

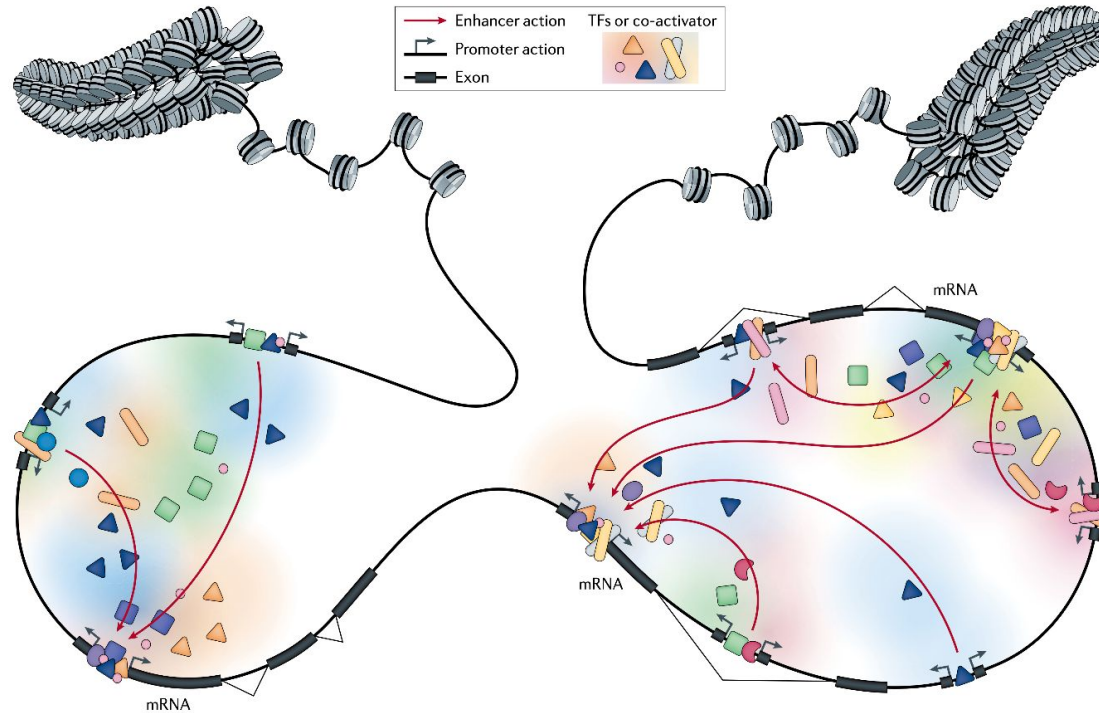


Decoding the Mechanistic Impact of Genetic Variation on Regulatory Sequences with Deep Learning

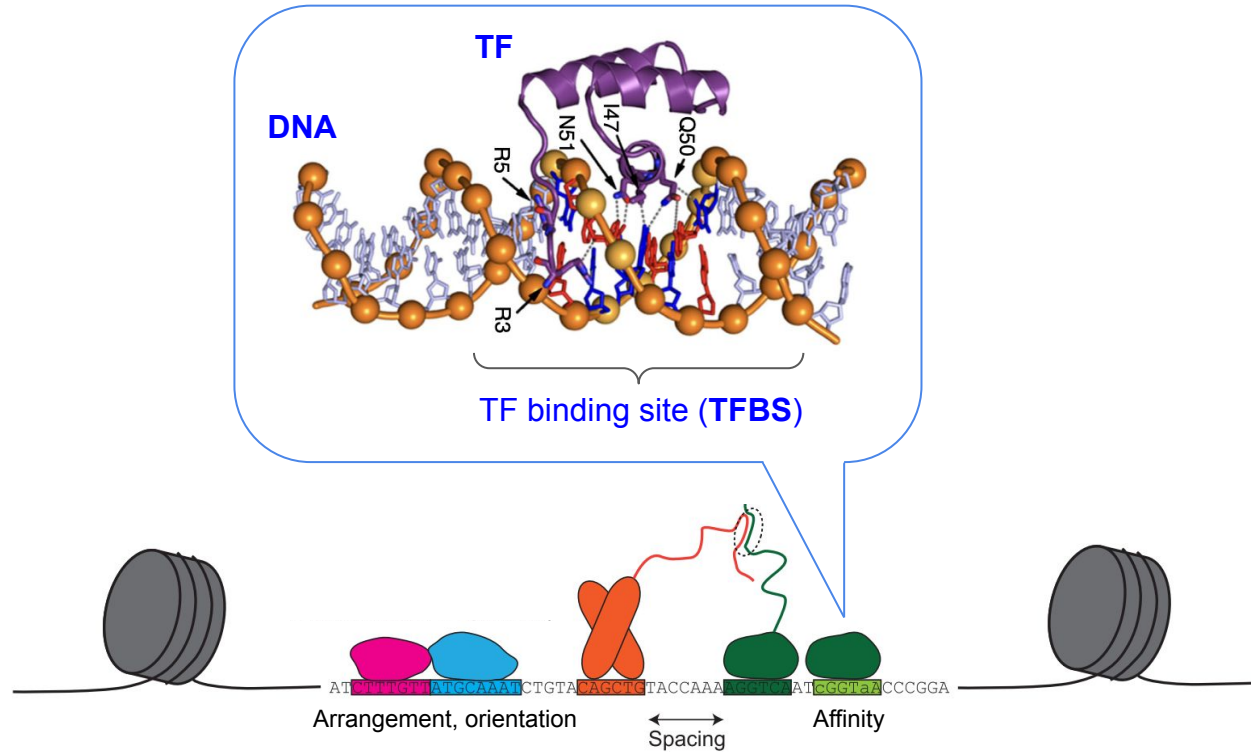
Presented by **Evan Seitz**

Koo Lab & Kinney Lab
Cold Spring Harbor Laboratory
April 27, 2025

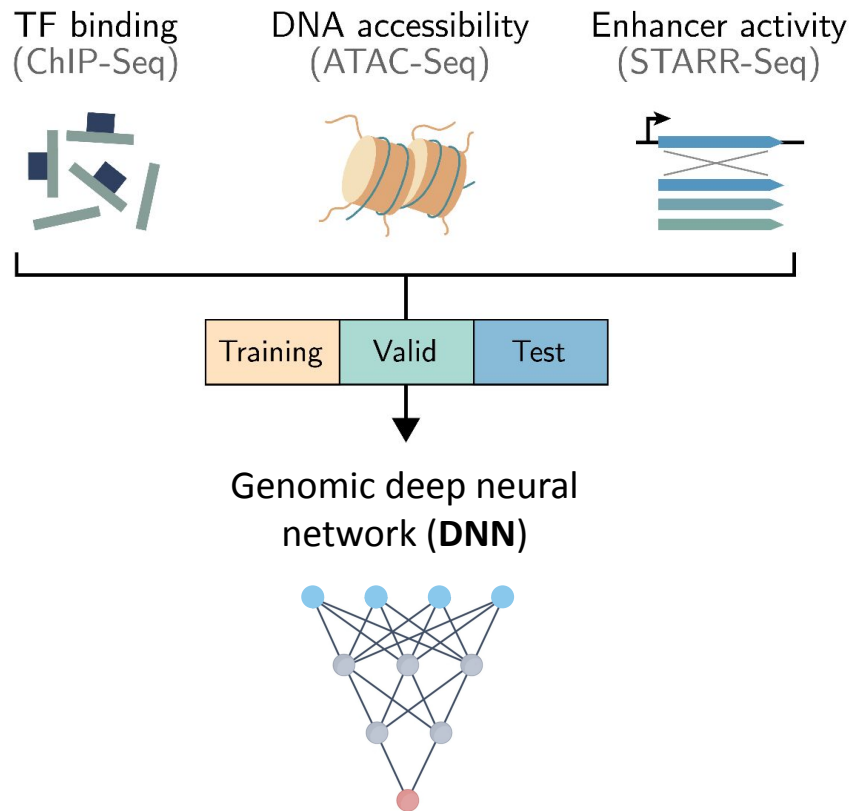
How cells control gene expression with precision



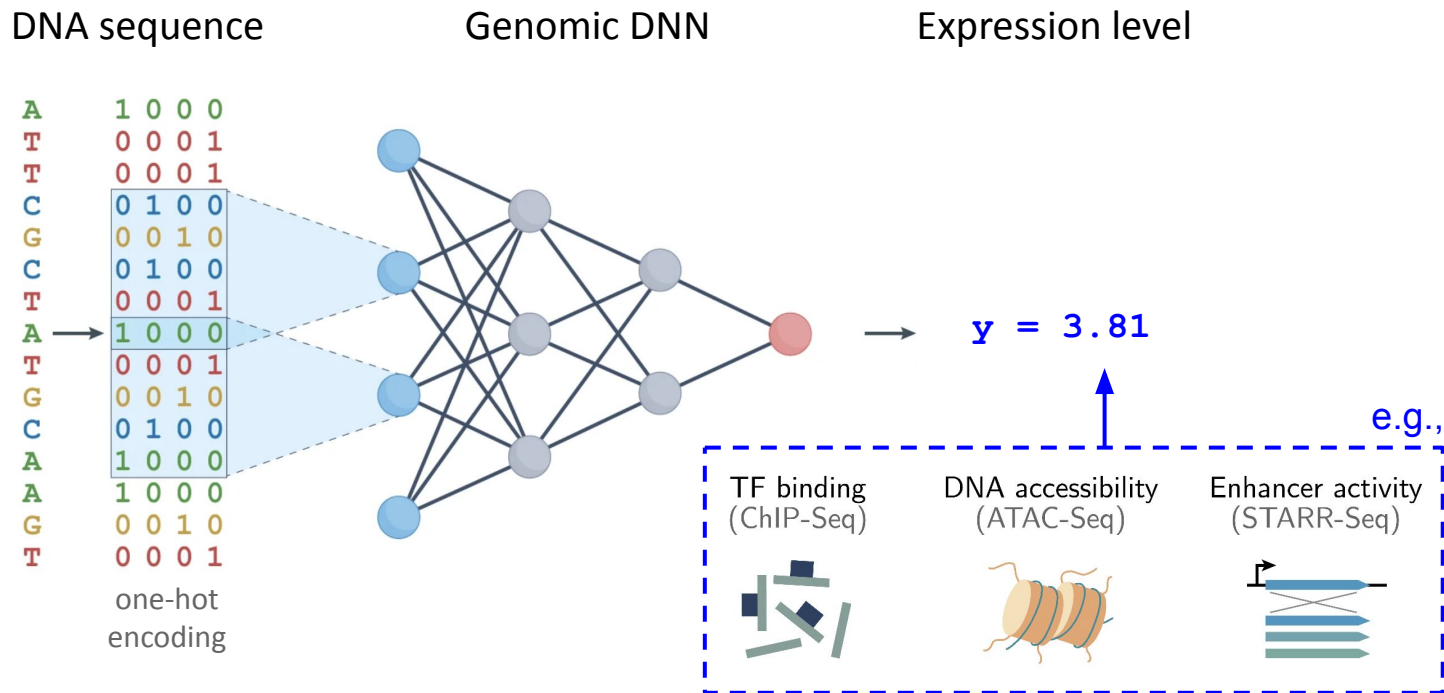
The cis-regulatory code remains a major unsolved problem

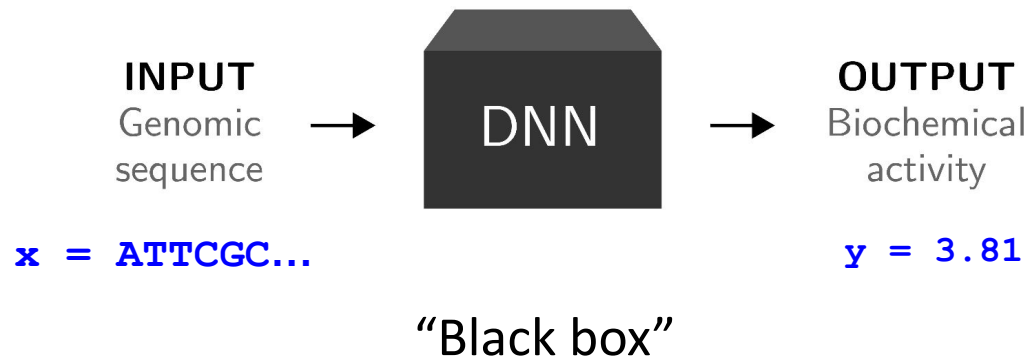


DNNs can learn the cis-regulatory code from functional genomics data

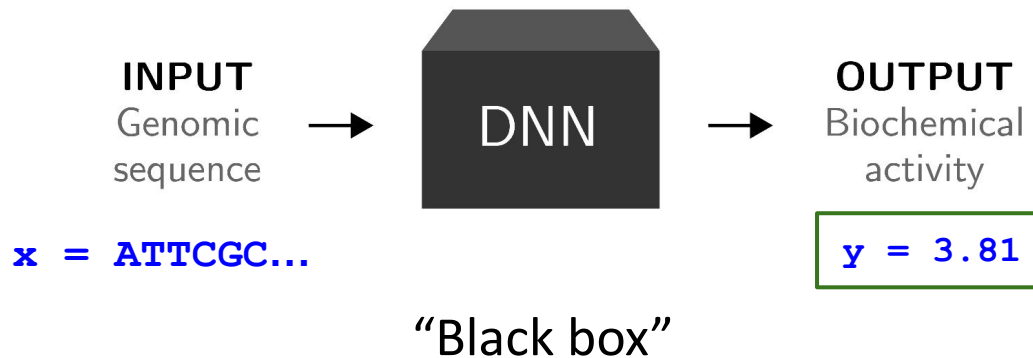


DNNs generalize to predict regulatory activity on new sequences



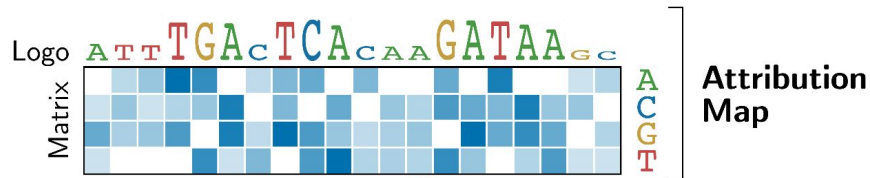
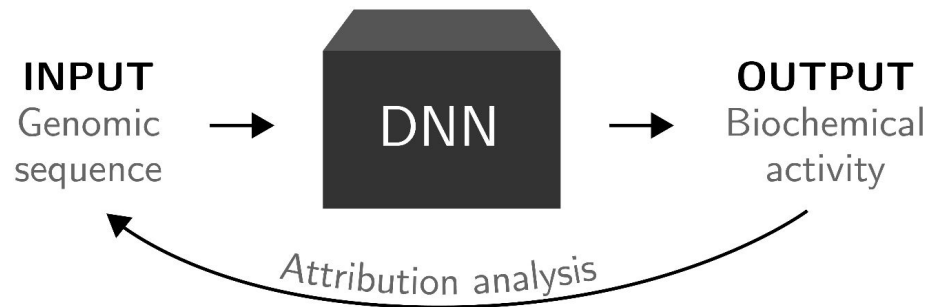


But we don't yet understand why the model makes its predictions

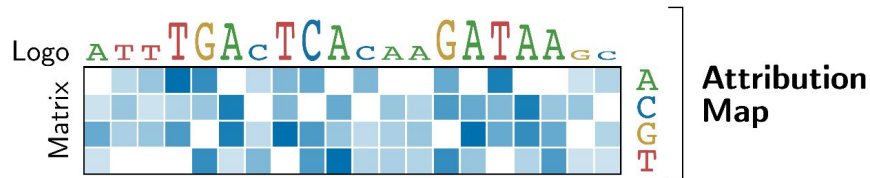
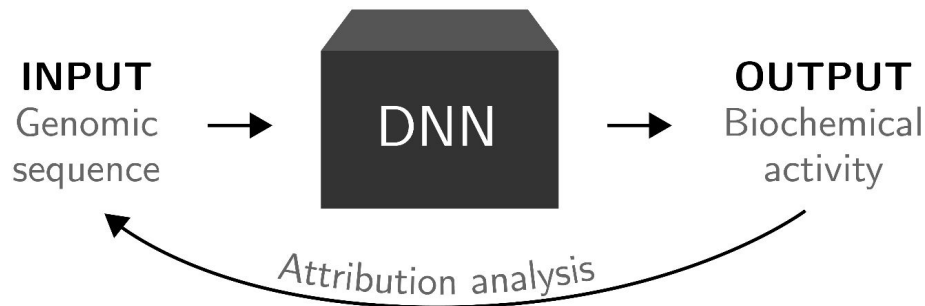


“Why does $y = 3.81$ for this sequence?”

But we don't yet understand why the model makes its predictions



Attribution maps help reveal which sequence features drive predictions



***In silico* mutagenesis (ISM)**

Zhou *et al.*, 2015

Saliency maps

Simonyan *et al.*, 2014

Integrated gradients

Sundararajan *et al.*, 2017

SmoothGrad

Smilkov *et al.*, 2017

DeepSHAP

Lundberg *et al.*, 2017

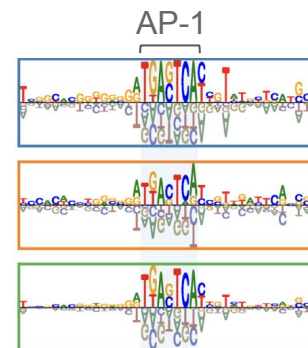
SQuID

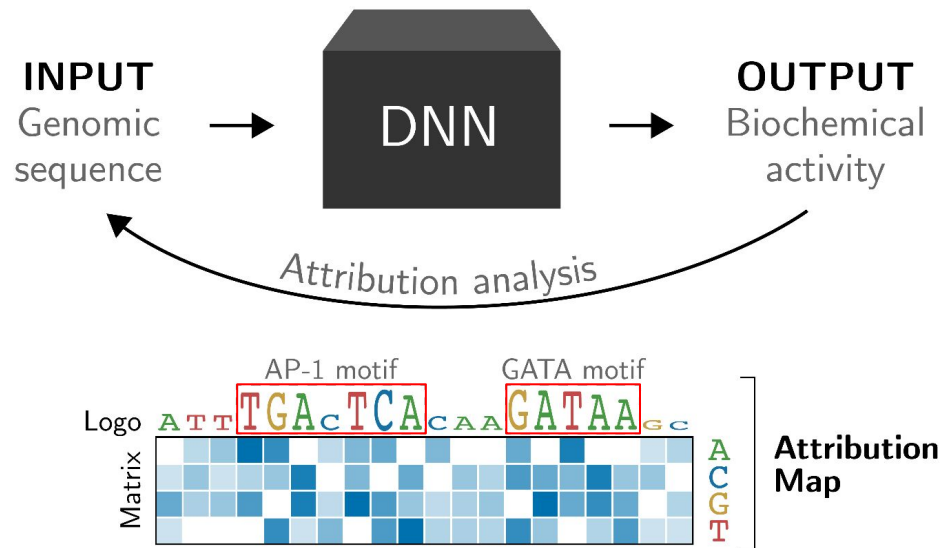
Seitz *et al.*, 2024

ISM

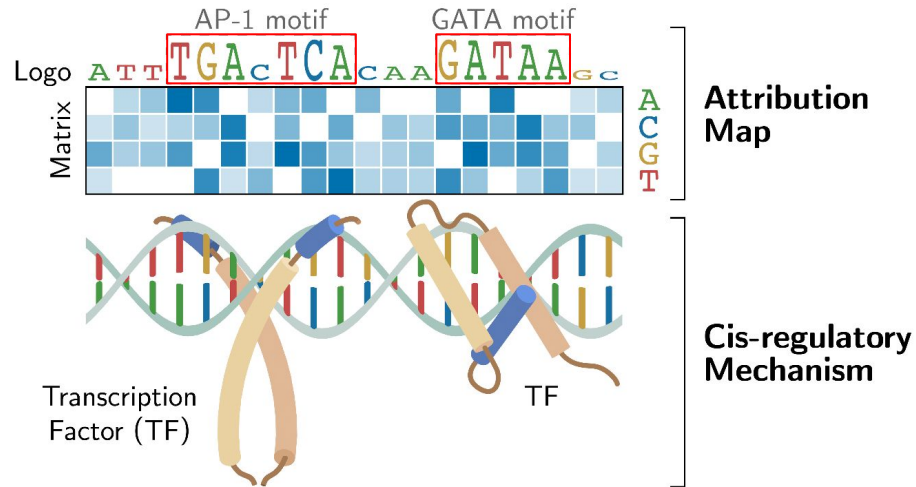
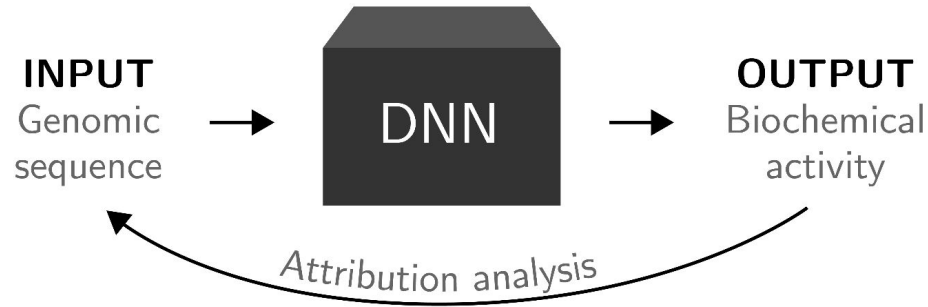
Saliency

SQuID





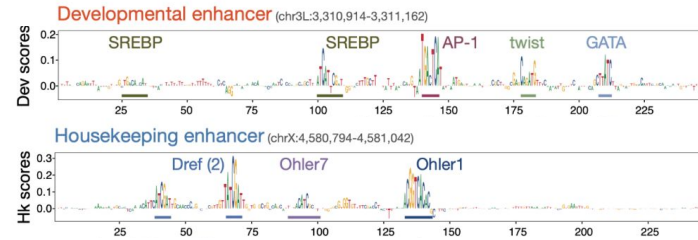
Attribution-based motifs reflect known TF binding sites



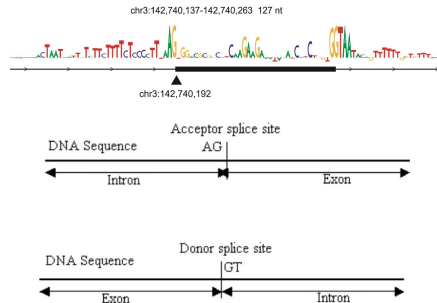
Attribution maps suggest a specific mechanistic hypothesis

Attribution has helped uncover a diversity of regulatory motifs

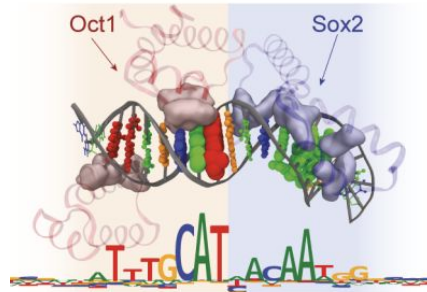
DeepSTARR



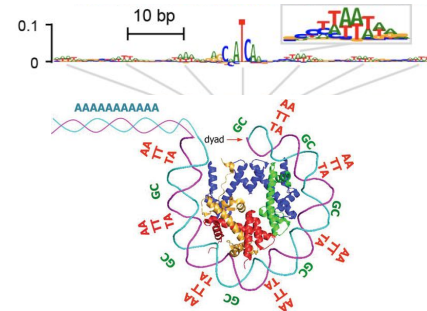
SpliceAI



BPNet



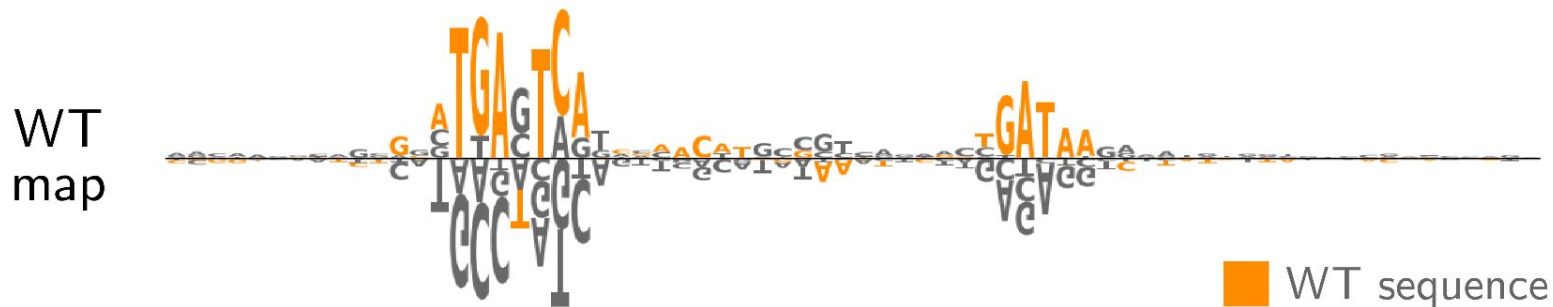
BPNet



Attribution maps don't capture how mechanisms change

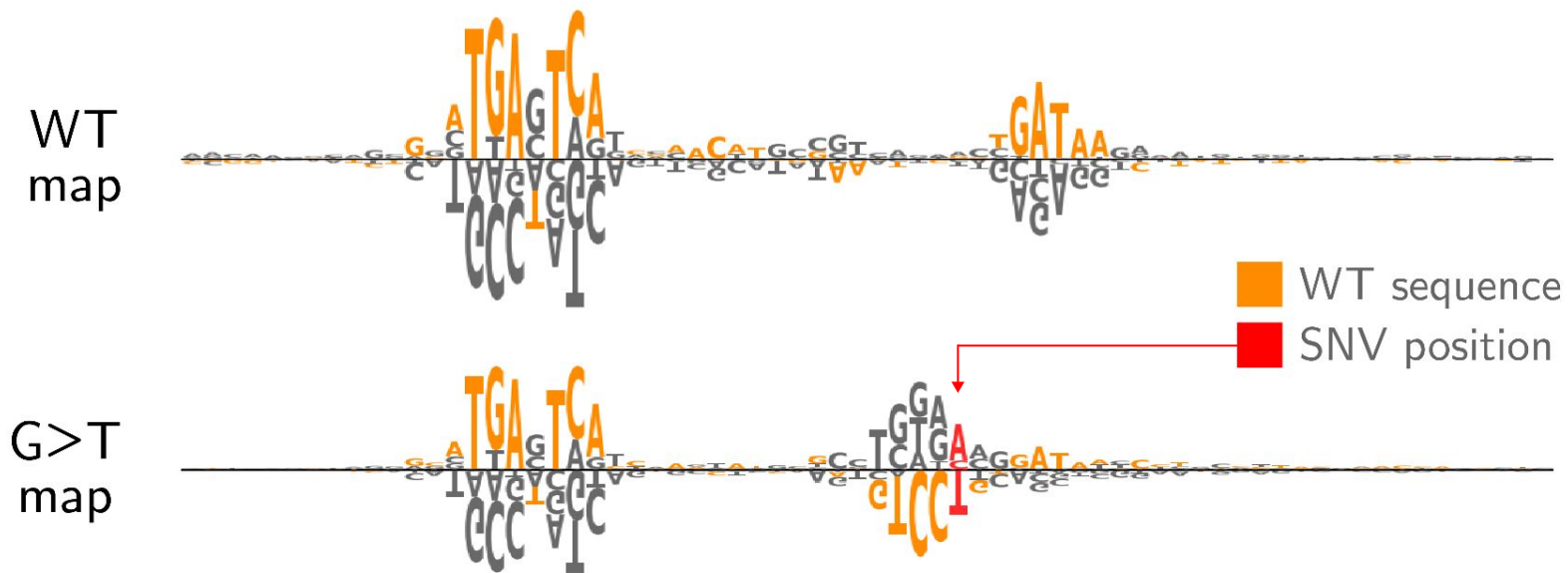


Attribution maps don't capture how mechanisms change

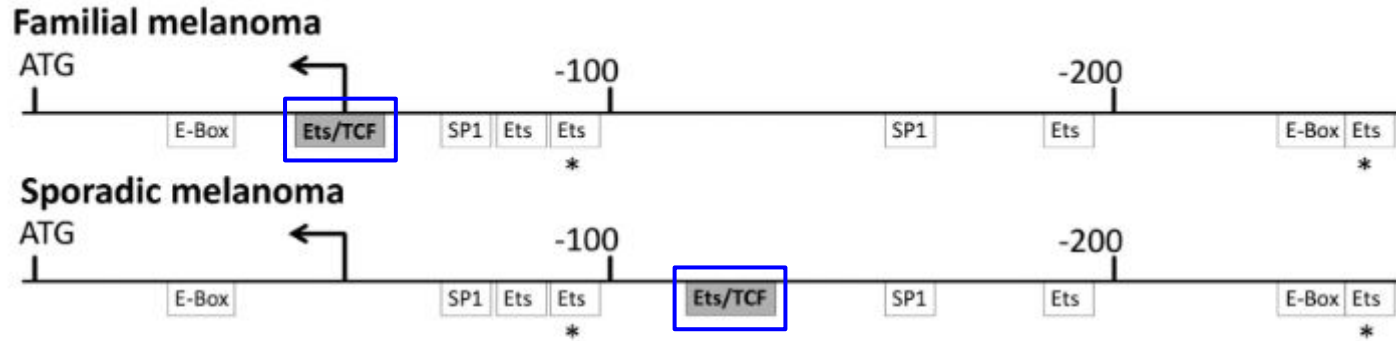


"What kinds of binding sites are allowed?"
"How will mutations change them?"

Key mutations can rewire the underlying regulatory mechanism

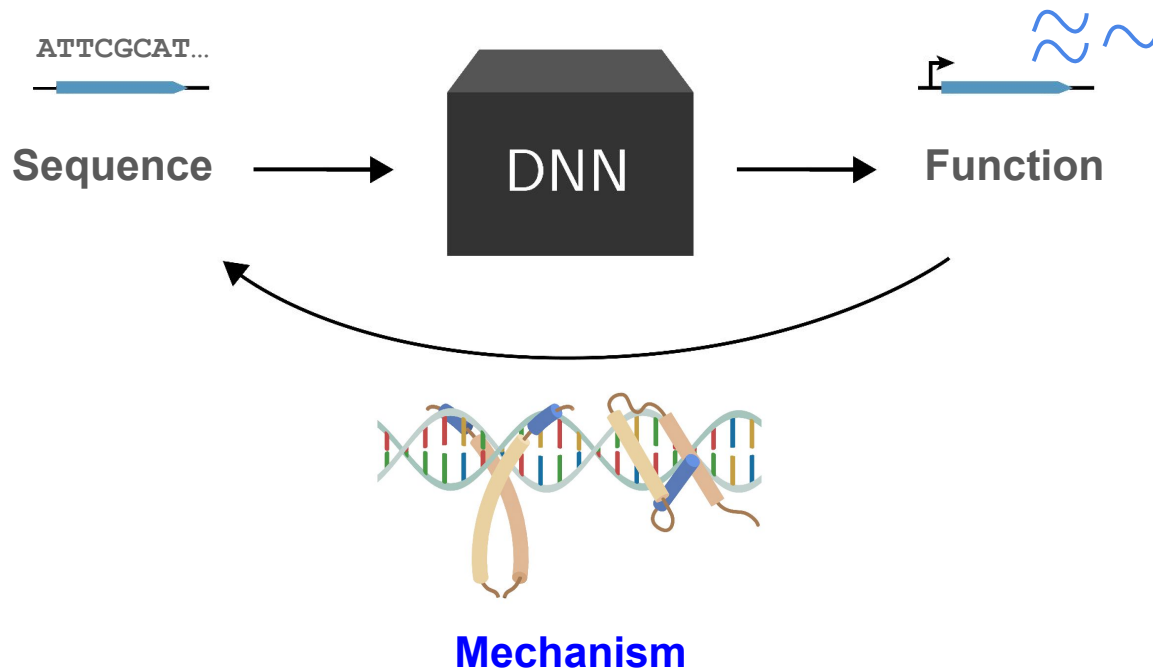


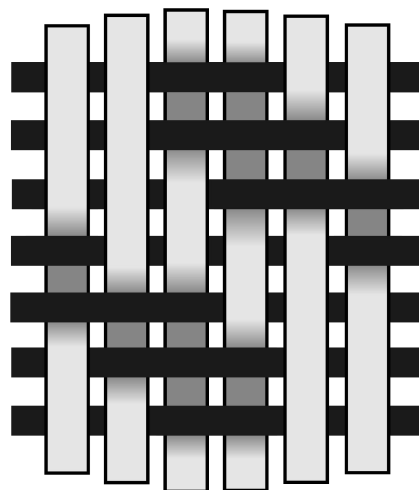
Regulatory mutations can drive evolution, disease, or dysfunction



In a study by Horn *et al.*, 70% of melanomas harbored one of two specific mutations in the promoter region of TERT

We need a systematic way to uncover sequence–mechanism relationships





SEAM

Systematic **E**xplanation of **A**tribution-based **M**echanisms

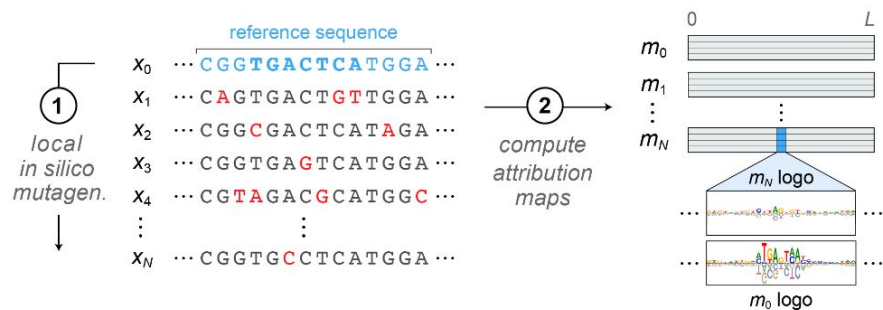
SEAM Framework



1. Generate in silico mutagenesis library

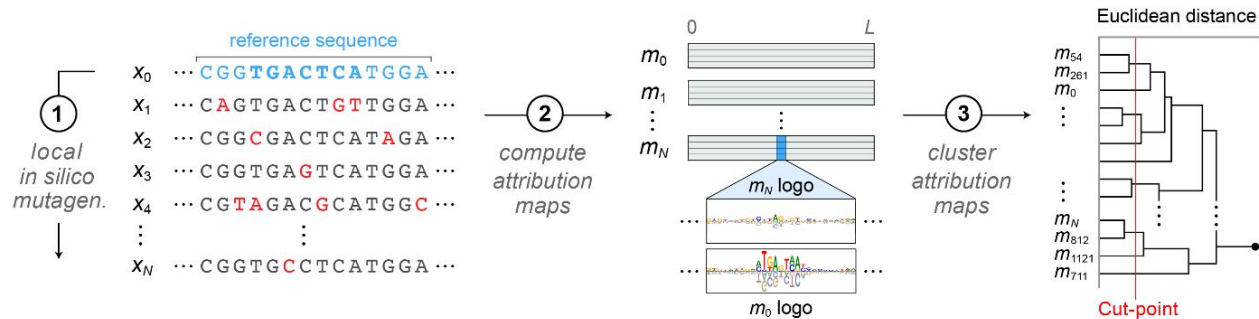
- Local library {10% mutation rate, 100k sequences} and others

SEAM Framework



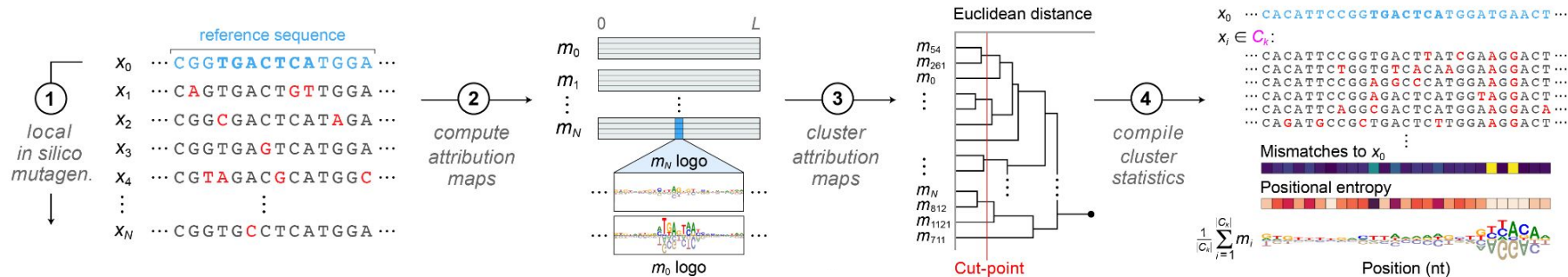
1. Generate in silico mutagenesis library
 - Local library {10% mutation rate, 100k sequences} and others
2. Compute attributions maps
 - DeepSHAP, Saliency Maps

SEAM Framework



1. Generate in silico mutagenesis library
 - Local library {10% mutation rate, 100k sequences} and others
2. Compute attributions maps
 - DeepSHAP, Saliency Maps
3. Cluster attribution maps
 - Hierarchical clustering

SEAM Framework



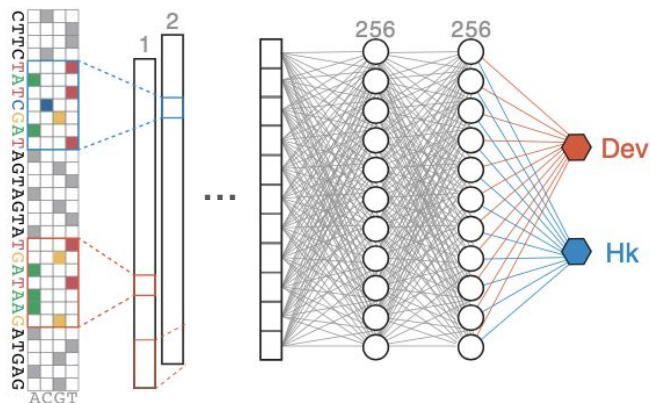
1. Generate in silico mutagenesis library
 - Local library {10% mutation rate, 100k sequences} and others
2. Compute attributions maps
 - DeepSHAP, Saliency Maps
3. Cluster attribution maps
 - Hierarchical clustering
4. Compile cluster statistics
 - Sequence summary matrix

SEAM Case Studies – Local libraries

1. DeepSTARR

- DNN prediction:** Enhancer activity measured using UMI-STARR-seq in *Drosophila* S2 cells

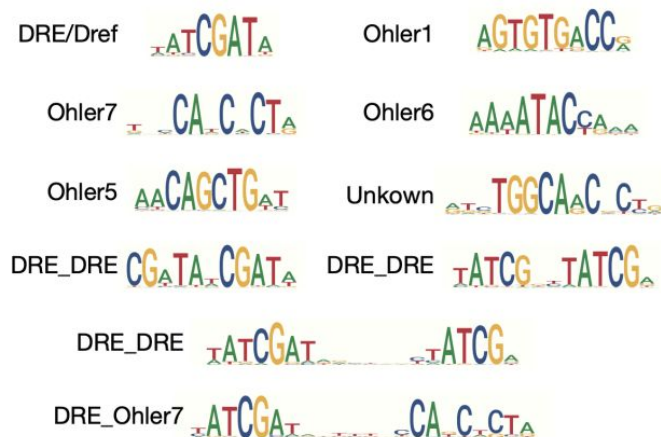
Model architecture



Input: 249-length sequence

Output: 2 scalars

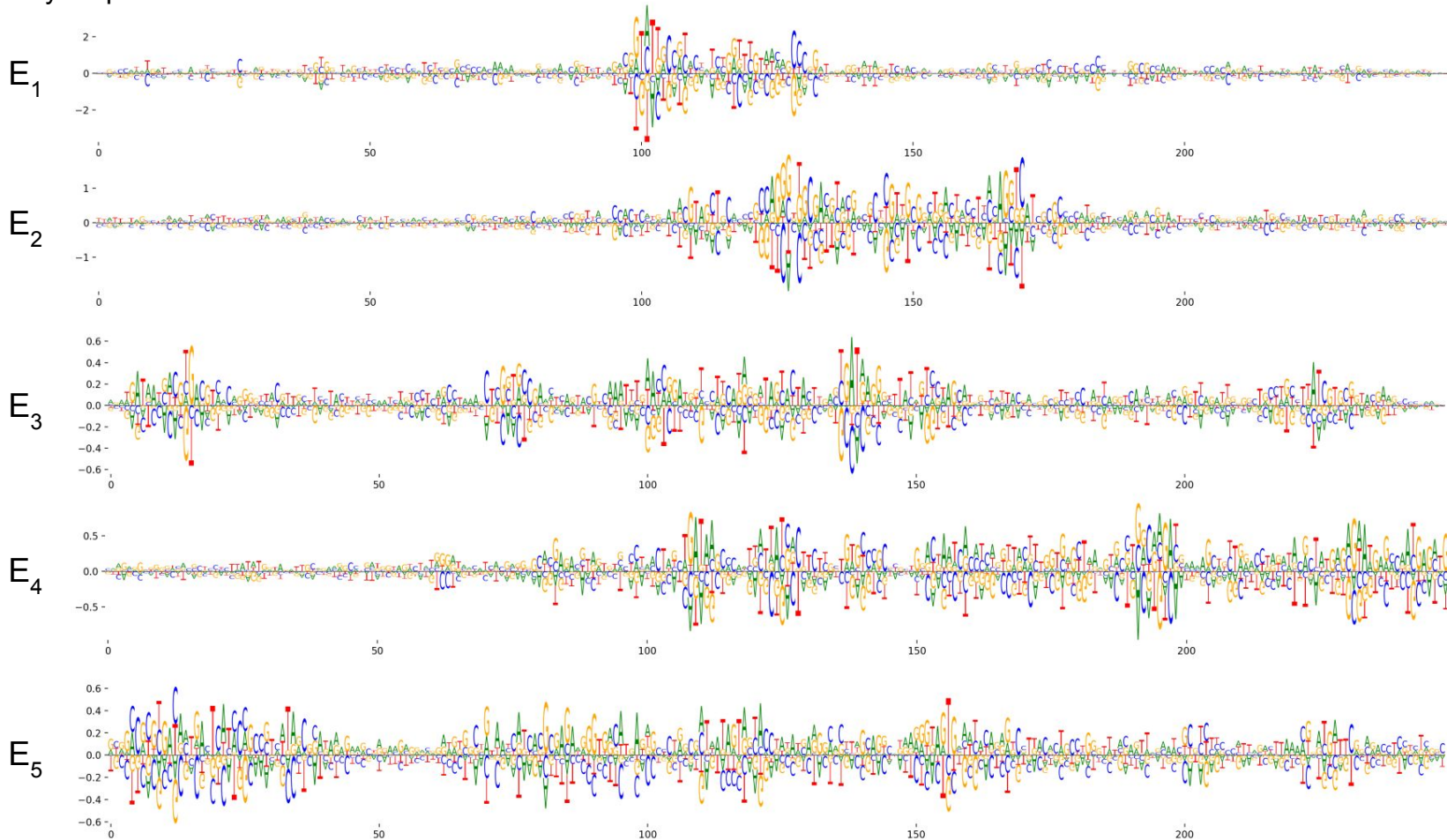
Genome-wide TFs



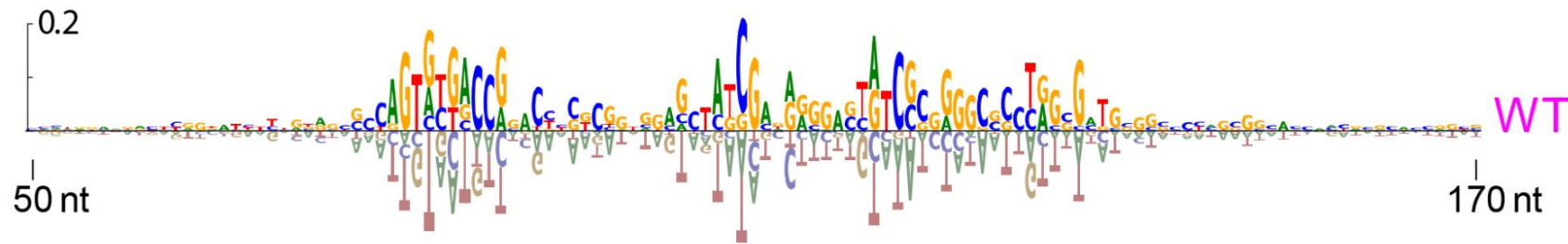
Drosophila enhancers (across the genome)

Saliency maps

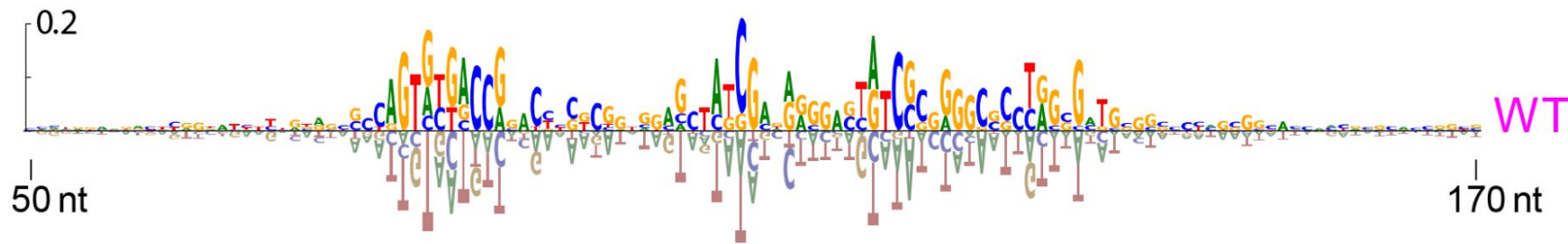
WT MAP



Drosophila enhancer
DeepSHAP

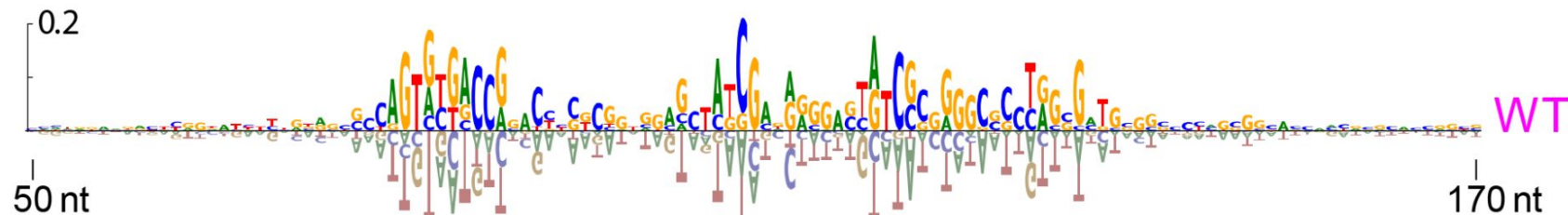


Drosophila enhancer
DeepSHAP



"Which motifs are present?"
"What is biological signal?"
"What is noise?"

Drosophila enhancer
DeepSHAP

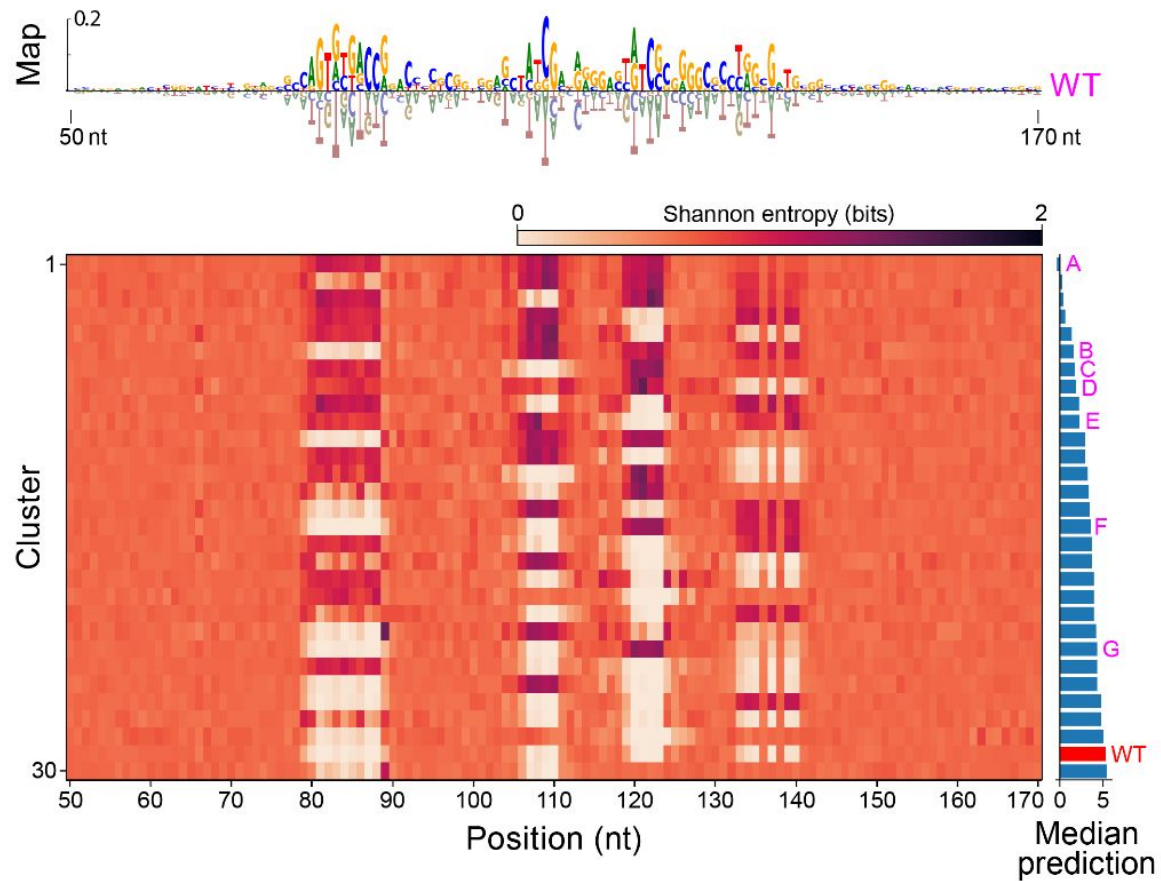


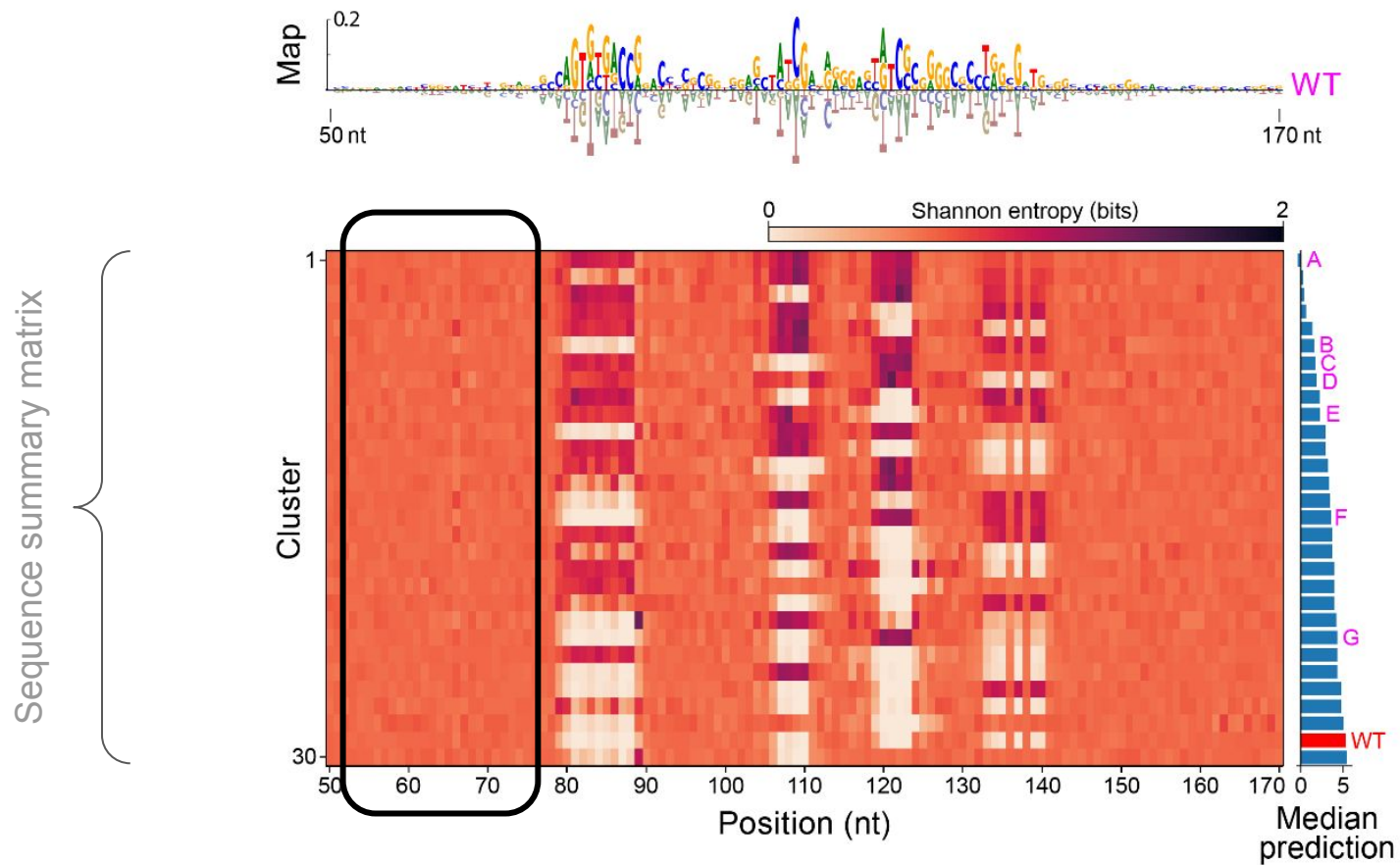
SEAM

AVG. CLUSTER 1

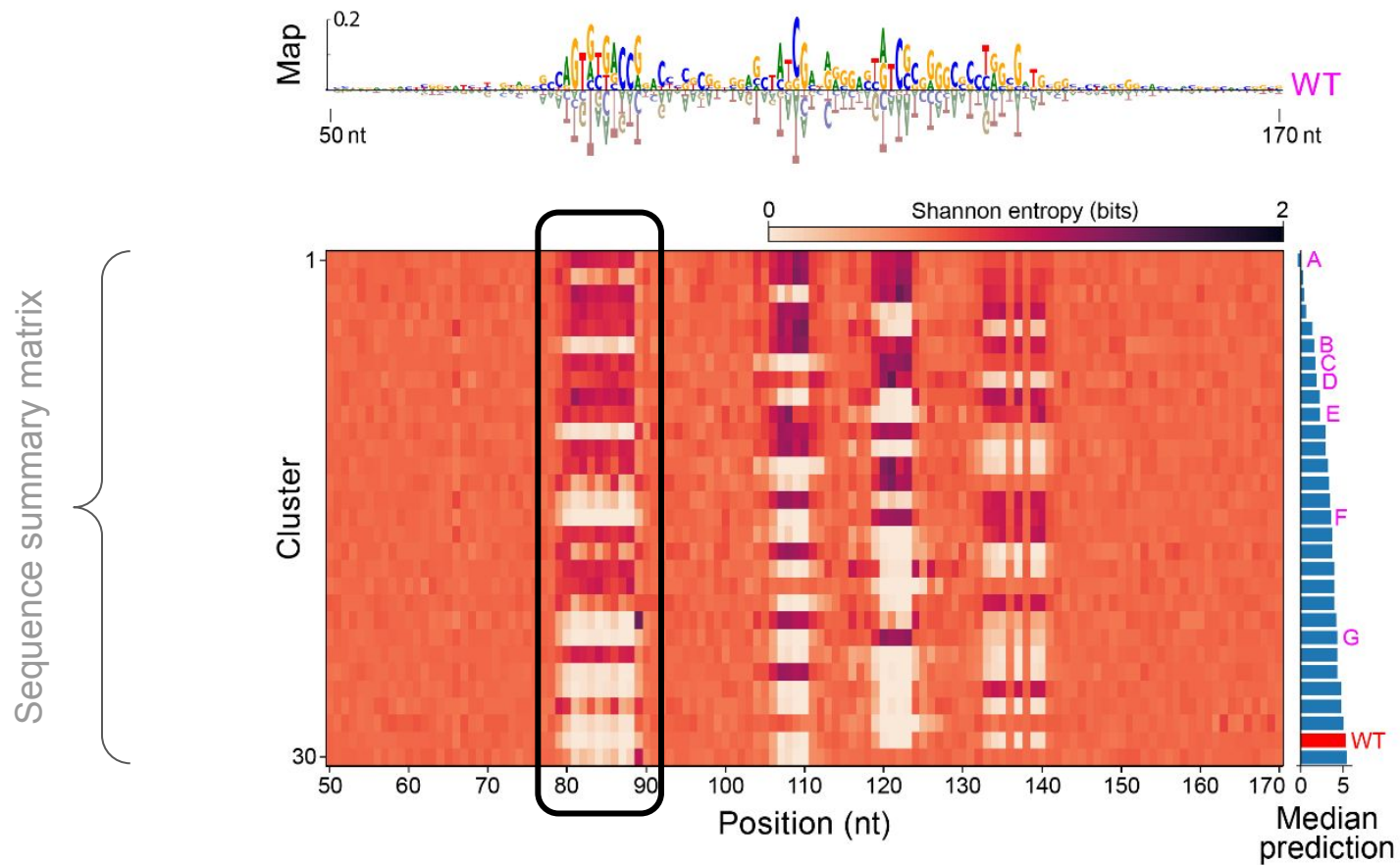


Sequence summary matrix

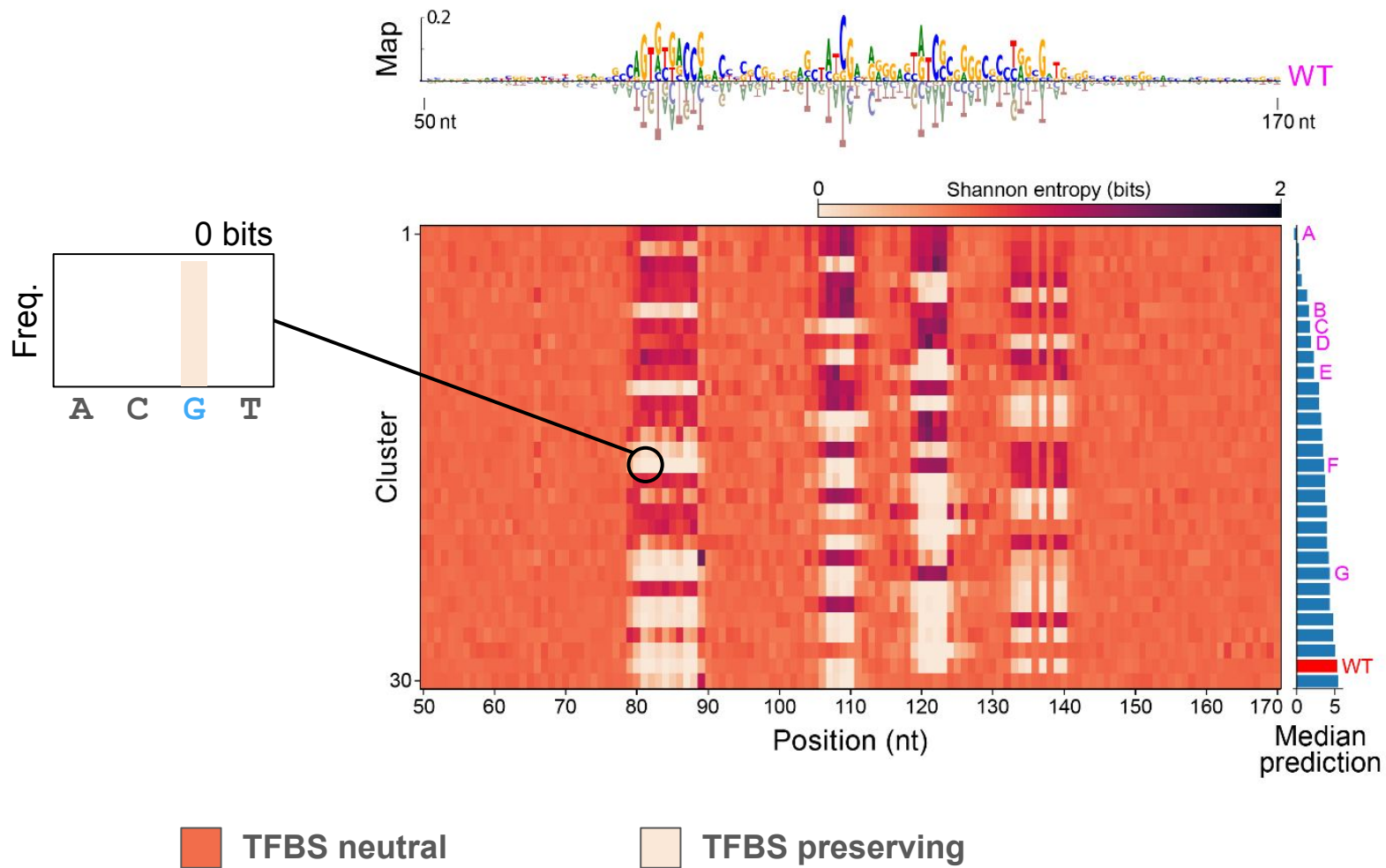


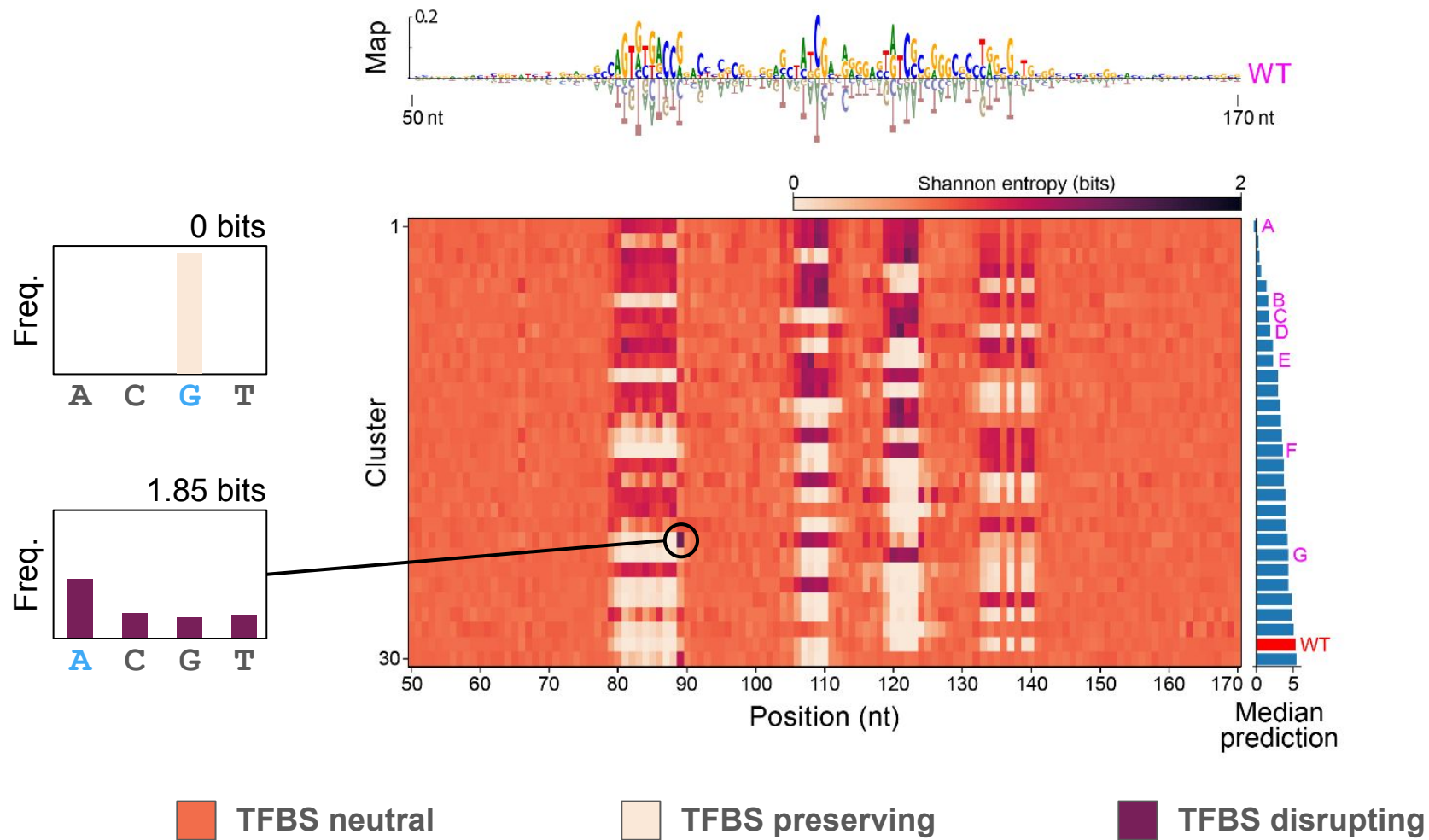


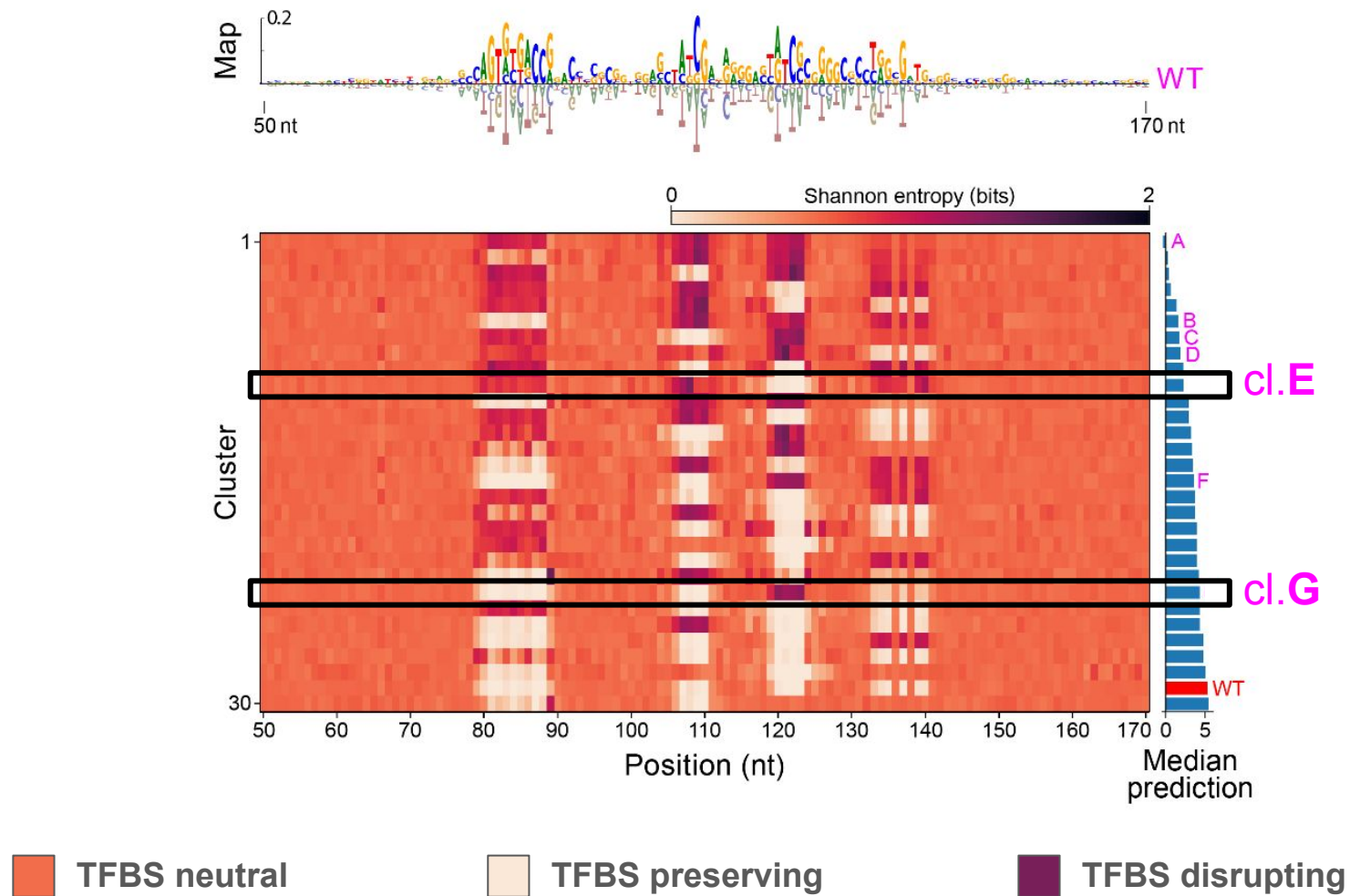
TFBS neutral → background entropy (~0.63 bits)

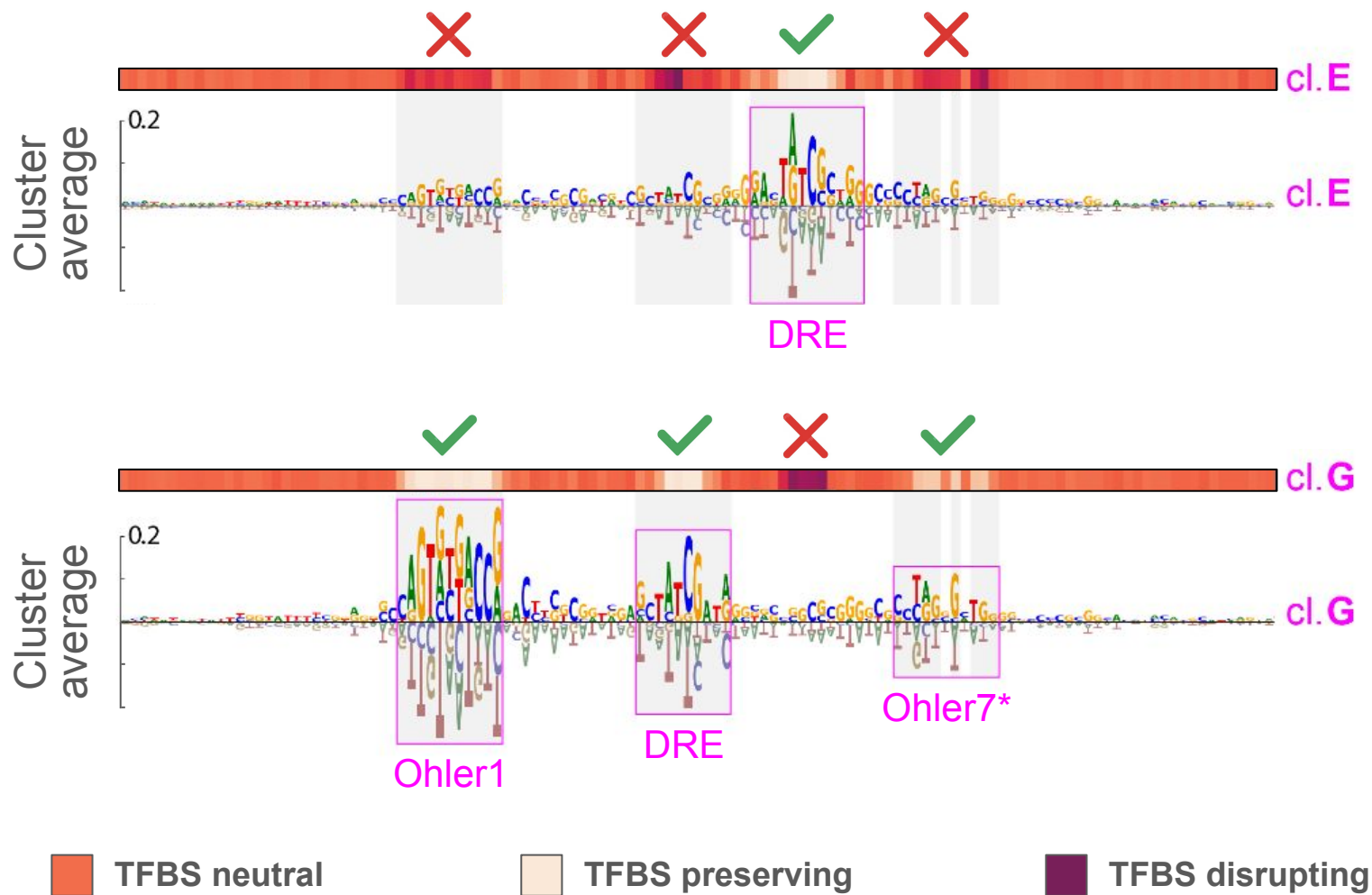


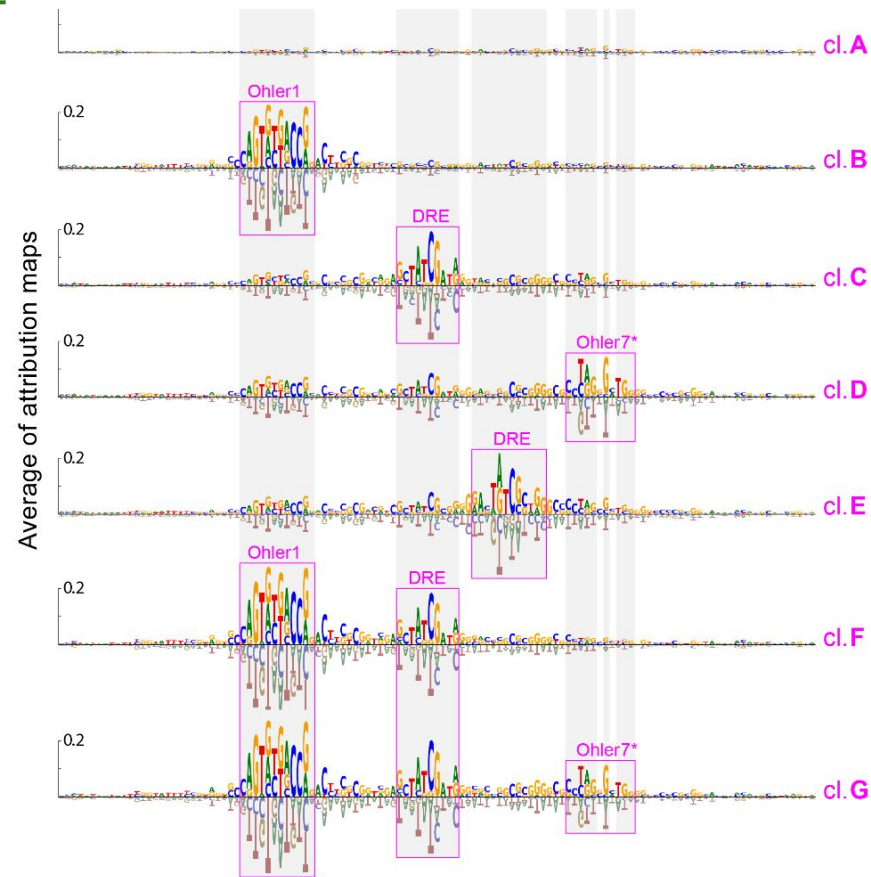
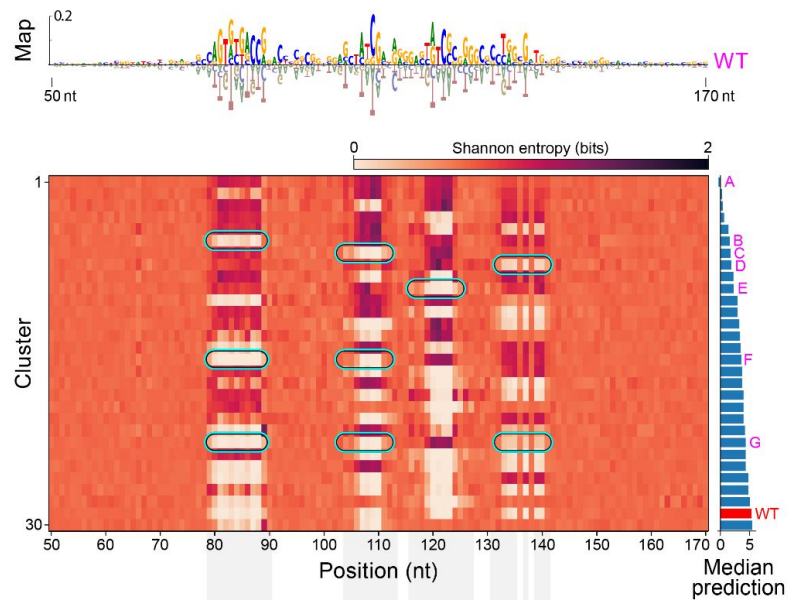
TFBS neutral → background entropy (~ 0.63 bits)

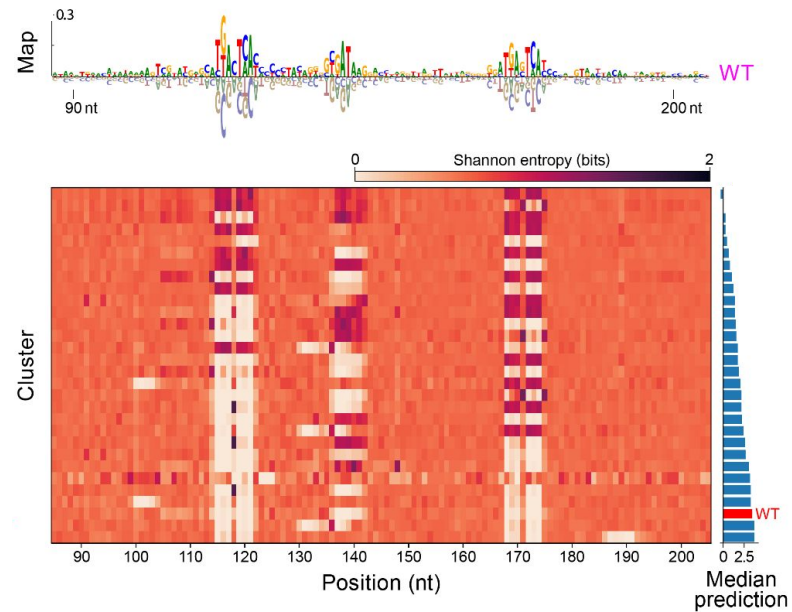












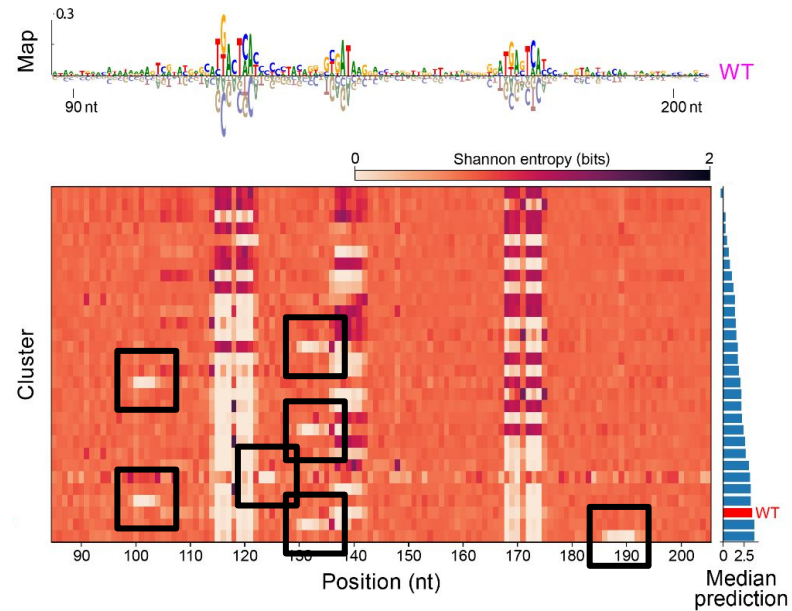
Different genomic locus



TFBS neutral

 **TFBS preserving**

■ TFBS disrupting



Different genomic locus



TFBS neutral

 TFBS preserving

■ **TFBS disrupting**

The top panel shows a genomic map of the WT region from 90 nt to 200 nt. The y-axis is labeled 'Map' with a scale from 0 to 0.3. The map displays various colored peaks representing different nucleotides. A red box highlights a specific region of the map. The bottom panel is a heatmap showing the distribution of mismatches to the WT sequence. The x-axis is labeled 'Position (nt)' and ranges from 90 to 200. The y-axis is labeled 'Cluster' and ranges from 0 to 2.5. A color scale at the top indicates the percentage of mismatches to the WT sequence, ranging from 0% (dark blue) to 100% (yellow). A red box highlights a specific region of the heatmap. A red bar at the bottom right indicates the 'Median prediction' for the WT sequence.

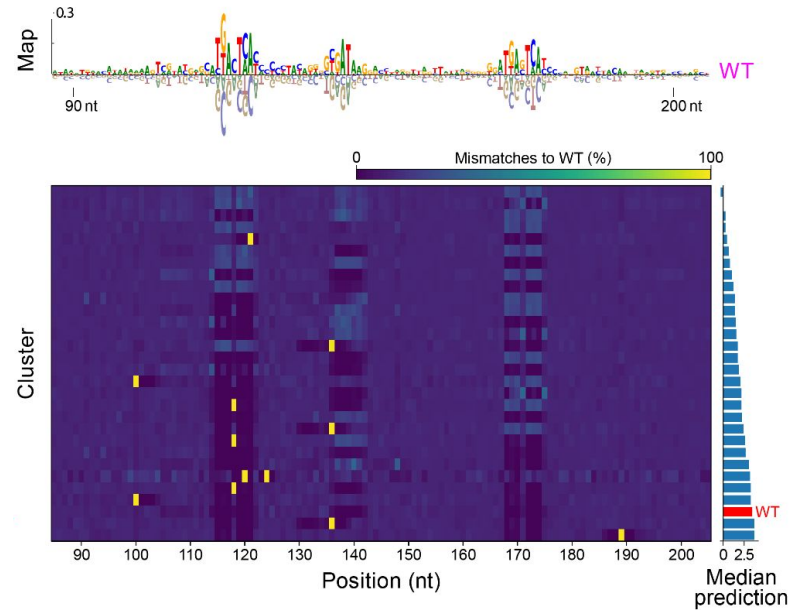
Colored by
Mismatches to WT (%)

The top panel shows a genomic map of the WT genome. The y-axis is labeled 'Map' and ranges from 0 to 0.3. The x-axis is labeled 'Position (nt)' and ranges from 90 to 200. The map displays a series of colored lines representing different genomic features, with a prominent peak around 120 nt. The label 'WT' is in the top right corner.

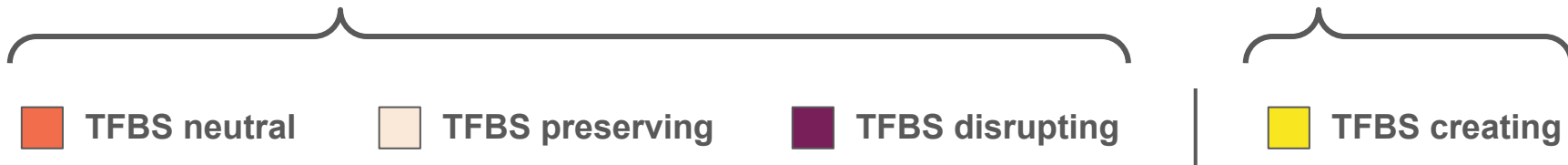
The bottom panel is a heatmap showing the distribution of mismatches to the WT genome. The y-axis is labeled 'Cluster' and the x-axis is labeled 'Position (nt)' and ranges from 90 to 200. A color scale at the top indicates 'Mismatches to WT (%)' from 0 (dark purple) to 100 (yellow). Several clusters are highlighted with pink circles, indicating regions of high mismatch. A vertical bar on the right side of the heatmap shows the 'Median prediction' for each cluster, with a color scale from 0 to 2.5. The label 'WT' is in the bottom right corner.

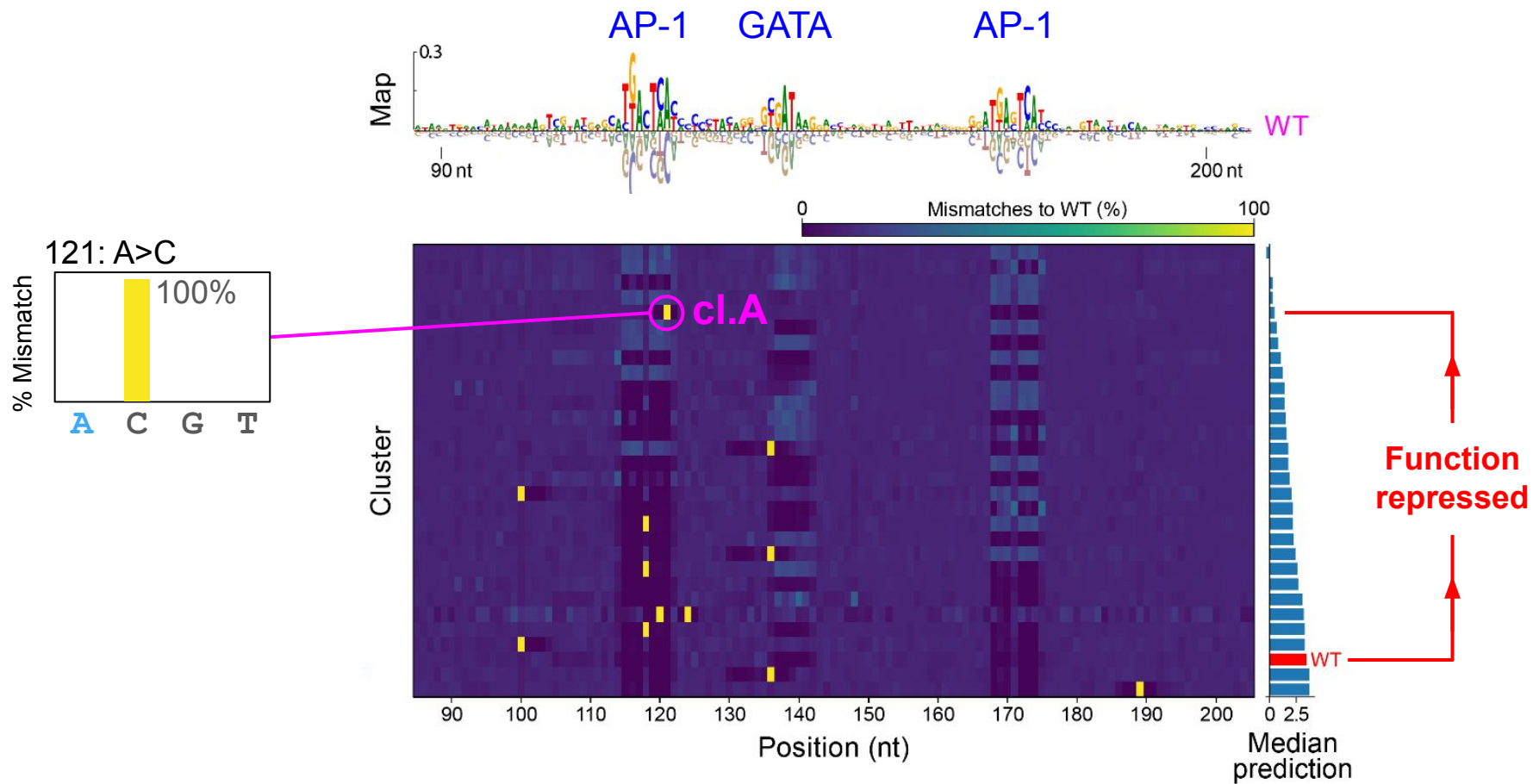
Colored by
Mismatches to WT (%)

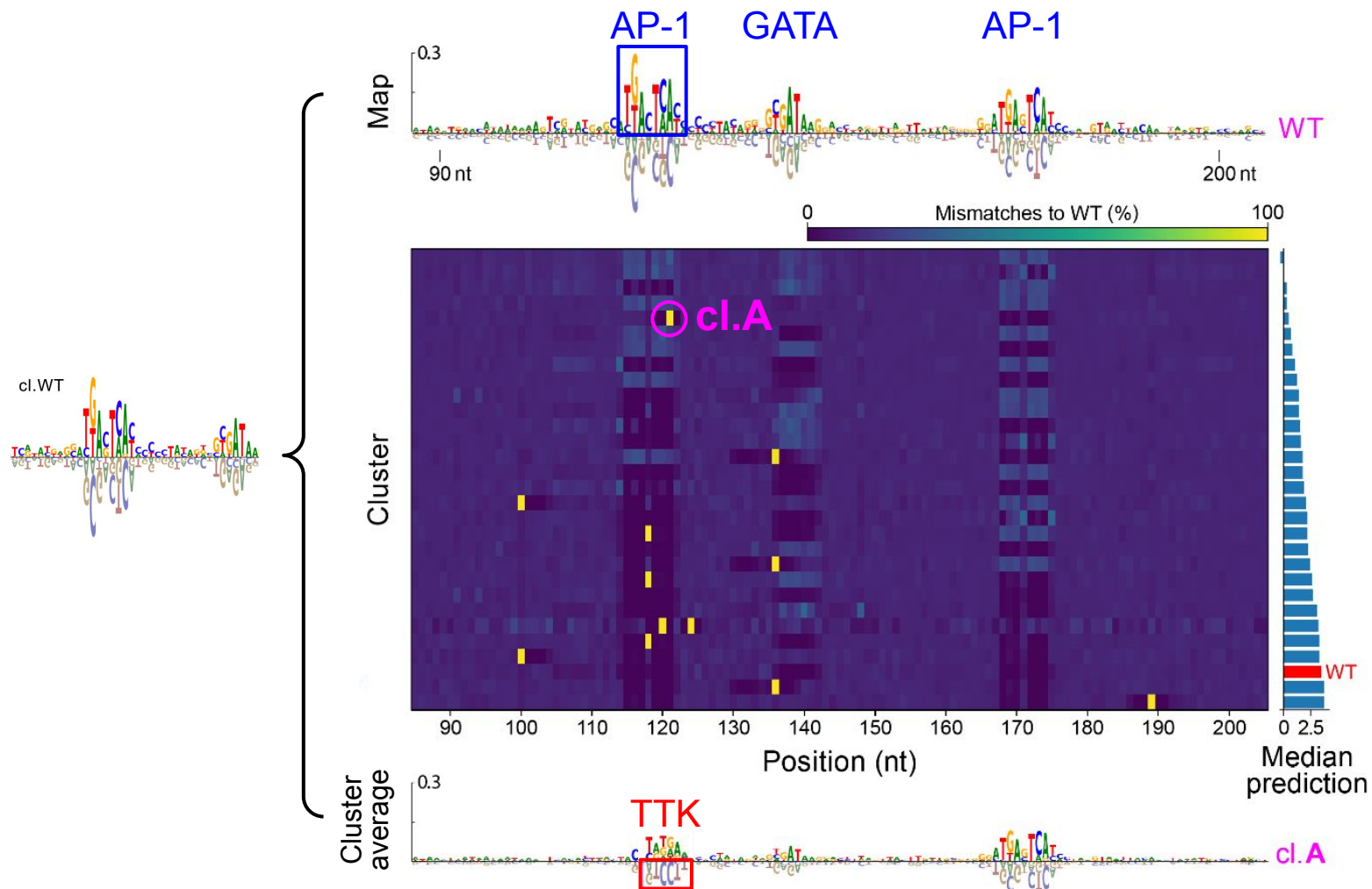
Colored by
Shannon entropy (bits)

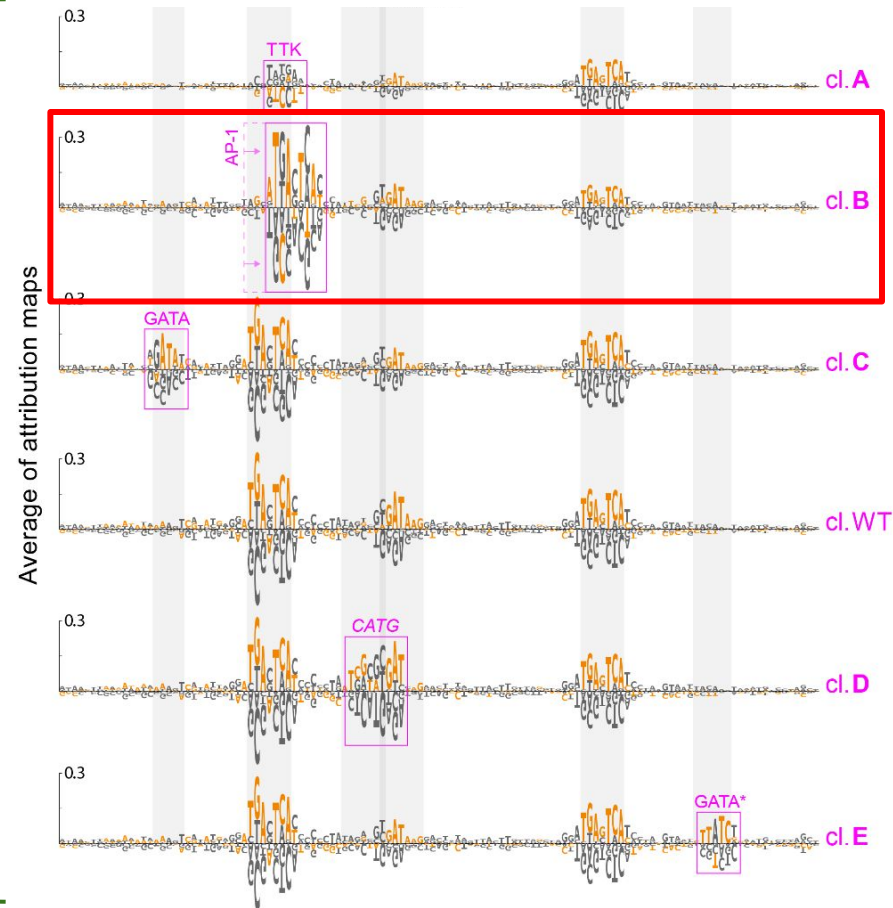
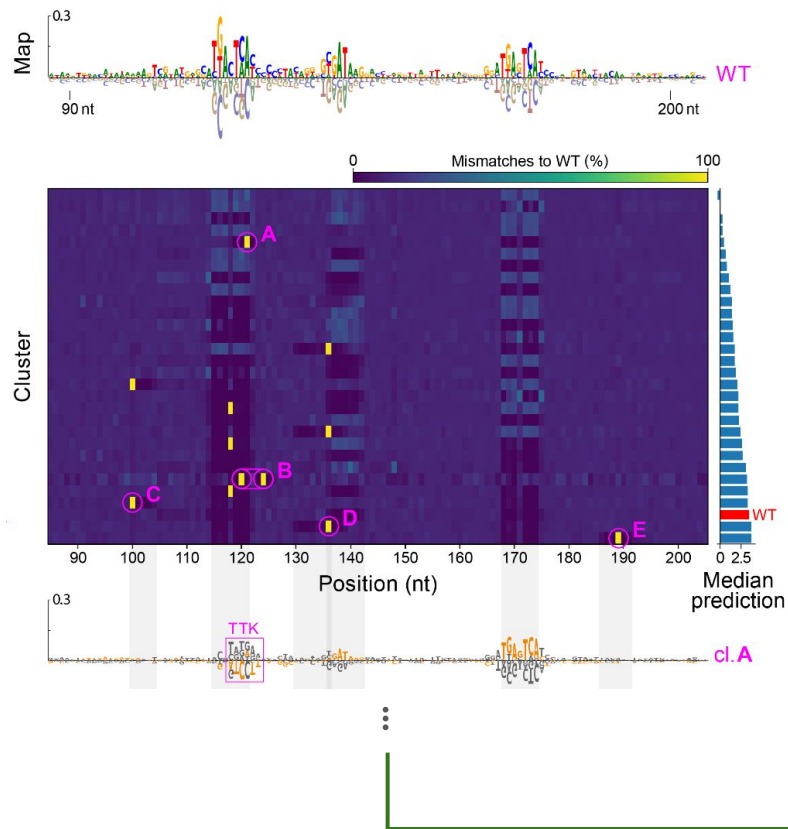


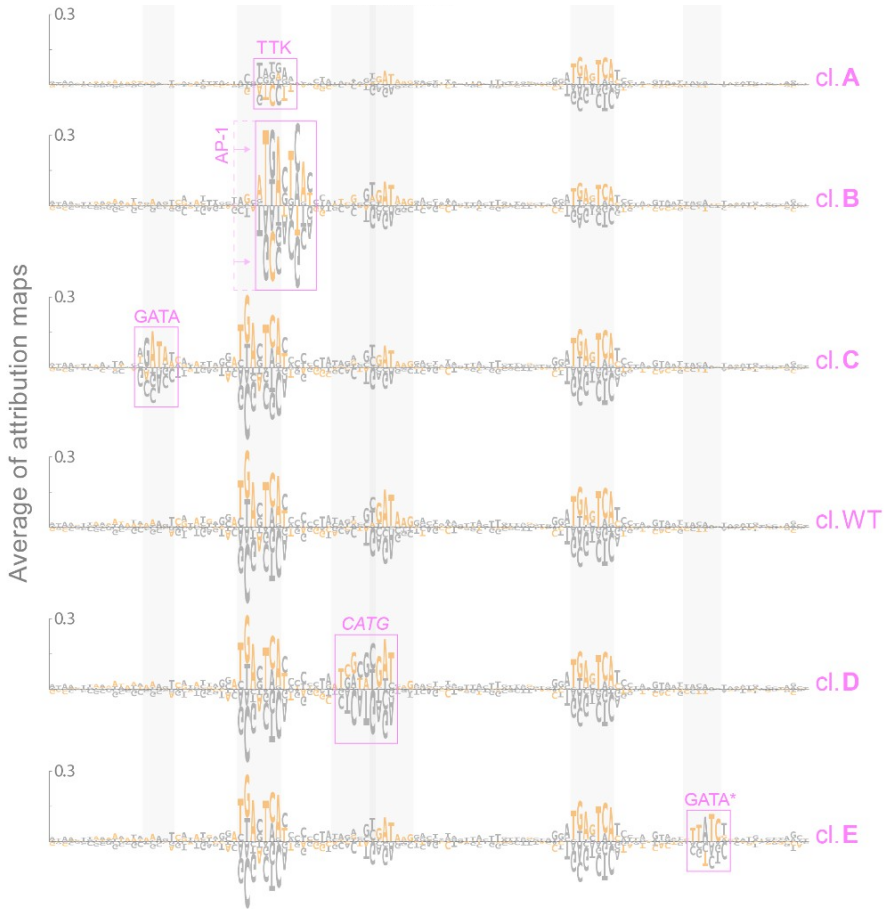
Colored by
Mismatches to WT (%)



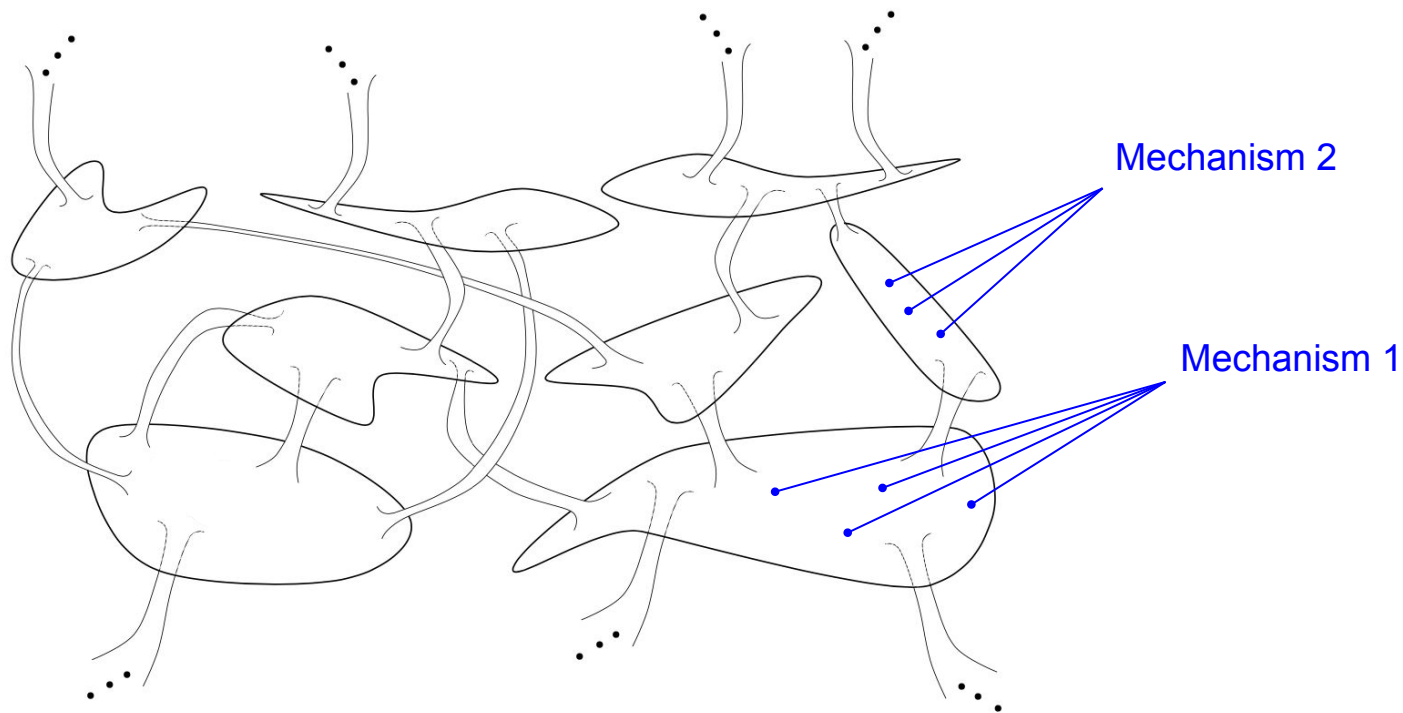




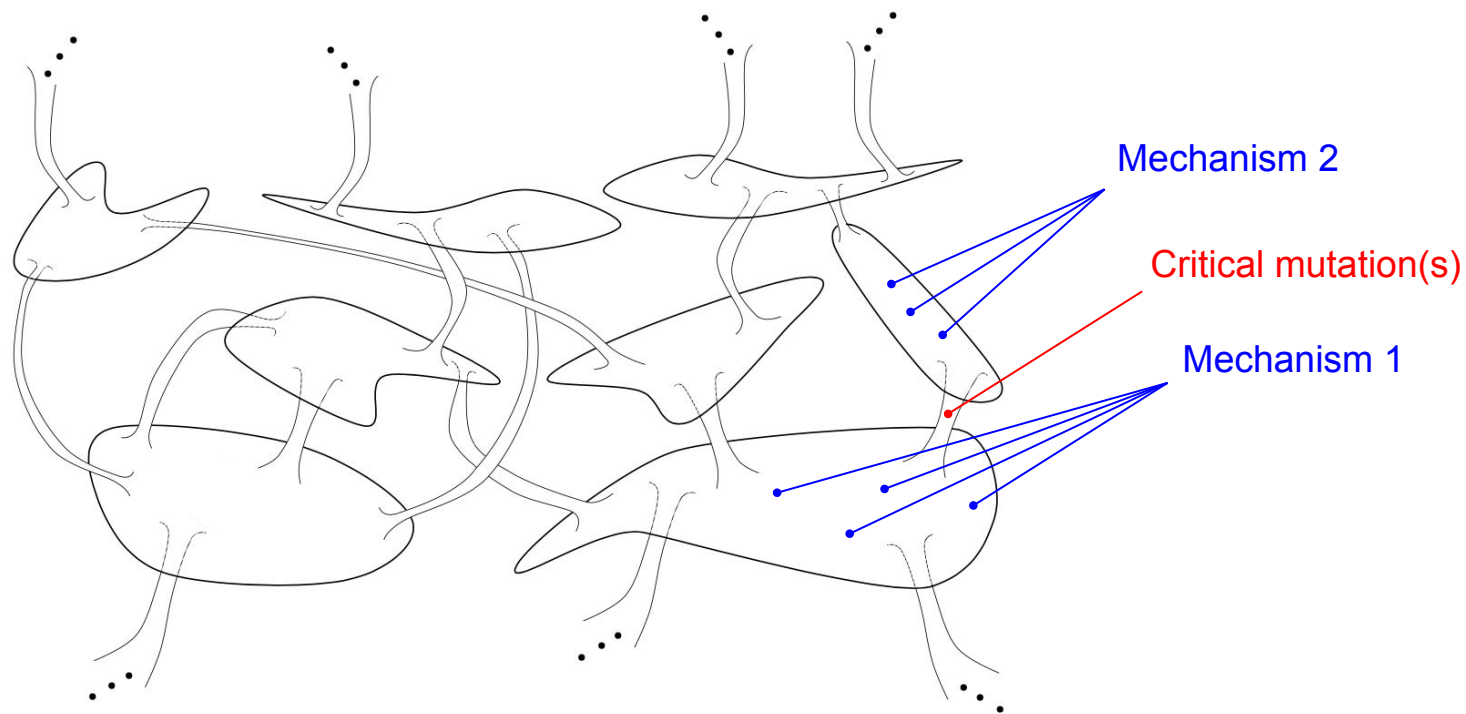




These results reflect the evolutionary dynamics of sequence space



These results reflect the evolutionary dynamics of sequence space



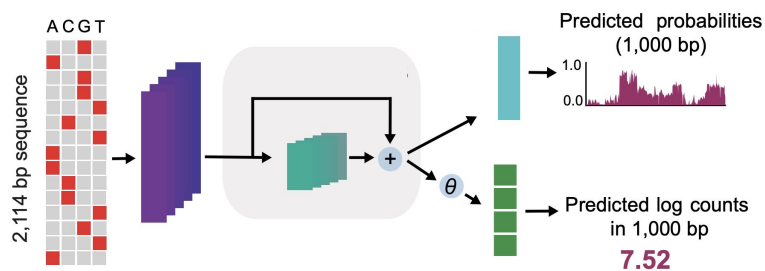
These findings are robust across genomic models and modalities

Year	Genomic DNN	Data modality	Genome
2024	ChromBPNet	ATAC-seq	Human
2024	CLIPNET	PRO-cap-seq	Human (population)
2021	DeepMEL2	ATAC-seq, ASCAVs	Human
2022	DeepSTARR	UMI-STARR-seq	Drosophila
2021	Enformer	ChIP-seq, DNase-seq, ATAC-seq...	Human, mouse
2024	ProCapNet	PRO-cap-seq	Human
2022	ResidualBind	ATAC-seq	Human
2019	SpliceAI	RNA-seq	Human

SEAM Case Studies – Local libraries

2. ChromBPNet

- **DNN prediction:** ATAC-seq profiles of chromatin accessibility in the THP-1 human cell line
- **Sequence:** PPIF promoter, cross-validated using Variant-EFFECTS

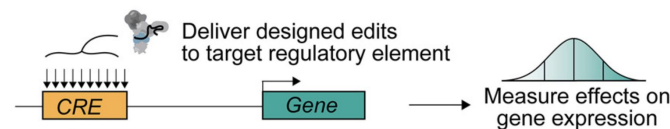


Input: 2114-length sequence

Output: {1000-length profile, scalar}

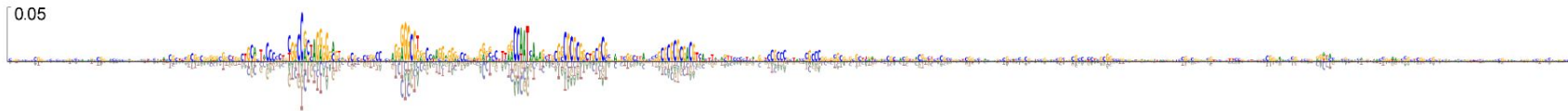
Variant-EFFECTS

Variant Effects From Flow-sorting Experiments with CRISPR Targeting Screens



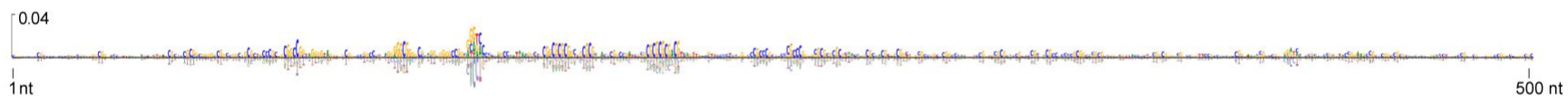
PPIF human promoter

WT



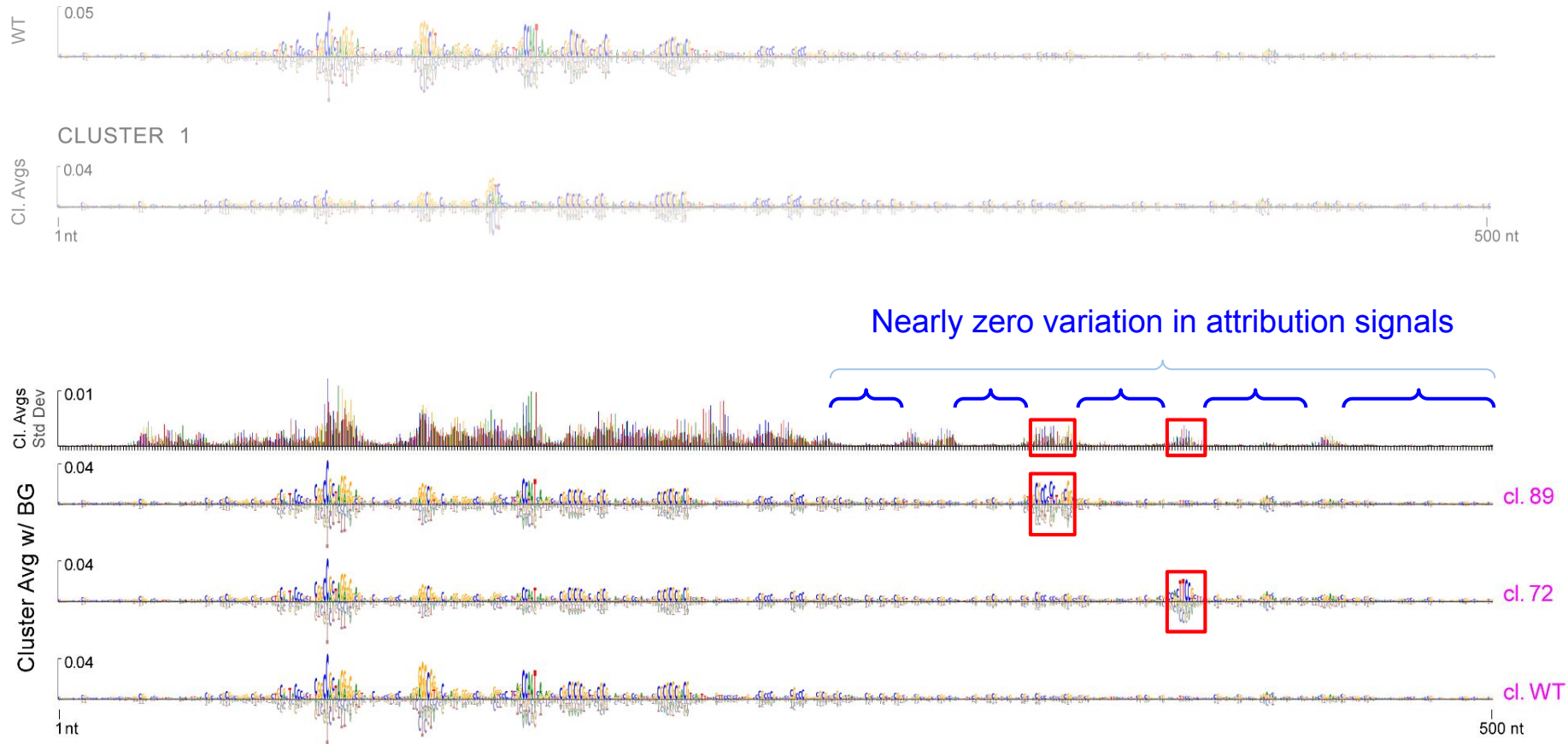
CLUSTER 1

Cl. Avgs

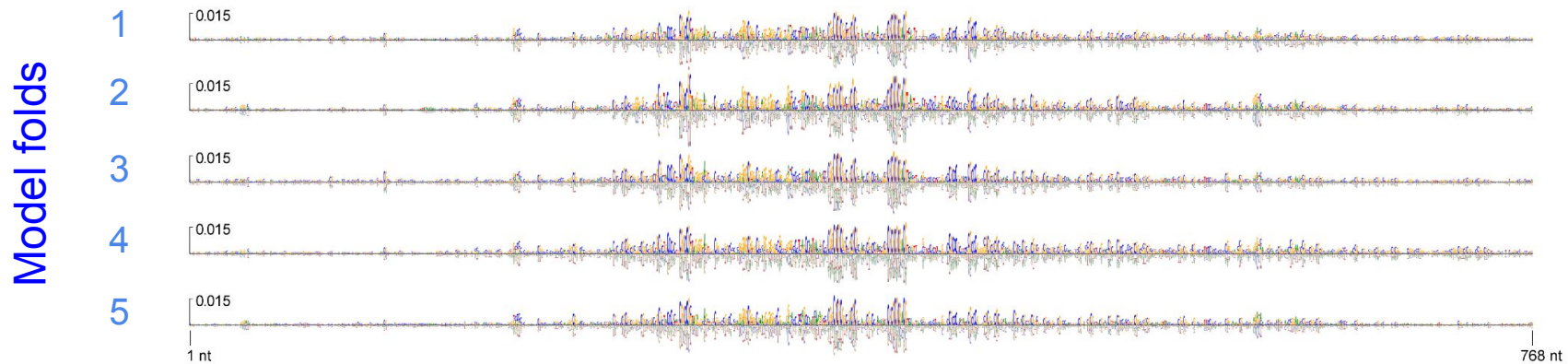




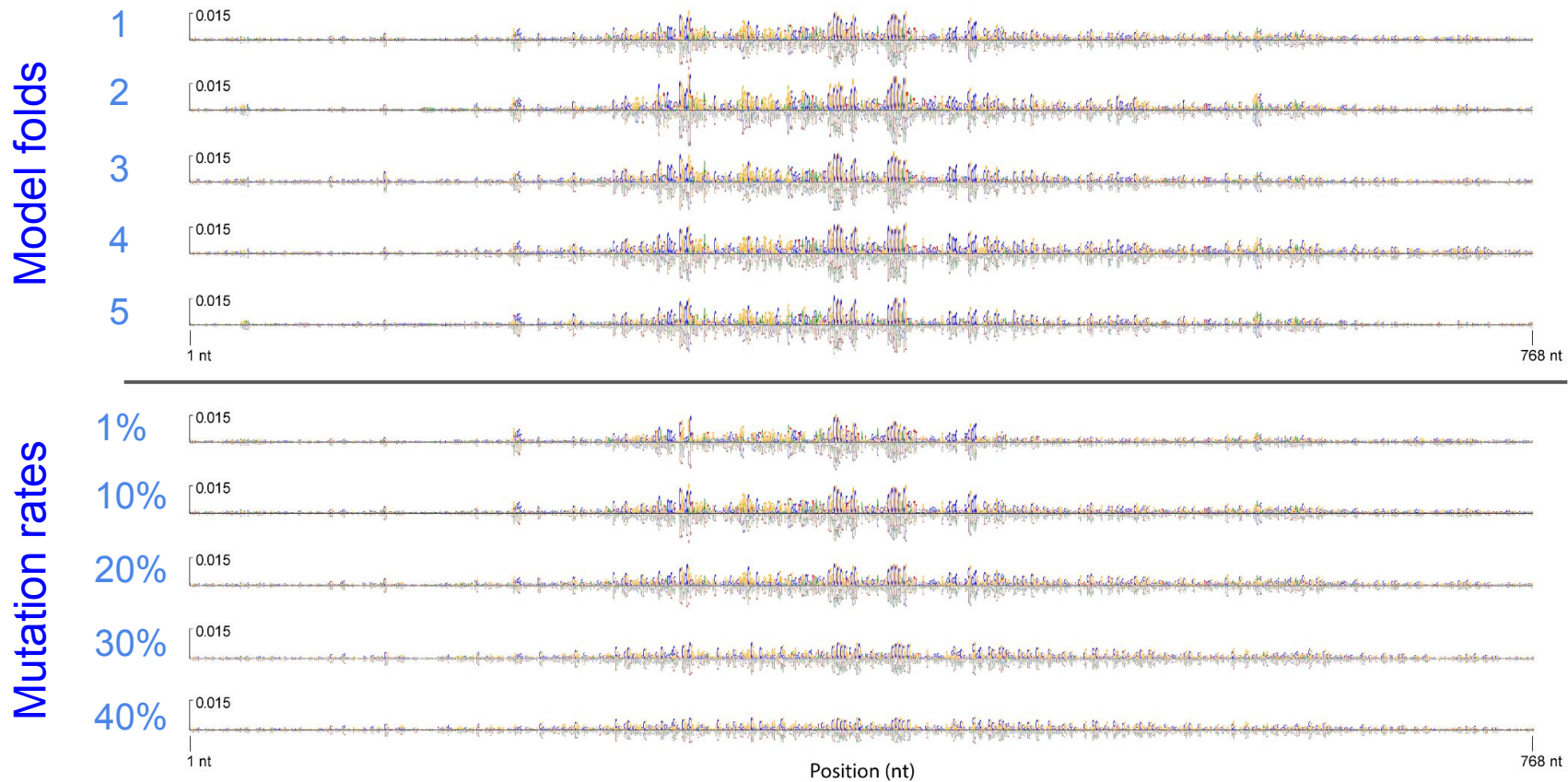
PPIF human promoter



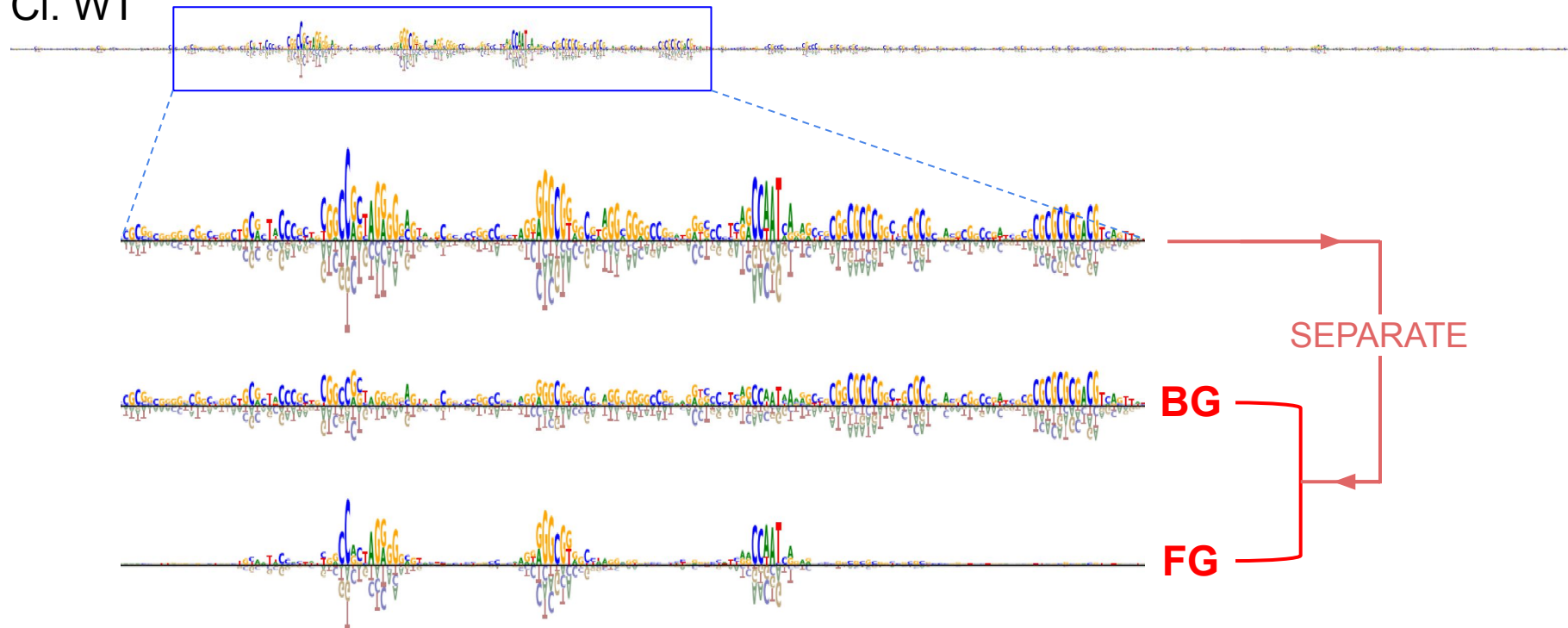
Background signals are robust to model fold and mutation rate



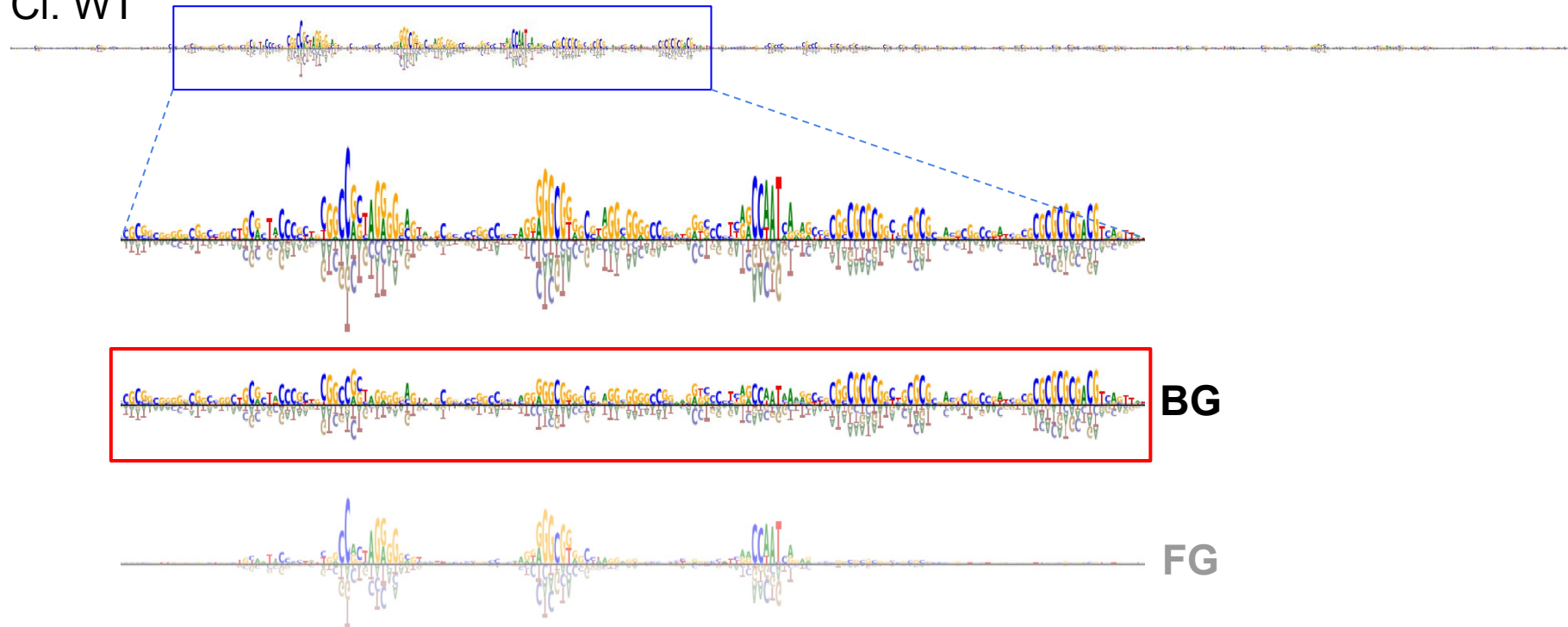
Background signals are robust to model fold and mutation rate



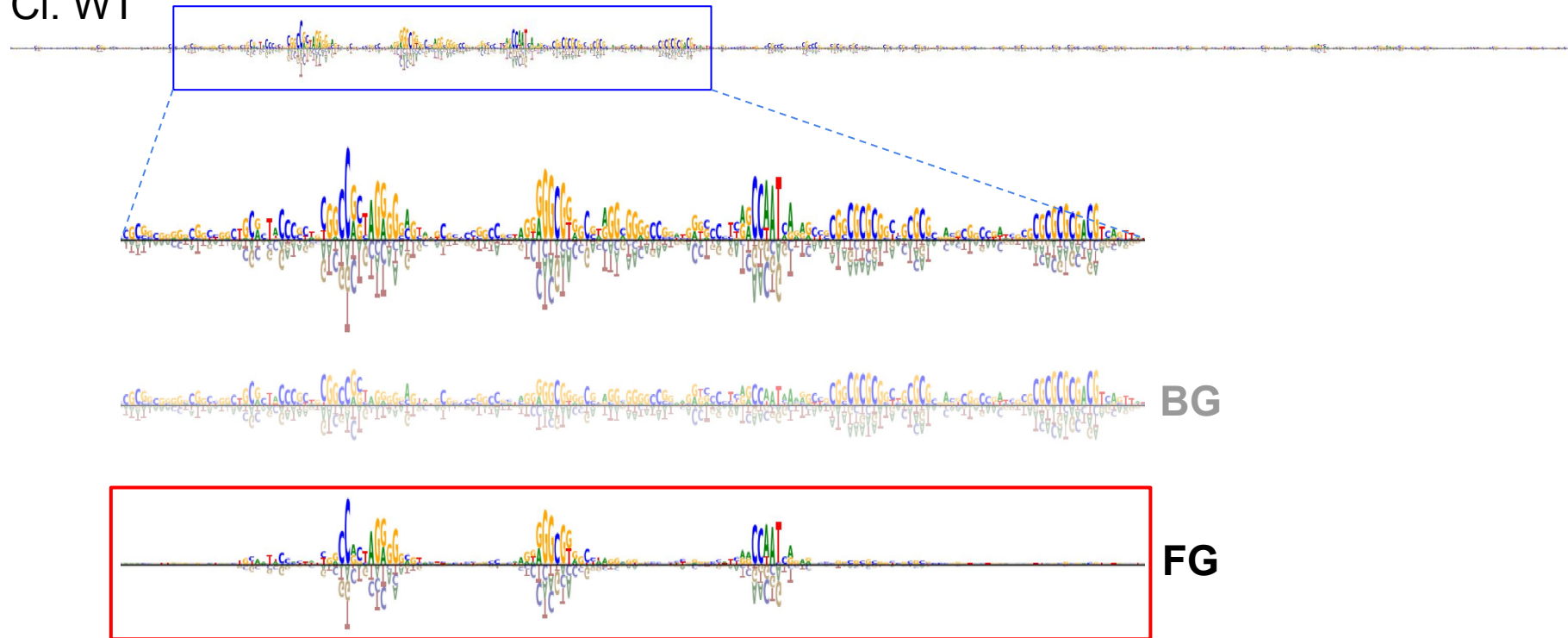
CI. WT



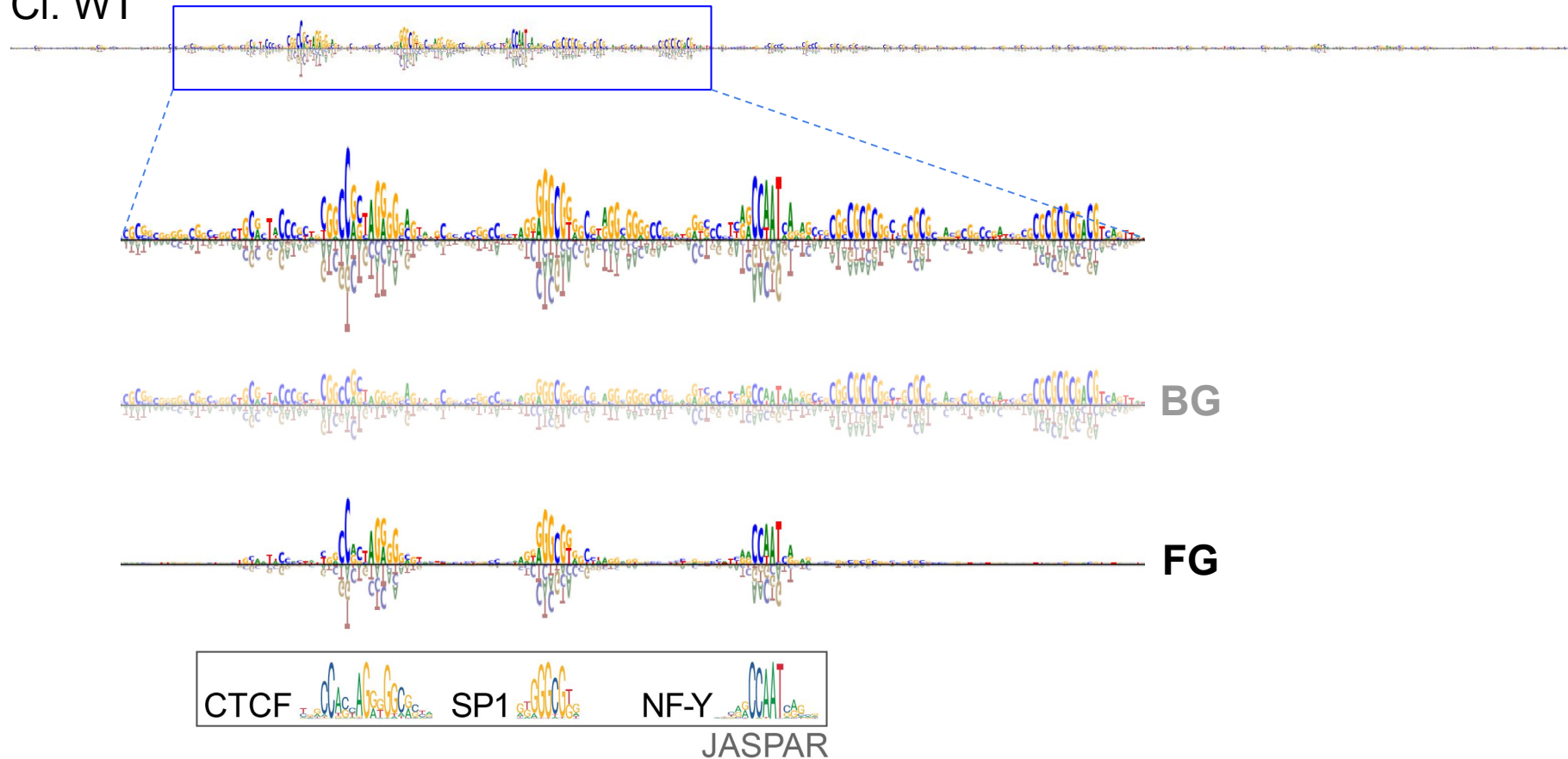
CI. WT



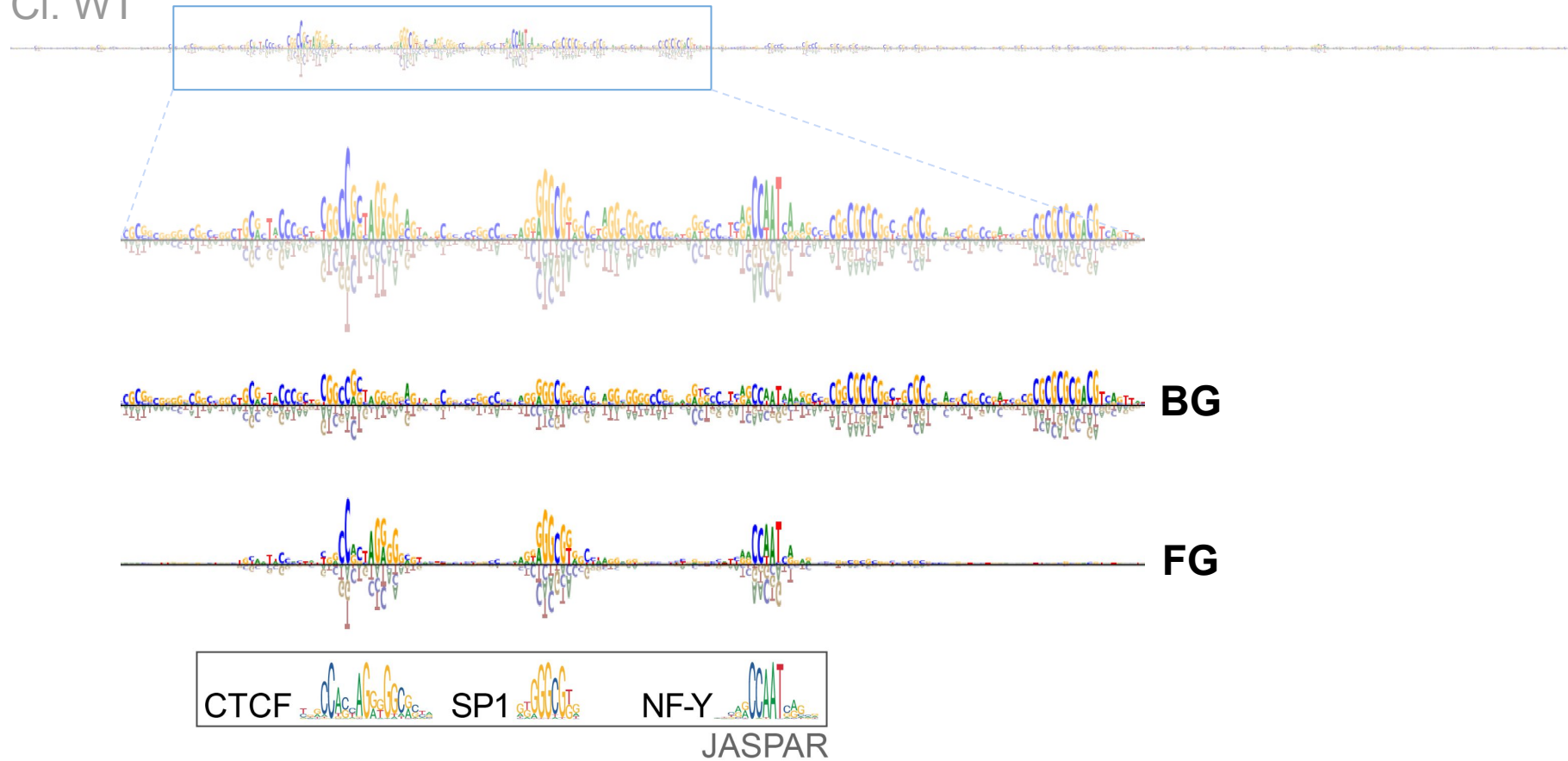
CI. WT



CI. WT



CI. WT

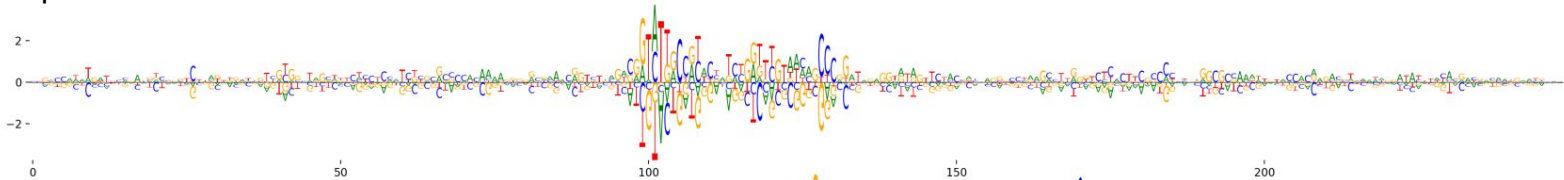


Recall: *Drosophila* enhancers

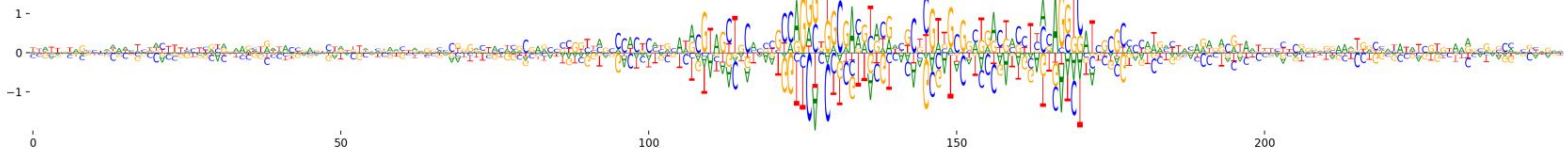
Saliency maps

WT MAP

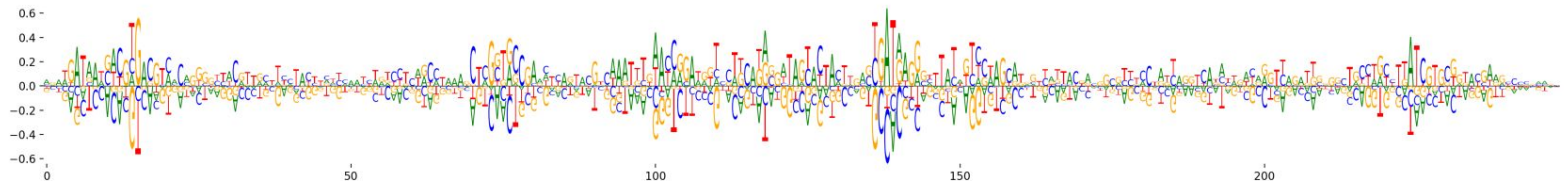
E_1



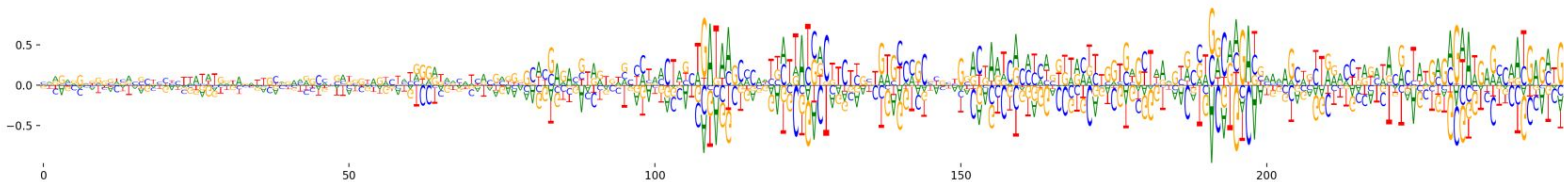
E_2



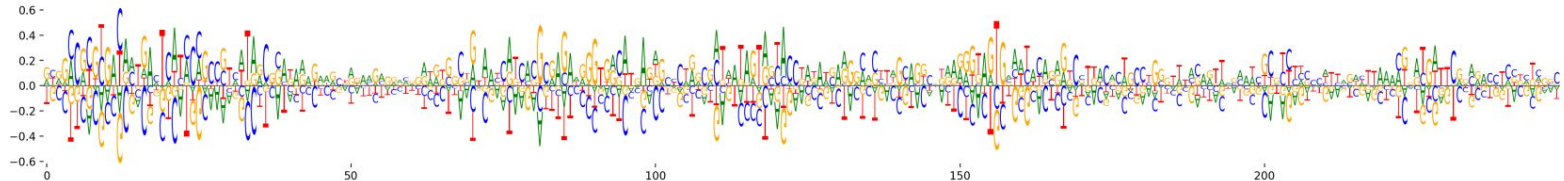
E_3



E_4

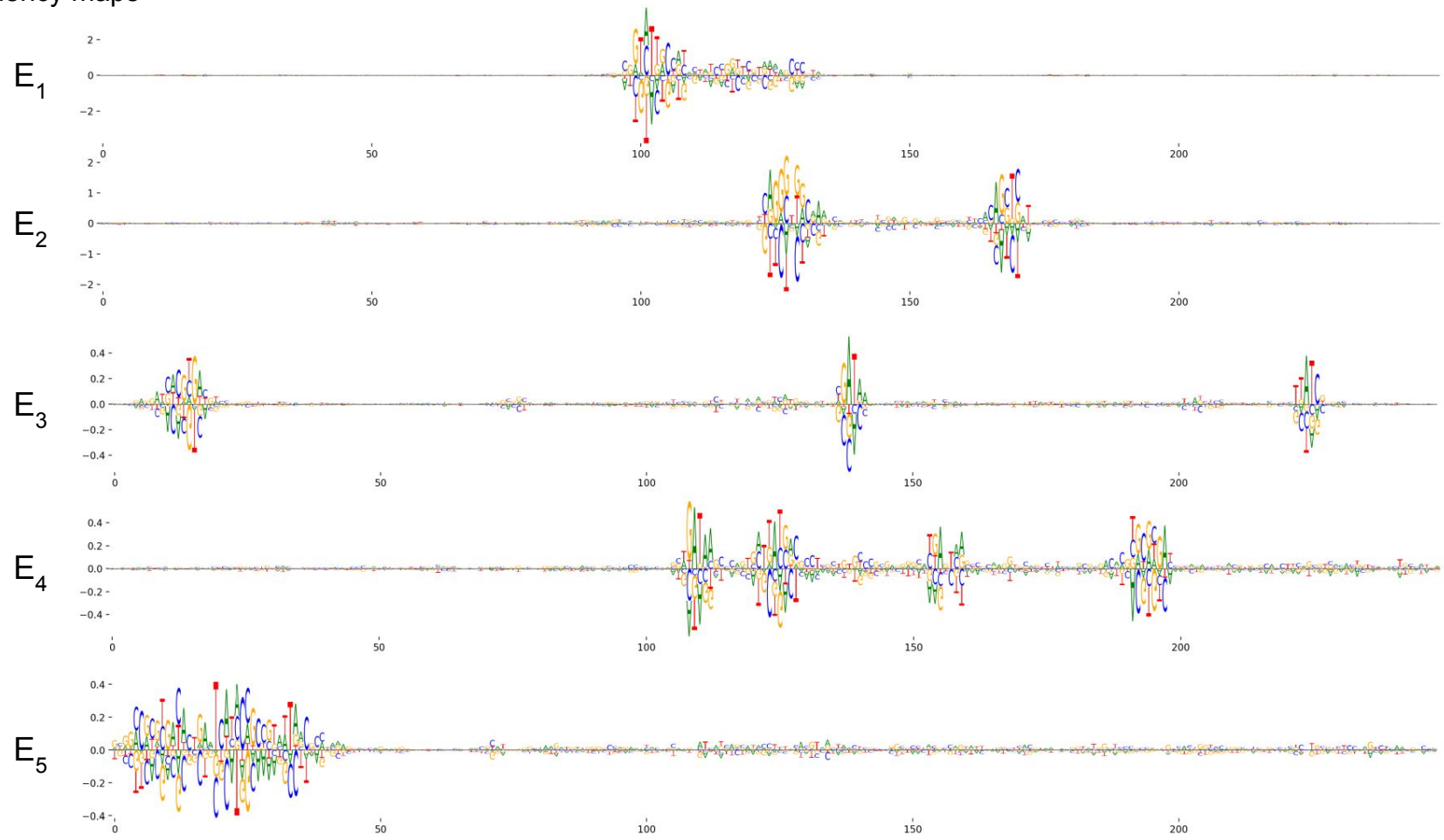


E_5



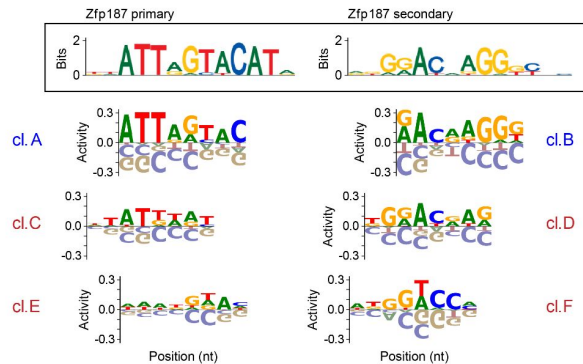
Recall: *Drosophila* enhancers
Saliency maps

WT CLUSTER – BG



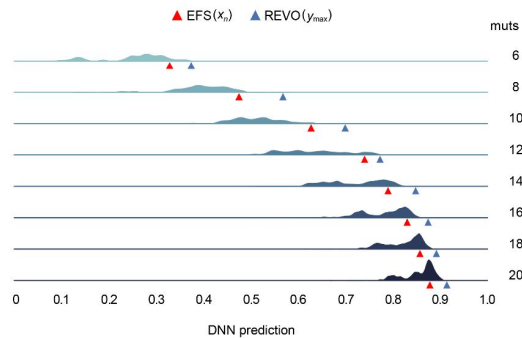
Combinatorial-complete library

Raw PBM data – Badis *et al.*, *Science*, 2009



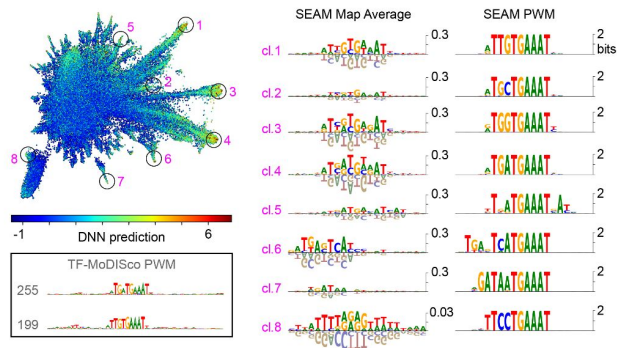
Group-optimized library

DeepMEL2 – Taskiran *et al.*, *Nature*, 2024

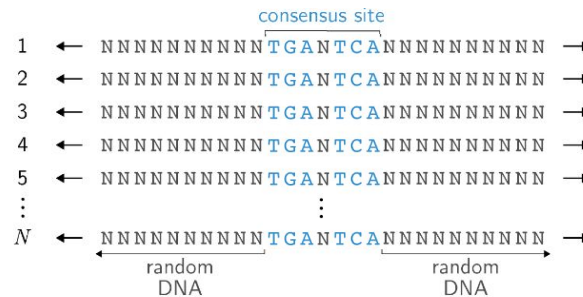


Global library

DeepSTARR



Global library construction



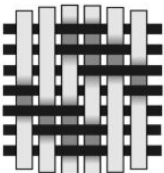
```
pip install seam-nn
```

- CPU/GPU-optimized
- Colab examples
- ReadTheDocs
- Dynamic GUI

README MIT license

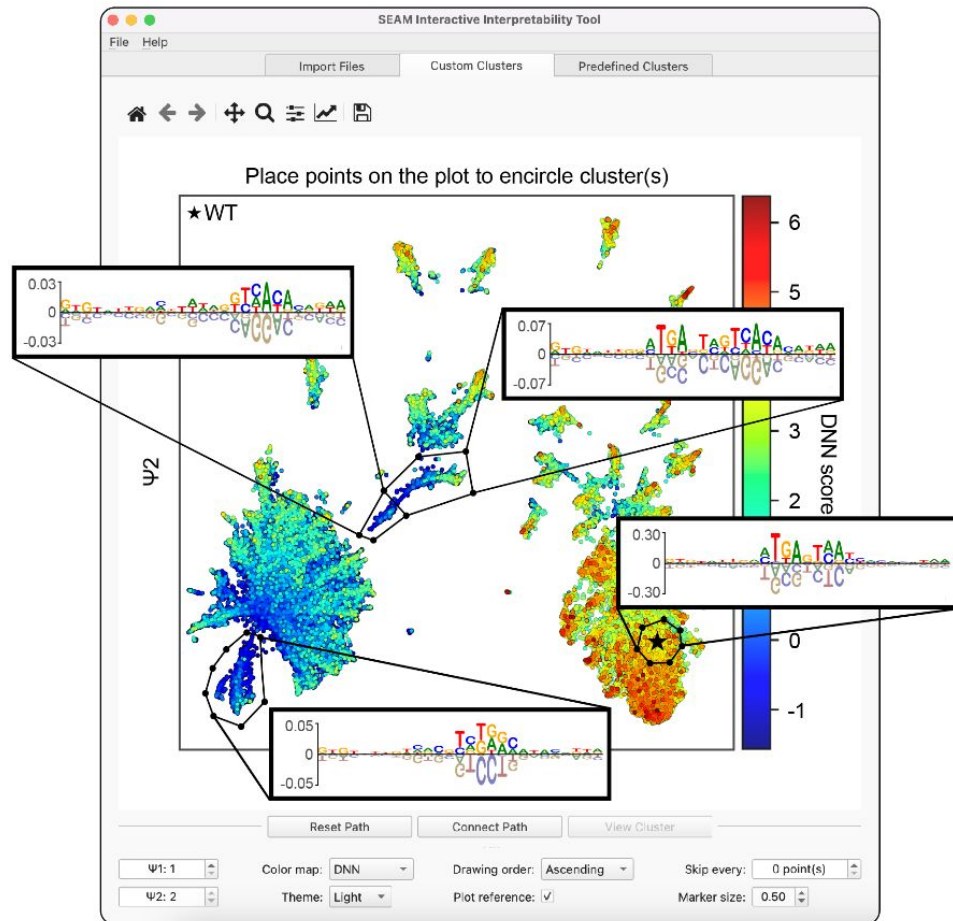
SEAM: systematic explanation of attribution-based mechanisms for regulatory genomics

pypi package 0.5.2 downloads 18k docs passing



SEAM

SEAM (Systematic Explanation of Attribution-based for Mechanisms) is a Python suite to use meta-explanations to interpret sequence-based deep learning models for regulatory genomics data. For installation instructions, tutorials, and documentation, please refer to the SEAM website, <https://seam-nn.readthedocs.io/>. For an extended discussion of this approach and its applications, please refer to our paper:



Peter Koo*
Justin Kinney*
David McCandlish
Kai Loell
Jack Desmarais
Leo Liu
Susan Friedrichs

Thank you!

