DECODING THE MECHANISTIC IMPACT OF GENETIC VARIATION ON REGULATORY SEQUENCES WITH DEEP LEARNING

Evan E. Seitz, David M. McCandlish, Justin B. Kinney & Peter K. Koo
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724, USA
{seitz,jkinney,koo}@cshl.edu

ABSTRACT

Non-coding DNA encodes complex cis-regulatory mechanisms that govern gene expression by orchestrating transcription factor binding within specific sequence contexts. While deep learning has advanced our understanding of these mechanisms, how genetic variation reconfigures them remains an open challenge. Here, we introduce SEAM, an AI-driven tool that systematically investigates how mutations reshape regulatory mechanisms. By mapping sequences into a mechanism space and clustering them based on shared features, SEAM reveals how specific mutations can reprogram regulatory DNA, driving mechanistic and functional diversity. SEAM highlights the remarkable evolvability of human regulatory elements, disentangles transcription factor-specific effects from broader sequence context, and provides a powerful framework for decoding the cis-regulatory code. By enabling systematic, unbiased exploration of reprogrammable mechanisms, SEAM illuminates evolutionary pathways and informs the rational design of synthetic sequences with tailored functions.

1 BACKGROUND

Deciphering how DNA sequences encode cis-regulatory mechanisms is a central challenge in biology. Cis-regulatory elements orchestrate gene expression by integrating multiple signals, including transcription factor binding sites (TFBS) and broader sequence context (Banerji et al., 1981; Spitz & Furlong, 2012; Rickels & Shilatifard, 2018). Features such as nucleotide composition, motif spacing, chromatin state, and sequence context collectively influence TFBS specificity, competition, and cooperativity (Barash et al., 2010; Shlyueva et al., 2014; Nagy & Nagy, 2020). Together, these elements form the cis-regulatory code, governing gene regulation across biological contexts.

Regulatory sequences evolve under selective pressures and genetic drift, with mutations altering TF binding, competitive landscapes, and TF interactions, leading to regulatory rewiring (Crutchfield & van Nimwegen, 2002; Levine, 2010). Understanding how genetic variation reshapes cis-regulatory mechanisms is crucial for linking genotype to phenotype, uncovering disease mechanisms, and designing synthetic sequences. However, existing tools struggle to systematically map the functional consequences of such mutations.

Deep neural networks (DNNs) have enabled accurate predictions of regulatory activity from DNA sequences (Zou et al., 2019; Koumakis, 2020; Avsec et al., 2021a;b). Post hoc explainability methods, such as attribution methods (Koo & Ploenzke, 2020; Novakovsky et al., 2022), assign importance scores to nucleotides, offering insights into motifs, syntax, and broader sequence integration (Zhou & Troyanskaya, 2015; de Almeida et al., 2022; Novakovsky et al., 2022). Experimental studies have validated attribution-based insights, yet existing approaches remain limited to analyzing individual sequences and lack a systematic framework to explore how mutations reconfigure regulatory mechanisms.

Here, we introduce SEAM (**S**ystematic **E**xplanation of **A**ttribution-based **M**echanisms), a computational framework leveraging deep learning to systematically investigate how mutations reshape

cis-regulatory mechanisms. SEAM maps regulatory sequences into a "mechanism space", clustering sequences based on shared regulatory logic through the lens of a DNN. This approach resolves complex attribution patterns, disentangles motif- and context-specific signals, and reveals key mutations driving mechanistic reprogramming. Applied to human and fly regulatory sequences, SEAM uncovers functional diversity and evolutionary pathways shaping regulatory DNA. SEAM provides a powerful platform for dissecting regulatory genomics, evolution, and disease.

2 SEAM: A FRAMEWORK FOR EXPLORING CIS-REGULATORY MECHANISMS

The sequence space of regulatory DNA is vast, with an astronomical number of possible mutations. Exhaustively sampling this space is computationally infeasible. To address this, SEAM employs partial random mutagenesis, introducing small, independent mutations to generate a synthetic library of variants. A trained sequence-to-activity DNN then maps each sequence to a "mechanism space" using an attribution method. SEAM clusters these mechanisms to reveal distinct regulatory strategies, while further sequence analysis within each cluster identifies shared mutations driving mechanistic shifts (Fig. 1a).

SEAM produces three key outputs: (1) the set of mechanisms for each cluster, including the assignment of each sequence to its respective cluster, revealing shared regulatory logic; (2) the predicted activity distribution for each cluster, summarizing functional variation introduced by mutations; and (3) the Mechanism Summary Matrix (MSM), a unique feature of SEAM that highlights sequence variations relative to the wild-type sequence. The MSM provides insights into shared mutations within clusters and identifies positions tolerant to genetic variation, offering a detailed view of how sequence variation influences regulatory mechanisms. Together, these outputs form a comprehensive toolkit for exploring the cis-regulatory rules driving functional and mechanistic diversity within regulatory DNA.

SEAM is a versatile framework that can be customized for different regulatory genomics applications by varying sequence libraries, attribution methods, and clustering strategies. In this study, SEAM primarily focused on "local" sequence libraries, described above, where 1–10% of nucleotides in each sequence were randomly altered. Attribution maps were generated using DeepSHAP (Lundberg & Lee, 2017), and hierarchical clustering (Ward, 1963) was applied to group sequences by shared attribution-based mechanisms (see Methods).

3 SEAM DISSECTS COMPLEX CIS-REGULATORY MECHANISMS

To demonstrate SEAM's ability to resolve regulatory mechanisms, we first applied it to DeepSTARR, a sequence-to-activity DNN trained to predict enhancer activity in Drosophila S2 cells (de Almeida et al., 2022). DeepSTARR has uncovered key cis-regulatory principles, such as TFBS-flanking dependencies and distance-dependent cooperative interactions, through attribution maps and in silico perturbations, with select mechanisms validated experimentally (de Almeida et al., 2022; Reiter et al., 2023). However, many attribution maps remain ambiguous, displaying diffuse signals or dense attribution clusters that obscure TFBS identification. Some enhancers exhibit tightly packed motifs (Fig. 1d), while others show weakly attributed regions that may correspond to low-affinity binding sites.

SEAM clarifies these complex patterns by applying partial random mutagenesis with a 10% local mutation rate, decomposing mechanisms into distinct subsets. Clustered attribution maps, visualized as sequence logos, provide higher-resolution views of functional motifs (Fig. 1e). SEAM's Mechanism Summary Matrix (MSM) further separates motifs across clusters (Fig. 1b), enabling precise segmentation of functional components (gray overlays in Fig. 1e). Covariance analysis of the MSM reveals combinatorial TFBS preferences across mechanisms (Fig. 1c). Notably, weak attribution patterns consistently reappear within specific clusters, supporting their role as biologically meaningful features, such as low-affinity binding sites.

SEAM's robustness was confirmed across multiple workflow variations, including hierarchical clustering parameters (Appendix Fig. 1), clustering algorithms (Appendix Fig. 2), mutation rates, library sizes (Appendix Fig. 3), and attribution methods (Appendix Fig. 4). These results establish SEAM

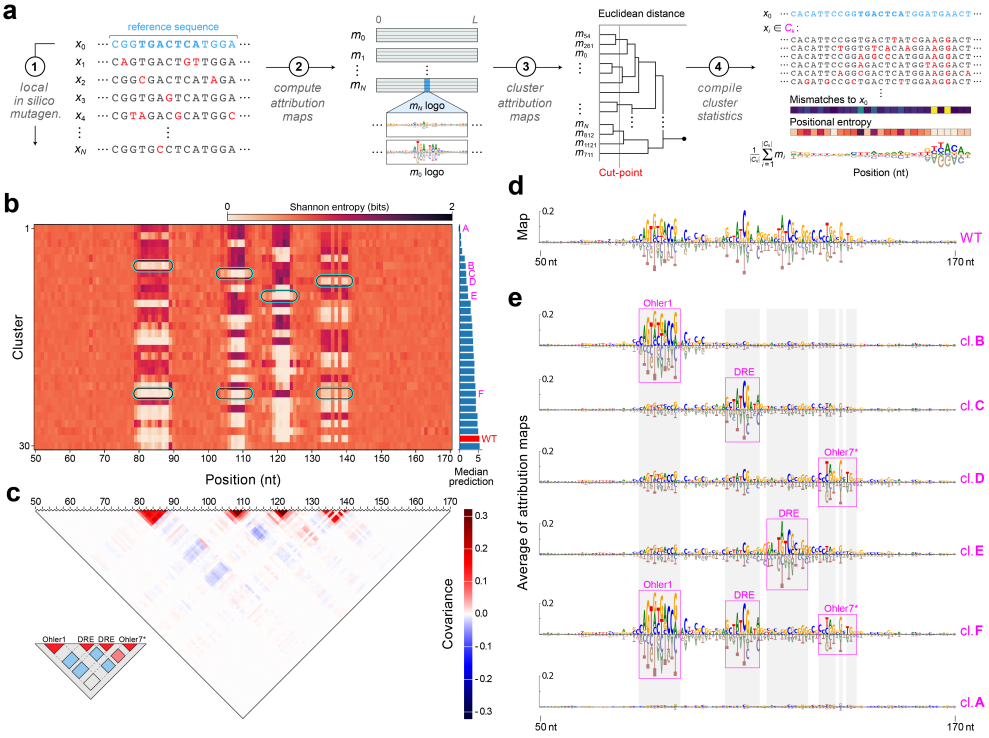


Figure 1: **Overview of SEAM. a**, Schematic of the SEAM meta-explainability framework. **b**, MSM based on positional Shannon entropy for a fly enhancer using DeepSTARR's Hk head, with median DNN prediction (right, bar plot) for each cluster. **c**, Heat map of the MSM covariance matrix highlights cooperative and competitive preferences between motifs. **d**,**e**, Sequence logos of the initial attribution map computed from the WT sequence (**d**) and the average of attribution maps in each cluster (**e**). Vertical gray bars represent the positions of entropy-based patterns in the MSM.

as a powerful tool for resolving regulatory motifs and mechanisms by leveraging counterfactual insights through sequence perturbations.

4 SEAM DISENTANGLES MOTIF- AND CONTEXT-DEPENDENT REGULATORY SIGNALS

We next applied SEAM to ChromBPNet, a sequence-to-activity DNN trained to predict chromatin accessibility from ATAC-seq data in THP-1 human cells (Brennan et al., 2022). The PPIF promoter, previously validated by VariantFlowFISH (Martyn et al., 2023), served as an ideal testbed. Using a 10% local mutagenesis library, SEAM uncovered diverse regulatory mechanisms, identifying distinct motif combinations across multiple clusters.

A key advantage of SEAM is its ability to disentangle TF-dependent signals—highly sensitive to mutagenesis—from context-specific signals, which remain robust to perturbation (see Methods). context-specific signals predominantly reflect diffuse, low-attribution patterns associated with GC content (Fig. 2a). Subtracting these signals from wild-type attribution maps sharpens and isolates TF motifs. Cluster-averaged attribution maps further denoise these motifs (Fig. 2b). For example, an NRF1 motif—previously obscured by GC-rich background—became clearly visible after SEAM's disentanglement (cl.80, Fig. 2b).

SEAM's findings were robust across ChromBPNet models trained on different data subsets and mutation rates (Appendix Fig. 5). Interestingly, context-specific signals varied substantially across loci and biological systems. For instance, Drosophila enhancers analyzed with DeepSTARR exhibited A/T-rich context-specific signals (Appendix Fig. 6), consistent with their known role as TF anten-

nas (Castellanos et al., 2020). These results highlight SEAM's ability to isolate regulatory signals, providing deeper insights into TF and context-specific mechanisms that shape cis-regulatory code.

5 Regulatory sequences are evolvable for diverse cis-regulatory mechanisms

SEAM also reveals how minimal mutations can unlock new regulatory functions. Using a 1% mutagenesis library, we applied SEAM to CLIPNET, a sequence-to-activity DNN trained to predict PRO-cap-seq profiles in human lymphoblastoid cells (He & Danko, 2024). We focused on the PIK3R3 promoter, a highly polymorphic locus (Gupta et al., 2020) linked to cancer (Zhou et al., 2012), to investigate how sequence variants regulate transcription initiation.

SEAM identified clusters with diverse transcriptional activities near wild-type levels (Fig. 2d), each predominantly driven by a common single-nucleotide mutation (Fig. 2c). For example, loss-of-function mechanisms emerged from a zinc finger motif near the transcription initiation site (cl.A, Fig. 2c,d,e), while conserved motifs like IRF, potentially under balancing selection, remained shared across clusters (Fig. 2e). SEAM also uncovered small mutation combinations that created new binding sites, such as a pairwise mutation forming a CAAT box (cl.C, Fig. 2c,e), which upregulated transcriptional activity (Fig. 2d) and reversed transcription direction (Fig. 2f). These results highlight the high evolvability of the PIK3R3 promoter, where specific mutations drive regulatory reprogramming, potentially contributing to disease predisposition.

To test the generality of these findings, we applied SEAM to ProCapNet (Cochran et al., 2024), another DNN predicting PRO-cap-seq profiles. At the MYC promoter, SEAM uncovered cryptic transcription start sites (TSSs), including a previously reported antisense TSS (cl.A, Appendix Fig. 7b), as well as novel cryptic motifs—such as alternate TATA and BRE/SP sites—that either relo-

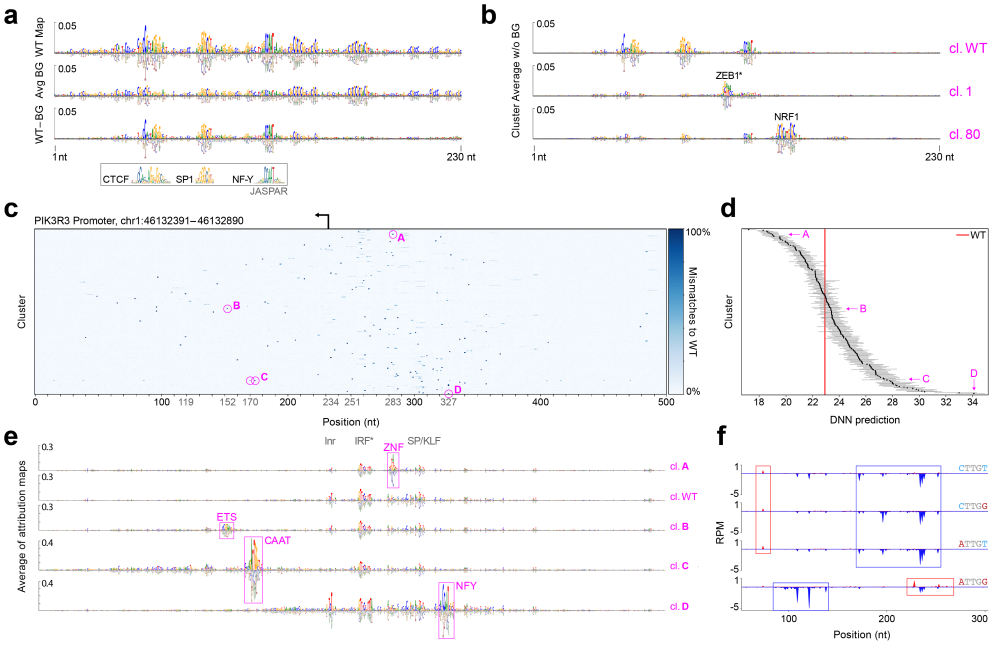


Figure 2: **SEAM captures diverse mechanisms using versatile sequence libraries.** **a**, Attribution logos for wild-type (WT) sequence, intra-cluster averaged backgrounds (BG), and WT sequence with background subtracted. Matched JASPAR motifs shown below. **b**, Representative meta-attribution maps for background adjusted clusters (cl). **c**, MSM for the PIK3R3 promoter, centered at the TSS, colored by percent mismatches to the WT sequence for each of the 200 clusters. **d**, DNN predictions for each cluster in the MSM. **e**, Background adjusted attribution maps for different clusters. **f**, profile predictions for a cryptic CAAT box in cluster C, where different perturbations (shown in red) were applied, highlighting changes in direction of transcription.

cated or reversed TSS direction (cl.E, Appendix Fig. 7b). Similarly, SEAM applied to DeepSTARR identified fly enhancer mutations that converted an activating AP-1 motif into a TTK repressor binding site, suppressing enhancer activity (cl.A, Appendix Fig. 8a). Another pairwise mutation shifted an AP-1 motif three nucleotides without significantly altering activity (cl.B, Appendix Fig. 8a). These results highlight the remarkable plasticity of cis-regulatory elements, where small mutations can reprogram regulatory sequences to drive diverse mechanisms and functional outcomes. This adaptability has profound implications for evolutionary biology and complex disease studies.

6 SEAM GENERALIZES ACROSS DIVERSE SEQUENCE LIBRARIES

SEAM is a versatile framework that generalizes across diverse sequence libraries, providing a unifying approach to decoding regulatory logic. Applying SEAM to combinatorial-complete libraries derived from experiments, such as Protein Binding Microarrays (PBMs), revealed both primary and secondary binding motifs for ZFP187, capturing previously-validated patterns while also uncovering insights into lower-affinity mechanisms (Appendix Fig. 9a,b). In group-optimized libraries, we developed Redirected Evolution (REVO), a motif-centric extension of in silico evolution (ISE), to enhance mechanistic diversity in regulatory design (see Methods). SEAM applied to REVO-optimized DeepMEL2 libraries revealed a broader range of motif types and regulatory mechanisms compared to standard ISE (Appendix Fig. 9c). Finally, in global libraries (see Methods), SEAM identified context-specific regulatory mechanisms in fly enhancers, including mutations that reprogram CREB/ATF binding sites into AP-1, GATA, or C2H2 zinc finger motifs with distinct functional consequences (Appendix Fig. 9d). These results demonstrate SEAM's adaptability across computational and experimental contexts, enabling deeper insights into cis-regulatory mechanisms.

7 DISCUSSION

SEAM represents a transformative advancement in understanding the cis-regulatory code, offering a scalable framework for extracting regulatory insights from attribution maps. Attribution maps traditionally highlight nucleotide importance in model predictions but are often constrained by noise, ambiguity, and an inability to capture broader regulatory logic. SEAM overcomes these limitations by systematically perturbing DNA sequences, clustering the resulting attribution maps, and generating meta-attribution maps that distill shared regulatory patterns. This approach transforms diffuse single-sequence analyses into structured, interpretable summaries, revealing critical insights into the cis-regulatory grammar that governs gene expression.

Unlike hypothesis-driven tools that catalog motifs from observed genomic data, SEAM is discovery-driven, systematically probing mutational landscapes to uncover novel regulatory mechanisms and interactions. By leveraging synthetic sequence libraries, SEAM reveals how genetic variation reshapes the cis-regulatory code, uncovering both motif- and context-specific mechanisms. This unbiased exploration provides a comprehensive understanding of regulatory complexity and highlights how small mutations can drive significant regulatory changes.

A defining strength of SEAM lies in its ability to disentangle motif-dependent and context-specific mechanisms, two fundamental drivers of regulatory activity. While motif-dependent mechanisms tied to well-characterized TFBS are well-understood, SEAM reveals context-specific signals—diffuse patterns often dismissed as noise—as critical modulators of regulatory outcomes. By isolating these features, SEAM uncovers previously hidden layers of complexity, advancing our understanding of how DNA sequences encode functional diversity and fine-tune gene regulation.

SEAM also highlights the remarkable evolvability of regulatory sequences, demonstrating how minimal sets of mutations can reprogram motifs, reverse transcriptional direction, or activate new pathways. For instance, at the PIK3R3 promoter, small sequence changes preserved overall function while enabling diverse regulatory mechanisms. These findings underscore how regulatory elements balance robustness with flexibility, supporting phenotypic diversity and evolutionary innovation.

By systematically exploring sequence variants, SEAM maps mutational landscapes through the lens of mechanisms learned by a DNN. This provides a powerful framework for understanding gene regulation and rationally designing synthetic sequences with precision, advancing applications in synthetic biology, precision medicine, and functional genomics.

A APPENDIX

METHODS

THE SEAM FRAMEWORK.

SEAM takes as input a sequence of interest, a specified attribution method, a clustering method, and a sequence-function oracle.

- Mutagenizer. An in silico sequence dataset is generated by sampling a library of N sequences using random partial mutagenesis of a sequence of interest. We modulate the size of the sequence-space region from which this library is drawn using two hyperparameters: the sequence that defines the region of interest, which has length L, and the mutation rate r. The resulting number of mutations in each individual sequence is a Poisson distributed random variable having mean Lr. The Mutagenizer class contains objects that apply the chosen mutagenesis strategy.
- Attributer. An attribution method is used to compute the attribution map for each sequence in the in silico sequence library.
- Clusterer. The N attribution maps are directly clustered or first embedded in a low-dimension space and then clustered.
- MetaExplainer. Attribution maps within each cluster are averaged to form meta-attribution maps representing a noise-reduced consensus of each mechanism. Sequence-function and sequence-mechanism relationships are compiled using the sequence and function statistics associated with attribution maps in each cluster. Sequence-mechanism relationships are further used for delimiting the positions of individual motifs and background separation.

ATTRIBUTION METHODS

Our analyses used attribution maps computed using a variety of methods, implemented as follows.

- In Silico Mutagenesis (ISM) scores were computed by evaluating the change in the scalar DNN prediction from the wild type prediction for every single nucleotide variant of the sequence of interest (Zhou & Troyanskaya, 2015).
- Empirical Mutagenesis scores were computed similarly to ISM scores, using experimental measurements in place of DNN predictions. In the case of PBM data, for every sequence in the combinatorial-complete library, SEAM generates an empirical mutagenesis map using the log2 fold change between the E scores of the reference sequence and a single nucleotide variant.
- Saliency Maps scores were computed by evaluating the gradient of the scalar DNN prediction at the sequence of interest with respect to the one-hot encoding of that sequence (Simonyan et al., 2014).
- SmoothGrad scores were computed by averaging Saliency Maps over 50 noisy encodings of the sequence of interest. Each noisy encoding was computed by adding Gaussian noise (mean zero, standard deviation 0.25) to each of the $4L$ matrix elements of the one-hot encoding for the sequence of interest (Smilkov et al., 2017).
- Integrated Gradients (IntGrad) scores were computed by interpolating between a baseline reference sequence and the sequence of interest. The gradient of the DNN's scalar prediction is integrated along the interpolated path, and the resulting attributions reflect the cumulative contribution of each nucleotide to the prediction (Sundararajan et al., 2017).
- DeepSHAP scores were computed by using a background set of reference sequences and evaluating the contribution of each nucleotide at each position by comparing the DNN's activations for the sequence of interest to its activations for the background (Lundberg & Lee, 2017). DeepSHAP hyperparameters were chosen to match those used in each DNN's original publication.

CLUSTERING METHODS

Our analyses used methods to generate clusters that were assigned directly on a distance matrix or indirectly on an embedded space.

- Hierarchical clustering was applied with Ward's linkage on the Euclidean distance matrix (Ward, 1963), computed directly from the attribution maps library. Ward's method minimizes the total within-cluster variance by iteratively merging clusters, which results in a hierarchical tree (dendrogram) that groups attribution maps based on their similarity in the original feature space. We cut the dendrogram at a defined level to select a specific number of highest-level clusters.
- Alternative to the direct clustering of attribution maps, SEAM can embed attribution maps before clustering them. For this purpose, PCA (Pearson, 1901), t-SNE (van der Maaten & Hinton, 2008), and UMAP (McInnes et al., 2020) embedding is currently integrated within the SEAM framework, with k-means (Lloyd, 1982) and DBSCAN (Ester et al., 1996) used for subsequent clustering on the embedded space (i.e., Appendix Fig. 2; Appendix Fig. 9d).

The number of clusters should be chosen based on the objective of the analysis. For building biophysical state models, the number of TFBSs, T, present at a locus informs the number of biophysical states, 2^T, and thus the number of clusters. For other studies, initially using a large cluster number may be advantageous to assess the overall amount and placement of mechanistic variation at a locus using the SEAM variability logo (e.g., Appendix Fig. 1). After this assessment, the number of clusters can be scaled down to focus on specific aspects of the dataset. In general, the number of clusters should not be so low that a cluster with zero entropy across all positions appears in the MSM.

BACKGROUND SEPARATION ANALYSIS

When using a local library, an approximately uniform set of backgrounds emerges in the attribution maps across all sequences. As the SEAM-derived Mechanism Summary Matrix (MSM), based on positional Shannon entropy, captures the sequence determinants driving TF motif activity per cluster, SEAM uses this information to separate the common background signal from TF-dependent motifs in each cluster. First, the background entropy of the sequence library is calculated using the mutation rate, r, used to generate the library, and the corresponding probability, p, of a position remaining unchanged, where $p = 1 - r$. For a sequence with c nucleotides, the background entropy, H_{BG}, is calculated as the entropy of the following distribution:

$$H_{\text{BG}} = -p \cdot \log_2(p) - (1-p) \cdot \log_2\left(\frac{1-p}{c-1}\right)$$

Next, an entropy threshold $H_0 = H_{\text{BG}}/2$ is set. For each averaged attribution map of a given cluster, indexed by k, the attribution values are set to zero for positions i in the associated row of the MSM where the positional sequence entropy is less than the threshold entropy, $H_{k,i} < H_0$. Repeating this operation across all clusters and averaging the result effectively captures the attribution background, while reintegrating background attribution values that were removed from each cluster based on the presence of cluster-specific TF motifs. Finally, the averaged attribution background is subtracted from each of the averaged attribution maps per cluster. Within a local library, the background is uniform up to a constant scaling factor. To avoid introducing excess background signal due to mismatched amplitudes, an additional per-cluster scaling factor is applied to the background before the background is subtracted. Subtraction isolates cluster-specific TF motifs by removing common background signals, thereby enhancing the specificity of the attribution maps for TF-dependent motifs within each cluster.

DEEP LEARNING MODELS

This study used five DNNs: DeepSTARR, CLIPNET, ProCapNet, ChromBPNet, and DeepMEL2. Here, we briefly describe each DNN and how that DNN was used in our study to compute attribution maps.

- DeepSTARR (de Almeida et al., 2022) predicts *Drosophila* enhancer activity as assayed by UMI-STARR-seq. DeepSTARR takes as input a DNA sequence of length 249 nt and

outputs two scalar-valued predictions for enhancer activity for developmental (Dev) and housekeeping (Hk) regulatory programs. The DeepSTARR parameters were retrained in TensorFlow as specified in the original release, and the resulting model was confirmed to recapitulate the published model using performance metrics and visualization of attribution maps.

- CLIPNET (He & Danko, 2024) predicts nucleotide-resolution transcription initiation profiles from a dataset consisting of matched precision run-on and 5'-capped (m^7G) RNA enrichment (PRO-cap) and individual heterozygous human genomes from 58 genetically distinct lymphoblastoid cell lines (LCLs). CLIPNET takes as input a DNA sequence of length 1,000 nt and outputs two predictions via a "profile" head and a "counts" head. The profile head predicts strand-specific PRO-cap coverage over the central 500 nt (500 for the plus strand concatenated with 500 for the minus strand), representing the predicted base-resolution profile of initiation. The counts head predicts the total PRO-cap signal across both strands. CLIPNET is an ensemble model comprising 9 structurally identical models, each trained with a distinct holdout set of chromosomes. Unless otherwise specified, SEAM analysis was performed by averaging predictions and attribution maps across all 9 folds. Attribution analysis in CLIPNET was conducted on two-hot encoded DNA sequences, where each nucleotide at a given position is represented as a sum of two one-hot encoded nucleotides, capturing the unphased diploid sequence. When applying DeepSHAP to two-hot encoded sequences, heterozygous positions can be seen as vectors between the two orthogonal features (alleles) in the input domain. DeepSHAP evaluates the function's behavior at this new composite point, reflecting the model's interpretation of the combined contribution from both alleles.
- ProCapNet (Cochran et al., 2024) predicts nucleotide-resolution transcription initiation profiles as measured by PRO-cap in human K562 cells. ProCapNet takes as input a homozygous DNA sequence of length 2,114 nt and generates two predictions via a profile head and a counts head. The profile head predicts nucleotide-resolution initiation activity across both strands within a central 1,000 nt region, while the counts head predicts the log-transformed total number of PRO-cap reads with 5' ends mapped within this region, summed across both strands. ProCapNet was trained using a 7-fold cross-validation scheme. Unless otherwise specified, SEAM analysis was performed by averaging predictions and attribution maps across all 7 folds. Profile head predictions were consolidated into a single explainable scalar following the approach used in the original publication.
- ChromBPNet (Brennan et al., 2022) predicts nucleotide-resolution chromatin accessibility profiles as measured by ATAC-seq in THP-1 cells. ChromBPNet takes as input a DNA sequence of length 2,048 nt and generates two predictions via a "profile" head and a "counts" head. The profile head predicts nucleotide-resolution ATAC-seq coverage within a central 1,000 nt region, while counts head predicts the natural log count of the aligned reads within this region. ChromBPNet was trained using a 5-fold cross-validation scheme. Unless otherwise specified, SEAM analysis was performed by averaging predictions and attribution maps across all 5 folds. Profile head predictions were consolidated into a single explainable scalar following the approach used in the original publication. The version of ChromBPNet used in this work is available at `10.5281/zenodo.10403551`.
- DeepMEL2 (Taskiran et al., 2024) predicts melanoma-specific chromatin accessibility as measured by ATAC-seq and allele-specific chromatin accessibility variants (ASCAVs). DeepMEL2 takes as input the forward and reverse DNA strands, each of length 500 nt, and outputs a vector of binarized predictions across 47 classes, each representing a melanoma cis-regulatory topic. Only class 16 (MEL) was used in this work. Contribution scores were generated for each strand separately, and following previous work (Taskiran et al., 2024), we averaged the contribution scores over both strands to visualize attribution maps.

SEQUENCE DESIGN METHODS

SEAM is a versatile framework that can be applied on a diversity of sequence libraries.

- Local libary. Random partial mutagenesis is applied to a genomic sequence of interest. The number of mutations in each individual sequence is a Poisson distributed random variable having mean Lr, where L is the sequence length and r is the mutation rate.

- Global library. Fixed genetic elements, such as putative TFBSs, are embedded at the same position across a library of completely random sequences.
- Combinatorial-complete library. A set of sequences that includes all possible combinations of nucleotides at specified positions, ensuring comprehensive coverage of sequence variation. In this analysis, SEAM was applied directly to experimental datasets that met this requirement using empirical mutagenesis maps.
- Group-optimized library. In this study, we used REVO (see below) to generate a library of heterozygous sequences anchored at a genomic sequence of interest. SEAM is not limited to REVO, and other optimization strategies could also be tried (e.g., BEAM search or genetic algorithms).

Our optimization methods evolved sequences by iteratively selecting mutations that maximize predictive changes in function.

- In silico evolution (ISE) is a hill-climbing algorithm that was adapted for DeepMEL2 optimization (de Boer et al., 2020). In the adapted protocol—called Evolved From Scratch (Taskiran et al., 2024) and repeated in our analysis—ISE starts from a random, GC-adjusted sequence. At each iteration of the evolution, represented by a partially mutated sequence, saturation mutagenesis is performed to generate all possible SNVs. The SNV with the highest positive change in prediction (for the selected class) is chosen. For the selected sequence with one new mutation, saturation mutagenesis is recalculated, with this procedure repeated until the initial random sequence accumulates t mutations. In their study (Taskiran et al., 2024), ISE typically produced a single optimized sequence after 10-15 iterations whose predicted activity correlated with in vitro luciferase reporter results. Attribution analysis showed that the optimized sequences corresponded to distinct mechanisms, influenced by the starting sequence context.
- Redirected Evolution (REVO) is an extension of ISE (starting from a random, GC-adjusted sequence) where in silico mutagenesis is performed using redirected evolution to identify optimized modes in sequence-function space. At the end of one round of ISE, the attribution map for the final sequence, having t accumulated mutations, is generated. Windows of length w_l slide across the attribution map to determine the top w_n non-overlapping regions, ranked by the sum of max attribution values at each position in each window. ISE is then rerun from the initial, random GC-corrected sequence in a branching structure, with each branch corresponding to a combination of the w_n regions, and the specified region(s) protected from mutations during the subsequent t iterations. This process is repeated T times, with redundant branches pruned in real time to mitigate computational repetition. By intentionally blocking the optimization of previously discovered sequence elements, this approach redirects the evolutionary process to explore new areas of sequence space, promoting diversity and the discovery of new sequences. All N sequences are used as inputs to SEAM, with N indeterminate due to the unpredictable nature of pruning based on repeated elements.

DATA AVAILABILITY

PBM data are available in the UniPROBE database (Newburger & Bulyk, 2008).

CODE AVAILABILITY

SEAM is an open-source Python package based on TensorFlow, and contains CPU and GPU optimized code for attribution analysis and clustering (Abadi et al., 2015). SEAM can be installed via pip (`https://pypi.org/project/seam-nn`) or GitHub (`https://github.com/evanseitz/seam-nn`). The GitHub repository contains links to running several examples from our analysis in Google Colab. Documentation for SQUID is provided on ReadTheDocs (`https://seam-nn.readthedocs.io`).

ACKNOWLEDGEMENTS

We thank Charles Danko, Adam He, Kai Loell, Jack Desmarais, and Zhihan Liu for helpful discussions. This work was supported in part by: the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory; NIH grants R01HG012131 (P.K.K, E.E.S., J.B.K., D.M.M.), R01HG011787 (J.B.K., E.E.S., D.M.M.), R01GM149921 (P.K.K.), R35GM133777 (J.B.K.), F32HG013265 (E.E.S.), and R35GM133613 (D.M.M.). Computations were performed using equipment supported by NIH grant S10OD028632.

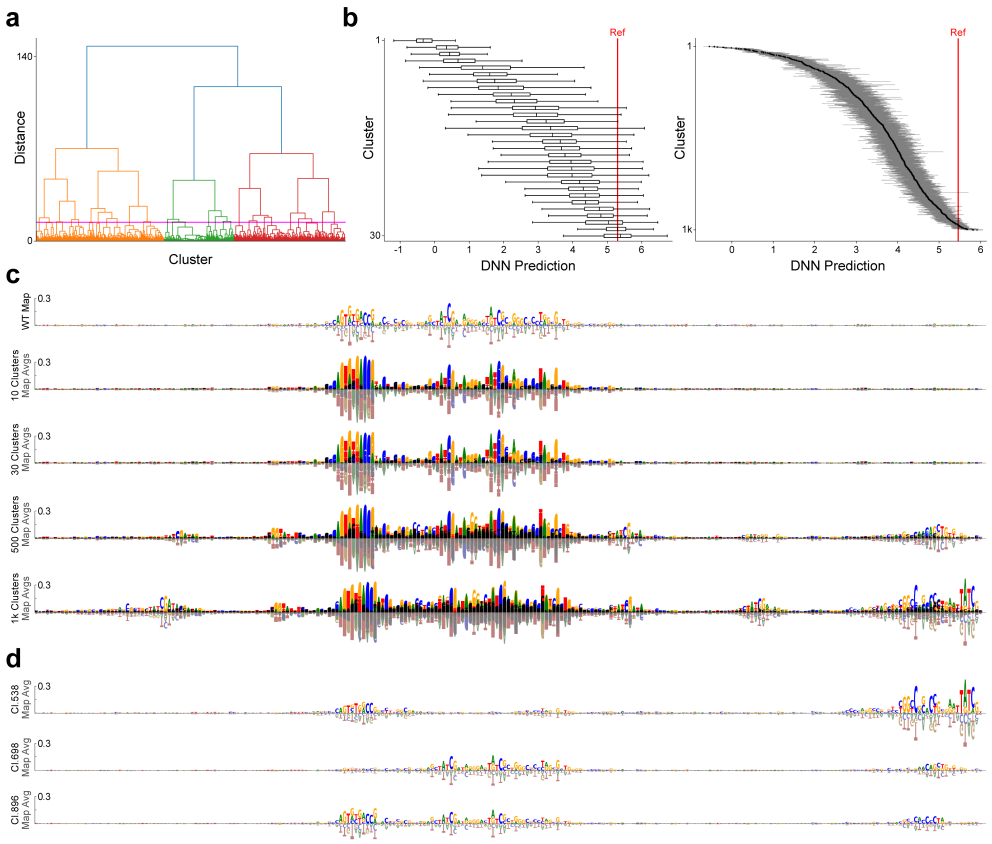
AUTHOR CONTRIBUTIONS

E.E.S. conceived the method, developed the software, and performed the analysis. E.E.S., D.M.M., J.B.K., and P.K.K. designed the study and wrote the paper. J.B.K. and P.K.K. supervised the research.

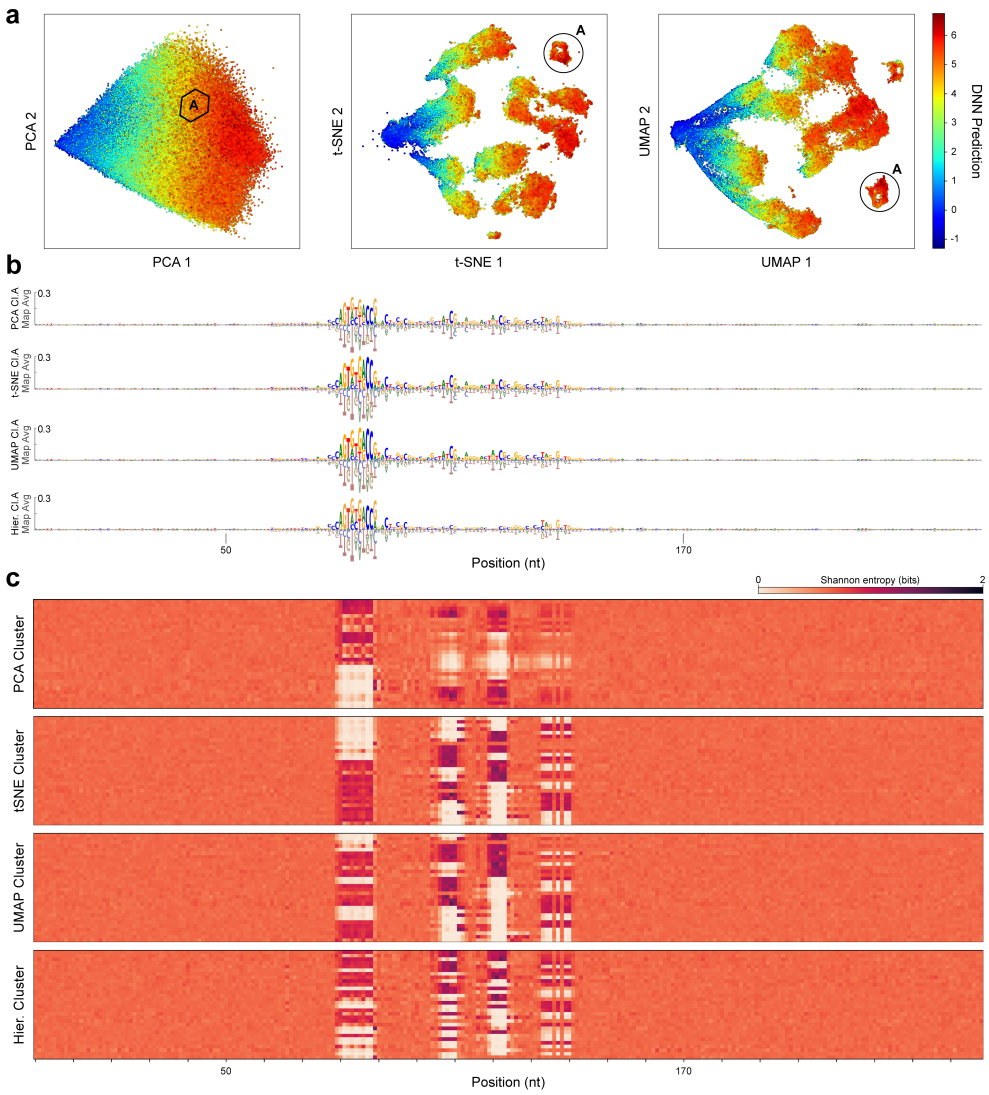
COMPETING INTERESTS

The authors declare no competing interests.

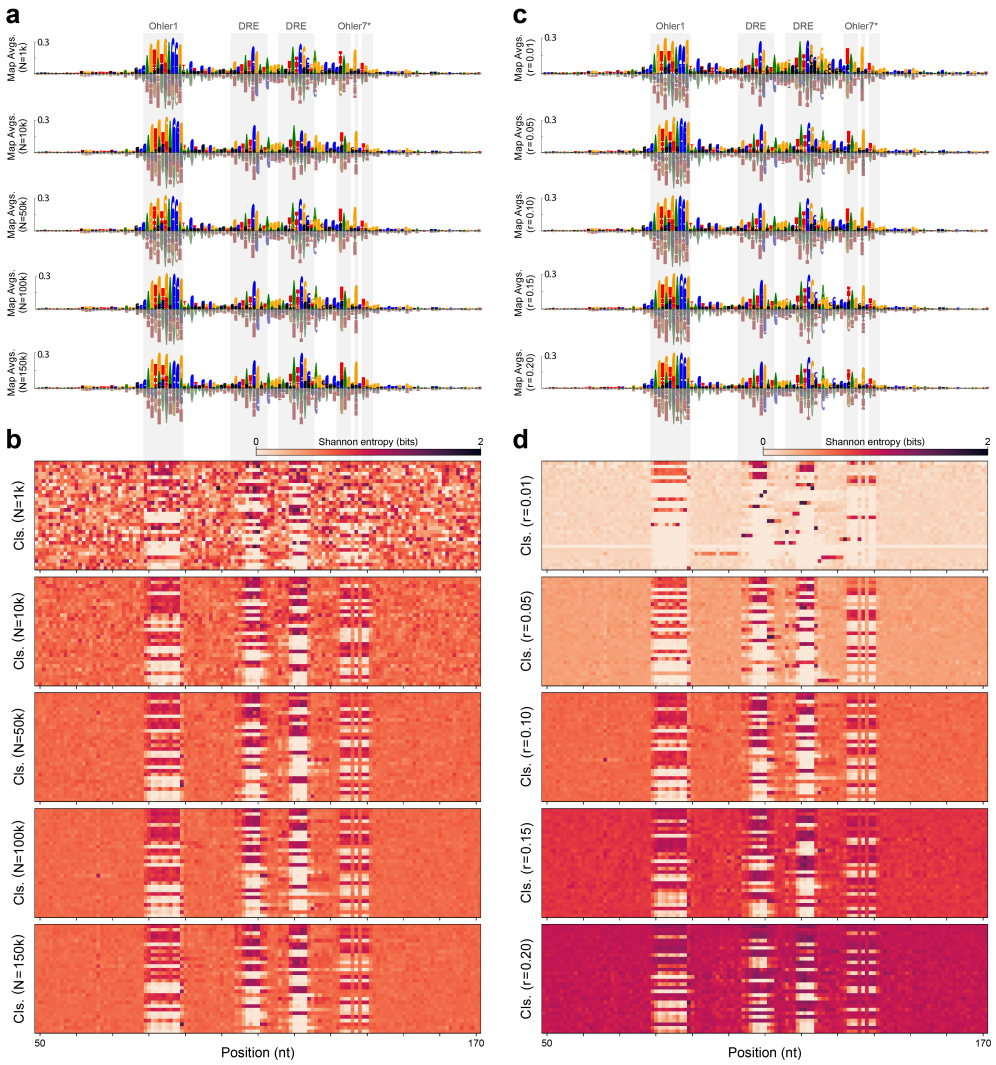
APPENDIX FIGURES



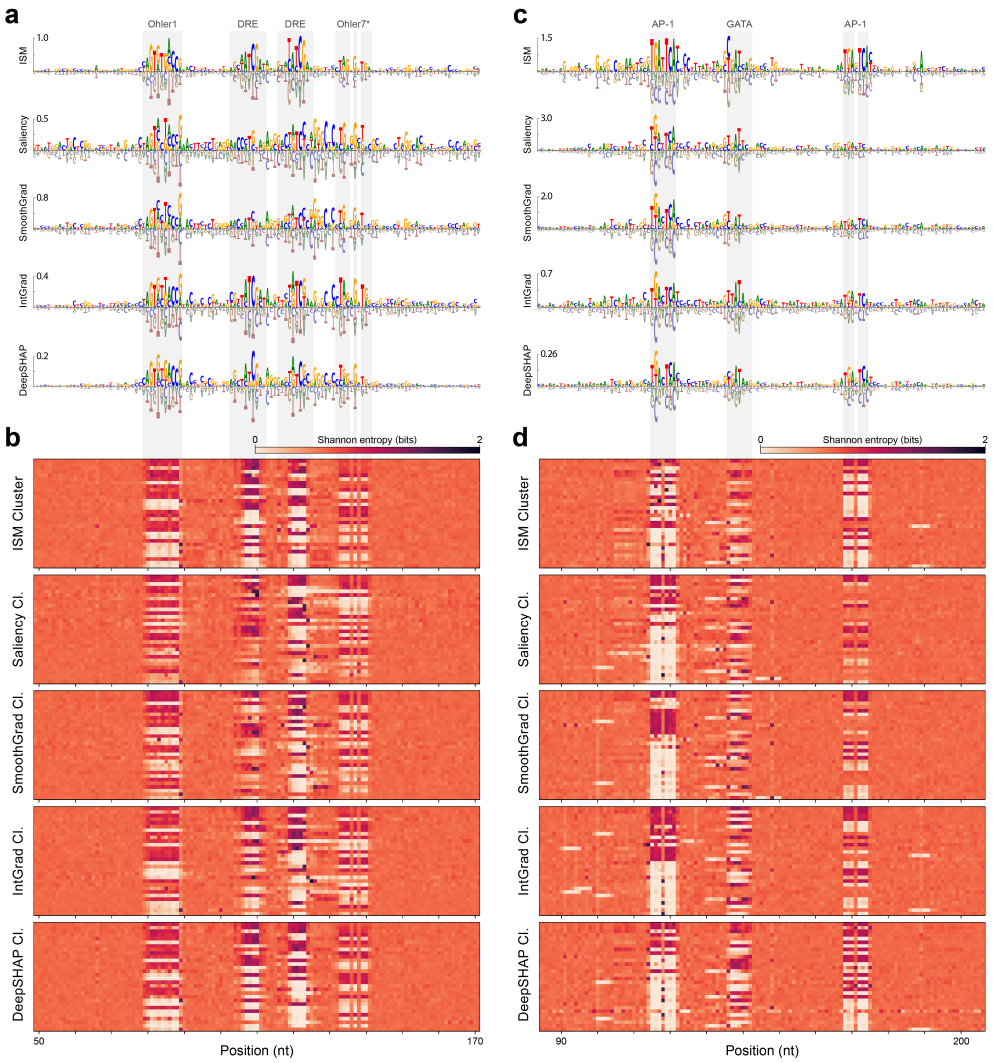
Appendix Figure 1. Impact of maximum cluster number on SEAM mechanistic insights. **a**, SEAM dendrogram for the DeepSTARR locus shown in Fig. 1, generated using hierarchical clustering with Ward's linkage on a Euclidean distance matrix. The pink horizontal line indicates the cut level for selecting the 30 highest-level clusters. **b**, Median DNN predictions for the 30 (left) and 1000 (right) highest-level clusters. Both plots demonstrate that the clusters span a dynamic range of DNN predictions, with finer granularity as the number of clusters increases. Lines represent the upper and lower quartiles. **c**, Comparison of the WT map to the overlay of all average maps from each cluster, as the maximum number of clusters increases from 10 to 1000. As seen with 500 and 1000 clusters, more mechanisms are identified. **d**, Examples of individual mechanisms obtained using 1000 clusters, revealing new instances of TFBSs not observed at the other cut levels shown.



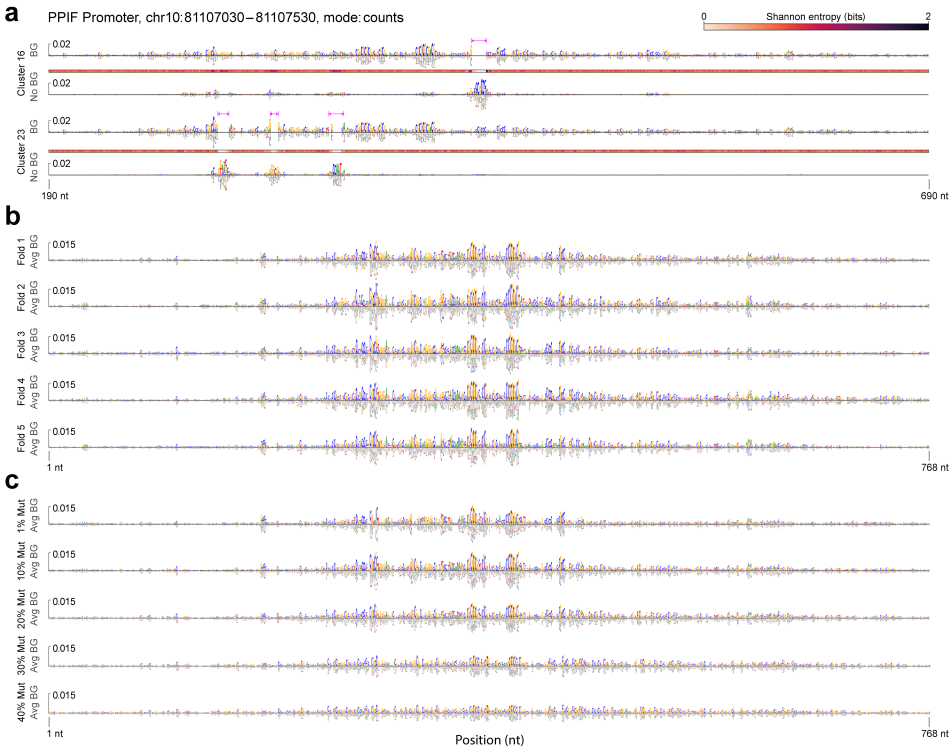
Appendix Figure 2. Impact of clustering method on SEAM mechanistic insights. **a**, Comparison of SEAM embeddings of attribution maps for the DeepSTARR locus shown in Fig. 1, generated using PCA, t-SNE, and UMAP. For each embedding, k-means was performed, resulting in 200 clusters. For each set of clusters in each embedding, the position of an example cluster A is encircled, corresponding to a mechanism with a highly similar visual appearance across all three embeddings. **b**, Sequence logo for the mechanism corresponding to cluster A, generated by averaging the attribution maps in cluster A for each embedding. **c**, Comparison of MSMs, based on positional Shannon entropy, for the three embeddings. The MSM created using hierarchical clustering, as shown in Fig. 1, is also shown for comparison. By visual inspection, PCA (using the two leading eigenvectors) with k-means produces the MSM with the lowest-resolution features, while hierarchical clustering produces the MSM with the highest-resolution features. Avg., average; Cl., cluster.



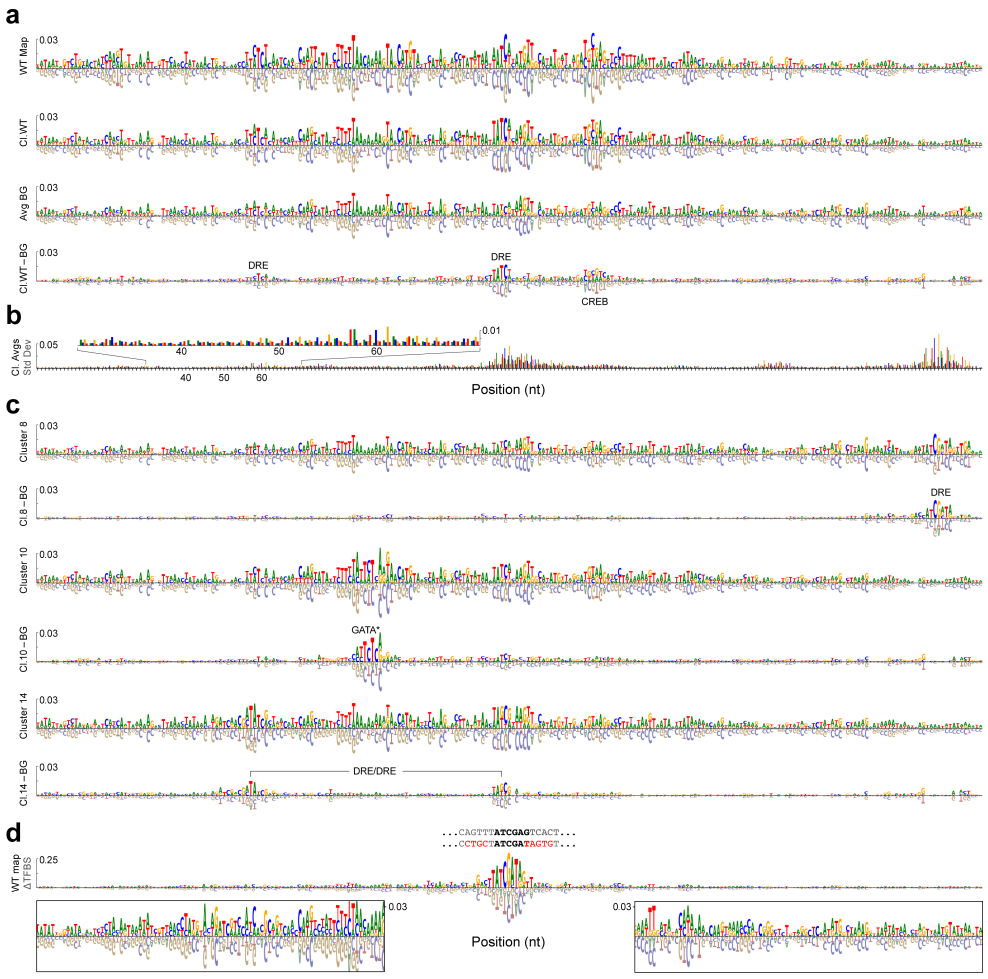
Appendix Figure 3. Impact of hyperparemeters on SEAM mechanistic insights. Comparison of SEAM outputs for the DeepSTARR locus shown in Fig. 1 as SEAM hyperparameters—including the size of the sequence library, N, and the mutation rate, r, used to generate the library—are varied. **a**, Comparison of SEAM variability logos (i.e., the overlay of all average attribution maps for 30 clusters generated by hierarchical clustering) as N is varied with a constant mutation rate, $r = 0.10$. **b**, Comparison of the corresponding MSMs based on positional Shannon entropy. **c**, Comparison of SEAM variability logos as r is varied with a constant library size, $N = 100,000$ sequences. **d**, Comparison of the corresponding MSMs based on positional Shannon entropy. By visual inspection, SEAM outputs are robust to library size and rate of partial mutagenesis. Comparing the MSM shown for $N = 100,000$ and $r = 0.10$, which is a replicate of the MSM shown in Fig. 1 generated with a different random mutagenesis seed, shows that SEAM outputs are also robust to random sequence generation. Avg., average; Cls., clusters.



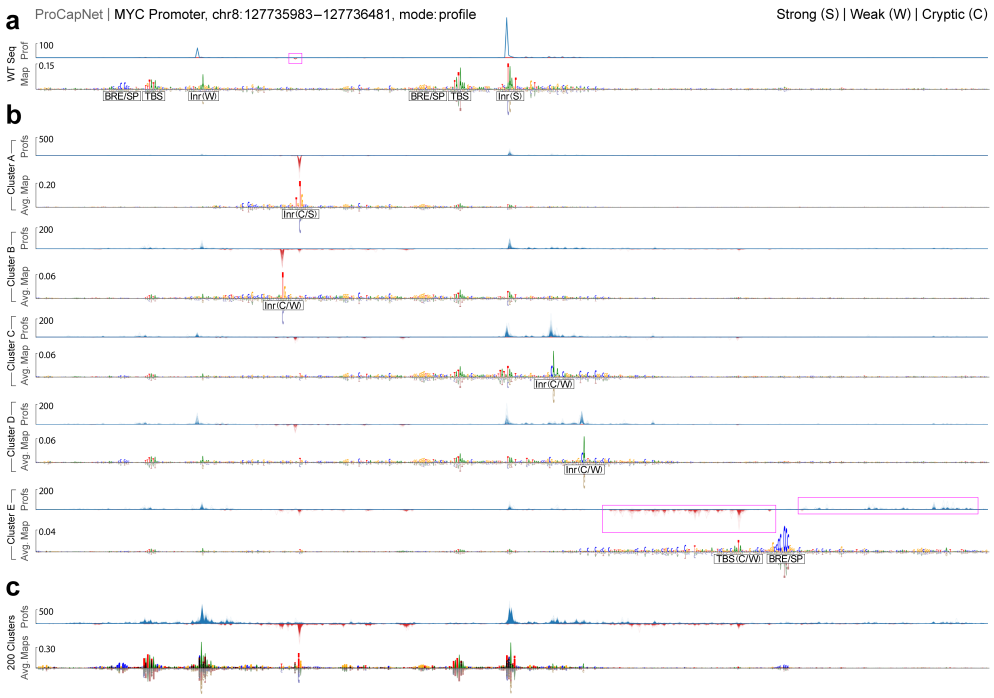
Appendix Figure 4. Impact of attribution method on SEAM mechanistic insights. **a**, Comparison of wild-type attribution maps for the DeepSTARR locus shown in Fig. 1, generated using in silico mutagenesis (ISM), Saliency Maps, SmoothGrad, Integrated Gradients (IntGrad), and DeepSHAP. Gray bars running vertically across the attribution maps align with entropy-biased regions in the corresponding MSMs, below. **b**, MSMs based on positional Shannon entropy generated by SEAM using different attribution methods for the same locus. Features in the MSM are consistent across attribution method, and identify locations of important motifs that can be difficult to discern in the wild-type map generated by each attribution method. **c,d**, Attribution maps and corresponding SEAM MSMs for another locus obtained from the DeepSTARR test set (index 22612) using the Dev head, with similar trends observed. Cl., cluster.



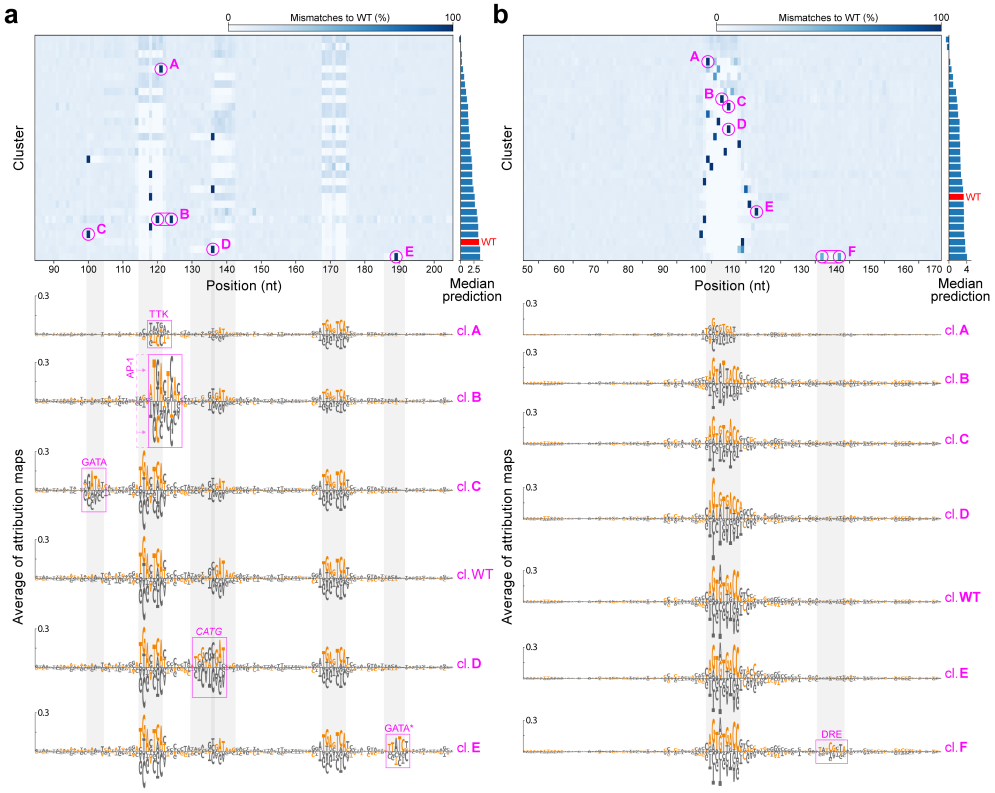
Appendix Figure 5. Impact of ChromBPNet model fold and mutation rate on SEAM background separation. **a**, Examples of attribution maps for intra-mechanism background (top) and foreground (bottom) for clusters 16 and 23. Intra-mechanism background is informed by the positional Shannon entropy over sequences in each cluster (middle). SEAM uses the thresholded positional Shannon entropy to mask out TF-dependent attributions in each map. The foreground attribution map is derived by subtracting the average attribution map of a given cluster from the average of all intra-mechanism background maps over all clusters. SEAM was run using the average of attribution maps over all folds, with sequences sampled using a 10% mutation rate. **b**, Results of running SEAM background separation independently on each of the first five ChromBPNet folds (10% mutation rate). **c**, Results of independently running SEAM background separation on fold 1 using datasets generated with different mutation rates. Avg., average; BG, background.



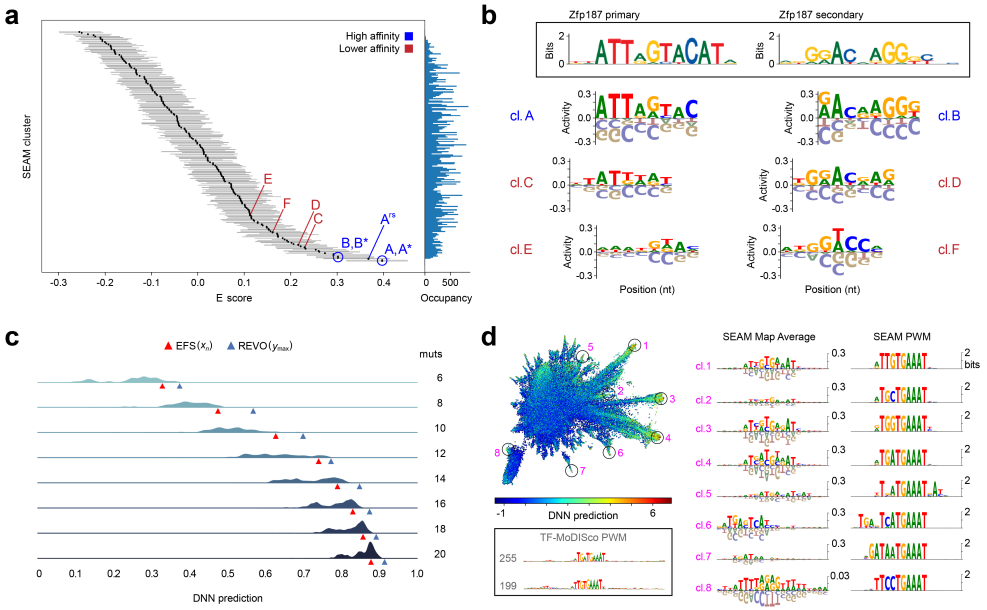
Appendix Figure 6. Background separation at a DeepSTARR enhancer. **a**, First row: WT attribution map from a DeepSTARR enhancer (test set index 4071). A low-affinity DRE TFBS, ATCGAG–with one mutation to the consensus TFBS, ATCGAT–is positioned near the center, resulting in an overall low Hk expression (-0.49). Seemingly spurious scores across the attribution map obscure the identification of TF motifs. Second row: Average attribution map for the WT cluster, showing SEAM's ability to denoise the WT mechanism by averaging over qualitatively similar attribution maps. Third row: Average of all intra-cluster backgrounds separated by SEAM, featuring AT-rich attribution signals across the enhancer. Fourth row: WT cluster after background removal, generated by subtracting the average background (row 3) from the averaged WT mechanism (row 2). Background separation reveals three previously obscured TF motifs. **b**, Standard deviation of attribution values over clusters, highlighting the static nature of the background on which the TF motifs reside across all clusters. **c**, Examples of background separation on other mechanisms discovered by SEAM at the same locus, revealing previously obfuscated TF motifs in each cluster. Cluster 14 recovers a higher-affinity DRE/DRE motif compared to WT that is not necessitated by CREB on the right-hand side. **d**, Attribution map was generated from the WT sequence after mutating the central DRE TFBS to match the consensus motif ATCGAT, including optimized flanking nucleotides. Compared to the WT, these 10 mutations increase Hk expression to 2.01, while substantially altering the attributed background context across the enhancer (see insets). This example highlights the sensitivity of background signals to coordinated mutations. WT, wild-type; Avg., average; Cl., cluster; BG, background.



Appendix Figure 7. SEAM analysis of mechanistic variation at ProCapNet MYC promoter. a, Predicted profile and attribution map for the WT sequence, demonstrating the TF motifs and profile peaks present, which are used for biased ablation experiments. **b**, Five (cl. A-E) of the 200 clusters generated using SEAM. For each panel, the predicted profiles associated with the attribution maps in the given cluster are overlaid on top, with the average of all attribution maps in the cluster displayed on the bottom. In cluster A, SEAM finds a stronger version of the previously-documented weak antisense initiator. In clusters B-D, SEAM finds previously-undiscovered weak antisense (cl. B) and sense (cl. C, D) initiators. In cluster E, SEAM discovers an alternative TATA and BRE/SP TFBS that substantially alters upstream and downstream attribution values, while reversing the direction of transcription (pink rectangles). **c**, Overlay of all predicted profiles for all clusters (top) and overlay of all average attribution maps per cluster (bottom). In this view, smeared motifs represent overlapping mechanisms that can be finely tuned over a broad range based on sequence context. Avg., average; Seq., sequence; Prof., profile.



Appendix Figure 8. SEAM identifies key mutations that change mechanism. **a**, MSM based on percent mismatches to wild type (WT) for a locus obtained from the DeepSTARR test set (index 22612) using the Dev head. In this representation, the influence of single nucleotide variations (SNVs) at specific positions is shown to largely govern many of the qualitatively distinct mechanism shifts away from the WT mechanism (cl. WT). Many of these SNVs drive the formation of poised motifs that can significantly increase, decrease, or fine-tune enhancer activity. For example, the SNV in cluster A effectively replaces the presence of the WT AP-1 motif with a poised TTK repressor, significantly driving down enhancer activity. Pairwise mutations are also observed to govern mechanism changes. As seen in cluster B, two mutations shift the wild-type AP-1 motif three nucleotides to the right while fine-tuning enhancer activity. **b**, At another locus from the DeepSTARR test set (index 22627) using the Hk head, SNVs predominantly create distinct conformations of a central Ohler1 motif. These results demonstrate the diversity of cis-regulatory syntax discovered by SEAM. Sequence logos for each averaged mechanism are colored based on deviations (gray) to the wild-type sequence (orange). WT, wild-type.



Appendix Figure 9. SEAM captures diverse mechanisms using versatile sequence libraries. a, Left: Box plots showing E scores for each of the 200 clusters produced by SEAM from the PBM ZFP187 dataset, ranked in ascending order by each cluster's median E score. Letters A-F label example clusters, with superscripts denoting reverse complements (*) or a register shift (rs). Right: Number of empirical mutagenesis maps (occupancy) in each cluster. **b**, Top inset: Information content sequence logos for the two alternative binding modes captured in the original PBM study of ZFP187 using the Seed-and-Wobble algorithm. Bottom: Sequence logos of averaged empirical mutagenesis maps corresponding to the example SEAM clusters labeled in the previous panel. **c**, Sequence design comparison of DeepMEL2 predicted activities for REVO designed sequences versus in silico evolution of a starting sequence EFS-6 from the original analysis. **d**, Global analysis of CREB/ATF binding mechanisms using the sequence TNNTGAAAT (Dev head). Top left: UMAP embedding of cropped attribution maps. Bottom left: Previously published TF-MoDISco results, highlighting two distinct CREB/ATF motifs with 255 and 199 supporting seqlets, respectively. Right: SEAM-derived meta-attribution maps and associated PWMs for encircled regions in the UMAP embedding.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

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

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, oct 2021b. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x. URL https://doi.org/10.1038/s41592-021-01252-x.

Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a -globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2):299–308, 1981. ISSN 0092-8674. doi: 10.1016/0092-8674(81)90413-X. doi: 10.1016/0092-8674(81)90413-X.

Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010. ISSN 1476-4687. doi: 10.1038/nature09000. URL https://doi.org/10.1038/nature09000.

Kaelan J. Brennan, Melanie Weilert, Sabrina Krueger, Anusri Pampari, Hsiao-Yun Liu, Ally W.H. Yang, Timothy R. Hughes, Christine A. Rushlow, Anshul Kundaje, and Julia Zeitlinger. Chromatin accessibility is a two-tier process regulated by transcription factor pioneering and enhancer activation. *bioRxiv*, 2022. doi: 10.1101/2022.12.20.520743.

Milagros Castellanos, Nivin Mothi, and Victor Muñoz. Eukaryotic transcription factors can track and control their target genes using dna antennas. *Nature Communications*, 11(1):540, 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-14217-8. URL https://doi.org/10.1038/s41467-019-14217-8.

Kelly Cochran, Melody Yin, Anika Mantripragada, Jacob Schreiber, Georgi K. Marinov, and Anshul Kundaje. Dissecting the cis-regulatory syntax of transcription initiation with deep learning. *bioRxiv*, 2024. doi: 10.1101/2024.05.28.596138.

James P. Crutchfield and Erik van Nimwegen. The evolutionary unfolding of complexity. In Laura F. Landweber and Erik Winfree (eds.), *Evolution as Computation*, pp. 67–94, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

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, may 2022.

Carl G. de Boer, Eeshit Dhaval Vaishnav, Ronen Sadeh, Esteban Luis Abeyta, Nir Friedman, and Aviv Regev. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology*, 38(1):56–65, 2020. ISSN 1546-1696. doi: 10.1038/s41587-019-0315-8. URL https://doi.org/10.1038/s41587-019-0315-8.

Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Hemant Gupta, Khyati Chandratre, Siddharth Sinha, Teng Huang, Xiaobing Wu, Jian Cui, Michael Q. Zhang, and San Ming Wang. Highly diversified core promoters in the human genome and their effects on gene expression and disease predisposition. *BMC Genomics*, 21(1):842, 2020. ISSN 1471-2164. doi: 10.1186/s12864-020-07222-5. URL https://doi.org/10.1186/s12864-020-07222-5.

Adam Y. He and Charles G. Danko. Dissection of core promoter syntax through single nucleotide resolution modeling of transcription initiation. *bioRxiv*, 2024. doi: 10.1101/2024.03.13.583868.

Peter K Koo and Matt Ploenzke. Deep learning for inferring transcription factor binding sites. *Current Opinion in Systems Biology*, 19:16–23, 2020.

Lefteris Koumakis. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, 18:1466–1473, Jun 2020. ISSN 2001-0370. doi: 10.1016/j.csbj.2020.06.017. URL https://doi.org/10.1016/j.csbj.2020.06.017. eCollection 2020.

Mike Levine. Transcriptional enhancers in animal development and evolution. *Current Biology*, 20(17):R754–R763, 2010. ISSN 0960-9822. doi: 10.1016/j.cub.2010.06.070. doi: 10.1016/j.cub.2010.06.070.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Gabriella E. Martyn, Michael T. Montgomery, Hank Jones, Katherine Guo, Benjamin R. Doughty, Johannes Linder, Ziwei Chen, Kelly Cochran, Kathryn A. Lawrence, Glen Munson, Anusri Pampari, Charles P. Fulco, David R. Kelley, Eric S. Lander, Anshul Kundaje, and Jesse M. Engreitz. Rewriting regulatory dna to dissect and reprogram gene expression. *bioRxiv*, 2023. doi: 10.1101/2023.12.20.572268.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Gergely Nagy and Laszlo Nagy. Motif grammar: The basis of the language of gene expression. *Computational and Structural Biotechnology Journal*, 18:2026–2032, Jul 2020. ISSN 2001-0370. doi: 10.1016/j.csbj.2020.07.007. URL https://doi.org/10.1016/j.csbj.2020.07.007. eCollection 2020.

Daniel E. Newburger and Martha L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Research*, 37(suppl$_1$) : $D77 - -D82, 102008. ISSN 0305 - 1048. doi$: .

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, oct 2022.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 10.1080/14786440109462720.

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. 10.1101/gr.277246.122. URL http://genome.cshlp.org/content/33/3/346.abstract.

Ryan Rickels and Ali Shilatifard. Enhancer logic and mechanics in development and disease. *Trends in Cell Biology*, 28(8):608–630, 2018. ISSN 0962-8924. 10.1016/j.tcb.2018.04.003. URL https://doi.org/10.1016/j.tcb.2018.04.003.

Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, Apr 2014. ISSN 1471-0064. 10.1038/nrg3682. URL https://doi.org/10.1038/nrg3682.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv*, 2017.

François Spitz and Eileen E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012. ISSN 1471-0064. 10.1038/nrg3207.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv*, 2017.

Ibrahim I. Taskiran, Katina I. Spanier, Hannah Dickmänken, Niklas Kempynck, Alexandra Pančíková, Eren Can Ekşi, Gert Hulselmans, Joy N. Ismail, Koen Theunis, Roel Vandepoel, Valerie Christiaens, David Mauduit, and Stein Aerts. Cell-type-directed design of synthetic enhancers. *Nature*, 626(7997):212–220, 2024. ISSN 1476-4687. 10.1038/s41586-023-06936-2.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. 10.1080/01621459.1963.10500845.

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

Jin Zhou, Geng Bo Chen, Yew Chung Tang, Rohit Anthony Sinha, Yonghui Wu, Chui Sun Yap, Guihua Wang, Junbo Hu, Xianmin Xia, Patrick Tan, Liang Kee Goh, and Paul Michael Yen. Genetic and bioinformatic analyses of the expression and function of pi3k regulatory subunit pik3r3 in an asian patient gastric cancer library. *BMC Medical Genomics*, 5(1):34, 2012. ISSN 1755-8794. 10.1186/1755-8794-5-34. URL https://doi.org/10.1186/1755-8794-5-34.

James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18, jan 2019. ISSN 1546-1718. 10.1038/s41588-018-0295-5. URL https://doi.org/10.1038/s41588-018-0295-5.