Learning In-Silico Maps of Transcription Factor Binding and Cooperativity Interactions

Ruben Solozabal*
MBZUAI
ruben.solozabal@mbzuai.ac.ae

Albert Baichorov*
MBZUAI
albert.baichorov@mbzuai.ac.ae

Tamir Avioz
Weizmann
tamir.avioz@weizmann.ac.il

Le Song
MBZUAI
le.song@mbzuai.ac.ae

Ariel Afek
Weizmann
ariel.afek@weizmann.ac.il

Martin Takáč
MBZUAI
martin.takac@mbzuai.ac.ae

Abstract

Genomic deep learning has rapidly advanced the prediction of transcription factor (TF) binding and other genomic profiles directly from DNA sequence, yet the biological mechanisms captured by these models remain largely unexplored. In this work, we investigate how state-of-the-art genomic models encode the regulatory logic underlying TF binding. We first systematically analyze the model-derived patterns for each TF, revealing that models frequently rely on broad contextual co-associations of motifs to predict a TF to bind. To quantify the dispersity of this association, we introduce the *Jaccard Overlap Score (JOS)*, which distinguishes concentrated recognition of canonical motifs from more distributed binding signatures. Next, we investigate TF–TF cooperativity through *in silico* knockout experiments, revealing pronounced self-dependence of key regulators and cell-type–specific cooperative grammars. Together, our results provide a mechanistic interpretation of genomic deep learning models, demonstrating both their ability to capture biologically meaningful combinatorial regulation and their reliance on contextual sequence features. Our code is publicly available at https://github.com/AlbertBay/EpiBinder.

1 Introduction

In recent years, deep learning has emerged as a powerful paradigm for modeling regulatory genomics, leading to remarkable progress in predicting transcription factor (TF) binding, gene expression, and chromatin profiling directly from DNA sequence [1, 2, 3, 4, 5]. Models such as DeepBind [1], DeepSEA [2], and more recent transformers [4] or state-space models [5] have achieved strong predictive performance across diverse genomic assays. These approaches have established the field of *genomic deep learning* as a new tool for large-scale functional genomics.

Despite this progress, most studies have primarily emphasized benchmark performance metrics, often without deeper investigation into what these models actually *learn* about the underlying biology. In particular, the ability of models to uncover biologically meaningful representations—such

*Both authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

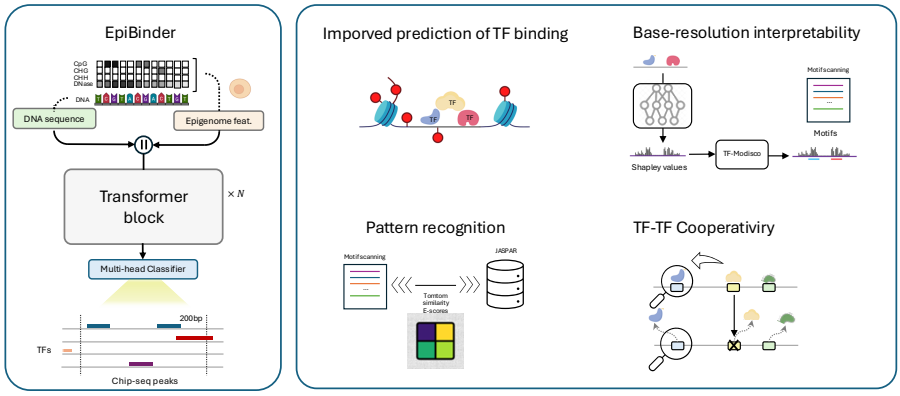

Figure 1: **Overview of the study design.** The schematic on the left illustrates the genomic deep learning framework *EpiBinder*, which integrates DNA sequence with regulatory context to predict transcription factor (TF) binding. On the right, we highlight the two downstream analyses explored in this work: **(i)** model-based interpretation of learned patterns, and **(ii)** *in silico* perturbation experiments to quantify TF–TF interactions.

as interactions between TFs, positional dependencies, or context-specific binding rules—remains broadly underexplored. As a result, while predictive accuracy continues to improve, the question of whether these models truly capture the logic of regulation regions is still largely unanswered.

In this work, we address this gap using *EpiBinder* [6], a multimodal genomic model that integrates DNA sequence with epigenetic context to achieve state-of-the-art results in TF binding prediction (Fig. 1). This model is used as a testbed to perform a mechanistic interpretation of patterns learned. By leveraging controlled perturbations and model-based interpretation, we aim to reveal how modern deep learning architectures encode cooperative TF interactions and evaluate their capacity to generalize beyond sequence-level pattern matching. Our results highlight both the strengths and limitations of current genomic deep learning models, providing a deeper understanding of how they encode transcriptional regulation.

2 Interpreting Genomic Models through Pattern Discovery

A central question motivating our study is whether genomic models truly recognize TFs binding motifs, or whether their predictions rely on other sequence-level signals to identify their presence. To address this, we designed an experiment aimed at systematically characterizing the patterns that such models search for when predicting a TF binding.

Experimental setting. In this experiment, we collect all patterns that the model recognizes along enhancer regulatory regions for each individual TF to bind. In order to do so, we interpret the model predictions to identify *seqlets*, or locally important subsequences, and subsequently cluster to consolidate motifs (see Methods). For each TF under study, we thus created an in-silico library that contains all identified patterns the model exhibits when detecting a particular TF. To assess the biological relevance of the discovered patterns, we performed motif similarity analysis. We computed pairwise E-score similarities within the discovered patterns against the Jaspar [7] database. This comparison allowed us to annotate the model-derived patterns with known TF binding motifs.

Results. In Fig. 2a we depict the resulting annotation map of model-derived patterns against the JASPAR database. In general, we observe that the models exhibit broad associations for most TFs, consistent with the view that genomic deep learning models capture a rich landscape of regulatory determinants. For most of the TFs in the study, the model consistently identifies the coexistence of several context-dependent motifs relevant for the prediction, while for others it highlights more narrow signatures. A notable example is CTCF, which the model appears to recognize almost exclusively through its well-defined binding motif.

In order to better quantify this effect, we introduced the *Jaccard Overlap Score (JOS)*. This score summarizes whether the patterns identified for a TF converge on a few motifs—ideally a single one—or instead represent a more diverse set of binding signatures. For each TF, we compared the

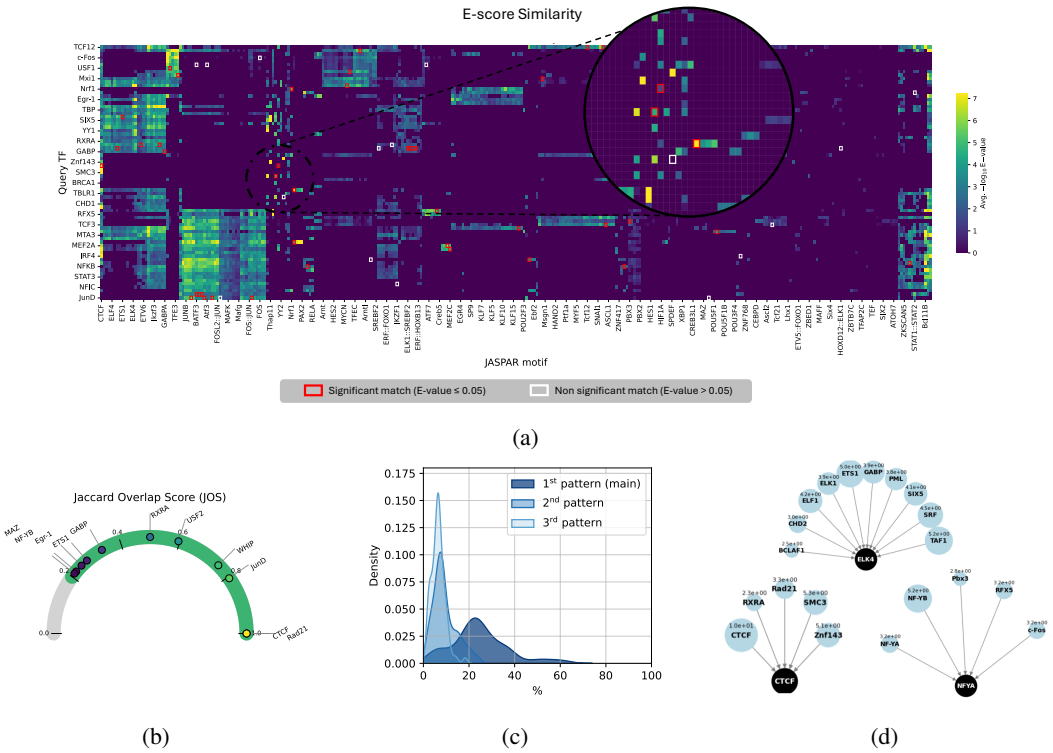

Figure 2: **Model-derived patterns reveal widespread contextual associations. (a)** Motif similarity between model-derived patterns and the JASPAR database reported as the Tomtom similarity score ($-\log_{10}$ of the E-value). Bright colors correspond to stronger similarity (lower E-values), while darker colors indicate low similarity. Red boxes highlight significant matches (E-value ≤ 0.05) where the query TF aligns particularly with its expected JASPAR motif, whereas white boxes mark cases where no strong correspondence was found. **(b)** Jaccard Overlap Score (JOS) summarizing diversity in the association for each TF. The score ranges from 0 (patterns exhibit wide associations with diverse motifs) to 1 (patterns converge on a single canonical motif). **(c)** Percentage of analyzed sequences containing the main (most frequent) pattern identified for each TF. **(d)** Similarity of the dominant pattern for each TF to canonical JASPAR motifs (in parentheses the similarity score).

sets of JASPAR matches for every pair of discovered patterns and computed their Jaccard index. The JOS ranges from 0 (patterns exhibit wide association with diverse motifs) to 1 (patterns converge on key individual motifs). As observed in Fig. 2b, CTCF emerges with one of the highest JOS values, indicating that the model consistently and almost uniquely associates CTCF with its canonical binding motif. Interestingly, RAD21 also displays a similarly high JOS, reflecting the fact that the model recognizes this cohesin component primarily through patterns that overlap with CTCF binding motifs. Together, these two cases illustrate how the JOS can capture concentrated recognition of master regulators, in contrast to other TFs where the patterns are more diverse and context-dependent.

Finally, we refined our analysis to focus on the main pattern (i.e., the most frequent motif) identified for each TF. Strikingly, for many TFs, the dominant pattern identified does not correspond to their canonical JASPAR motif. In Fig. 2c, we report the percentage of sequences in which the main pattern is present, revealing that several TFs are primarily recognized through non-canonical motifs. Instead, for many TFs the dominant patterns often show higher similarity to motifs of other factors such as CTCF, NFYA, or JUND (Fig. 2d). This finding reflects the prominent role of these factors as architectural and regulatory hubs. The enrichment of these shared motifs in place of canonical ones indicates that the models are not limited to capturing the intrinsic determinants of individual TF binding but also leverage broader co-association signals and contextual features of regulatory regions. Such cases highlight the capacity of genomic deep learning models to incorporate partners beyond canonical motif recognition.

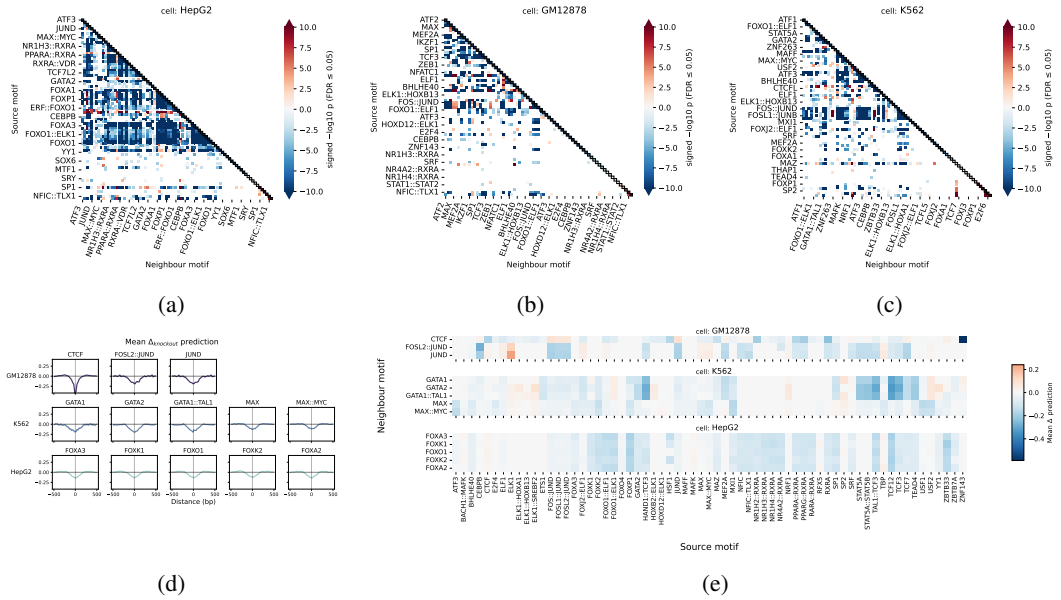

Figure 3: **In silico knockout experiments reveal cell-type–specific TF-TF cooperativity.** Heatmap of TF–TF cooperativity scores for **(a)** GM12878, **(b)** HepG2, and **(c)** K562 cell-lines. Each cell shows the signed $-\log_{10}(p)$ value from a two-sided one-sample t-test of the knockout effect relative to baseline. Off-diagonal entries (i, j) correspond to the effect of knocking out target motif j in the presence of source motif i, while diagonal entries (i, i) represent the self-effect of knocking out motif i alone. Positive values (red) indicate that knockout increases the model prediction, whereas negative values (blue) indicate a decrease; with color intensity reflecting statistical significance.**(d)** Average knockout effect as a function of the distance. **(e)** Absolute knockout effect for the most sensitive TFs per each cell-line.

3 Context-dependent TF-TF Interactions

In our second experiment, we address how these patterns can interact in a combinatorial or cooperative manner. To probe whether our model captures such context-dependent interactions, we next designed a series of in silico perturbation experiments aimed at quantifying the influence that TFs exert on one another.

Experimental setting. To investigate TF-TF cooperativity, we designed in silico knockout experiments. For each locus, we first evaluated the predicted binding probability of the TF of interest under the unperturbed sequence. We then simulated knockouts by masking motif contributions corresponding either to the TF itself or to an adjacent TF ("neighbor"), and re-computed binding probabilities. This procedure allows us to quantify the direct self-dependence of a TF as well as its cooperative or competitive interactions with nearby factors.

Results. We summarized the significance of the perturbations in Fig. 3a. Each heatmap entry shows the signed $-\log_{10}(p)$ value from a one-sample t-test, measuring whether knockout-induced changes in predicted binding differ from the baseline. Diagonal entries capture the self-dependence of each TF, while off-diagonal entries reflect how perturbing one TF influences the binding of another. Strong negative values along the diagonal indicate that removing a TF motif substantially reduces its own predicted binding, whereas significant off-diagonal values reveal cooperative or competitive dependencies. Overall, most TFs displayed a clear reduction in binding when their motif was randomized, although a few showed little or no recognition by the model. The broadest knockout effects were observed in HepG2, where several hepatocyte master regulators exerted strong influence on neighboring TFs, consistent with the cooperative enhancer grammar characteristic of hepatocytes.

When considering absolute knockout effects across cell types, the most influential TFs consistently had **architectural or pioneer-like functions** (see Fig. 3b,3c). In GM12878, CTCF and AP-1 components (FOSL2::JUND, JUND) dominated as loss of CTCF broadly weakened neighboring motifs, reflecting its role in chromatin insulation and loop anchoring. In K562, the most impactful TFs were GATA1, GATA2, MAX, and MAX::MYC. GATA proteins act as lineage-defining erythroid

regulators with pioneering activity that facilitates co-binding. In HepG2, consistently with [8], FOX-family members (FOXA2, FOXA3, FOXK1, FOXK2, FOXO1) showed the strongest effects, consistent with their role as pioneer factors (FOXA) and as components of enhancer or repressor complexes (FOXO, FOXK) that shape cooperative occupancy. Together, these results indicate that the model detects cooperativity effects for TFs that establish the regulatory context.

Conclusions. In this work, we show that deep genomic models capture a rich landscape of regulatory determinants when predicting the regulatory landscape that usually extends beyond canonical motifs. Moreover, these models are also able to capture cooperative interactions, particularly those that shape the regulatory landscape, either by providing architectural anchoring (e.g., CTCF) or by acting as pioneer factors (e.g., FOX, GATA), upon which the binding of other factors depends.

References

[1] 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.

[2] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

[3] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

[4] 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.

[5] Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of machine learning research*, 235:43632, 2024.

[6] Ruben Solozabal, Albert Baichorov, Tamir Avioz, Le Song, Martin Takáč, and Ariel Afek. Epibinder: a multimodal deep learning model at base-resolution to analyze in vivo transcription factor-dna binding. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*.

[7] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1):D91–D94, 2004.

[8] E Christopher Partridge, Surya B Chhetri, Jeremy W Prokop, Ryne C Ramaker, Camden S Jansen, Say-Tar Goh, Mark Mackiewicz, Kimberly M Newberry, Laurel A Brandsmeier, Sarah K Meadows, et al. Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature*, 583(7818):720–728, 2020.

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

[10] 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.

[11] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202–W208, 2009.

[12] Tianshun Gao and Jiang Qian. Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic acids research*, 48(D1):D58–D64, 2020.

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

Supplementary Material

A Jaccard Overlap Score (JOS)

For each transcription factor (TF) t, let P_t denote the set of model-derived patterns and $\mathcal{J}(p)$ the set of significant JASPAR matches (q-value ≤ 0.05) for pattern $p \in P_t$. To assess whether the model learns a diverse repertoire of motifs or collapses onto a few canonical motifs, we define the *Jaccard Overlap Score (JOS)* as

$$\mathrm{JOS}_t \;=\; 1 - \frac{1}{|P_t|(|P_t|-1)} \sum_{\substack{p_i \neq p_j \\ p_i, p_j \in P_t}} \frac{|\mathcal{J}(p_i) \cap \mathcal{J}(p_j)|}{|\mathcal{J}(p_i) \cup \mathcal{J}(p_j)|}.$$

Here, the inner term is the Jaccard index of the JASPAR match sets for patterns p_i and p_j. The JOS ranges from 0 to 1: values close to 0 indicate that patterns map to distinct and non-overlapping motifs (wide and diverse associations), while values close to 1 indicate that the patterns converge on the same motif or a very restricted set of motifs (concentrated recognition).

B Statistical testing for motif cooperativity effects

For each pair of motifs (i, j) we considered the change in model prediction after in silico transcription factor knockout. Specifically, we defined

$$\Delta_{ij} = \begin{cases} knockout_{target} - baseline, & i \neq j, \\ knockout_{source} - baseline, & i = j, \end{cases}$$

where $baseline$ is the unperturbed model prediction, $knockout_{target}$ the prediction after knockout of the target motif j, and $knockout_{source}$ the prediction after knockout of the source motif i.

For each motif pair (i, j) we obtained a collection of values $\{\Delta_{ij}^{(k)}\}_{k=1}^{n}$ across genomic loci. To assess whether the average effect was significantly different from zero, we applied a two-sided one-sample t-test:

$$H_0 : \mu_{ij} = 0 \qquad H_A : \mu_{ij} \neq 0,$$

with test statistic

$$t_{ij} = \frac{\bar{\Delta}_{ij}}{s_{ij}/\sqrt{n}},$$

where $\bar{\Delta}_{ij}$ is the sample mean, s_{ij} the sample standard deviation, and n the number of loci contributing to that cell. The p-value was computed from a Student t distribution with $n - 1$ degrees of freedom.

Edge cases were handled conservatively: if $n < 2$, if the variance $s_{ij}^2 = 0$, or if numerical evaluation failed, the test returned $p = 1.0$ (no evidence against H_0).

To encode both effect direction and significance in the visualization, we reported the *signed* $-\log_{10}$ p-value:

$$score_{ij} = \mathrm{sign}\big(\bar{\Delta}_{ij}\big)\ \big(-\log_{10} p_{ij}\big).$$

Finally, to account for multiple hypothesis testing across all motif pairs, we applied the Benjamini–Hochberg procedure to control the false discovery rate (FDR). Cells with FDR-adjusted q-values greater than 0.05 were masked in the heatmap.

C Methods

C.1 Pattern Collection and Motif Similarity Analysis

For each cell line, we randomly selected a 1000 set of genomic regulatory regions to probe the internal representations of the models. On these sequences, we computed base-resolution importance scores using DeepSHAP [9], which provide per-nucleotide attributions quantifying the contribution of each base to the model's prediction for a given transcription factor (TF). These attribution maps highlight locally important subsequences that the model relies on when predicting TF binding.

To extract interpretable patterns from the noisy attribution profiles, we applied TF-MoDISco [10]. This algorithm first identifies short subsequences of consistently high importance ("seqlets") and then clusters them into non-redundant motif-like patterns. For each TF in each cell line, this procedure yielded a collection of recurrent sequence patterns that represent the key signals used by the model.

Finally, to annotate and assess the biological relevance of the discovered patterns, we used Tomtom [11] to compare them against the JASPAR database of known TF binding motifs [7]. Tomtom computes similarity scores between model-derived patterns and reference motifs, allowing us to link the patterns to canonical TF binding sites or, in some cases, to context motifs that may reflect cooperative or architectural factors.

C.2 Construction of the TF–TF Cooperativity Dataset

To study context-dependent TF interactions, we constructed a dataset integrating sequence and regulatory context. We began with cell-type–specific enhancer annotations from EnhancerAtlas [12], restricting to regions relevant to each of the three studied cell lines. From these enhancers, we used ChIP-seq peak calls provided in [2] to identify candidate binding regions. Because ChIP-seq peaks are broad and do not provide precise motif positions, we refined them using FIMO scans. For "all" enhancers, we searched for motif instances consistent with the ChIP-seq–labeled TF. In cases where FIMO reported motif complexes, we retained the match if at least one component corresponded to the annotated TF.

Using this set of localized motifs, we enumerated all possible TF pairs within a genomic distance of $\leq$ 500 bp. We defined directionality by assigning the first TF as $source$ and the second as $target\ neighbour$. Centering each $source$ motif, we constructed a genomic profile for the size of the model context window, which for EpiBinder is 1000 bp to query prediction on the model.

To quantify cooperative contributions, we generated two knockout variants for every TF–TF pair:

1. $baseline$, the prediction for the unmodified sequence.
2. $knockout_{source}$, created by replacing source motif nucleotides with random bases.
3. $knockout_{neighbour}$, created by replacing neighbouting motif nucleotides with random bases.

Randomization was implemented by sampling nucleotides uniformly from $\{A, C, G, T\}$, thereby eliminating motif-specific signals.

D Tools

- **TF-MoDISco** [10]: applied to cluster high-importance subsequences ("seqlets") identified from attributions scores into motif-like patterns. We use the default false discovery rate (FDR $\leq$ 0.05).
- **MEME Suite / Tomtom** [11]: employed to measure similarity between model-derived motifs and known binding motifs from the JASPAR database [7]. Tomtom computes statistical similarity scores and provides motif annotations. We use it setting a minimum overlap of 5 bp, and a significance threshold $q \leq 0.01$.
- **FIMO** (part of MEME Suite) [13]: used to refine broad ChIP-seq peaks by scanning for precise motif instances, ensuring alignment between experimentally observed binding and sequence-level motif coordinates. We applied a significance threshold of $p \leq 1 \times 10^{-4}$.