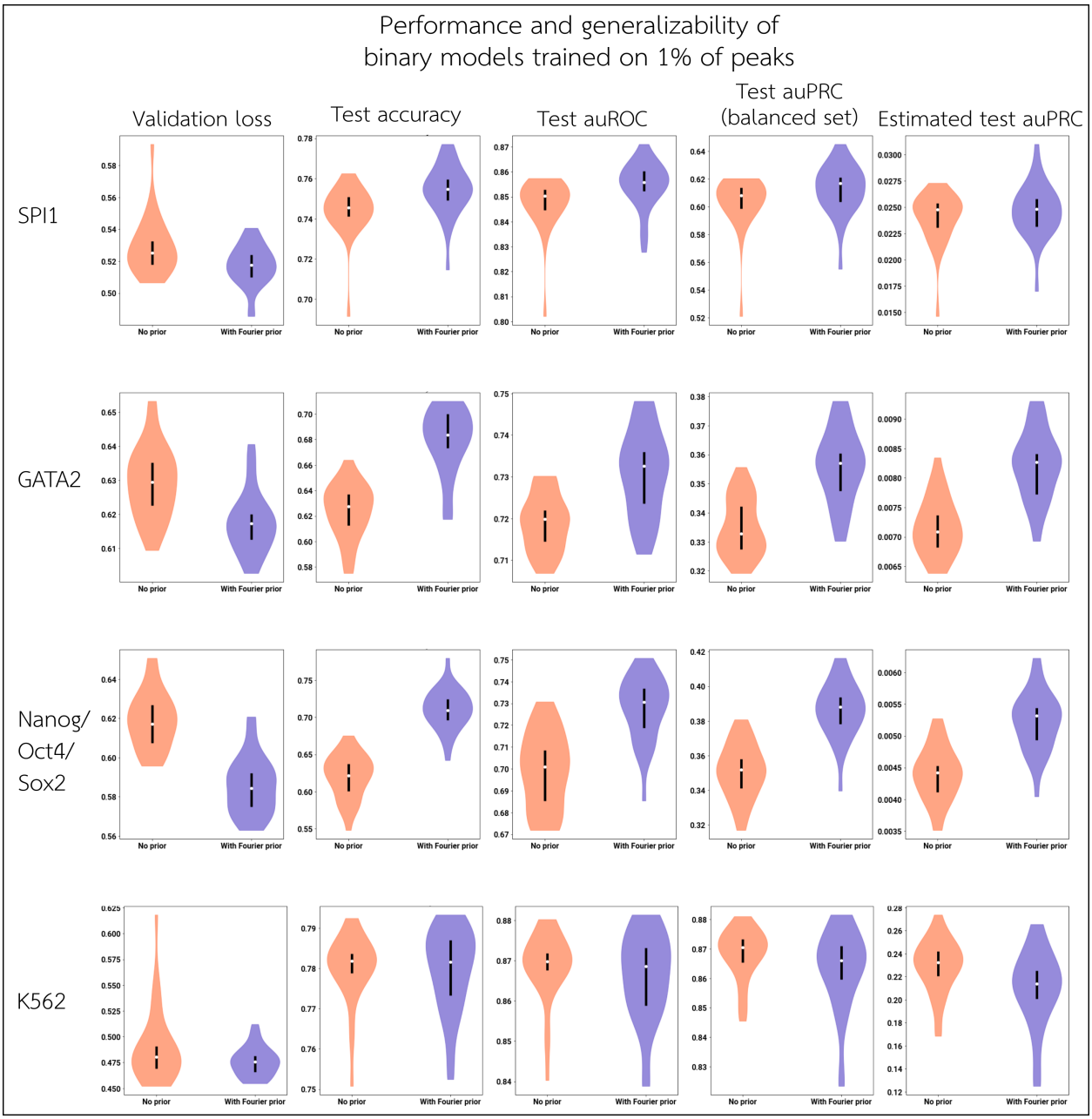

Figure S12: Validation loss and test-set performance of binary models on sparse training sets. For each dataset, we consider the validation- and test-set performance of models trained with and without the Fourier-based prior (on only 1% of the training set), over 30 random initializations each. Validation loss is computed over all positive examples in the validation set and a equal-sized sample of negative validation examples. Test accuracy, test auROC, and test auPRC are computed over all positive examples in the test set and a equal-sized sample of test negative examples. "Estimated test auPRC" is an estimated measure of auPRC on the full test set without subsampling the negative examples, and is computed by artificially inflating the false positive rate (Supplementary Methods Sec. 2.5)

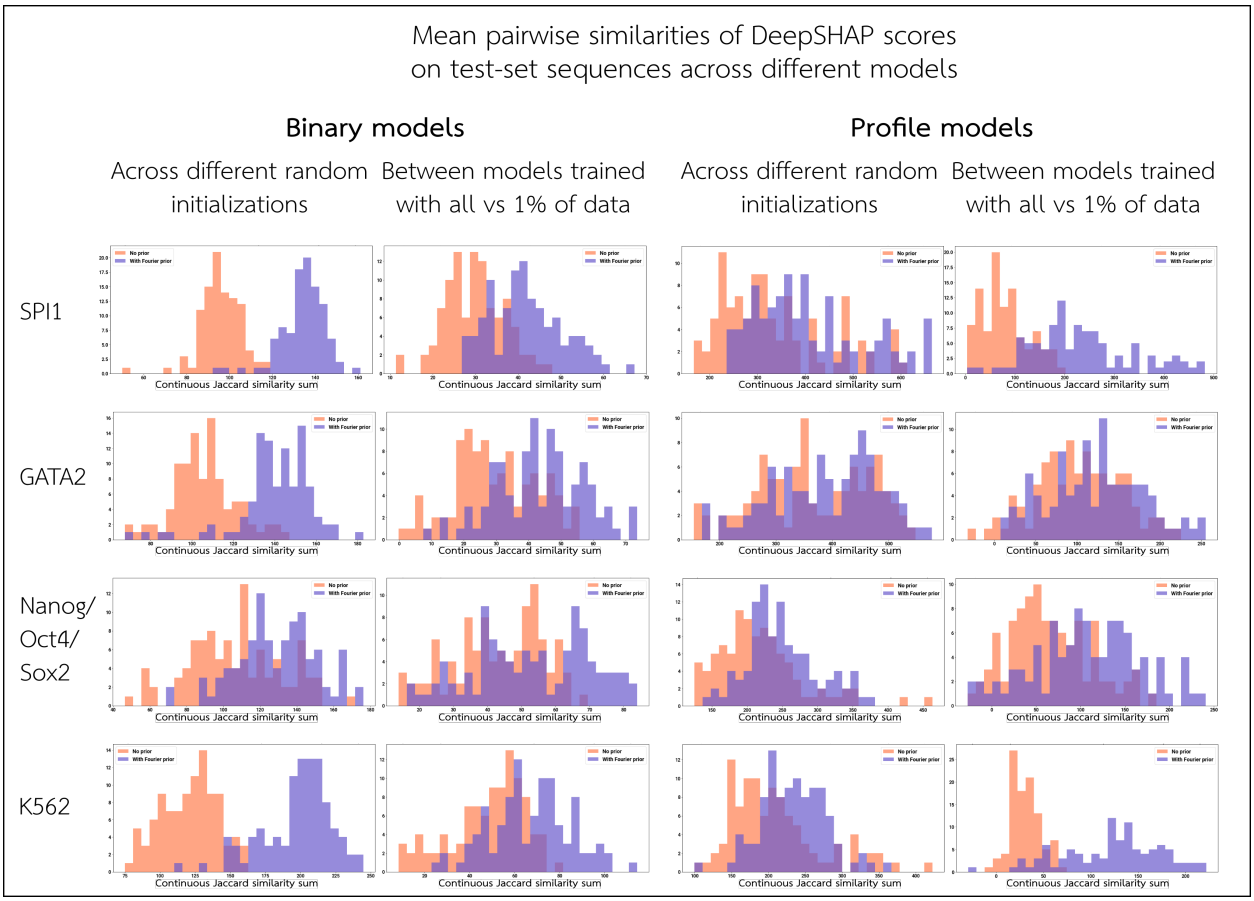

Figure S13: Stability of DeepSHAP scores across different models on test-set sequences. For each architecture and dataset, we sample 100 peak sequences from the test set and compute the DeepSHAP attributions for the sequence in multiple models. For each sequence, we compute the pairwise similarity of the attributions between 30 random initializations (left), or between the top 5 models trained with all of the data versus the top 5 trained with only 1% of the data (right). We quantify attribution similarity by computing the continuous Jaccard score at each base, and summing the scores across the sequence.

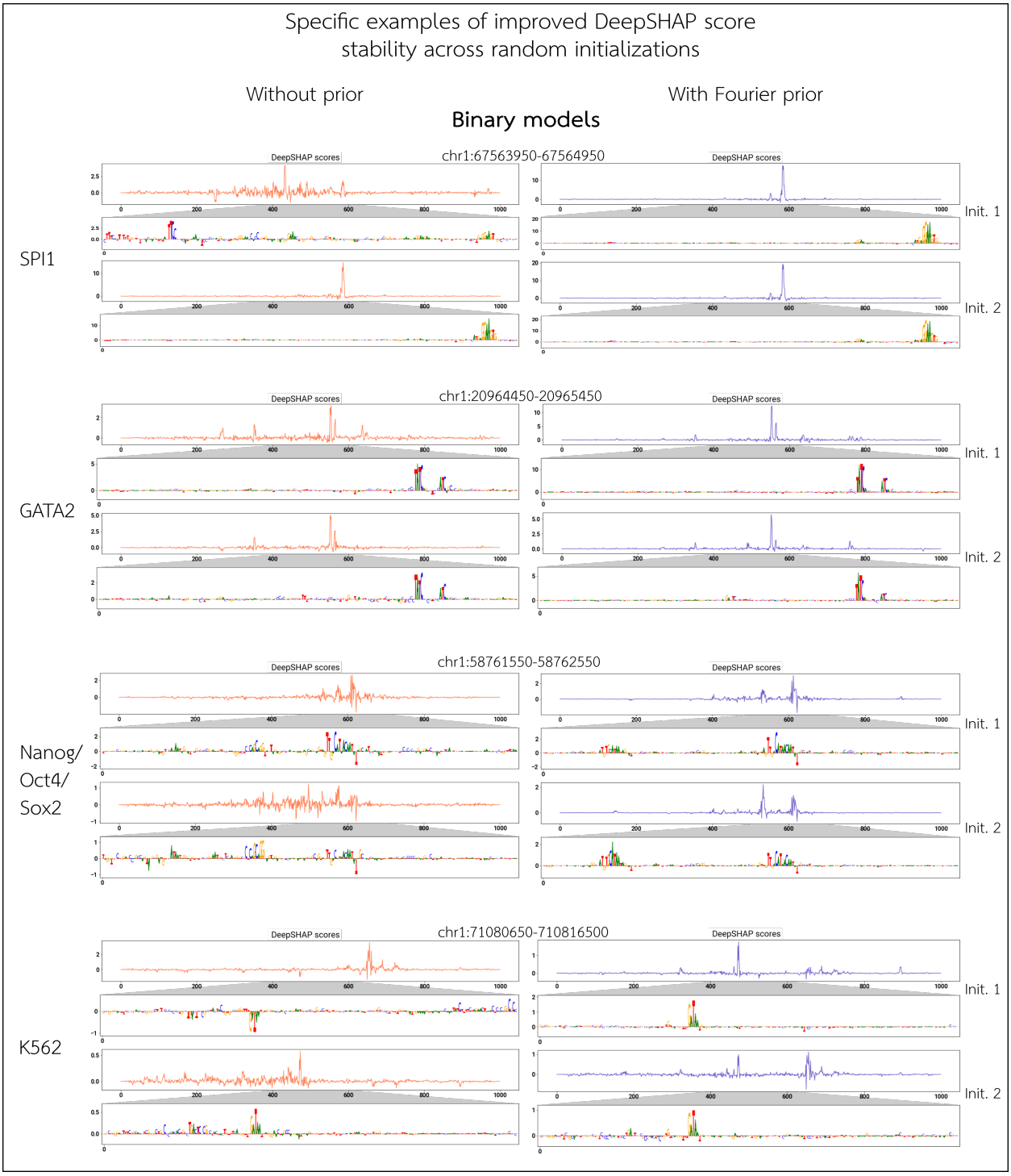

Figure S14: Specific examples of DeepSHAP attribution stability across different random initializations (binary models). For each binary model, we show an example of the DeepSHAP attributions on a test-set sequence between a pair of models of different random initializations, comparing models trained with versus without the Fourier-based prior.

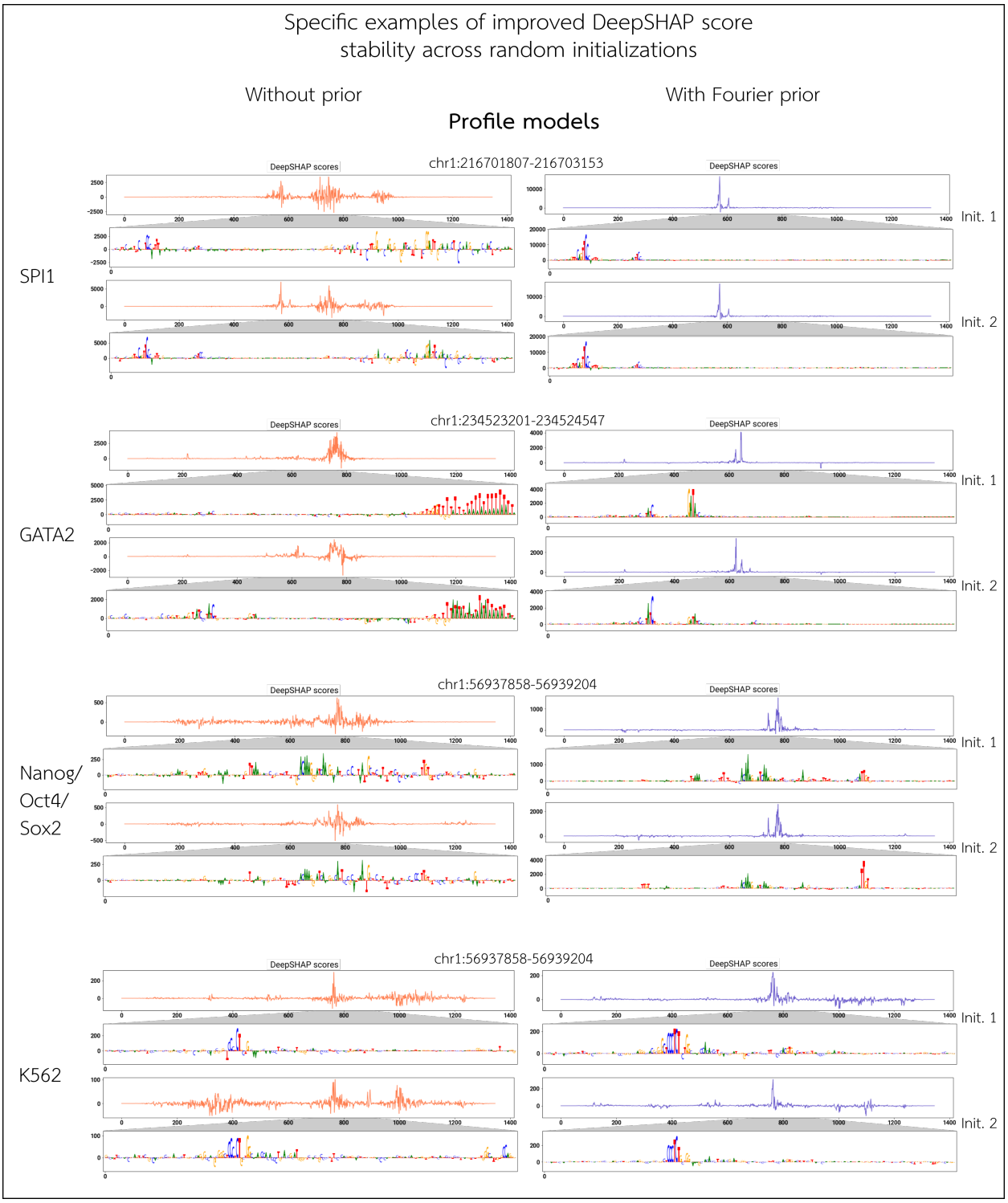

Figure S15: Specific examples of DeepSHAP attribution stability across different random initializations (profile models). For each profile model, we show an example of the DeepSHAP attributions on a test-set sequence between a pair of models of different random initializations, comparing models trained with versus without the Fourier-based prior.

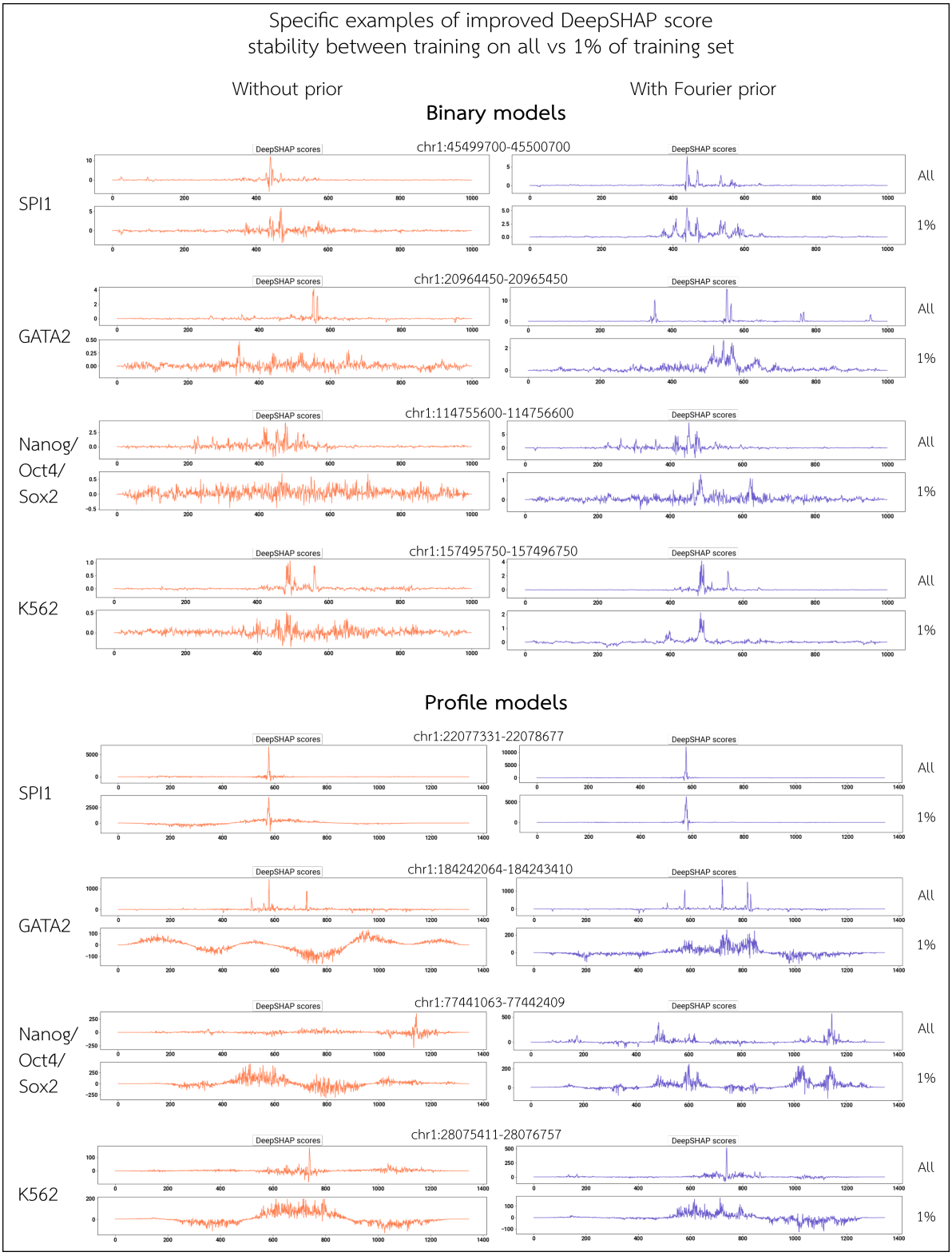

Figure S16: Specific examples of DeepSHAP attribution stability between training on all versus 1% of the training set. For each architecture and dataset, we show an example of the DeepSHAP attributions on a test-set sequence between a model trained on all and only 1% of the training set, comparing models trained with versus without the Fourier-based prior.

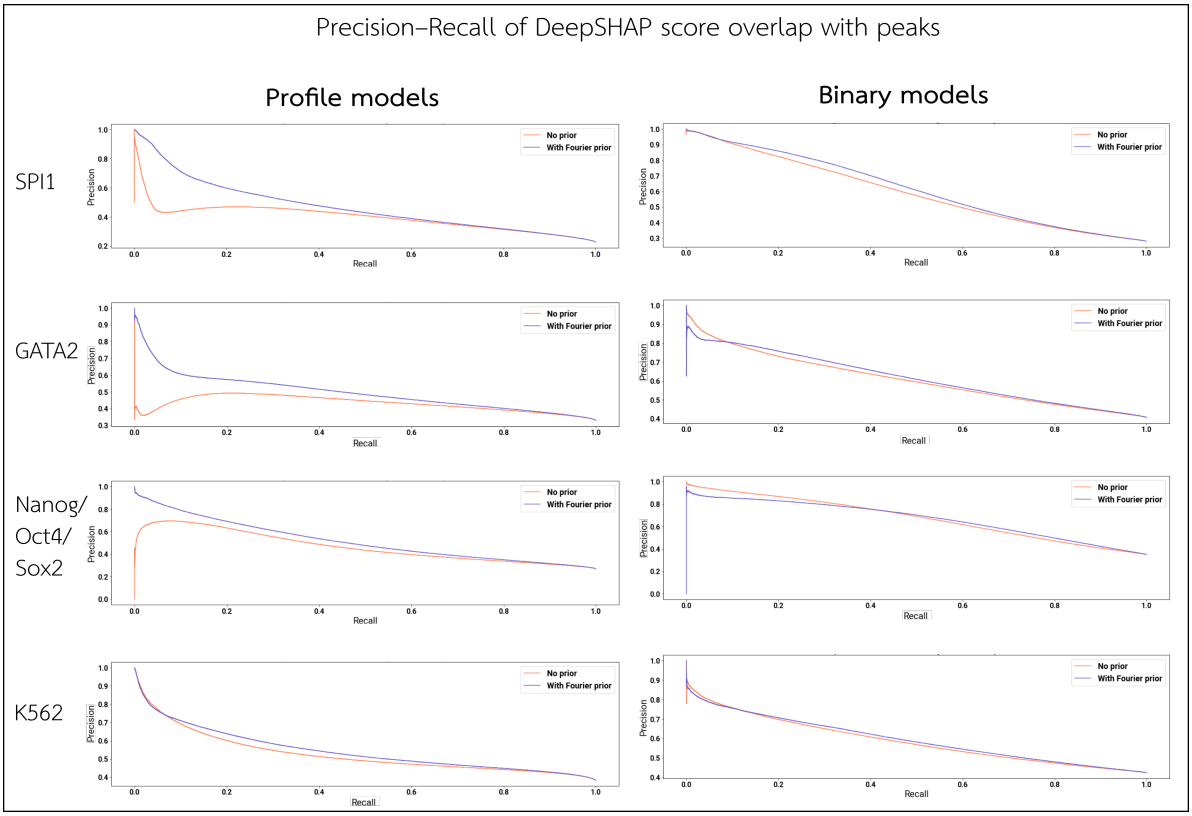

Figure S17: Precision–recall of importance-ranked base overlap with called peaks. For each architecture and dataset, we sample 1000 peak sequences from the test set and compute the DeepSHAP attributions. We rank bases in descending order of total importance, and generate a precision–recall curve by treating the set of bases that overlap an underlying ChIP-seq or DNase-seq peak as "positives". See Table 3 in the main text for the corresponding auPRC values.

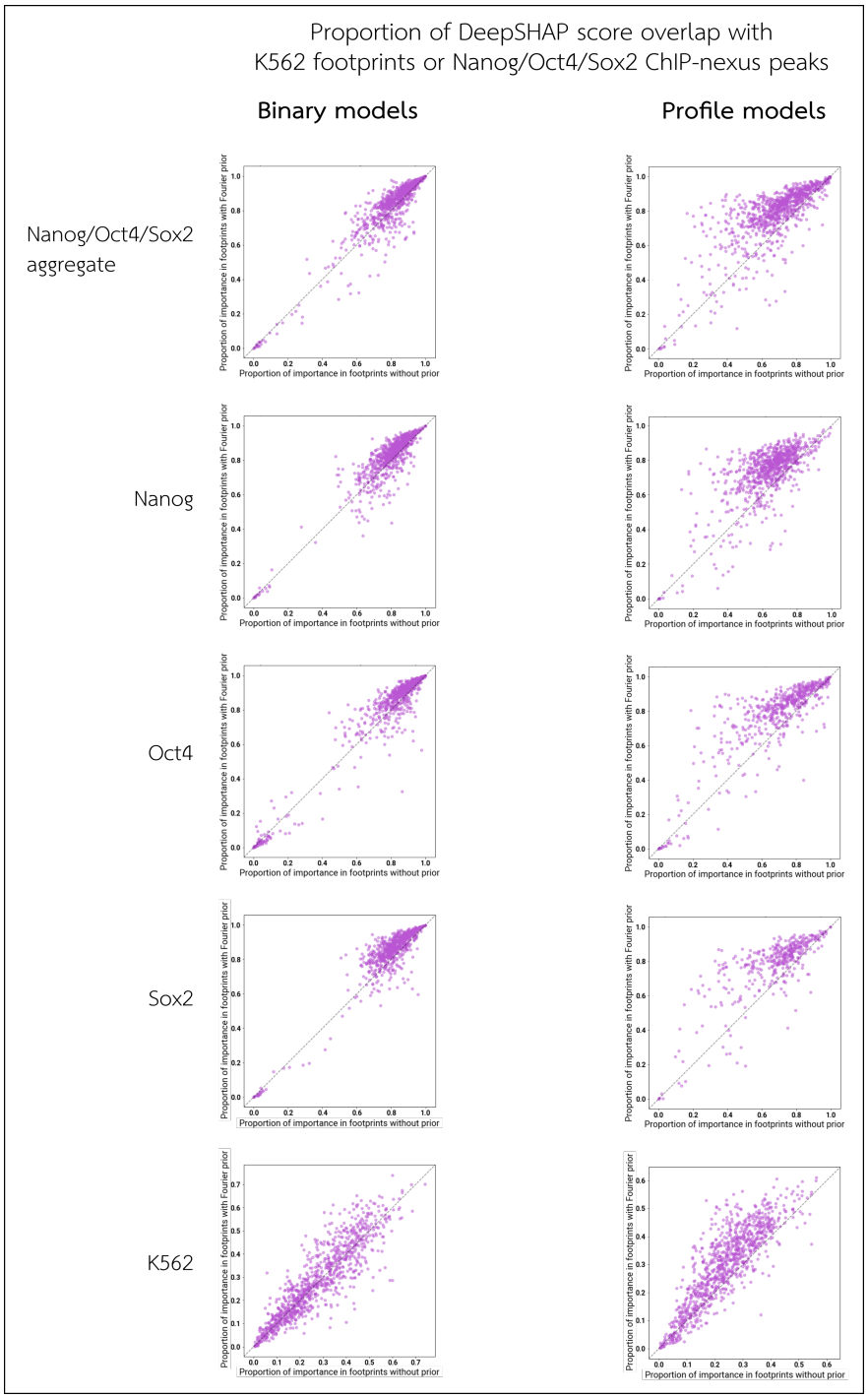

Figure S18: Fraction of importance in ChIP-nexus peaks or DNase footprints. For Nanog/Oct4/Sox2 TF ChIP-seq or K562 DNase-seq models, we sample 1000 peak sequences from the test set and compute the fraction of total DeepSHAP importance that overlaps a Nanog/Oct4/Sox2 ChIP-nexus peak or K562 footprint. For each input sequence, we compare the proportion of attribution by magnitude overlapping a ChIP-nexus peak or footprint when a model is trained with versus without the Fourier-based prior. For Nanog/Oct4/Sox2 models, we show the importance overlap using the total importance across all three tasks (in which case we use the union of ChIP-nexus peaks to define overlapping regions), as well the overlap using the importance scores of each individual task (in which case we use only the ChIP-nexus peaks of the corresponding task). See Table 4 in the main text for the corresponding average overlap values.

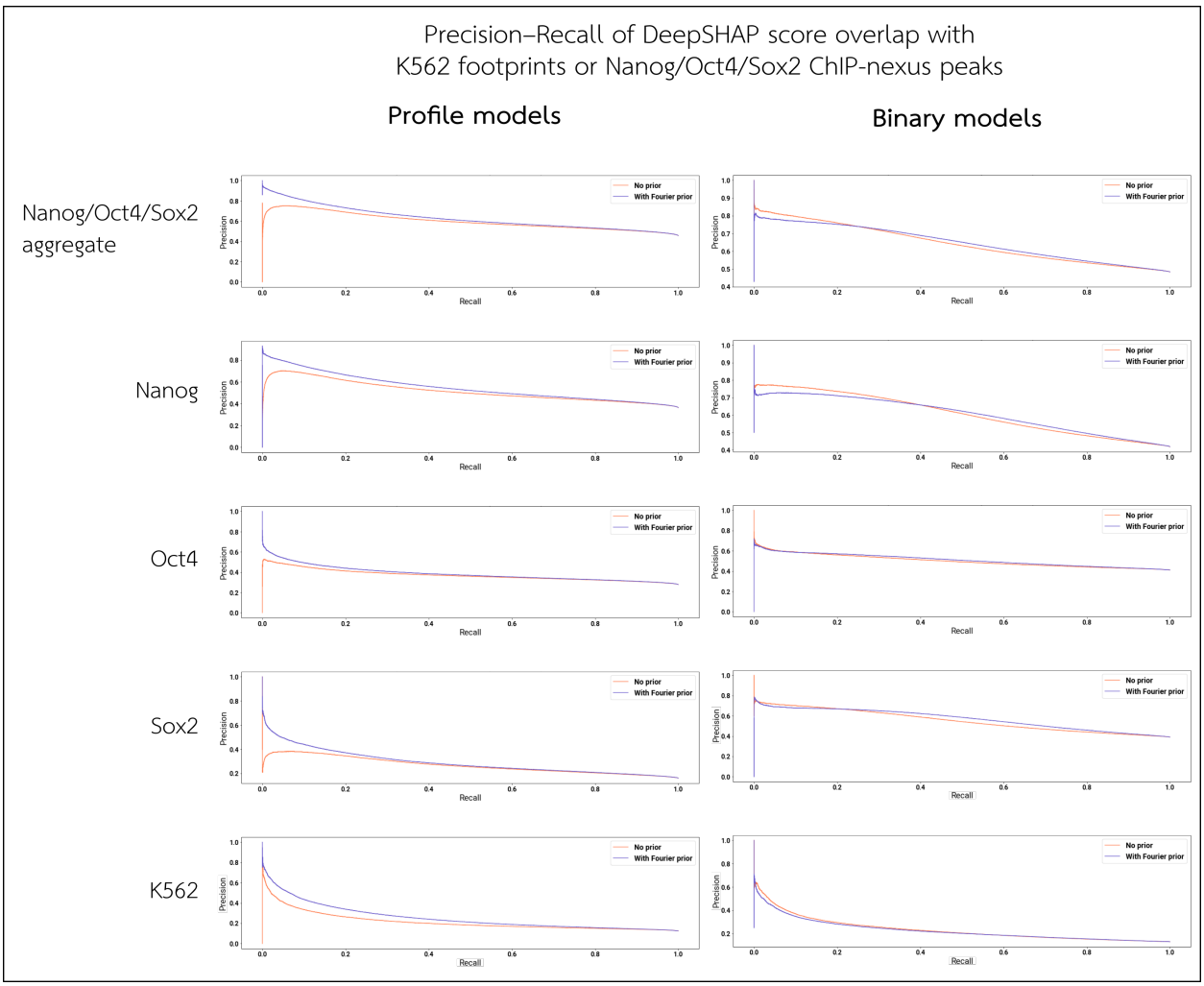

Figure S19: Precision–recall of important bases overlapping with ChIP-nexus peaks or DNAse footprints. For models trained to predict Nanog/Oct4/Sox2 TF ChIP-seq or K562 DNase-seq, we sample 1000 peak sequences from the test set and compute the DeepSHAP importance. We rank bases in descending order of total importance and generate a precision–recall curve by treating the set of bases that overlap a Nanog/Oct4/Sox2 ChIP-nexus peak or K562 DNase-seq footprint as "positives". For Nanog/Oct4/Sox2 models, we show these curves both using a ranking based on the total importance across all three tasks (in which case we use the union of ChIP-nexus peaks over the three tasks for the labels), as well as a ranking based on the importance scores of each individual task (in which case we use only the ChIP-nexus peaks of the corresponding task). See Table 4 in the main text for the corresponding auPRC values.

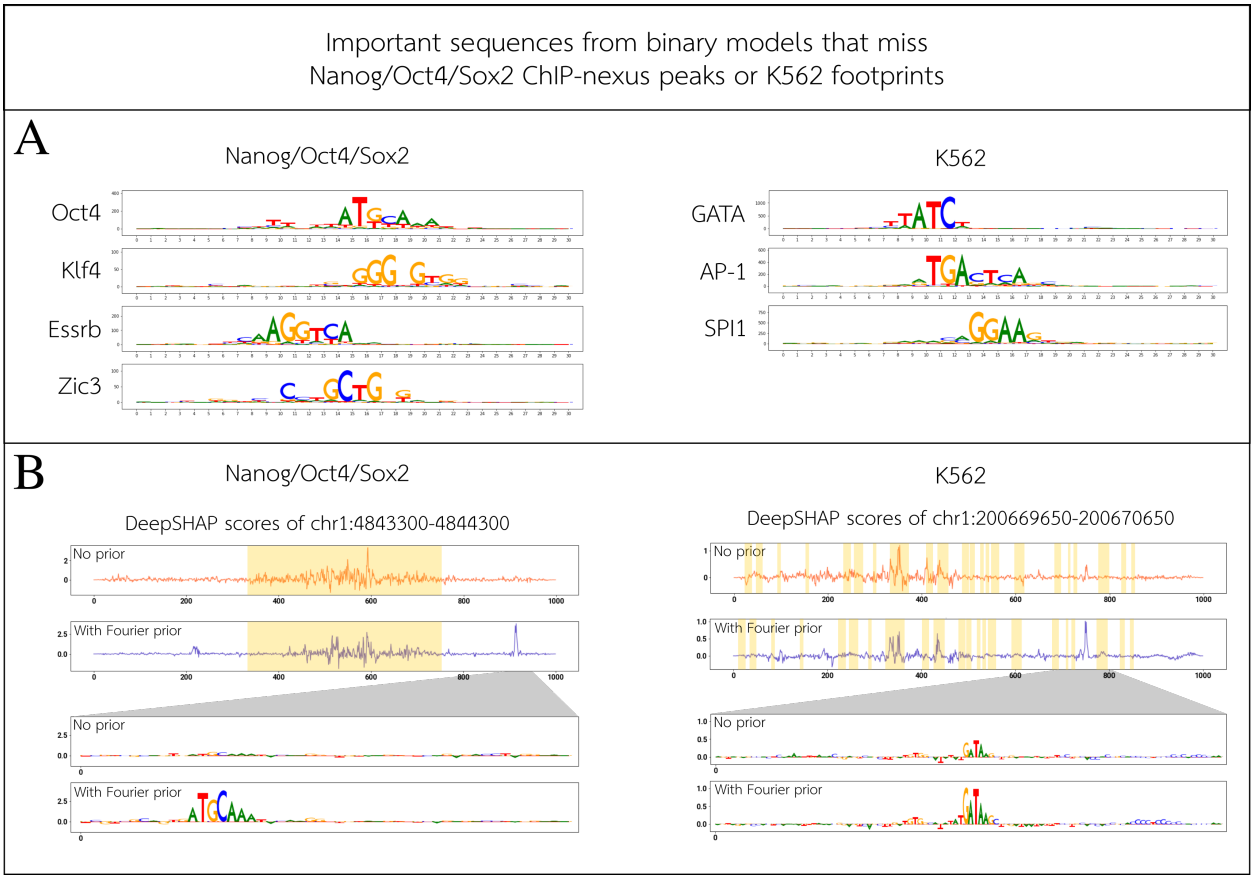

Figure S20: Motifs in important regions outside of ChIP-nexus peaks or footprints. For the Nanog/Oct4/Sox2 TF ChIP-seq and K562 DNase-seq binary models, we examine regions where models trained with the Fourier-based prior place high DeepSHAP importance, yet do not overlap Nanog/Oct4/Sox2 ChIP-nexus peaks or K562 footprints, respectively. **A)** We cluster these regions using the TF-MoDISco clustering algorithm, and show the PWMs of top motif clusters, along with annotations of relevant TFs that are associated to each motif. **B)** We show some specific examples where a model trained with the Fourier-based prior places higher importance (relative to the model trained without the prior) outside of a ChIP-nexus peak or K562 footprint. Yellow shading denotes the location of the ChIP-nexus peak (left) or K562 footprint (right). This illustrates the Fourier-based prior's highlighting biologically relevant motifs outside of peak or footprint regions; the model trained without the prior, on the other hand, identifies these motifs more weakly/noisily, or not at all. Several mechanisms exist by which secondary motifs outside the central peak region can nonetheless assist TF binding within the peak region (e.g. via cofactor interactions [3] or 1D sliding [4]). A binary prediction model would be correct in identifying such motifs as being predictive of peak strength. Note that in contrast to binary models, profile models are less likely to detect such secondary motifs, as these motifs contribute to peak strength without contributing to the shape of the peak itself (peak shape is primarily dictated by motifs lying within the central peak region).

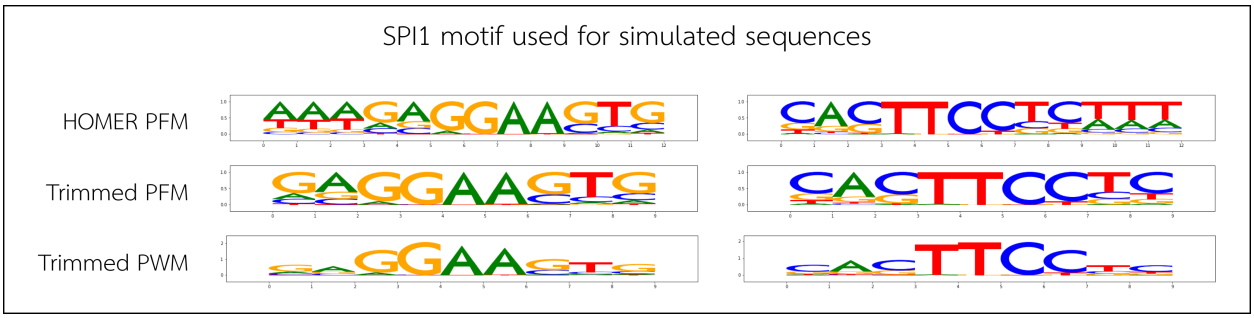

Figure S21: SPI1 motif used in constructing simulated sequences. We show the SPI1 motif (left) and its reverse complement (right) used to simulate SPI1-binding sequences for models trained on simulated data. Shown here are (from top to bottom): the Position Frequency Matrix (PFM) of the top motif identified by running HOMER 2 on IDR-thresholded peaks for SPI1 (Supplementary Methods Sec. 2.8); the trimmed motif PFM after removing flanks with low information content; and a PWM of the trimmed motif derived from weighting the PFM by information content.

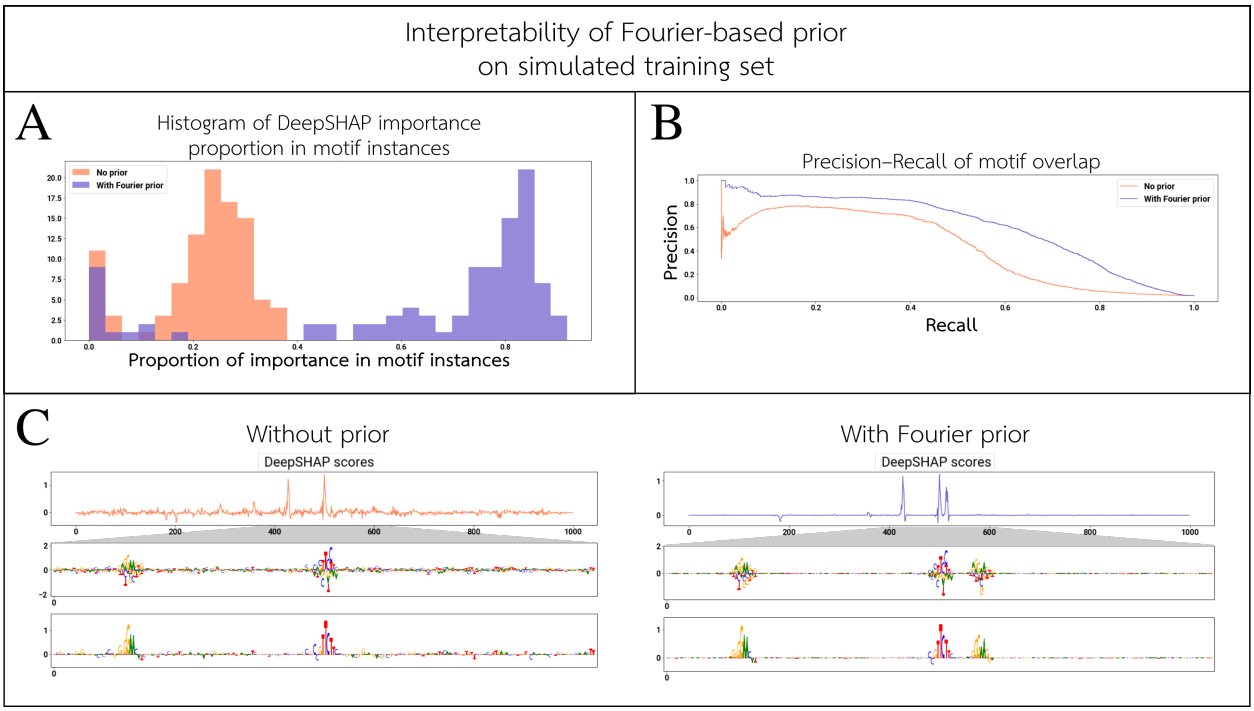

Figure S22: Attributions in simulated motif instances. We examine binary models trained to predict single-task SPI1 binding on simulated sequences. On a random sample of 100 motif-containing sequences, we compare models trained with versus without the Fourier-based prior by computing: **A)** the proportion of DeepSHAP importance overlying a motif instance; and **B)** the precision–recall of importance-ranked bases overlying motif instances. **C)** We also select an example sequence, and show the DeepSHAP attributions; the model trained with the Fourier-based prior cleanly highlights all three motif instances. Of the two zoomed-in attribution score tracks in Panel **C**, the upper track represents the hypothetical importance scores (i.e. the importance that would be given to each of the four bases at each position, even if that base were not in the input sequence—see Supplementary Methods Sec. 3), and the lower track represents the actual importance scores. See Supplementary Figure S21 for the SPI1 PWM used in the simulations. Note that the central motif instance is in the reverse-complement orientation.

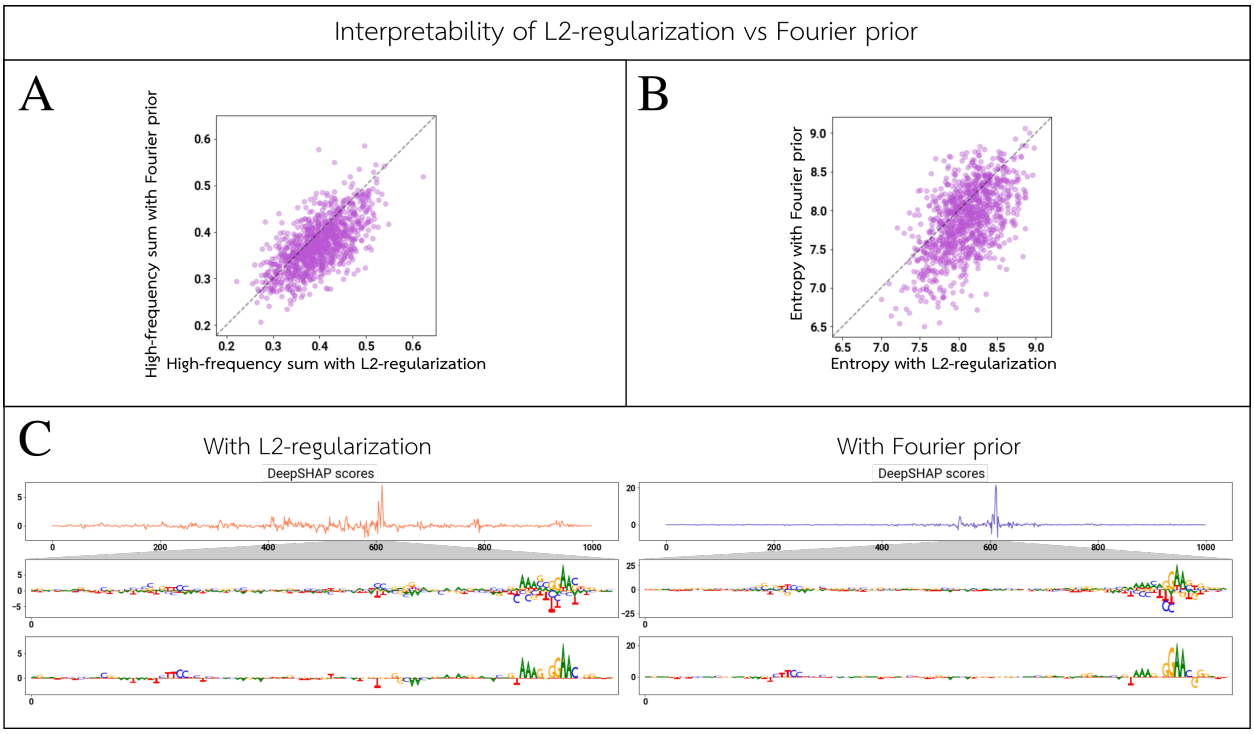

Figure S23: Interpretability of attributions with L2-regularization versus Fourier-based prior. For a SPI1 binary model, we show the DeepSHAP attributions for 1000 randomly selected peak sequences from the test set. We compare the sum of normalized high-frequency Fourier components and the Shannon entropy for each sequence, between a model trained with L2-regularization (i.e. weight decay) and a model trained with the Fourier-based prior. For a single selected test peak, we also display the value of the DeepSHAP importance for the bases present along the entire input sequence, as well as a zoomed-in view of a region close to the peaks summit. We also show the "hypothetical" attributions along each input sequence (i.e. the importance that would be given to each of the four bases at each position, even if that base were not in the input sequence—see Supplementary Methods Sec. 3).

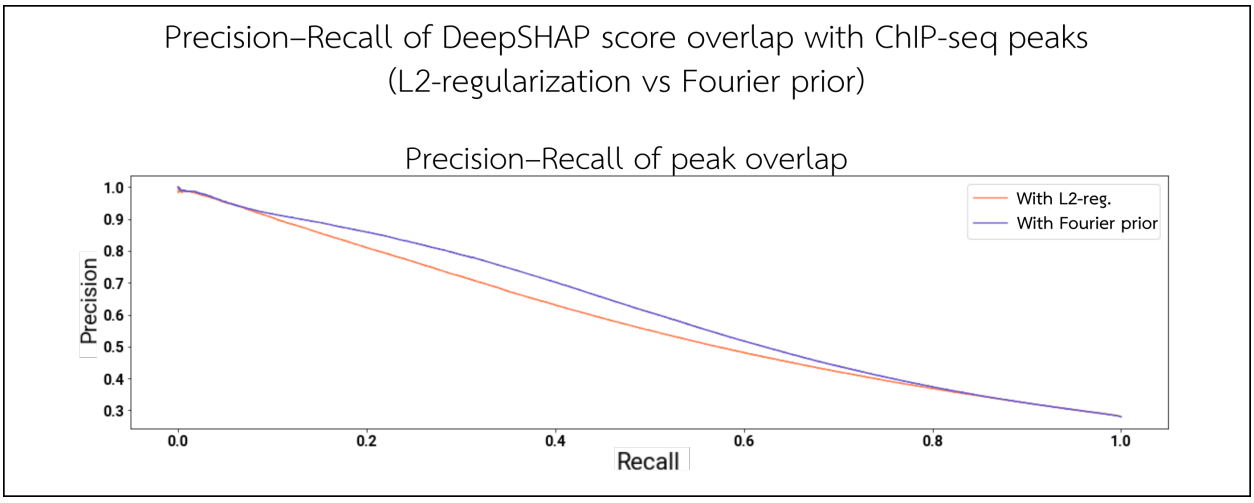

Figure S24: Precision–recall of important bases overlapping with ChIP-seq peaks (L2-regularization versus Fourier-based prior). For a SPI1 binary model, we sample 1000 peak sequences from the test set and compute the DeepSHAP attributions. We rank bases in descending order of total importance and generate a precision–recall curve by treating the set of bases that overlap a SPI1 ChIP-seq peak as "positives". Shown here are the precision–recall curves for models trained with L2-regularization versus the Fourier-based prior.

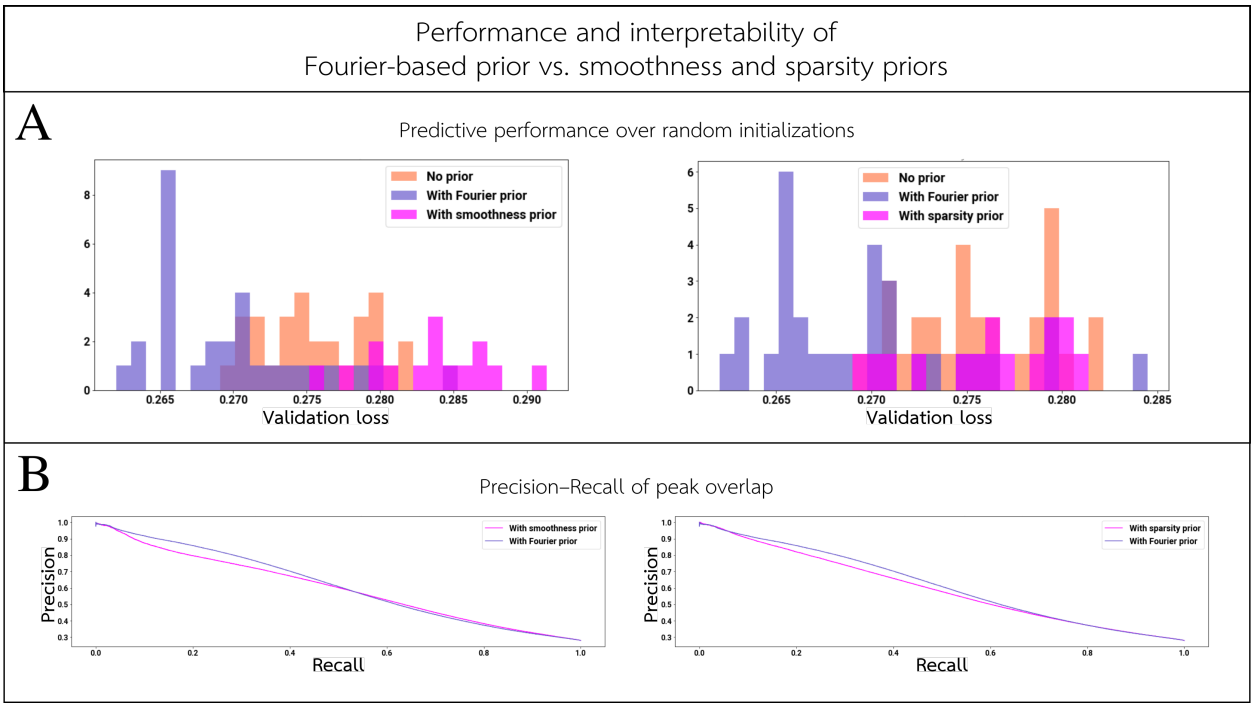

Figure S25: Comparison of predictive performance and interpretability between the Fourier-based prior and alternative attribution prior formulations. **A)** On SPI1 binary models trained with the smoothness prior or sparsity prior [5], we show the distribution of the predictive performance (as measured by validation-set prediction loss) compared to models trained with the Fourier-based prior or with no prior at all, over several random initializations. The predictive performance of the smoothness and sparsity priors is significantly worse than the Fourier-based prior. Models trained with the smoothness prior have worse performance than models trained with no prior at all. **B)** We tune the prior loss weights for the smoothness and sparsity priors, and select the models with the lowest validation loss (Supplementary Methods Sec. 2.7). In these models, the smoothness and sparsity priors have poor interpretability compared to the Fourier-based prior, as measured by the auPRC of importance overlap with ChIP-seq peaks.

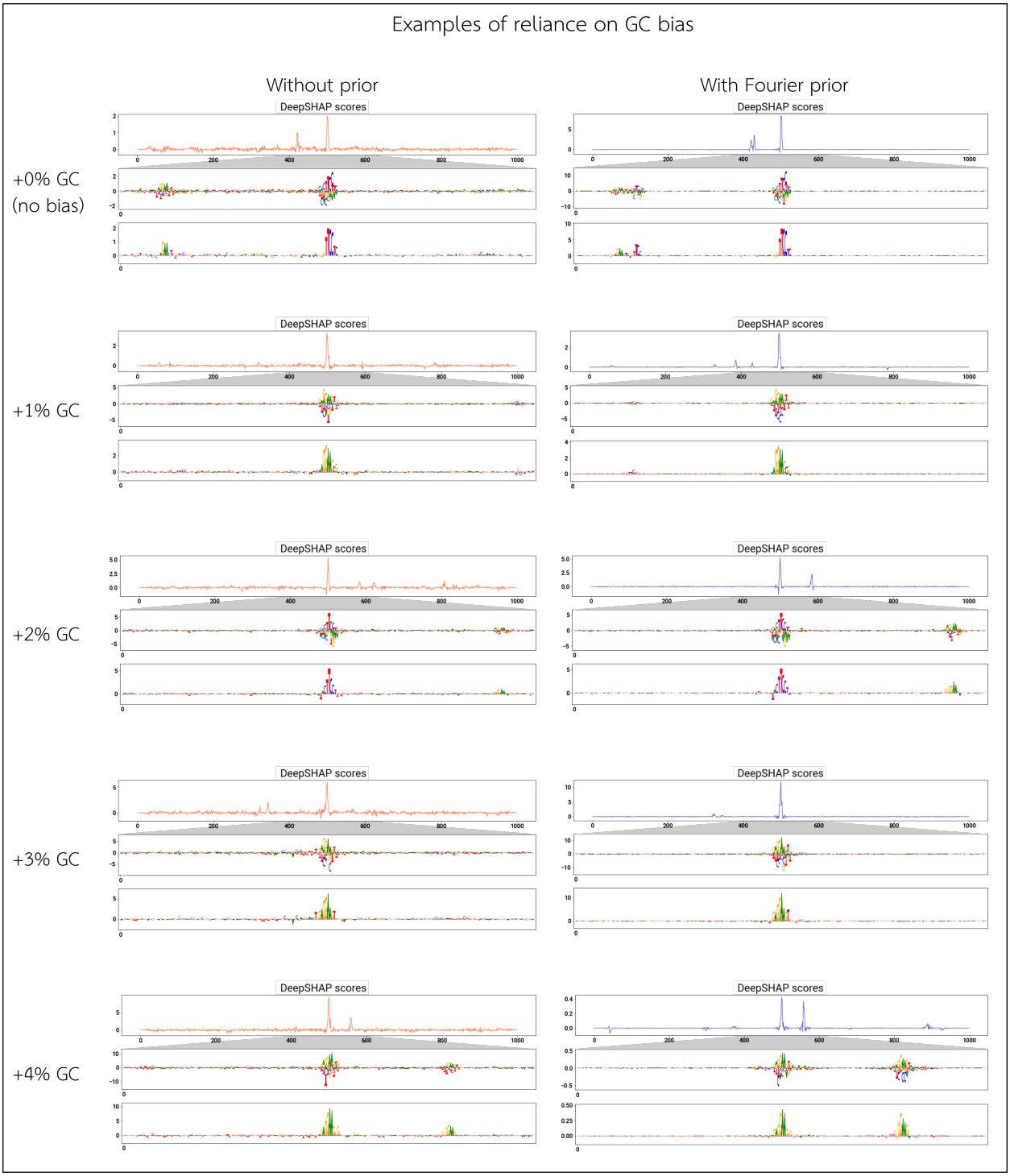

Figure S26: Specific examples of attributions under various levels of GC content. For each level of GC bias, we select a sampled input sequence and examine the DeepSHAP attributions. GC content ranges from +0% to +4%: a GC bias of $+x\%$ means the probability of G or C in the background of a positive sequence is $(50 + x)\%$, while the background of a negative sequence has a G/C probability of 50% (i.e. no bias at all). We display the value of the DeepSHAP importance for the bases present along the entire input sequence, as well as a zoomed-in view of the region surrounding the central motif. We also show the "hypothetical" attributions along each input sequence (i.e. the importance that would be given to each of the four bases at each position, even if that base were not in the input sequence—see Supplementary Methods Sec. 3).

References

[1] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Amr Alexandari, Sabrina Krueger, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *bioRxiv*, page 737981, Aug 2019. doi: 10.1101/737981. URL `https://www.biorxiv.org/content/10.1101/737981v1`.

[2] David R. Kelley, Jasper Snoek, and John L. Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, Jul 2016. ISSN 15495469. doi: 10.1101/gr.200535.115.

[3] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. The Human Transcription Factors, Feb 2018. ISSN 10974172.

[4] David M. Suter. Transcription Factors and DNA Play Hide and Seek, Jun 2020. ISSN 18793088.

[5] Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Learning Explainable Models Using Attribution Priors. Jun 2019. URL `http://arxiv.org/abs/1906.10670`.