Uncovering motif interactions from convolutional-attention networks for genomics

Rohan S. Ghotra
Partners for the Future Program
Cold Spring Harbor Laboratory
Syosset High School

Nicholas K. Lee
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory

Peter Koo
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
koo@cshl.edu

Abstract

A major goal of computational genomics is to understand how sequence patterns, called motifs, interact to regulate gene expression. In principle, convolution-attention networks (CANs) should provide an inductive bias to infer motif interactions; convolutions can capture motifs while self-attention learns their interactions. However, it is unclear the extent to which this is true in practice. Here we perform an empirical study on synthetic data to test the efficacy of uncovering motif interactions in CANs. We find that irrespective of design choice, interpreting local attention (i.e. on an individual sequence basis) is noisy, leading to many false positive motif interactions. To address this issue, we propose Global Interactions via Filter Activity Correlations (GLIFAC). GLIFAC robustly uncovers motif interactions across a wide spectrum of model choices. This work provides guidance on design choices for CANs that lead to better interpretability for regulatory genomics without sacrificing generalization performance.

1 Introduction

Model interpretability is key to translating the powerful prediction performance of deep neural networks (DNNs) into scientific discovery. However, due to the enormous numbers of parameters in modern DNNs, individual parameters are often not meaningful on their own. It remains unclear which design choices provide the right inductive biases to learn robust and interpretable features. For regulatory genomics, the goal often is to use a machine learning model to take DNA sequences as input and predict a regulatory function, such as transcription factor binding [1], which are typically measured experimentally via high-throughput experiments.

Deep convolutional neural networks (CNNs) have indeed demonstrated state-of-the-art performance at these kinds of tasks [2, 3]. Post hoc interpretation via visualizing first layer filters and attribution maps have demonstrated that CNNs make decisions based on learning biologically meaningful sequence patterns [4, 5, 6], called motifs. However, attribution maps provide an anecdotal glimpse into the decision making process but can be quite noisy [7], making it a challenge to discover generalizable patterns beyond a first-order approximation, i.e. single nucleotide perturbations. On the other hand, first layer filters can also learn to detect motifs, though the extent that they do depends strongly on the design choices, such as the activation function [8] or the max-pooling after the first layer [9]. Many simpler, traditional computational methods can also discover similar motifs; so it remains a mystery as to why CNNs make better predictions? One hypothesis is that they are able to not only learn motifs

NeurIPS 2021 AI for Science Workshop.

but also how they interact, the so-called regulatory code [10]. Uncovering motif interactions is a major goal in regulatory genomics as it has a direct impact on understanding mechanisms of gene regulation and can help to design functional regulatory sequences.

In principle, motif interactions would be encoded in deeper layers of a CNN, but it remains difficult to interpret deeper layer filters because their activations are complex – we don't know whether a filter that captures a motif is 'active' and motif interactions occur at different scales, information that is spread across different layers of a CNN. By contrast, self-attention, which is a key component of the transformer [11], provides a strong inductive bias to learn long-range interactions. Thus, it offers a natural avenue to efficiently capture elusive motif-motif interactions at any scale within a single layer. Indeed, previous convolution-attention hybrid networks have demonstrated improved performance versus pure CNNs at regulatory genomic prediction tasks [12, 13]. While there has been effort previously at interpreting the attention maps to identify motif interactions [12], it remains unclear whether there exists design principles that result in more interpretable attention maps.

Here, we perform a systematic, empirical study to investigate how design choices of convolution-attention networks impact the interpretability of attention maps. As part of this study, we propose a new method to interpret motif interactions, significantly improving upon previous efforts. This work provides an avenue to narrow the infinite spectrum of architecture choices when designing interpretable convolution-attention hybrid networks for discovery of motifs and their interactions, and it proposes a new computational approach to extract this information from them.

2 Experimental overview

To assess the ability to recover motif-motif interactions, we generated synthetic dataset that comprises a multi-task binary classification. Briefly, 100,000 synthetic sequences embedded with 1 to 5 non-overlapping binding sites, selected with replacement from a pool of 12 known motifs, each of which represents a different task; the labels of some motifs encode complex regulatory logic consisting of positive and negative interactions (see Appendix A). The sequences were split into training, validation, and testing sets, with a 0.7, 0.1, 0.2 distribution, respectively. Thus, a model must learn to recognize TF binding site motifs as well as their interactions (if any) to be successful at this task.

Attention maps are only intrinsically interpretable if the convolutional layer learns robust and interpretable motif representations. The extent that first layer filters learn motif representations was previously explored for various convolutional-hybrid models, including those that incorporate self-attention [14]. Building upon this work, we designed a baseline architecture with minimal components: convolutional layer, max-pooling, multi-head attention (MHA), dense layer, and output layer. We explored variations of this baseline model, using a nomenclature that corresponds to design choices. For instance, CNN4-ReLU-norm represents a baseline CNN with a max-pool of 4, ReLU activations in first layer filters, and normalization, which includes batch normalization in the convolutional layer and layer normalization prior to MHA. Additionally, we trained the model proposed as part of SATORI [12], a model and methodology to extract motif interactions from local attention maps. We trained each model on the synthetic dataset with 10 different random inializations. See Appendix A for details of models and training procedure.

3 Local attention is noisy, leading to many false positive interactions

The 3 main challenges to using attention maps for model interpretability in genomics are: (1) motif representations must be identifiable for each filter, (2) interacting position pairs must be identifiable from an attention map, and (3) the active filter(s) within each attended position must be identifiable. Classification performance for all tested models was high (AUPR of 0.95-0.97) and they all yielded high coverage of ground truth motifs in first layer filters. Here, we focus on issues (2) and (3).

Identifiability of significant attention. To test the efficacy of extracting learned motif interactions from attention maps, we must first identify which values in the attention map constitute a significant interaction; the attention map is considered the softmax of the key-query matrix products, an $L \times L$ matrix that corresponds to position-position interactions. Previously, a method called SATORI employed a threshold of 0.1 [12], above which is considered to represent a significant attention value. By visualizing the attention distributions for each test sequence, we found that the distributions can be

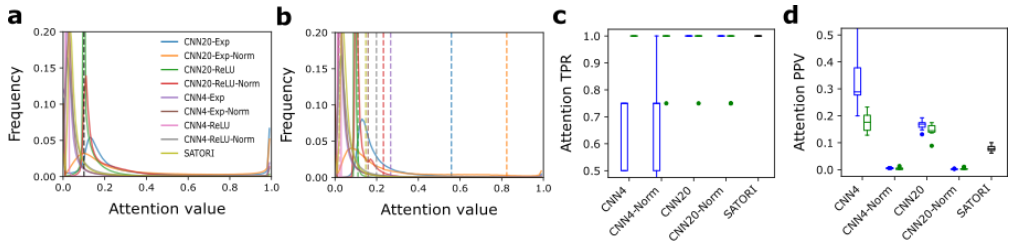

Figure 1: Local attention analysis. Distributions of MHA attention matrix values across (**a**) test sequences and (**b**) dinucleotide shuffled sequences (i.e. null distribution). The 0.1 threshold is shown as a dashed black line in (**a**) and the 0.05 significance cutoffs are shown in (**b**) with colors corresponding to the model. (**c**) True positive rate and (**d**) positive predictive value of filter interactions extracted using local attention. Green and blue box-plots represent models with exponential and ReLU first layer activations, respectively.

quite variable from model to model. Moreover, while the distributions largely appear bi-modal, there is a positive tail, which makes it hard to manually select a suitable threshold. One solution to automate this choice is to establish a statistical test, where the significance of attention values can be compared to a null distribution. We define a null distribution by shuffling sequences to break coherent motifs while maintaining the same (di-)nucleotide frequencies. Surprisingly, the attention distributions for shuffled sequences are only slightly shifted from the distributions for actual sequences. Thus, if we were to establish a threshold based on an empirical p-value cutoff, we may unintentionally miss many motif interactions. Together, this highlights the difficulty of defining an optimal threshold to identify attended position pairs. Even though it remains a challenge to define an "optimal" threshold, it may turn out that identifying motif-motif interactions is robust to the choice of threshold.

Identifiability of active filters. Assuming that the attention maps can identify interacting positions, the next challenge is identifying which filters within the feature maps at those attended positions are interacting. Previously, Satori asserted that all combinations of "active" filters at attended positions are interactions; a filter is deemed to be active if its activity is greater than 50% of the filter's maximum possible activation. This definition of filter activity follows from the standard approach to identify which (sub-)sequences to include in an activation-based alignment for filter visualization [15, 16].

To determine the efficacy of the full pipeline to uncover filter-filter interactions from local attention maps, we conducted a series of experiments assuming significant attention is above a threshold of 0.1 and filter-filter interactions pairs identified through filter 'activity', i.e. Satori's method. The true-positive rate (TPR), which measures the proportion of ground truth interactions that were identified, was near perfect in most models, indicating each ground truth interaction was extracted at least once. However, the average positive predictive value (PPV), which measures the proportion of correct interactions identified, was quite low across all models (Fig. 1). Upon further inspection, we found that one major source of false positives can be explained by seemingly "noisy" filters that have a statistically significant match to a motif in a database of known motifs using Tomtom [17], a motif comparison search tool, but not to any ground truth motif; we call this a random hit. This issue with Tomtom has been documented previously [9, 12]. Filters that don't learn ground truth motifs, but have significant Tomtom matches, are frequently "active", thus causing false positives when position interactions are converted to filter interactions. Deciphering random hits by Tomtom is not straightforward in practice due to a lack of ground truth with *in vivo* data.

To test how sensitive these results are to the choice of attention threshold, we systematically varied the threshold from 0.0 to 0.9 (see Appendix B). We observed that while PPV may increase with a higher threshold, it is accompanied by a decrease in TPR; as the specificity increased, false positives decreased, but ground truth interactions were also lost. Moreover, a variable threshold based on statistical significance from an empirical null distribution also proved ineffective (see Appendix B). Together, this suggests that the core problem with interpreting local attention maps is not the attention threshold, but rather the conversion from position interactions to filter interactions.

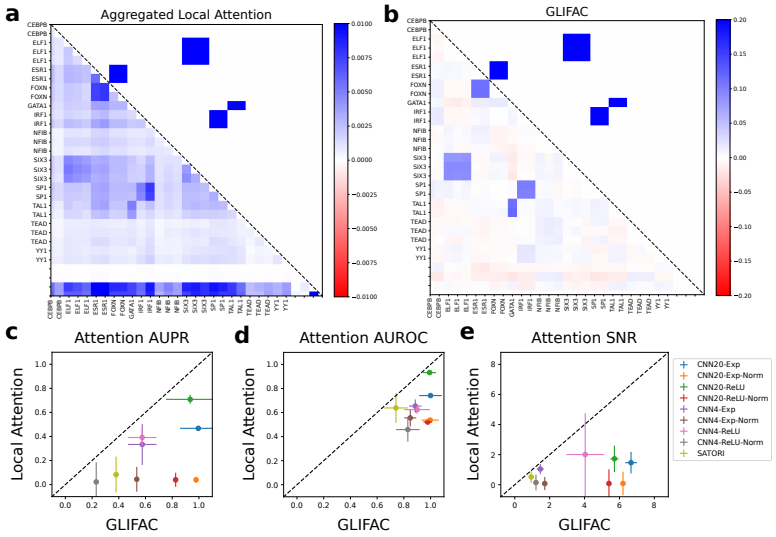

Figure 2: Comparing GLIFAC to aggregated local attention. Example interaction maps generated by CNN40-ReLU using (**a**) aggregated local attention (**b**) and GLIFAC; the upper triangle of each interaction map represents ground truth interactions. Comparison of attention (**c**) AUPR, (**d**) AUROC, and (**e**) SNR in aggregated local attention and GLIFAC generated interaction maps.

4 Global attention yields robust interaction maps

To resolve the limitations of local attention maps, we propose Global Interactions via Filter Activity Correlations (GLIFAC). GLIFAC leverages the fact that attended position pairs should share similar activation patterns across sequences for interacting filters while other non-active filters will be more-or-less random. GLIFAC seeks to exploit the correlations in filter activity, and as a result, it does not require an arbitrary definition of filter activity. By design, GLIFAC should be less susceptible to spurious filter activations compared to local attention. For more details of GLIFAC, see Appendix C.

An alternative strategy to increase the accuracy of local attention is to directly aggregate the extracted filter interactions into a global summary (Fig. 2a). A visual comparison of this summary for local attention and GLIFAC demonstrates that local attention-based interactions have now become more identifiable, though there still remains a high background level of false positive interactions. Moreover, correlations are an effective strategy to find motif interactions (Fig. 2b).

To benchmark the efficacy of GLIFAC, we created 2 distributions of the motif interaction scores (counts for aggregated local attention and correlation value for GLIFAC) – a distribution for ground truth motif interactions (positive) and another for all other motif pairs (negative) – and quantified their separation using the AUROC, AUPR, and signal-to-noise ratio (SNR), which is defined as the ratio between the averaged distributions. Strikingly, we find that GLIFAC consistently leads to a better characterization of motif interactions, with less false positives compared to local attention (Fig. 2c-e). Interestingly, a global view of local attention-based interactions can capture motif interactions better than the noisy interactions proposed by individual sequences – though SNR remains low as a result of the high rate of false positive interactions from local attention. To test the robustness of GLIFAC on the choice of threshold used to identify significant attention values, we performed a sensitivity analysis in Appendix D. We found that indeed GLIFAC provides a global view of filter interactions that is statistically robust. We also explored different architectural choices that improve GLIFAC in Appendix E.

5 Conclusion

By exploring methods to uncover motif interactions from convolution-attention networks, we found that local attention maps are noisy, leading to many false positive interactions. Thus, local attention

with current model designs remain unreliable for use with *in vivo* data. The efficacy and robustness of identifying motif interactions can be improved by aggregating information across local attention maps and further gains can be achieved through statistical relationships like correlation as was demonstrated with GLIFAC. Indeed, other global interpretability methods have found success in genomics, such as clustering attribution maps [18] and via *in silico* experiments [19, 5]; here we demonstrate its utility with attention maps. While global interactions are very informative, it comes with a trade-off of losing information about specific motif interactions within a given sequence, which may be important to dissect mechanisms of cis-regulation as it is often context dependent. Nevertheless, this work provides a significant advance towards highly expressive, interpretable models for regulatory genomics.

References

[1] Peter K Koo and Matt Ploenzke. Deep learning for inferring transcription factor binding sites. *Current Opinion in Systems Biology*, 19, 2020.

[2] DR Kelley, YA Reshef, Bileschi M, D Belanger, CY McLean, and J. Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–50, 2018.

[3] K Jaganathan, S K Panagiotopoulou, J F McRae, S F Darbandi, D Knowles, Y I Li, J A Kosmicki, J Arbelaez, W Cui, G B Schwartz, and E D Chow. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–48, 2019.

[4] Alexandra Maslova, Ricardo N Ramirez, Ke Ma, Hugo Schmutz, Chendi Wang, Curtis Fox, Bernard Ng, Christophe Benoist, Sara Mostafavi, et al. Deep learning of immune cell differentiation. *Proceedings of the National Academy of Sciences*, 117(41):25655–25666, 2020.

[5] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021.

[6] Zeynep Kalender Atak, Ibrahim Ihsan Taskiran, Jonas Demeulemeester, Christopher Flerin, David Mauduit, Liesbeth Minnoye, Gert Hulselmans, Valerie Christiaens, Ghanem-Elias Ghanem, Jasper Wouters, et al. Interpretation of allele-specific chromatin accessibility using cell state–aware deep learning. *Genome Research*, 31(6):1082–1096, 2021.

[7] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[8] Peter K Koo and Matt Ploenzke. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature Machine Intelligence*, 3(3):258–266, 2021.

[9] P K Koo and S R Eddy. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Computational Biology*, 15(12), 2019.

[10] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[12] Fahad Ullah and Asa Ben-Hur. A self-attention model for inferring cooperativity between regulatory features. *Nucleic Acids Research*, 49(13):e77–e77, 2021.

[13] Jiawei Li, Yuqian Pu, Jijun Tang, Quan Zou, and Fei Guo. Deepatt: a hybrid category attention neural network for identifying functional effects of dna sequences. *Briefings in Bioinformatics*, 22(3):bbaa159, 2021.

[14] Rohan Ghotra, Nicholas Keone Lee, Rohit Tripathy, and Peter K Koo. Designing interpretable convolution-based hybrid networks for genomics. *bioRxiv*, 2021.

[15] D. R. Kelley, J Snoek, and J. L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–9, 2016.

[16] B Alipanahi, A Delong, M T Weirauch, and B J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–8, 2015.

[17] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):1–9, 2007.

[18] A Shrikumar, K Tian, A Shcherbina, Z Avsec, A Banerjee, M Sharmin, S. Nair, and A Kundaje. Tf-modisco v0. 4.4. 2-alpha. *arXiv*, page 1811.00416, 2018.

[19] Peter K Koo, Antonio Majdandzic, Matthew Ploenzke, Praveen Anand, and Steffan B Paul. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Computational Biology*, 17(5):e1008925, 2021.

[20] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, page 1412.6980, 2014.

[21] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9, pages 249–256, 2010.

A Experiment details

Synthetic dataset with motif regulatory grammars. 100,000 randomly generated 200-nt DNA sequences were embedded with 1 to 5 non-overlapping binding sites, selected with replacement from a pool of 12 known motifs (CEBPB, ELF1, ESR1, FOXN3, GATA1, IRF1, NFIB, SIX3, SP1, TAL1, TEAD1, YY1). A spacing of at least 1-nt was maintained between all embedded motifs and sequence ends. The labels of some motifs encoded complex regulatory logic consisting of positive and negative interactions. Positively interacting TFs can only bind if both motifs in the interaction pair are present (i.e. an "and" gate); on the other hand, negatively interacting TFs bind normally, but are inhibited if both motifs are present (i.e. an "xor" gate). The motif pairs, ELF1/SIX3 and TAL1/GATA1, exhibited positive interactions, and the motif pairs, ESR1/FOXN3 and IRF1/SP1, exhibited negative interactions. All other TFs (CEBPB, NFIB, TEAD, and YY1) bound to their motifs independently. These regulatory prediction tasks are typically framed as a binary classification with one-hot encoded DNA as inputs and binary labels indicative of a regulatory function as targets; here, each output label signifies the binding of a ground truth motif.

Models. All models follow the base architecture: a convolutional layer with activation and pooling, dropout with a rate 0.1, a multi-head self-attention layer, a dense layer, dropout with a rate of 0.5, followed by an output layer. Batch normalization (BN) was optionally included before convolutional activations, and layer normalization was optionally incorporated before (LN) / after (LN2) the MHA layer. Unless otherwise specified, all models employed 32 filters (19 kernel size), 8 attention heads, 32 attention vector size, and 512 dense layer units. The base models used in GLIFAC and SATORI comparisons were designed as such:

- **CNN4-ReLU**: Conv, ReLU activation, pool size 4, MHA, Dense
- **CNN4-ReLU-Norm**: Conv, BN, ReLU activation, pool size 4, LN, MHA, Dense
- **CNN4-Exp**: Conv, exponential activation, pool size 4, MHA, Dense
- **CNN4-Exp-Norm**: Conv, BN, exponential activation, pool size 4, LN, MHA, Dense
- **CNN20-ReLU**: Conv, ReLU activation, pool size 20, MHA, Dense
- **CNN20-ReLU-Norm**: Conv, BN, ReLU activation, pool size 20, LN, MHA, Dense
- **CNN20-Exp**: Conv, exponential activation, pool size 20, MHA, Dense
- **CNN20-Exp-Norm**: Conv, BN, exponential activation, pool size 20, LN, MHA, Dense
- **SATORI**: Conv, BN, softplus activation, pool size 6, MHA, LN2, Dense

Training. We uniformly trained each model by minimizing the binary cross-entropy loss function with mini-batch stochastic gradient descent (100 sequences) for 100 epochs with Adam updates using default parameters [20]. We decayed the learning rate which started at 0.001, and when the area under the precision recall curve did not improve for 5 epochs, the learning rate was decayed by a factor 0.3. All reported performance metrics are drawn from the test set using the model parameters which yielded the highest performance AUPR on the validation set. Each model was trained 10 times with different random initializations according to Ref. [21].

B Sensitivity analysis of local attention to attention threshold

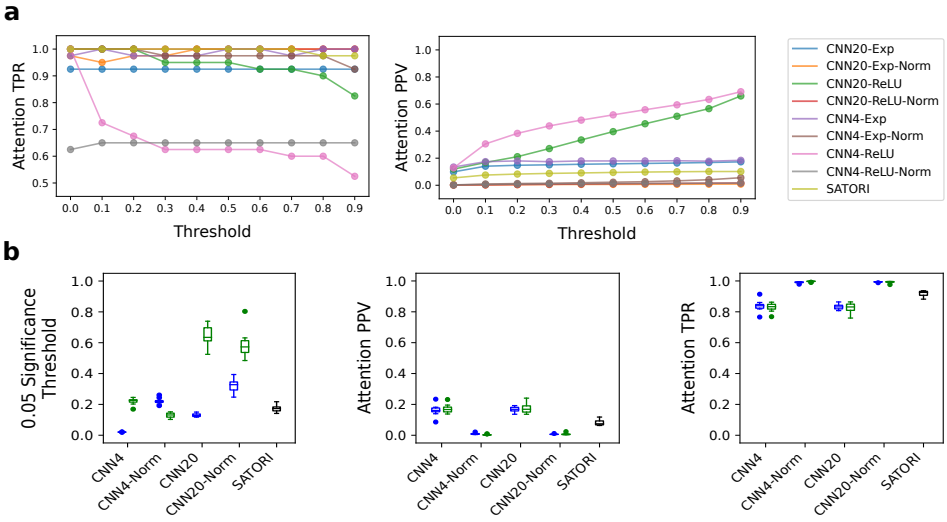

Figure 3: Sensitivity analysis of local attention. (**a**) Plots of true positive rate (TPR, *left*) and positive predictive value (PPV, *right*) of motif interactions determined by local attention, at varying thresholds. (**b**) Box plots of the observed threshold at the 0.05 significance p-value of the empirically determine null distribution (*left*); PPV (*middle*) and TPR (*right*) of local attention at these significance-based thresholds. Green and blue box-plots represent models with exponential and ReLU first layer activations, respectively.

We performed a sensitivity analysis on local attention to determine how choice of attention threshold impacts the quality of interactions extracted. After training, SATORI's local attention technique was applied with thresholds ranging from 0.0 to 0.9, with 0.1 increments. Most models had similar PPVs and TPRs for all thresholds (Fig. 3). We observed a noticeable increase in PPV for CNN20-ReLU and CNN4-ReLU for higher thresholds, however the TPR also dropped, portraying the trade-off between accuracy and coverage, which are arguably of equal importance. We also perform the same analysis as illustrated in Fig. 1c-d, however with an attention distribution determined by using the 0.05 p-value of the empirical null distribution; we observe no noticeable improvements in PPV or TPR.

C GLIFAC

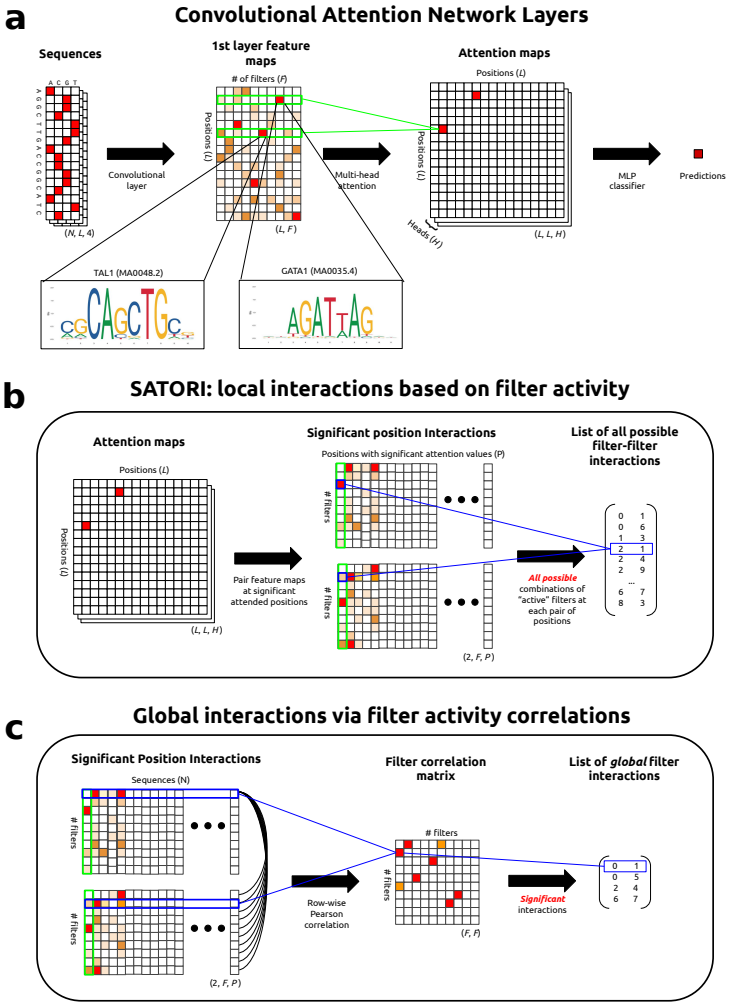

Figure 4: Attention-based interpretability methods. (**a**) Architecture of the convolution-attention network; one-hot encoded sequences are processed through a convolutional layer; the resulting feature maps are pooled and passed through a multi-head self-attention layer, followed by a dense and output layer. (**b**) SATORI's local attention method; position interactions are obtained by computing attention matrices for all test sequences, and isolating attention values greater than 0.1; all combinations of active filters at attended positions are considered statistically significant filter interactions. (**c**) Our proposed Global Interactions via Filter Activity Correlations; after feature map (position) interactions are isolated, the interactivity between all combinations of filters is determined by computing the correlation between each filter's activations at all attended positions.

GLIFAC Technical details. Attention maps heads are aggregated to a single attention map by taking the maximum value across heads. Significant attention values are identified by a threshold of 0.1, above which yield a list of position pairs for each sequence. Given position pairs identified via significant attention, we concatenate the feature maps of each position separately across the entire test set. We also add feature maps from random position pairs, resulting in two sets of concatenated feature maps that consists of 50% positive interactions and 50% no interactions; we found this step necessary to ensure the Pearson correlation was able to align "activation velocities". Using the correlation values between each filter combination, a correlation matrix is constructed where all filters in the first position set are correlated against all filters in the second position set (Fig. 4).

D GLIFAC is robust across all attention thresholds

We perform a sensitivity analysis for GLIFAC on the choice of threshold for significant attention values. We systematically change the attention threshold from 0.1 to 0.9 in 0.1 increments and montiro the performance of recovering motif interactions. In general, the attention AUPR and attention AUROC change very little with threshold, demonstrating GLIFACs robustness to choice of threshold (Fig. 5). Although the SNR appears to change substantially, all SNR values remain above 2.0, a value above which increases contrast between significant and insignificant interactions, but the ability to identify interactions largely remains the same.

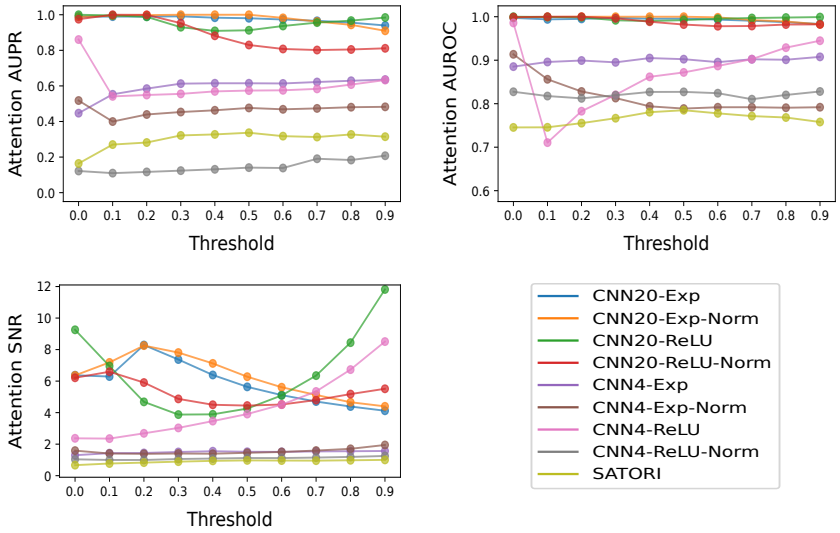

Figure 5: Sensitivity analysis of GLIFAC. Plots of AUROC, AUPR, and SNR of motif interactions determined by GLIFAC, at varying attention thresholds.

E Network architecture influences the interpretability of GLIFAC correlation maps

Using the base models CNN10-ReLU and CNN10-Exp (with layer norm), we systematically varied several key components of the model, while keeping all other settings constant, to determine the impact of design choices on GLIFAC. The following hyperparameters were examined:

- Batch Normalization - enabled, disabled
- Normalization - None, BN (batch normalization after CNN), LN (layer normalization before MHA), LN2 (layer normalization after MHA), and BN+LN
- Pooling Size - 2, 5, 10, 20, 40
- Number of Attention Heads - 1, 2, 4, 8, 16
- Query/Key Vector Depth - 32, 64, 128, 256, 512
- Dense Layer Size - 64, 128, 256, 512, 1024

We uniformly train each model with 10 different random initializations, and monitored the classification AUPR, as well as the attention AUPR, AUROC, and SNR (with an attention threshold of 0.1). We find that in nearly all models, exponential activation-based models outperform ReLU-based models for all measured statistics (Fig. 6). Moreover, their performance is more consistent and generally more robust to changes in architecture, further corroborating it as a more favorable design choice for interpretable models in regulatory genomics [8]. We note, that the SNR is higher for ReLU in the best models, but this is essentially negligible, as any SNR greater than 2.0 indicates that significant interactions are already easily discernable. Pooling size was another critical factor for improving attention interpretability, presumably from learning a slightly better motif representations [9]. Both

batch normalization and layer normalization were detrimental to the efficacy of GLIFAC with the proposed baseline models. We note that they may be critical to training deeper attention models, such as transformers [11]; this was not explored here. Models with few attention heads generally had more interpretable correlation maps. We note that it may be more beneficial for *in vivo* data where variations of regulatory grammars may be dependent on sequence context or across cell types. Vector size had little to no effect on correlation matrix interpretability. Interestingly, we noticed a slight improvement in models with smaller dense layers, but this was followed by a decrease in classification performance.

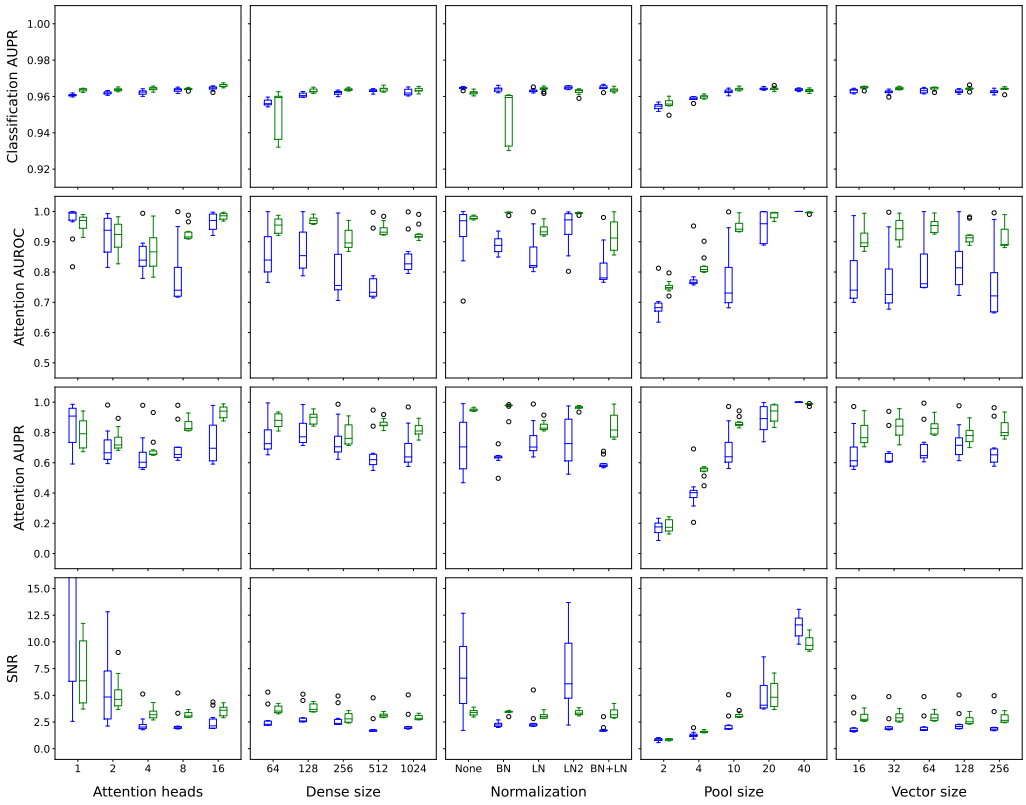

Figure 6: Efficacy of GLIFAC in different network architectures. Classification performance (AUPR) and attention interpretability (AUROC, AUPR, SNR) were measured over ten trials for each design variation. Green and blue box-plots represent models with exponential and ReLU first layer activations, respectively.