Explainable AI–Guided Virtual Experiments Reveal How DNA Sequence Context Shapes Gene Regulation

Sophia Chen
Farragut High School
Knoxville, TN, USA

David M McCandlish
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory

Justin B Kinney
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory

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

Abstract

Deciphering the cis-regulatory code, the rules by which DNA sequence governs gene regulation, is a central challenge in biology with wide-ranging implications for understanding disease mechanisms and engineering DNA for synthetic biology and therapeutic applications. Deep learning models consistently achieve state-of-the-art performance in predicting regulatory activity from DNA sequence, but their black-box nature limits mechanistic insight. Post hoc interpretability tools have identified important sequence motifs corresponding to transcription factor (TF) binding sites, yet the quantitative contribution of surrounding sequence context remains poorly understood. Here, we treat a high-performing sequence-to-function model as a virtual experimental platform, pairing explainable AI with large-scale in silico motif-context swap experiments to quantify the relative contributions of TF motifs and surrounding sequence context to the model's predicted enhancer activity. Using attribution maps, we identify and localize motif instances, then systematically transplant identical motif syntax between different sequence contexts and measure changes in predicted activity to estimate each component's effect. Surprisingly, we find that sequence context plays an outsized role compared to motifs, sometimes accounting for most of the predicted activity. Context effects are most pronounced in housekeeping gene programs, where motifs modestly tune a baseline set by sequence context, whereas developmental programs show stronger motif-driven regulation. Our results motivate a paradigm shift from motif-centric models toward quantitative motif–context frameworks that treat sequence context as an active component of the cis-regulatory code rather than a passive scaffold.

1 Introduction

Gene regulation coordinates when, where, and to what extent genes are expressed, enabling essential processes from cell division and tissue development to environmental response [1, 2]. Much of this control is mediated by transcription factors (TFs), proteins that recognize specific DNA sequence motifs – short patterns typically 6-20 nucleotides long – and influence transcriptional output. The set of sequence-encoded rules by which motifs and other sequence features combine to control activity is referred to as the cis-regulatory code [3]. This code governs a wide range of functional outcomes, including chromatin accessibility [4], chromatin conformation [5], and gene expression [6]. Deciphering it remains a central challenge in genomics, with broad implications for understanding disease, predicting the effects of genetic variation, and designing synthetic regulatory sequences for biotechnology and therapeutic applications.

For decades, efforts to decipher the cis-regulatory code have centered on cataloging TF motifs. High-throughput sequencing assays such as ChIP-seq [7], ATAC-seq [8], and massively parallel reporter assays (MPRAs [9]) have been instrumental in mapping functional sequence elements across the genome. Yet these approaches typically highlight broad regions rather than pinpointing the specific nucleotides that drive activity. Computational motif discovery methods refine these maps by detecting enriched sequence patterns [10, 11], but such methods are agnostic to whether the motifs they detect are functionally active. By contrast, sequence-to-function deep neural networks (DNNs) can directly learn sequence patterns that are predictive of functional activity [12–15]. These models can identify functional motifs, including weak or partial sites, and capture how they act in combination with surrounding sequence context [16–20]. As a result, sequence-to-function DNNs now underpin many tasks in regulatory genomics, from variant effect prediction to the design of synthetic regulatory elements [6, 21, 22].

Despite these advances, genomic DNNs have not yet resolved the cis-regulatory code. A central challenge is interpretability: most DNNs function as black boxes, making it difficult to extract mechanistic rules they have learned. In genomics, post hoc interpretation methods attempt to address this by identifying sequence positions most influential for predictions [23, 24]. Common approaches—including in silico mutagenesis [25], saliency maps [26], integrated gradients [27], and DeepSHAP/DeepLIFT [28, 29]—often recover patterns resembling known motifs (Fig. 1). While reinforcing the motif-centric view of regulation, they offer limited insight into the role of surrounding sequence context.

A complementary perspective comes from virtual experiments. A trained DNN can be viewed as a global function approximator [30], a surrogate for the biological experiment that maps sequence to functional readout. This enables scalable counterfactual experiments performed in silico that would be costly or impractical in the lab [31–33]. Counterfactual predictions highlight a direct causal link between the perturbed element and model predictions, providing insights into the biology through the lens of the DNN. Global importance analysis (GIA), for instance, tests motif sufficiency by embedding motifs into compositionally matched background sequences and measuring predicted activity [33]. This framework has revealed the effects of flanking nucleotides, motif–motif interactions, and distance-dependent binding relationships [17–20]. In this paradigm, a motif (or set of motifs) is deemed sufficient if its insertion into an otherwise neutral background produces strong activity.

This motif sufficiency framework, however, rests on a key assumption: that sequence context does not itself carry functional information [34–36]. However, it is well known that TF binding is influenced by a host of other factors, including local DNA shape [37], low affinity binding sites [20], chromatin accessibility [38], nucleosome positioning [39, 40], and 3D chromatin structure [41]. Moreover, transplanting motifs into new genomic contexts often fails to reproduce endogenous activity [42–44]. Further evidence is observed in attribution maps of state-of-the-art genomic DNNs, which often reveal distributed but low signal outside canonical motifs. These observations suggest that motif syntax and sequence context jointly shape regulatory output, but their relative contributions remain incompletely quantified.

Here, we treat a high-performing DNN trained to map DNA sequence to enhancer activity as a virtual experimental platform to quantify the contribution of sequence context to regulatory activity predictions. This computational approach enables systematic analysis across billions of sequence pairs that would be prohibitively expensive and time-consuming to test experimentally. Using attribution maps, we identify the arrangement of motif instances and treat the remaining sequence as the surrounding DNA context. We then run counterfactual motif-context swaps, moving identical motifs into new contexts or placing new motifs into the same context, and measure changes in predicted activity. We find that context interventions can shift predictions as much as, and sometimes more than, motif interventions. A compositional readout then separates motif syntax from background features, revealing structured context signals that help explain these effects. These results call for a paradigm shift from motif-centric models to quantitative motif–context frameworks that treat background sequence as an active determinant of enhancer activity rather than a passive scaffold.

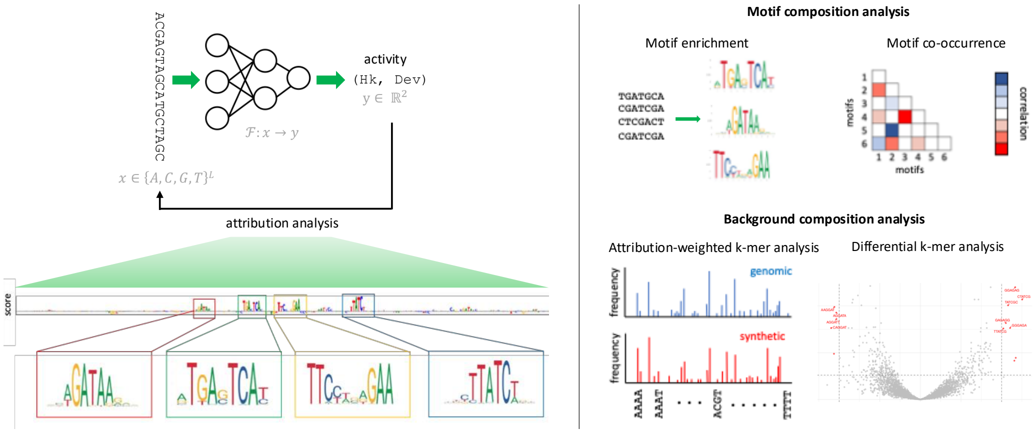

Figure 1: Analysis overview. **Left:** Sequence-to-function deep learning workflow. The DeepSTARR model takes 249-bp DNA sequences as input and outputs predicted enhancer activities under housekeeping (Hk) and developmental (Dev) promoter conditions, given by STARR-seq measurements. Post hoc attribution methods generate base-resolution importance maps, which can be visualized as sequence logos to highlight motif features. **Right:** Compositional analysis framework. Motif composition is assessed through enrichment and co-occurrence of recurring motifs identified from attribution maps. Background composition is characterized using attribution-weighted k-mer spectra and differential k-mer analysis.

2 Quantifying the contribution of motif syntax on model predictions

To systematically assess how sequence context influences regulatory activity, we treated a DeepSTARR model, a convolutional neural network trained on *Drosophila melanogaster* enhancer STARR-seq (Self-Transcribing Active Regulatory Region sequencing) data [18], as a virtual experimental platform to run counterfactual perturbation experiments. DeepSTARR takes 249-nt sequences as input and outputs two quantitative values corresponding to enhancer activity measured under developmental and housekeeping promoter contexts (see Methods in Appendix B for details).

We first asked whether motif syntax alone is sufficient to drive enhancer activity. For 5,000 sequences randomly sampled from predicted activity bins, we identified motif syntax using SEAM [45]. SEAM uncovers the repertoire of cis-regulatory mechanisms accessible within a *local region of sequence space* around a sequence anchor by applying partial random mutagenesis, computing attribution maps across the resulting library, and clustering these maps. A key property we exploit is SEAM's ability to *disentangle* attribution signal into components that are sensitive to mutagenesis (motif-driven "foreground") versus components that are robust to mutagenesis (context-driven "background") (see Methods, Appendix B). Foreground maps retain positional and strength information of motifs, enabling syntax-level analysis (placement, spacing, orientation). We summarized recurring foreground patterns with TF-MoDISco-lite [46, 47], localized instances with FIMO-lite [46], and matched them to JASPAR TF motifs [48] using Tomtom-lite [49], with manual curation to align *Drosophila* motif annotations reported in the original DeepSTARR study. This yielded binary masks that delineate foreground motif positions from background context, enabling controlled motif–context swap interventions.

To test motif sufficiency, we randomly sampled 100 sequences along with corresponding SEAM-separated foregrounds and transplanted each sequence's complete motif syntax, preserving native arrangement, into dinucleotide-shuffled versions of the same sequence. This procedure disrupts endogenous context while retaining overall nucleotide and dinucleotide composition. For each wild-type sequence, ten shuffled backgrounds were generated, and predictions were averaged across replicates—an established GIA-style marginalization step to isolate the effect size of the embedded motif patterns while reducing contributions from spurious motifs introduced by shuffling (Fig. 2A).

Our results revealed that motif syntax embedded in shuffled backgrounds yielded substantially lower predicted activities compared to the same motifs in their endogenous contexts, even when transplanted from high-activity sequences (Fig. 2B). Similar observations were obtained using an

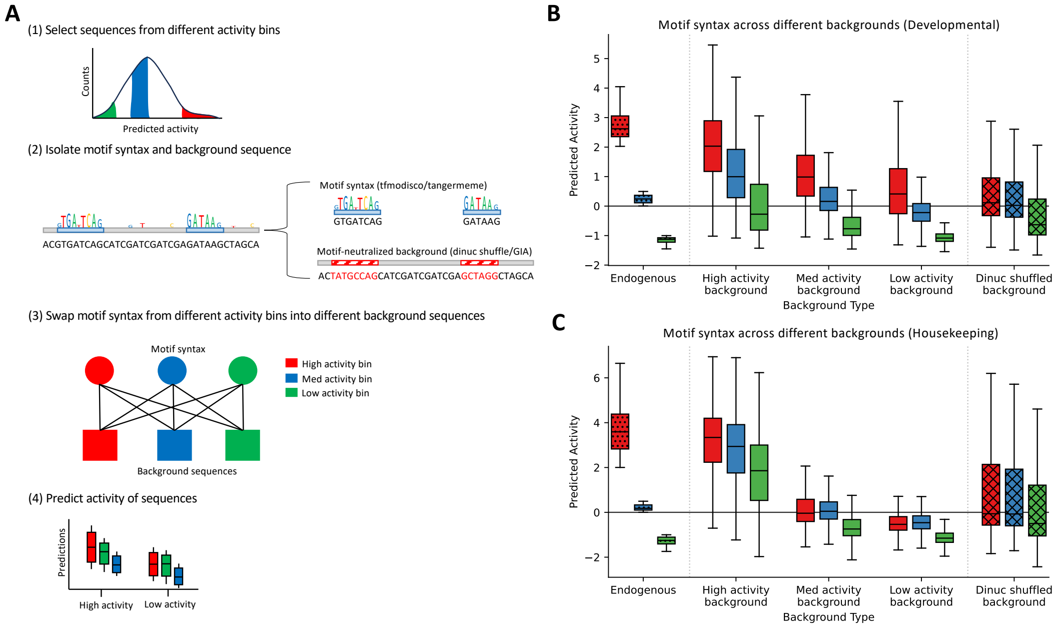

Figure 2: Virtual motif-context swap experiments reveal strong sequence context effects. **(A)** Schematic of the motif-context swap experiment. 100 sequences were randomly sampled from different predicted activity bins (high, medium, low). For each sequence, motif syntax was isolated using attribution-derived binary masks, and background context was isolated by neutralizing motifs by replacing masked positions with dinucleotide-shuffled sequences. Motif syntax from a given sequence was then transplanted into background contexts drawn from different activity bins, and enhancer activities were predicted with DeepSTARR. **(B-C)** Predicted activity of identical motif syntax placed in different sequence contexts. Boxplot show predicted activities of motif syntax in endogenous sequence, in high-, medium-, and low-activity backgrounds, and in a dinucleotide-shuffled control background for developmental (**B**) and housekeeping (**C**) regulatory programs.

alternative approach based on attribution thresholds rather than motif discovery with TF-MoDISco (Fig. 5, Appendix A). Together, these in silico experiments indicate that motif syntax alone is insufficient to reproduce wild-type activity in the model, implying a substantial learned dependence on sequence context.

3 Quantifying the dependency of motif syntax and sequence context

The insufficiency of motif syntax in shuffled backgrounds raises a critical question: how much does sequence context itself contribute to the model's regulatory output? To address this, we performed controlled motif–context swap experiments using sequences from the three predicted activity bins (Fig. 2A). Motif syntax was held constant while backgrounds were systematically exchanged between bins. Backgrounds were generated by neutralizing all motif positions: masked bases were replaced with nucleotides drawn from a dinucleotide-shuffled version of the same sequence, thereby preserving local composition while removing specific motif instances. Ten independent shuffles were generated per background, and predictions were averaged to marginalize residual contributions from unintended motifs.

These counterfactuals revealed striking context dependence. High-activity motif syntax placed into high-activity backgrounds retained strong activity, but when embedded in low-activity backgrounds its activity dropped sharply (Fig. 2B). Conversely, low-activity motif syntax transplanted into high-activity backgrounds gained substantial activity. Thus, background context alone can modulate enhancer output as strongly as, and in some cases more strongly than, motif syntax. The balance between motif and background effects differed by regulatory program: in developmental enhancers, motifs tuned activity around a moderate baseline set by context, whereas in housekeeping enhancers, background context often dominated, suggesting a greater reliance on broad sequence features in housekeeping gene regulation.

Beyond these program-specific trends, we observed asymmetric compatibility between motifs and backgrounds. High-activity motifs yielded higher activities in their endogenous contexts, with reduced activity in other high-activity backgrounds, consistent with co-adaptation of activating elements and their contexts [50]. By contrast, low-activity motif syntax was more interchangeable across backgrounds: embedding it in different low-activity backgrounds produced similarly repressive predictions. Importantly, these effects cannot be explained by simple nucleotide composition. When low-activity motifs were placed into dinucleotide-shuffled sequences, repression was lost and activity increased. This suggests that repression can be achieved more flexibly, whereas activation depends on tighter coordination between motifs and context.

Together, these findings reveal that sequence context contributes substantially more to regulatory activity than typically appreciated, highlighting it as a key component of the cis-regulatory code alongside motif syntax. To dissect the compositional basis of motif syntax and background context contributions, we next examine each component independently.

4 Compositional analysis of foreground attributions

To characterize the sequence features captured in the motif syntax, we analyzed SEAM-foreground attribution maps using TF-MoDISco-lite [46]. SEAM separates model attributions into motif-related (“foreground”) and broader sequence-context (“background”) components, allowing us to focus on motif arrangements without confounding from diffuse contextual signals. Because foreground attributions retain both positional and importance information, this approach captures subtleties in motif placement and strength that are often missed by conventional motif scans. The resulting position weight matrices were matched to known motifs in the JASPAR database, with manual curation to align with *Drosophila* naming conventions and the motifs reported in the original DeepSTARR study.

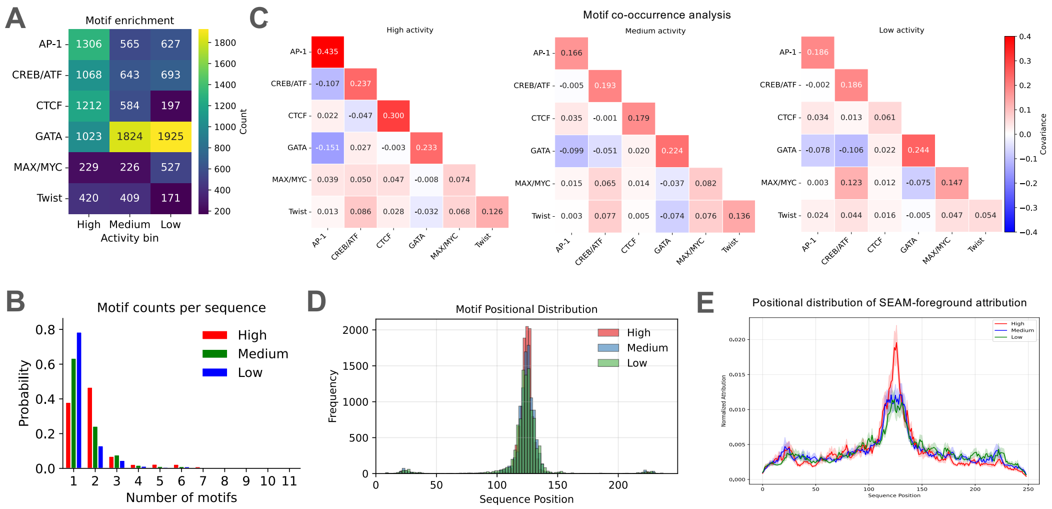

Figure 3: Motif composition analysis of enhancers under the developmental regulatory program. **(A)** Motif enrichment of the top 10,000 motif hits across 5,000 sequences per activity bin. Motifs were discovered using TF-MoDISco-lite applied to SEAM-foreground attribution maps (integrated gradients) and localized with FIMO-lite. Bars indicate total counts of each motif type after removal of redundant calls. **(B)** Histogram of motif counts per sequence for sequences from different activity bins. **(C)** Motif co-occurrence, showing covariance of motif-motif counts within the same sequence across bins. **(D)** Positional distribution of motif hits within the 249-nt input window, stratified by activity bin. **(E)** Position-wise normalized attribution profiles from SEAM-foreground maps. Attribution scores at each position were divided by the sum of absolute attributions per sequence and averaged across sequences within each bin.

Motif enrichment. We quantified motif composition in 5,000 sequences per activity bin (high, medium, low). Motifs such as AP-1, CREB/ATF, and CTCF were enriched in the high-activity bin, whereas GATA showed greater enrichment in the low-activity bin (Fig. 3A). Other motifs, including dMax/dMyc and Twist, exhibited only modest differences across bins. High-activity sequences also showed a higher motif burden (more motif hits per sequence) than medium- or low-activity sequences (Fig. 3B). Together, these results suggest that most motifs are shared across bins, but differences in their relative abundance and overall motif burden contribute to the observed activity differences.

Motif co-occurrence. We examined pairwise motif co-occurrence (via a covariance analysis) to observe whether motif pairs are differential across activity bins (Fig. 3B). Homotypic pairs (AP-1–AP-1, CREB/ATF–CREB/ATF, CTCF–CTCF) were enriched in the high-activity bin, whereas the heterotypic AP-1–CREB/ATF pair was depleted in the high-activity bin. GATA–GATA showed no appreciable difference across bins, while GATA–CREB/ATF was enriched in the low-activity bin. Overall, most pairs were shared across bins; differences reflect shifts in the relative frequency of a small subset of pairs rather than wholesale rewiring. These compositional patterns suggest that certain pairwise motif arrangements are associated with different activity levels, but the co-occurrence analysis alone does not establish whether these differences drive the observed separation between bins.

Motif positional bias. Next, we examined the positional distribution of motifs within the 249-nt input window. Across all activity bins, motifs were strongly concentrated near the sequence center, with only minor enrichment toward the $5'$ and $3'$ ends (Fig. 3C). Foreground attribution signals showed the same central bias, indicating that centrally located motifs contribute disproportionately to predicted enhancer activity (Fig. 3D).

Housekeeping program. In housekeeping enhancers, CTCF showed little enrichment, in contrast to the developmental program (Fig. 6, Appendix A). Instead, DRE/DREF, Ohler1, and Ohler2 were enriched in the high-activity bin, whereas Ohler7 was slightly enriched in the low-activity bin. Pairwise co-occurrence patterns showed no major differences across bins, with only minor variations in specific pairs. Motif burden was also similar across bins. Positional distributions were broadly similar to those in developmental enhancers, with strong central concentration but a broader spread. Taken together, the minimal compositional differences in motif enrichment, co-occurrence, and burden across bins suggest that motif syntax alone provides limited discriminatory power between activity levels in housekeeping programs.

Overall, the same core set of motifs was present across all activity bins, with shifts in relative abundance and co-occurrence. Low-activity sequences contained many of the same motifs as high-activity sequences but at different enrichments or in different pairings. Motifs were generally concentrated near the sequence center, indicating that syntax differences occur within a narrow local window rather than across broad positional ranges. Thus, while subtle compositional shifts in motif usage are evident, these alone are insufficient to fully account for the observed activity separation between bins. This findings suggest that background sequence context plays a substantial role alongside motif syntax in determining predicted enhancer activity.

5 Compositional analysis of background attributions

Our motif–context swap experiments revealed strong background effects despite only subtle differences in motif composition across activity bins, motivating us to identify the background features underlying these effects. We therefore analyzed SEAM-background attribution maps using an attribution-weighted k-mer approach (see Methods in Appendix B). By weighting k-mers according to their attribution scores, we capture the relative importance of sequence features without discarding weaker but biologically meaningful signals. This extends the framework of Majdandžić et al.[51], who computed k-mer spectra from high-attribution positions using a threshold-based approach, by instead applying continuous attribution weighting across all positions. This methodological advance avoids the information loss inherent in hard thresholding and enables more sensitive detection of background compositional differences across activity bins at the 6-mer level. Using this approach, we found that high- and medium-activity backgrounds display more structured, skewed k-mer distributions, whereas low-activity backgrounds show a more uniform, high-entropy spectrum (Fig. 4A).

To pinpoint specific k-mers enriched in different activity bins, we performed differential k-mer analysis on attribution-weighted spectra. For each bin, we generated replicate spectra by sampling

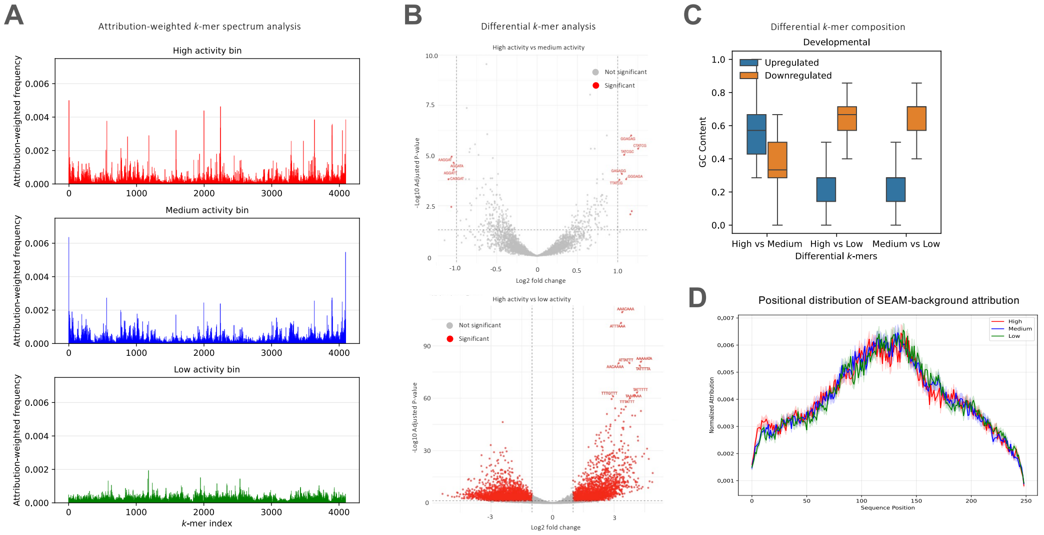

Figure 4: Background composition analysis for enhancers under the developmental regulatory program. **(A)** Attribution-weighted k-mer spectra comparing sequence composition across activity bins. Frequencies were weighted by attribution scores from SEAM-background maps to highlight contextual features contributing to predicted activity. **(B)** Differential k-mer analysis using DESeq2. Volcano plots show significant k-mers for high- vs. medium-activity bins (top) and high- vs. low-activity bins (bottom). Points represent individual k-mers, colored by statistical significance and fold-change direction. **(C)** GC content of differentially enriched k-mers. Boxplots compare GC content between upregulated and downregulated k-mers for each bin comparison. **(D)** Positional distribution of background attributions. Absolute value of attribution scores from SEAM-background maps were normalized by dividing each position's attribution by the sum of absolute attributions across the sequence.

250 sequences at a time and repeating the procedure 20 times, then discretized the weighted counts and analyzed them with DESeq2 [52], a statistical method originally developed for identifying differentially expressed genes in RNA-seq data. This framework is well-suited because attribution-weighted k-mer counts behave like RNA-seq data: discrete, overdispersed, and subject to variability in library size. Comparing high- and medium-activity backgrounds revealed only a modest set of differentially enriched k-mers (Fig. 4B). By contrast, high- versus low-activity comparisons produced a substantially larger set of differentially enriched k-mers, indicating that low-activity backgrounds differ more markedly in their 6-mer composition.

GC-content analysis of differentially enriched 6-mers revealed a consistent polarity: low-activity sequences were enriched for GC-rich k-mers, whereas high-activity sequences were enriched for AT-rich ones, most prominently in high versus low comparisons (Fig. 4C). To examine how these background features are organized within sequences, we assessed the spatial distribution of SEAM-background attributions across the 249-nt input window. As with motifs in SEAM-foreground maps, background attributions were concentrated near the sequence center, though in a broader and more diffuse profile (Fig. 4D). This pattern indicates that background context effects are most pronounced locally, in close proximity to motifs.

We observed similar trends in housekeeping enhancers, with the distinction that high-activity sequences displayed a more pronounced k-mer spectrum (Fig. 7, Appendix A). These results indicate that background context contains non-random, activity-specific sequence patterns that extend beyond simple base composition. The enrichment of AT-rich k-mer in high-activity backgrounds and GC-rich k-mer in low-activity backgrounds suggests that local sequence arrangements, rather than overall nucleotide content, underlie these effects. Given the concentration of background attributions near sequence centers, these patterns are consistent with local sequence features that could influence DNA accessibility, protein-DNA interactions, or other context-dependent regulatory mechanisms.

6 Conclusion

Using explainable AI and large-scale virtual experiments, we dissected the respective contributions of motif syntax and background context to *predicted* enhancer activity. Because controlled perturbations at this scale are infeasible in the lab, we treated DeepSTARR as a virtual experimental platform, enabling billions of paired counterfactual motif–context swaps. By directly quantifying how predictions change when identical motifs are placed into different backgrounds, we uncovered context effects that have been largely overlooked in regulatory genomics. While motif syntax has traditionally been the focus of mechanistic studies, our results challenge this motif-centric view: background sequence can rival or exceed motif effects in shaping predicted activity, elevating context from a passive scaffold to a fundamental component of the cis-regulatory code. Compositional analyses indicate that these effects arise from specific local sequence arrangements—particularly AT-rich k-mers in high-activity backgrounds and GC-rich k-mers in low-activity backgrounds—rather than simple base composition, and may reflect features related to DNA shape, nucleosome positioning, or other chromatin-mediated mechanisms.

Our analyses showed that placing identical motifs into different backgrounds often resulted in large changes in predicted activity, with background alone sometimes tuning activity across the entire dynamic range. This influence could not be explained by base composition alone and was consistent across developmental and housekeeping enhancers, though the relative importance of motifs versus background differed by program. We further found evidence that motifs and their native contexts may be co-adapted, with endogenous backgrounds supporting slightly higher activity than even other high-activity contexts, and that this specificity is particularly important for activating regulation while repressive regulation operates more flexibly across compatible contexts.

These findings have several implications. First, they highlight that motif activity is inherently context-dependent, suggesting that computational and laboratory-based experimental assays that place motifs in randomized sequence contexts may yield more biologically relevant measurements when using activity-matched or biologically realistic backgrounds rather than randomized shuffles. Second, they provide a framework for incorporating context effects into sequence design strategies, potentially improving the rational design of synthetic enhancers and regulatory elements. Third, they suggest that background features may reflect higher-order sequence properties such as DNA shape or chromatin accessibility that merit targeted investigation in future work. Finally, while our experiments were performed with the DeepSTARR model on *Drosophila* STARR-seq data, testing whether these effects generalize to other models, assays, and species will be essential for understanding the broader applicability of these conclusions.

This study is computational and all effect sizes reflect what DeepSTARR learned from *Drosophila* STARR-seq; therefore our conclusions pertain to the model's *predicted* enhancer activity and may not generalize to other assays or species. As with any surrogate modeling approach, DeepSTARR may have learned spurious associations rather than true causal relationships if the training data contained unmeasured confounders, and our counterfactual scenarios may involve sequence distributions the model has never encountered. However, DeepSTARR's predictions and attribution-derived features have been experimentally validated in prior work, supporting the use of attribution-guided counterfactuals as a meaningful lens on regulatory mechanisms. Additionally, if important regulatory patterns are not captured in attribution maps, this could artificially inflate the apparent importance of sequence context. We mitigated this concern by using SEAM, which calculates attribution maps across a local region of sequence space rather than from single sequences, ensuring more robust and complete pattern detection.

Together, this work demonstrates that explainable AI, coupled with counterfactual virtual experiments, can quantify how sequence context modulates regulatory predictions. While the field has long recognized that features like GC content influence regulatory activity, these have typically been treated as nuisance parameters to be controlled for rather than as integral regulatory elements, with motifs remaining the primary focus of mechanistic investigation. Our results necessitate a paradigm shift from motif-centric models to integrated motif–context models that regard sequence context (surrounding each motif) as an essential component of the cis-regulatory code rather than a passive scaffold. By highlighting the functional importance of sequence context in addition to motifs, our findings open new directions for both dissecting the principles of gene regulation and designing synthetic regulatory sequences with greater precision.

References

[1] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9):613–626, 2012.

[2] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, 2014.

[3] Seungsoo Kim and Joanna Wysocka. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular cell*, 83(3):373–392, 2023.

[4] Anusri Pampari, Anna Shcherbina, Evgeny Z Kvon, Michael Kosicki, Surag Nair, Soumya Kundu, Arwa S Kathiria, Viviana I Risca, Kristiina Kuningas, Kaur Alasoo, et al. Chrombpnet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *BioRxiv*, pages 2024–12, 2025.

[5] Jian Zhou. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nature genetics*, 54(5):725–734, 2022.

[6] Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, et al. Alphagenome: advancing regulatory variant effect prediction with a unified dna sequence model. *bioRxiv*, pages 2025–06, 2025.

[7] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[8] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, 2013.

[9] Justin B Kinney and David M McCandlish. Massively parallel assays and quantitative sequence–function relationships. *Annual review of genomics and human genetics*, 20:99–127, 2019.

[10] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[11] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[12] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[13] Kathleen M Chen, Aaron K Wong, Olga G Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, 54(7):940–949, 2022.

[14] Kishore Jaganathan, Nicole Ersaro, Gherman Novakovsky, Yuchuan Wang, Terena James, Jeremy Schwartzentruber, Petko Fiziev, Irfahan Kassam, Fan Cao, Johann Hawe, et al. Predicting expression-altering promoter mutations with deep learning. *Science*, page eads7373, 2025.

[15] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, 57(4):949–961, 2025.

[16] 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.

[17] 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.

[18] Bernardo P de Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers. *Nature genetics*, 54(5):613–624, 2022.

[19] Shushan Toneyan, Ziqi Tang, and Peter K Koo. Evaluating deep learning for predicting epigenomic profiles. *Nature machine intelligence*, 4(12):1088–1100, 2022.

[20] Connor A Horton, Amr M Alexandari, Michael GB Hayes, Emil Marklund, Julia M Schaepe, Arjun K Aditham, Nilay Shah, Peter H Suzuki, Avanti Shrikumar, Ariel Afek, et al. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science*, 381(6664):eadd1250, 2023.

[21] Seppe De Winter, Vasileios Konstantakos, and Stein Aerts. Modelling and design of transcriptional enhancers. *Nature Reviews Bioengineering*, pages 1–16, 2025.

[22] Ksenia Sokolova, Kathleen M Chen, Yun Hao, Jian Zhou, and Olga G Troyanskaya. Deep learning sequence models for transcriptional regulation. *Annual Review of Genomics and Human Genetics*, 25(1):105–122, 2024.

[23] Peter K Koo and Matt Ploenzke. Deep learning for inferring transcription factor binding sites. *Current opinion in systems biology*, 19:16–23, 2020.

[24] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

[25] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[26] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[28] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[29] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[30] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[31] Shushan Toneyan and Peter K Koo. Interpreting cis-regulatory interactions from large-scale deep neural networks for genomics. *bioRxiv*, 2023.

[32] Evan E Seitz, David M McCandlish, Justin B Kinney, and Peter K Koo. Interpreting cis-regulatory mechanisms from genomic deep neural networks using surrogate models. *bioRxiv*, 2023.

[33] 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.

[34] Biswajyoti Sahu, Tuomo Hartonen, Päivi Pihlajamaa, Bei Wei, Kashyap Dave, Fangjie Zhu, Eevi Kaasinen, Katja Lidschreiber, Michael Lidschreiber, Carsten O Daub, et al. Sequence determinants of human gene regulatory elements. *Nature genetics*, 54(3):283–294, 2022.

[35] Gabriella E Martyn, Michael T Montgomery, Hank Jones, Katherine Guo, Benjamin R Doughty, Johannes Linder, Deepa Bisht, Fan Xia, Xiangmeng S Cai, Ziwei Chen, et al. Rewriting regulatory dna to dissect and reprogram gene expression. *Cell*, 188(12):3349–3366, 2025.

[36] Carl G de Boer and Jussi Taipale. Hold out the genome: a roadmap to solving the cis-regulatory code. *Nature*, 625(7993):41–50, 2024.

[37] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

[38] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[39] Cizhong Jiang and B Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, 2009.

[40] Leonid A Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences*, 107(52):22534–22539, 2010.

[41] Seungsoo Kim and Jay Shendure. Mechanisms of interplay between transcription factors and the 3d genome. *Molecular cell*, 76(2):306–319, 2019.

[42] Lisa A Johnson, Ying Zhao, Krista Golden, and Scott Barolo. Reverse-engineering a transcriptional enhancer: a case study in drosophila. *Tissue Engineering Part A*, 14(9):1549–1559, 2008.

[43] Muhammad A Zabidi, Cosmas D Arnold, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank, and Alexander Stark. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540):556–559, 2015.

[44] Franziska Reiter, Bernardo P de Almeida, and Alexander Stark. Enhancers display constrained sequence flexibility and context-specific modulation of motif function. *Genome Research*, 33(3):346–358, 2023.

[45] Evan Seitz, David M McCandlish, Justin Kinney, and Peter K Koo. Decoding the mechanistic impact of genetic variation on regulatory sequences with deep learning. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*.

[46] Jacob Schreiber. tangermeme: A toolkit for understanding cis-regulatory logic using deep learning models. *bioRxiv*, pages 2025–08, 2025.

[47] Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416*, 2018.

[48] Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, et al. Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182, 2024.

[49] Jacob Schreiber. Tomtom-lite: Accelerating tomtom enables large-scale and real-time motif similarity scoring. *bioRxiv*, pages 2025–05, 2025.

[50] Michael Z Ludwig, Casey Bergman, Nipam H Patel, and Martin Kreitman. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769):564–567, 2000.

[51] Antonio Majdandzic, Chandana Rajesh, Ziqi Tang, Shushan Toneyan, Ethan L Labelson, Rohit K Tripathy, and Peter K Koo. Selecting deep neural networks that yield consistent attribution-based interpretations for genomics. In *Machine Learning in Computational Biology*, pages 131–149. PMLR, 2022.

[52] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[53] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

A Additional Figures

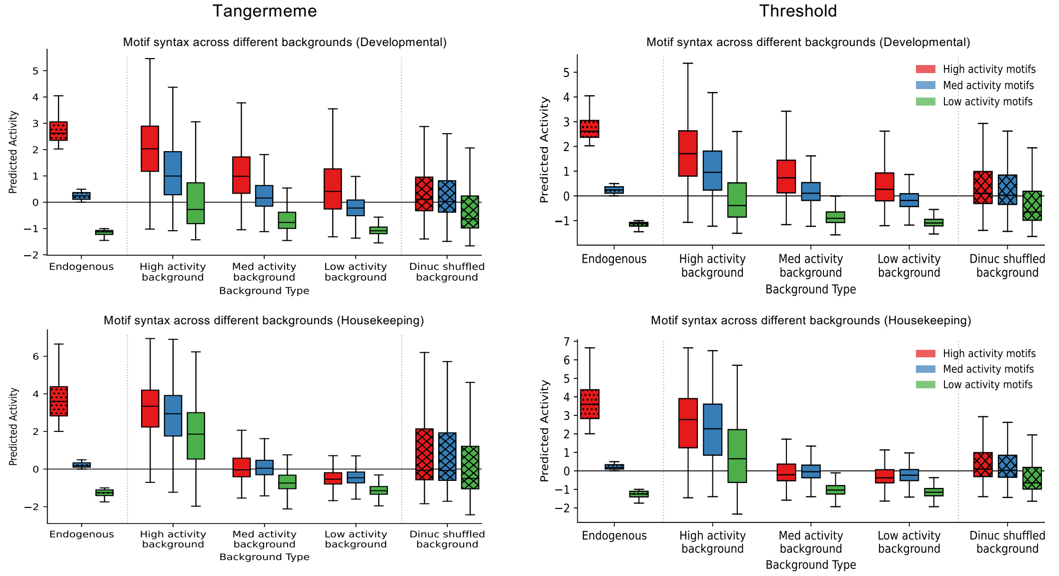

Figure 5: Comparison of motif syntax selection methods. Panel layout: *left* = Tangermeme; *right* = attribution-thresholding (90th percentile of the per-sequence maximum attribution). *Top row* = developmental program; *bottom row* = housekeeping program. Each panel shows box plots of predicted activity for identical motif syntax placed into five sequence contexts: endogenous, high-activity backgrounds, medium-activity backgrounds, low-activity backgrounds, and a dinucleotide-shuffled endogenous control. Results from the two methods are highly similar; conclusions about motif syntax are unchanged whether motifs are selected by Tangermeme or by attribution-thresholding.

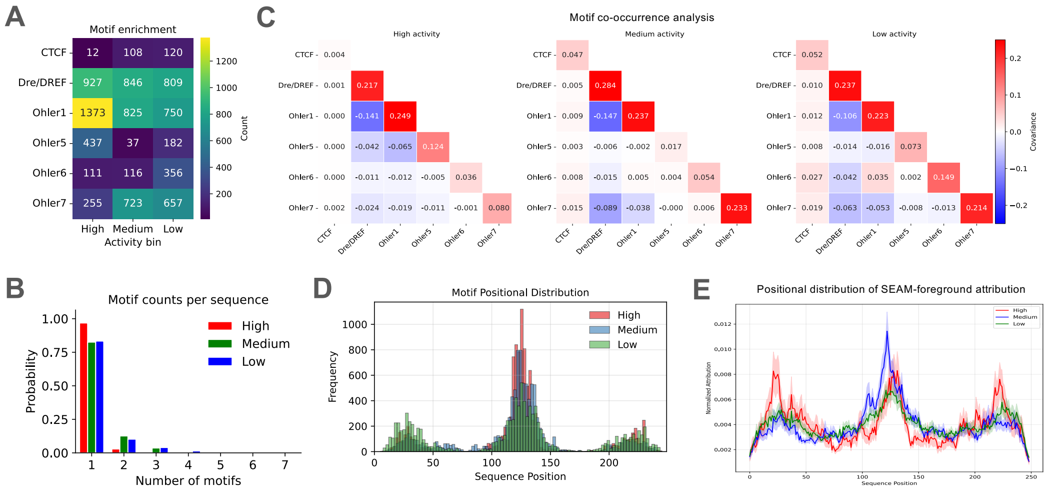

Figure 6: Motif composition analysis of enhancers under the housekeeping regulatory program. **(A)** Motif enrichment of the top 10,000 motif hits across 5,000 sequences per activity bin. Motifs were discovered using TF-MoDISco-lite applied to SEAM-foreground attribution maps (integrated gradients) and localized with FIMO-lite. Bars indicate total counts of each motif type after removal of redundant calls. **(B)** Histogram of motif counts per sequence for sequences from different activity bins. **(C)** Motif co-occurrence, showing covariance of motif-motif counts within the same sequence across bins. **(D)** Positional distribution of motif hits within the 249-nt input window, stratified by activity bin. **(E)** Position-wise normalized attribution profiles from SEAM-foreground maps. Attribution scores at each position were divided by the sum of absolute attributions per sequence and averaged across sequences within each bin.

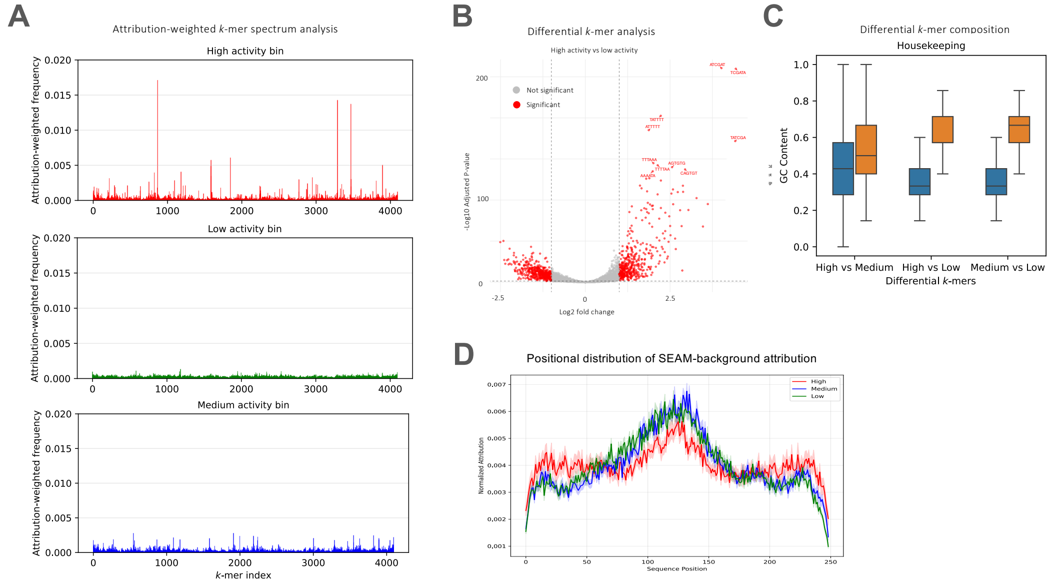

Figure 7: Background composition analysis for enhancers under housekeeping gene program. **(A)** Attribution-weighted k-mer spectra comparing sequence composition across activity bins. k-mer frequencies were weighted by attribution scores from SEAM-background maps to identify context features contributing to predicted activity. **(B)** Volcano plot for differential k-mer analysis identifies statistically significant k-mers from high-activity versus low-activity bins via DESeq2. Points represent individual k-mers colored by significance and fold-change direction. **(C)** GC content distribution of differentially enriched k-mers. Boxes compare GC content between upregulated and downregulated k-mers for each activity bin comparison. **(D)** Position-wise normalized attribution scores from SEAM-background maps across activity bins. Scores were normalized by dividing each position's attribution by the sum of absolute attributions across all positions in each sequence.

B Methods

B.1 Models and data

We employed the DeepSTARR model [18], a convolutional neural network trained to predict quantitative enhancer activity from DNA sequence. DeepSTARR was originally trained on *Drosophila melanogaster* UMI-STARR-seq data comprising $\sim$500,000 synthetic candidate enhancers of length 249 bp, inserted into a reporter construct and assayed in both embryonic and S2 cell contexts. Each sequence was assigned two quantitative labels: developmental activity (measured in embryos) and housekeeping activity (measured in S2 cells). Labels represent normalized enhancer activity as $\log_2$ fold-change relative to plasmid input.

The network architecture follows a multi-task convolutional design: four convolutional layers with max-pooling extract motif-level and higher-order sequence features, followed by two fully connected layers that integrate these features into two regression outputs corresponding to the developmental and housekeeping activities. Training was performed in the original study using mean squared error loss and held-out test sets to ensure generalization across sequences.

In this study, we used the publicly released pretrained DeepSTARR model weights without additional fine-tuning [18]. All input sequences were one-hot encoded over 249 nucleotides, and model outputs for the developmental and housekeeping tasks were used directly as quantitative activity predictions. Where classification was required, sequences were stratified into activity bins based on these model predictions.

B.2 Sequence selection based on predicted activity

Predicted enhancer activities were obtained directly from the pretrained DeepSTARR model outputs. Based on these quantitative predictions, sequences were stratified into three bins: high activity (predicted $\log_2$ fold-change > 2.0), medium/neutral activity (0.0–0.5), and low/negative activity ($\log_2$ fold-change $<$ -1.0). These thresholds were applied separately to developmental and housekeeping predictions, and all downstream analyses used the resulting binned sets. 5,000 sequences were randomly selected from each bin.

B.3 Attribution analysis

We computed nucleotide-level attribution maps using Integrated Gradients (IntGrad) [27]. For a given sequence x, IntGrad estimates the contribution of each nucleotide by integrating the gradients of the model output with respect to the input along an interpolation path from a baseline input x' to x. We used a dinucleotide-shuffled version of each sequence as its baseline, thereby preserving local composition while disrupting regulatory motifs. Linear interpolation between the baseline and the original sequence was performed in 20 evenly spaced steps, and gradients were accumulated across this path to yield the final attribution score for each nucleotide. Attribution maps were computed separately for developmental and housekeeping outputs of DeepSTARR.

B.4 SEAM analysis

SEAM (Systematic Explanation of Attribution-based Mechanisms) combines partial random mutagenesis with attribution analysis to probe the repertoire of cis-regulatory mechanisms within a local neighborhood of sequence space [45]. A central feature is the separation of motif-related "foreground" signal—features that are sensitive to partial random mutagenesis—from broader sequence-context "background" signal that remains robust under the same perturbations. Foreground attribution maps thus capture motif syntax (placement, orientation, strength), whereas background maps highlight contextual signals based on local sequence properties. For each input sequence, SEAM generates a local library of mutated variants, computes attribution maps for all variants, and clusters these maps to segregate foreground and background components. This workflow enables scalable dissection of motif syntax and context effects.

In our study, we optimized SEAM to improve computational efficiency while preserving the interpretability of background maps. We systematically evaluated clustering algorithms, cluster numbers, and the number of mutagenized sequences used per library. We found that using 10,000 mutated sequences per wild-type sequence, combined with k-means clustering into 30 clusters, faithfully reproduced the background attribution maps obtained from the original implementation (which used 100,000 sequences with hierarchical clustering). See Appendix C for additional details.

We then applied this optimized SEAM workflow to 5,000 sequences from each activity bin, sampling 4,000 from the training set and 1,000 from the test set. For each sequence, a local library was generated with a 10% mutation rate, attribution scores were computed for all variants, and k-means clustering was performed. Background separation was applied to each cluster to yield both foreground and background attribution maps. The background maps were averaged across all clusters to produce a representative background attribution map per sequence.

B.5 Foreground composition analysis

Motif discovery analysis. We applied TF-MoDISco from the tangermeme package [46] to the pooled set of sequences and foreground attribution maps across all activity bins. To increase coverage of regulatory features, we augmented the dataset by including foreground maps from the top 10 SEAM clusters ranked by average model prediction. Each sequence was paired with 10 distinct foreground attribution maps drawn from its local library, yielding 10,000 augmented sequences per bin and 30,000 total.

TF-MoDISco was run with a window size of 20 and flank size of 5, with all other parameters set to module defaults. Seqlets were initially clustered into 30 groups using k-means, requiring a minimum of 30 seqlets per cluster. Each cluster was iteratively refined by aligning seqlets to core regions, defined as contiguous positions of at least 6 nucleotides with attribution scores above the 20th percentile of local scores. Core regions were aligned across seqlets, and consensus boundaries (median start and end positions) were used to trim seqlets. Refinement was repeated for three iterations.

Cluster quality was assessed by coherence, defined as the average Pearson correlation across all pairwise seqlet comparisons within a cluster, and cluster similarity, defined as the Pearson correlation between consensus matrices. Clusters with coherence below 0.25 and at least 60 seqlets were split, retaining new subclusters if coherence improved. Clusters with similarity greater than 0.75 were merged. From each final cluster, a contribution weight matrix (CWM), summarizing the average attribution patterns across seqlets, and a position weight matrix (PWM), capturing nucleotide frequency profiles, were generated. The final clusters for the developmental and housekeeping programs yielded 44 and 64 motifs, respectively.

Motif identification. The resulting PWMs were queried against the JASPAR vertebrate motif database [48] to assess similarity with known transcription factor binding motifs. Motif matching was performed using Tomtom-lite [49], run with 1,000 nearest targets for each query motif and default parameters for all other settings. In addition, cluster PWMs from TF-MoDISco were manually compared to motifs reported in the original DeepSTARR analysis to ensure consistency with previously identified regulatory features.

Motif localization analysis. Motif CWMs were scanned separately across attribution maps in each bin and their reverse complements using FIMO-lite [49] to locate motif hits. For each motif, a background probability distribution was computed and used to calculate similarity scores between the motif and each scanned window. Hits were defined using a p-value threshold of 0.01, with multiple testing correction applied to report significant sites at a q-value threshold of 0.05. For analyses of motif enrichment and positioning, the top 10,000 hits by p-value were used as a representative foreground for each bin.

Merging redundant motifs. TF-MoDISco often returns multiple highly similar PWMs for the same TF family (e.g., $>$10 GATA-like clusters in the developmental program), which leads to double counting during scanning and inflates per-sequence hit counts (some sequences showed $\geq$40 calls). To address this, we collapsed redundant calls post hoc. Within each motif label, any two hits whose intervals overlapped by at least 80% of the shorter interval (on either strand) were merged and treated as a single site, retaining the higher-scoring instance. After collapsing, per-sequence counts fell to a plausible range (maximum 11), and all enrichment and positional analyses used this deduplicated set.

B.6 Motif–Context swap experiment

To quantify the relative contributions of motif syntax and surrounding sequence context, we performed motif–context swap experiments. 100 sequences were sampled from the test set of each bin. For each sequence, we defined a foreground (motif) mask by identifying motif hits (see Foreground composition analysis) and extending each hit by two nucleotides on either side. A complementary background mask was defined as all positions not included in the foreground mask.

Background-only sequences were generated by replacing nucleotides at the foreground mask positions with bases drawn from a dinucleotide-shuffled version of the entire sequence. In this way, motif instances were neutralized while preserving overall nucleotide and dinucleotide composition. Ten independent dinucleotide-shuffled versions were created per sequence, producing ten background-only replicates.

Motif–context swaps were then performed in an all-by-all design. For each source sequence, the complete motif syntax (foreground positions defined by its mask) was transplanted into the background-only sequences of every target sequence across all three activity bins. For each target background, the transplant was repeated across its ten dinucleotide-shuffled replicates, and DeepSTARR predictions were averaged to marginalize over residual contributions from spurious motifs introduced by shuffling.

In total, this yielded 90,000 motif–context swaps (300 sources × 300 targets), each evaluated with 10 shuffled replicates. Predicted activities were computed separately for the developmental and housekeeping outputs of DeepSTARR.

B.7 Background composition analysis

Attribution-weighted k-mer analysis. To identify background sequence features underlying activity differences, we extended the attribution consistency framework of Majdandžić et al. [51]. Their method quantified reliability of attribution maps across models by first selecting highly attributed regions using a local threshold (90th percentile of attributions within each sequence), and then comparing the k-mer spectra of these regions against an uninformative prior using Kullback–Leibler divergence.

In our extension, rather than discarding lower-attribution bases with a threshold, we weighted every possible k-mer by its attribution signal. Specifically, for each k-mer occurrence within a sequence, we summed the absolute attributions of its constituent nucleotides with optional separation of attribution scores into positive and negative contributions. These attribution sums were then aggregated across all instances of that k-mer, and normalized by the total absolute attribution over the sequence, yielding a weighted k-mer spectrum. This approach preserves weaker but biologically meaningful background features that would otherwise be lost (see Appendix C for more details).

We applied this method at the 6-mer level to generate weighted k-mer spectra for each sequence, producing 5,000 spectra per bin. These spectra were then compared across activity bins to assess systematic differences in background composition.

Differential k-mer analysis. To test whether background sequence composition differed systematically across activity bins, we performed a differential k-mer analysis. Because k-mer spectra derived from individual 249 nt sequences were too sparse to support statistical testing, we pooled sequences within each activity bin and applied k-means clustering ($k = 3$). From each cluster, we sampled 250 sequences to generate sufficiently large and homogeneous groups, which yielded 20 replicate pools per activity bin for downstream analysis. UMAP [53] visualizations confirmed that this pooling strategy produced tight clusters while maintaining diversity across replicates (data not shown).

Count matrices were generated for k ranging from 4 to 8, with each matrix containing k-mer counts across the 20 replicate pools per bin. We adapted DESeq2, originally developed for differential gene expression analysis, to test for k-mers enriched between bins. To avoid bias from automatic outlier handling, both replacement of outlying values and default filtering were disabled.

Differentially enriched k-mers were assigned to the bin in which they were most upregulated. To focus on robust signals, we removed k-mers with low overall abundance, filtering out those below the 20th percentile of counts across all sequences. For each set of bin-specific k-mers, we then calculated GC content distributions and compared these across bins.

C SEAM optimization analysis

To apply SEAM across thousands of sequences, we benchmarked configurations that preserve SEAM-background map fidelity while improving computational efficiency. We ran SEAM on 15 representative sequences spanning activity levels and varied three hyperparameters: (i) size of the local mutational library, (ii) clustering method, and (iii) number of clusters. As a high-fidelity reference, we used background attribution maps produced with a 100,000-variant local library and hierarchical clustering, default settings for SEAM. For each setting, we computed the Spearman correlation between its background map and the reference (and recorded runtime) to identify an efficient operating point for large-scale analyses.

C.1 Optimizing SEAM clustering

To choose a clustering strategy that preserves SEAM background-map fidelity while scaling to thousands of sequences, we fixed the local library size at 10,000 variants and compared four methods on 15 representative sequences spanning activity levels. For each sequence, attribution maps from the local library were clustered; cluster-level background maps were then averaged to produce a single background attribution map. Fidelity was quantified as the Spearman correlation to a high-fidelity reference generated with a 100,000-variant library and hierarchical clustering. We report the distribution of correlations across sequences and, for stochastic methods, assessed stability across random initializations.

We evaluated:

- **Hierarchical (Ward's linkage).** Clusters are merged to minimize within-cluster variance; a dendrogram cut yields a preset number of clusters (50 in the reference configuration).
- **Leiden on a k-nearest-neighbor graph.** A 10-NN graph on flattened attribution maps (no self-edges) is partitioned by modularity optimization, producing a data-driven number of clusters.
- **k-means in the original feature space.** k-means++ initialization with 50 clusters on flattened attribution maps; UMAP was used only for visualization.
- **k-means on low-dimensional embeddings.** Attribution maps were embedded with UMAP or t-SNE (default parameters), followed by k-means with 50 clusters.

Across methods, fidelity to the reference was uniformly high (Fig. 8), indicating that SEAM's background separation is robust to clustering choice at this library size. k-means in the original feature space achieved the highest median correlation (0.989) with the tightest spread across sequences and parallelizes cleanly with simple, reproducible hyperparameters. Embedding-based k-means and Leiden were comparable but offered no fidelity or stability advantage. We therefore adopted k-means in the original feature space for subsequent large-scale analyses.

C.2 Optimizing number of clusters

With k-means selected, we next chose the cluster count. Holding the local library size at 10,000 variants and all other settings fixed, we evaluated $k \in \{10, 20, 30, 40\}$ on 15 representative sequences. For each k, cluster-level background maps were averaged to a single background attribution map per sequence and compared to the high-fidelity reference via Spearman correlation.

Fidelity was high across all k values (Fig. 8B). A modest improvement in median correlation and reduced spread was observed around $k = 30$ (average Spearman's $\rho = 0.989$), balancing fidelity and computational cost. We therefore used 30 clusters in subsequent analyses.

C.3 Optimizing number of sequences

The initial SEAM implementation used a local library of 100,000 sequences, which presented a significant challenge for large-scale analysis. To identify a library size that offered sufficient coverage of the regulatory sequence space while maintaining efficiency, we evaluated local libraries with the following numbers of sequences: 100, 500, 1000, 5000, 10,000, 25,000, 50,000, 75,000, and 100,000. For each, both attribution accuracy and computational runtime were measured.

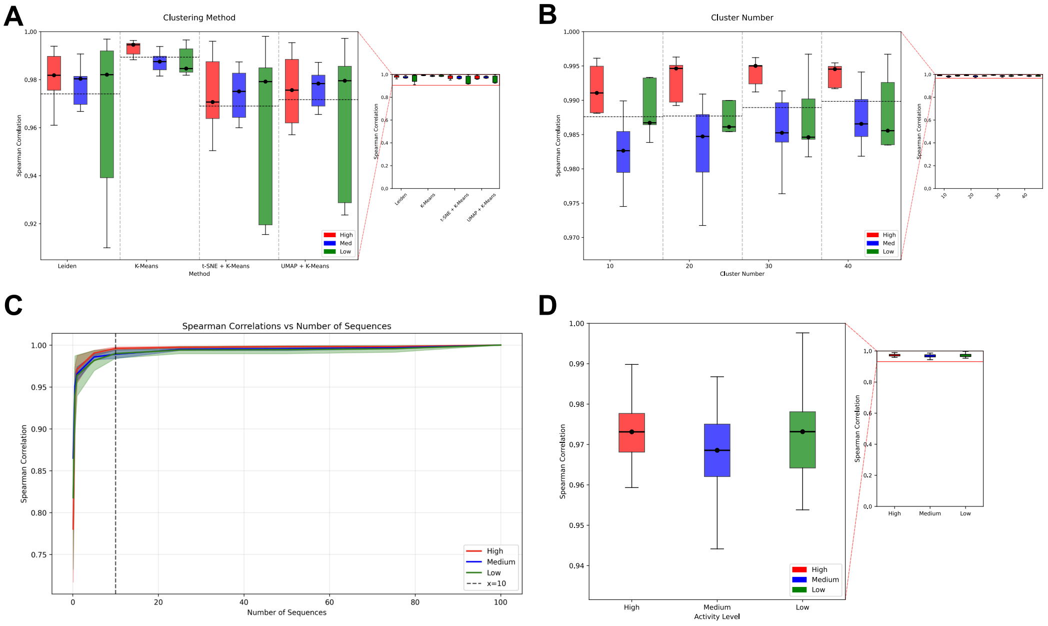

Figure 8: SEAM background attribution map optimization for scalability. Fidelity is the Spearman correlation between SEAM background attribution maps from each configuration and a high-fidelity reference (100,000-variant local library with hierarchical clustering). **(A)** Boxplots across clustering methods on the same 10,000-variant library; methods were comparable, with k-means providing the most stable alignment to the reference. **(B)** Boxplots across numbers of clusters (10–40) for k-means; higher k modestly improves robustness (reduced spread), with overall high fidelity across settings. **(C)** Spearman correlation versus library size; performance plateaus at ∼10,000 variants, well below the original 100,000. **(D)** Final operating point (k-means, 30 clusters, 10,000-variant library) shows high correspondence to the reference across activity bins. Optimization targeted background attribution fidelity; foreground maps were not used for tuning.

Fidelity increased with library size and plateaued at ∼10,000 variants (Fig. 8C). A 10,000-variant library achieved an average Spearman correlation of 0.992 to the reference while reducing average computation time by ∼24-fold. Smaller libraries degraded fidelity with limited additional speed gains. We therefore adopted a 10,000-variant local library in the final SEAM configuration.

Final configuration. The selected operating point (k-means (original feature space) with $k = 30$ and a 10,000 variant local library) achieves uniformly high fidelity across sequences in both regulatory programs; when stratified by activity bin, Spearman correlations to the reference cluster near 0.975 with tight dispersion, confirming suitability for large-scale analyses (Fig. 8D).

D Attribution-weighted k-mer analysis

To better assess background sequence features, we extended the attribution consistency framework of Majdandžić et al.[51]. This framework was originally designed for model selection in sequence-to-function prediction tasks, where the objective is to train DNNs on DNA sequence and evaluate not only predictive accuracy but also the reliability of attribution maps used for interpretation. They benchmarked consistency metrics on two representative tasks: (i) a synthetic binary classification task, where positive sequences were embedded with multiple instances of a small set of "core" motifs and negatives with other background motifs, and (ii) an experimental task, where models predict chromatin accessibility profiles from ATAC-seq data in human cell lines (see [51] for details).

In their approach, attribution consistency was quantified by identifying highly attributed positions (90th percentile within each sequence), aggregating the k-mer spectra from those positions, and comparing this distribution to an uninformative prior using Kullback–Leibler divergence (KLD). Models whose attribution maps produced motif-enriched k-mer distributions were judged to yield more interpretable explanations.

We modified this procedure by replacing the thresholding step with an attribution-weighted k-mer spectrum. Instead of discarding lower-attribution bases, each base contributes to the k-mer score in proportion to its attribution magnitude, thereby retaining weaker but biologically relevant context. Benchmarking on the same tasks as Majdandžić et al., we found that attribution-weighting provided a stronger indicator of model quality. On the synthetic motif-embedding dataset, where ground-truth motif positions are known, the weighted metric correlated more strongly with attribution signal-to-noise ratio (Pearson $r = 0.880$ vs. 0.587, Fig. 9A). On the chromatin accessibility task, where no ground truth exists, the weighted metric stratified models with comparable predictive performance but differing attribution quality, with higher-scoring models yielding clearer motif structure and reduced spurious background (Figs. 9B–C).

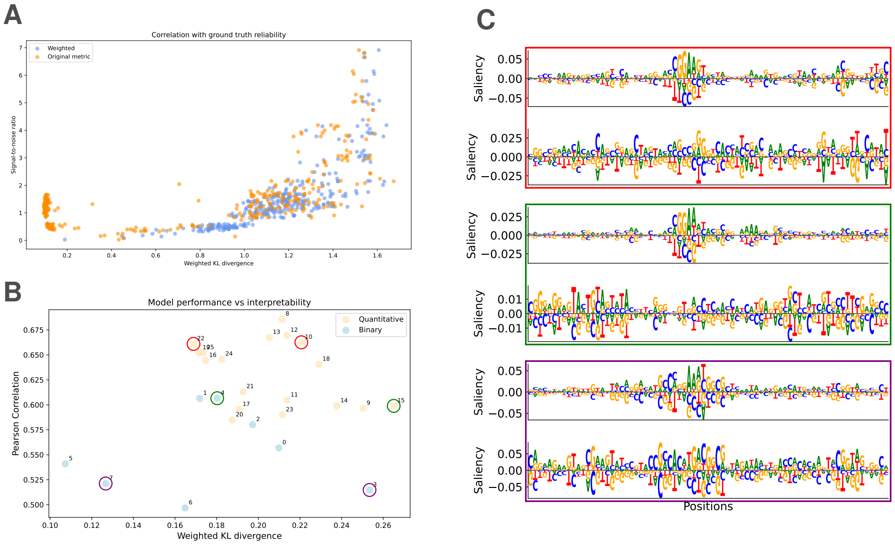

Figure 9: Efficacy of attribution-weighted metric in assessing model reliability. Models were trained on **(A)** a synthetic binary classification task in which sequences contained either core motifs (positives) or background motifs (negatives), and **(B-C)** an experimental task predicting chromatin accessibility from ATAC-seq data. **(A)** On synthetic data, the weighted KL-divergence metric correlates more strongly with ground-truth attribution signal-to-noise ratio (SNR) than the threshold-based method. **(B)** On experimental data, models with similar predictive accuracy can be separated by the attribution-weighted k-mer metric, indicating that accuracy alone does not guarantee interpretability. **(C)** Models identified as reliable by the weighted metric produce attribution maps with clearer motif structure and less spurious signal.

In the main text of this paper, we use attribution-weighting solely to compute k-mer spectra for sequence analysis. Its benchmarking as a model-consistency metric is included here in the appendix to validate the approach and demonstrate continuity with prior work.