A mechanistically interpretable neural-network architecture for discovery of regulatory genomics

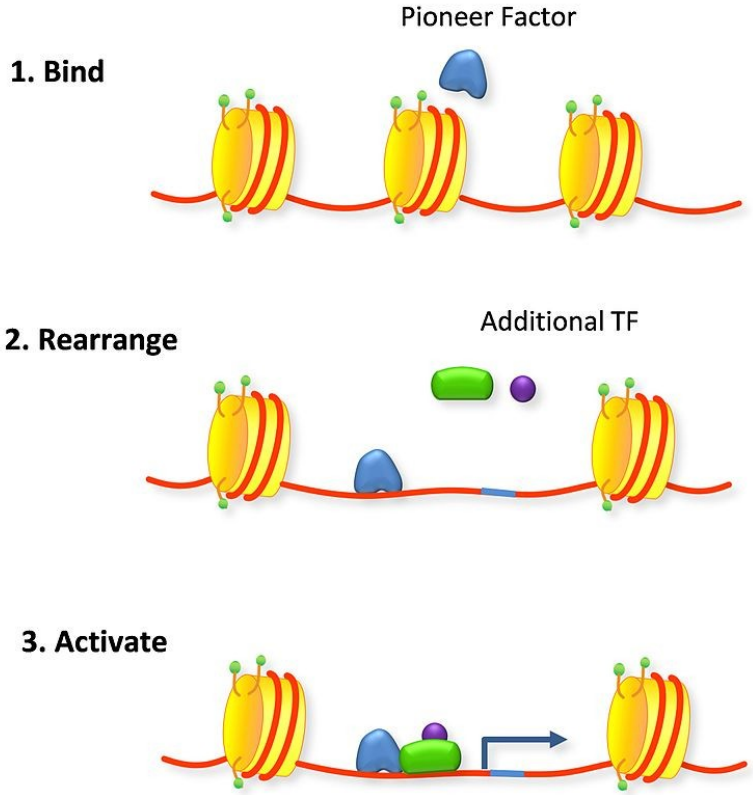
Alex M. Tseng, Gökcen Eraslan, Nathaniel Diamant, Tommaso Biancalani, Gabriele Scalia

MLGenX
11 May 2024

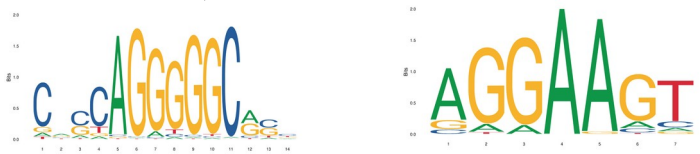
Biology Research | AI Development

Genentech
A Member of the Roche Group

Introduction to regulatory genomics



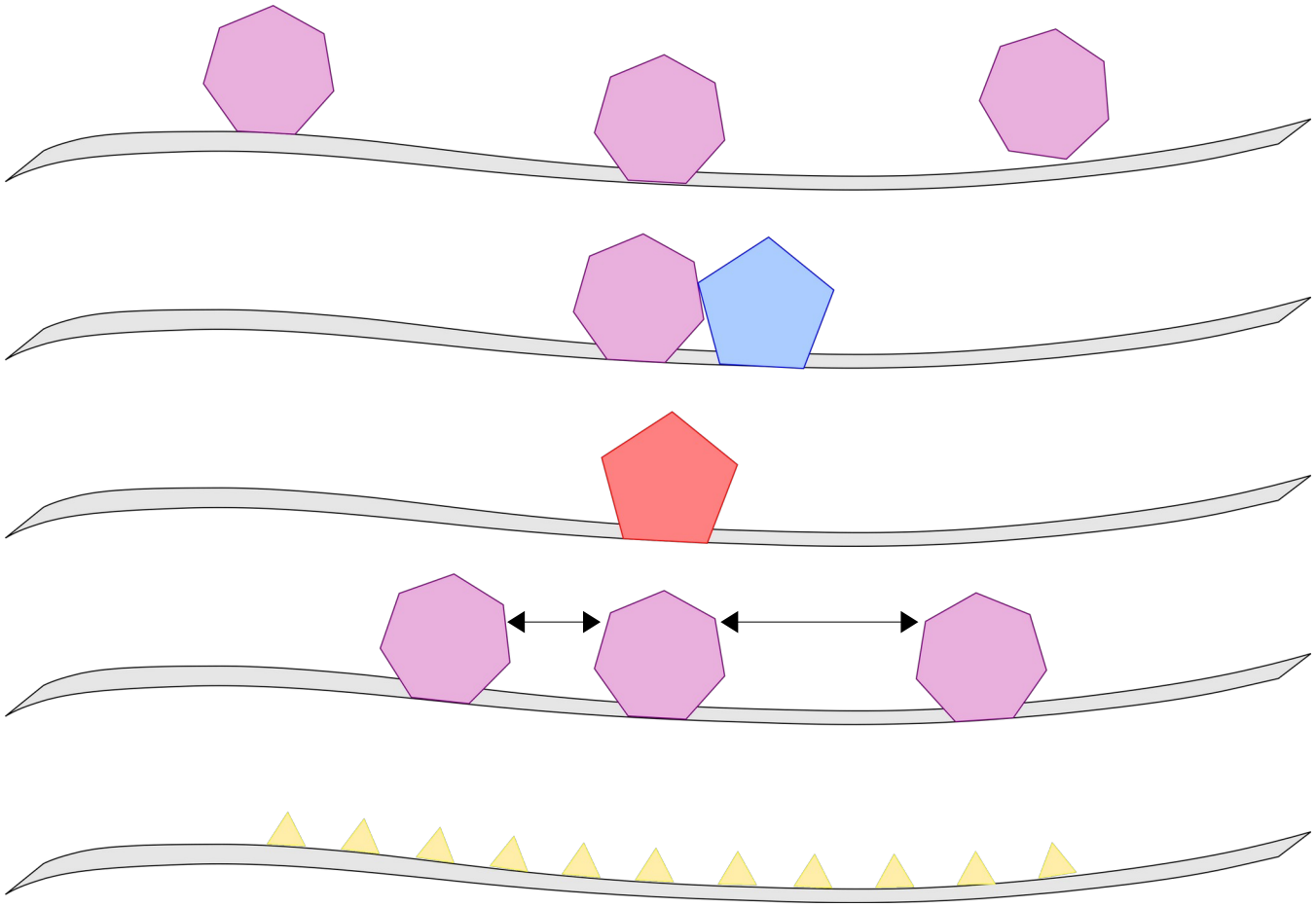
Pioneer factors and transcription factors



Motifs

JASPAR 2020: doi: 10.1093/nar/gkz1001

Motifs endow function through a complex and system-dependent syntax and grammar



Strong vs weak binding sites

Cooperative co-factors

Competition or inhibition

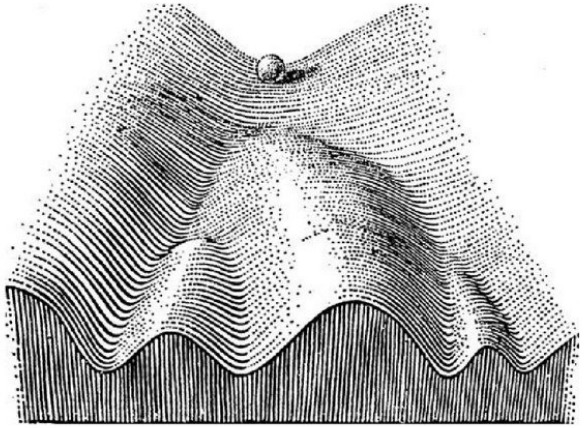
Spacing preferences

Chromatin modifications

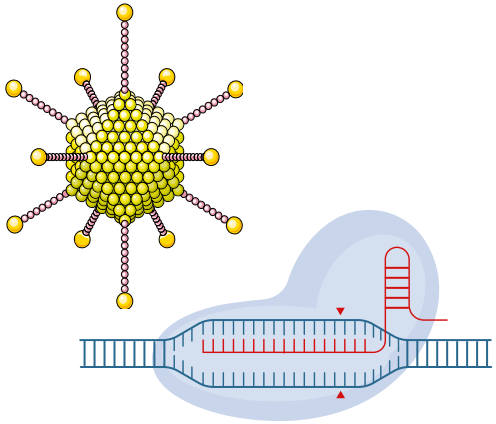
Understanding how proteins bind to DNA is important for:

Image by Svenska Mässan, distributed under CC BY 2.0

Disease diagnosis


Waddington, 1957

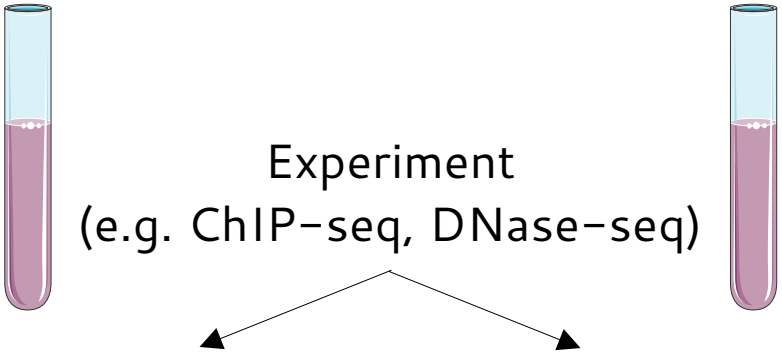
Cell reprogramming


BioIcons

Therapy design

Parker *et. al.*, 2009

A major challenge: it is hard to extract motifs and their syntax from experimental data



"active" sequences

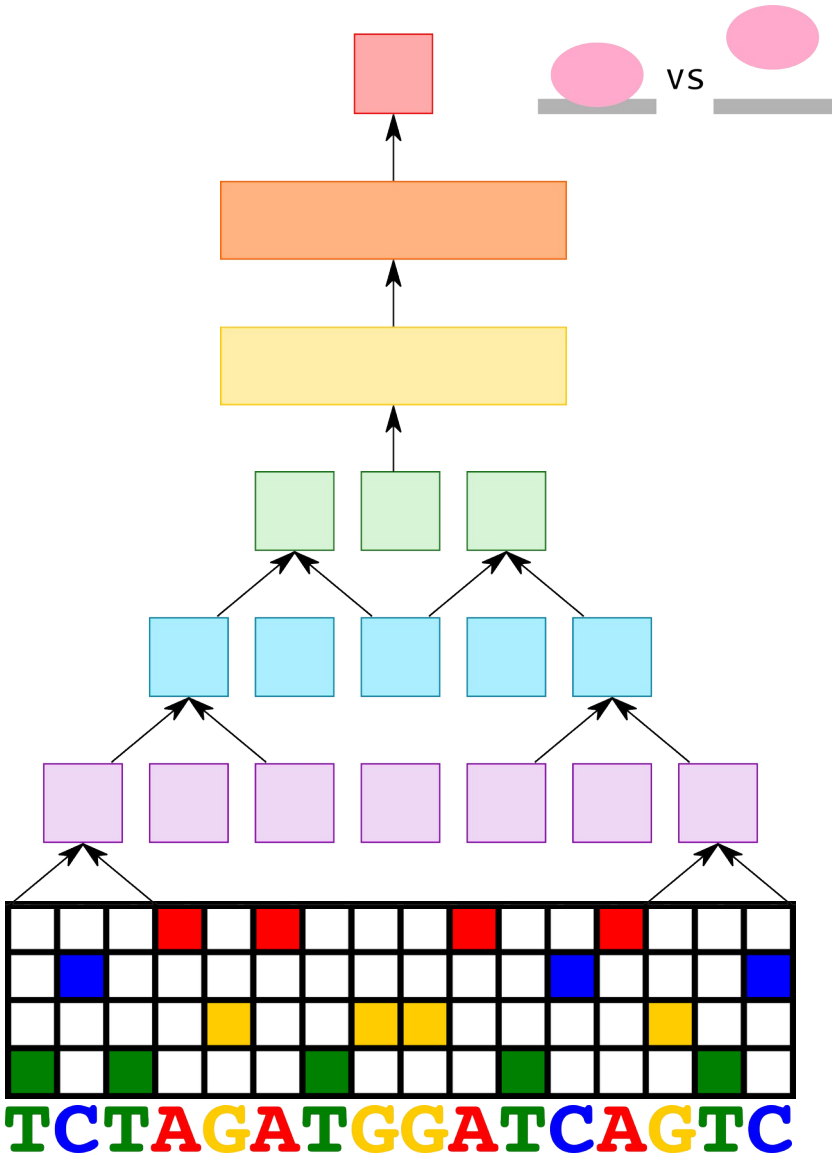
```
ATTTATGAGAAACAGTTAAACTTAATACTGACTTATAACC
CAAGTATGACCGAGTGCACGGTTCACGTAGATGTTGCCAC
AATTCCAACTGTTTGTGAGGGACGCCCTGTCCGCCGCTCG
TCCGCTCCATAAGGTAGGGGGACCAGACTGGCTGATACTC
AGTTGGCGGCCGATCGCCTCGTGATGCGCCCGACCCTCTA
                    ...
CGGCCTAACTGGCTTTCAAGGAACCTGGCAAGTGTAACTA
GGGCCGGCGCATGCATAGTTGAACATAATGGAATTGCAAT
GGGTTCGCGATTGATGACATGCTTATTCACACTCGTCACA
AATGGAATTCTGAGATACGTTCGTAGGCCATCTTTACTGT
CCAAAAGATTGATGCGTTCAGAACATACACTGAAGGGTAA
```

"inactive" sequences

```
GGTTCGACAAATATTTTCCATGGACTGGAAATTGCCGTAT
ACCGGCTAGTCTCGTTGCCAATCCGGGCCACCTTGCGCAG
CATTAGATTACCGTGCAATCCTGTCTGTCCGTTATTTTAT
GACAACTCCAGTGTAGTAGGTGACACCGAATAAGCCGAAG
AGCCTCGAATAAAGGTTTTGACTCTCCATGGTATAGAGAG
                    ...
TTGCTTTATGGTACATTTTAAGGCAAGATTACTGTCTTCC
CGATCCGGGCTTCTTAGTATTCGCCCGGCCGCGCAGTAAC
ATCAGAACCCATCGGAGGTTGTAGGTCGCCTTAACATATT
AGAATAGGTGTCAATGGCGATTTAAGAGACGCCGGCGACA
TGATTTTTGCATCACTAACCGTCTGCCCAGTGAGTCAGAG
```

What are the motifs (and syntax) which distinguish the "active" sequences from the "inactive" sequences?

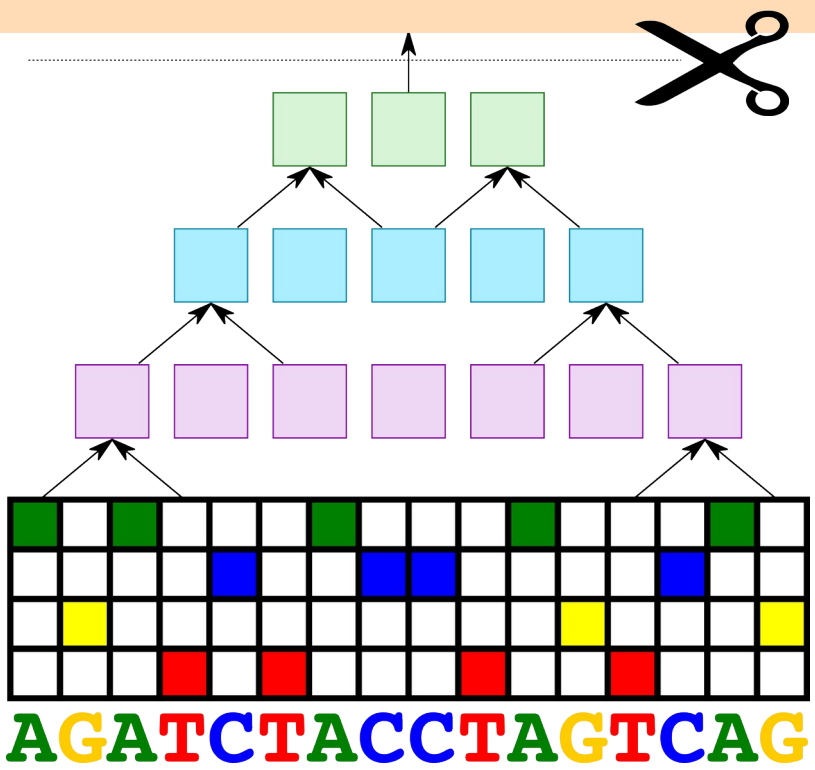
Learning regulatory genomics with neural networks



Experimentally determined molecular label

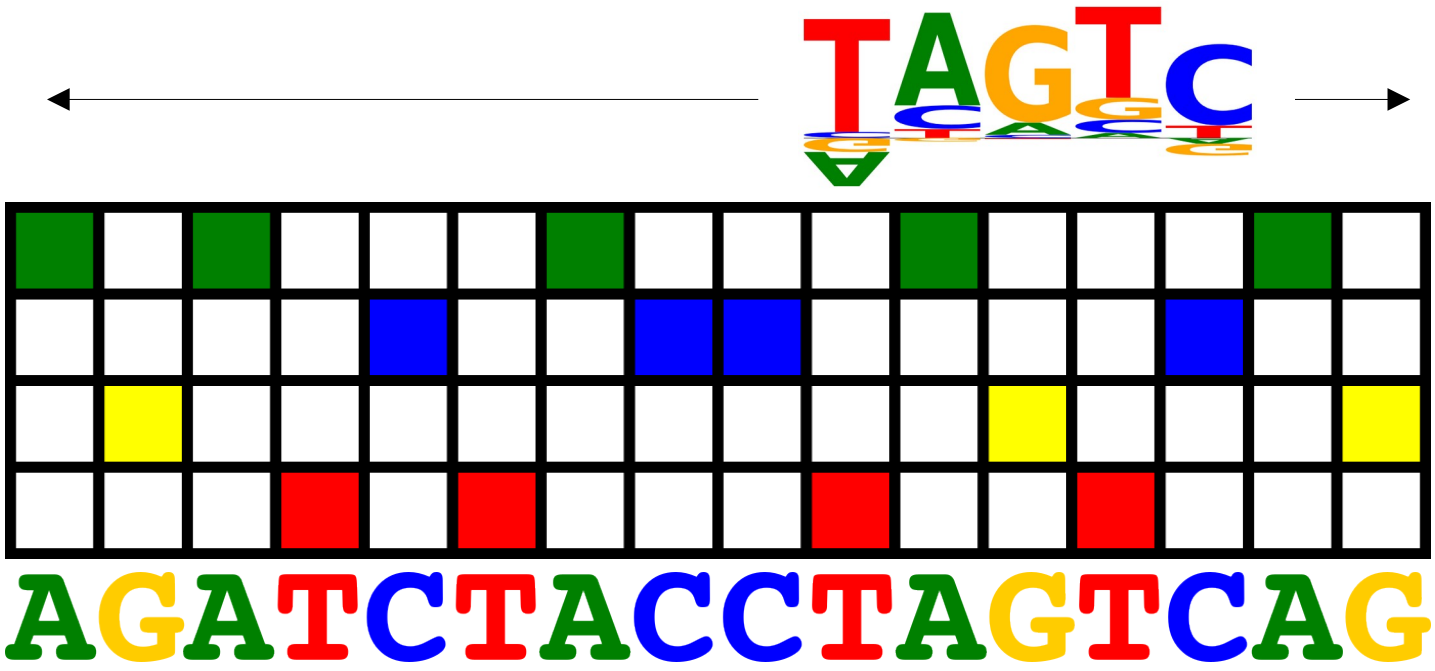
There are many flavors of architectures, all tweaks on a classic CNN

LSTM	Dilated convolutions	Transformers
Quang *et. al.*, 2016	Avsec *et. al.*, 2021	Avsec *et. al.*, 2021

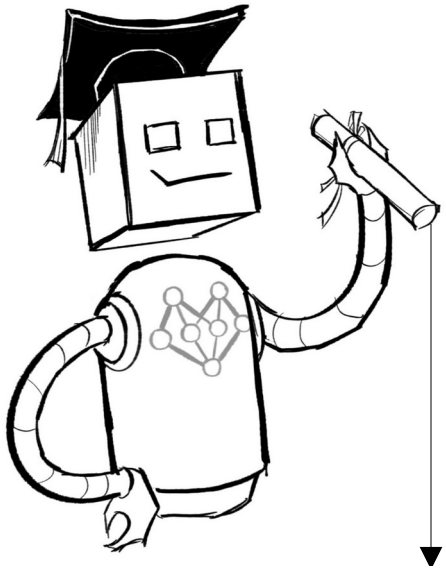


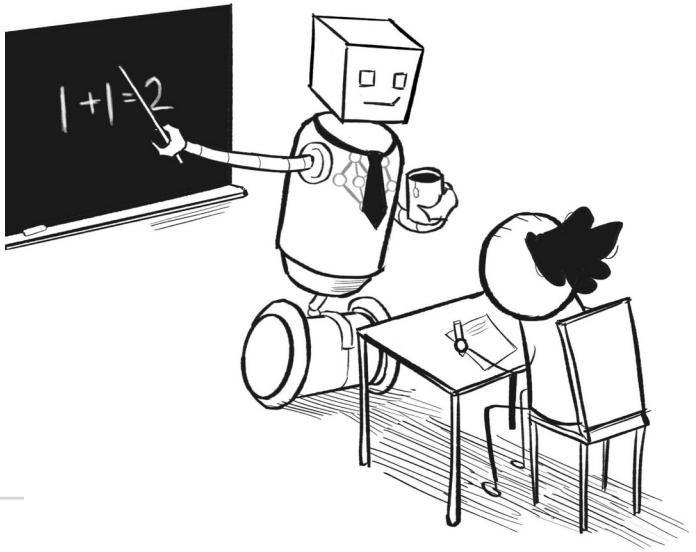
Why convolutions?

A	-0.33	0.90	0.12	0.05	-0.06
C	0.06	0.23	0.03	0.13	0.80
G	-0.14	-0.04	0.95	0.21	-0.12
T	1.11	0.09	-0.01	0.89	0.12



Deep neural networks have become regulatory-biology experts





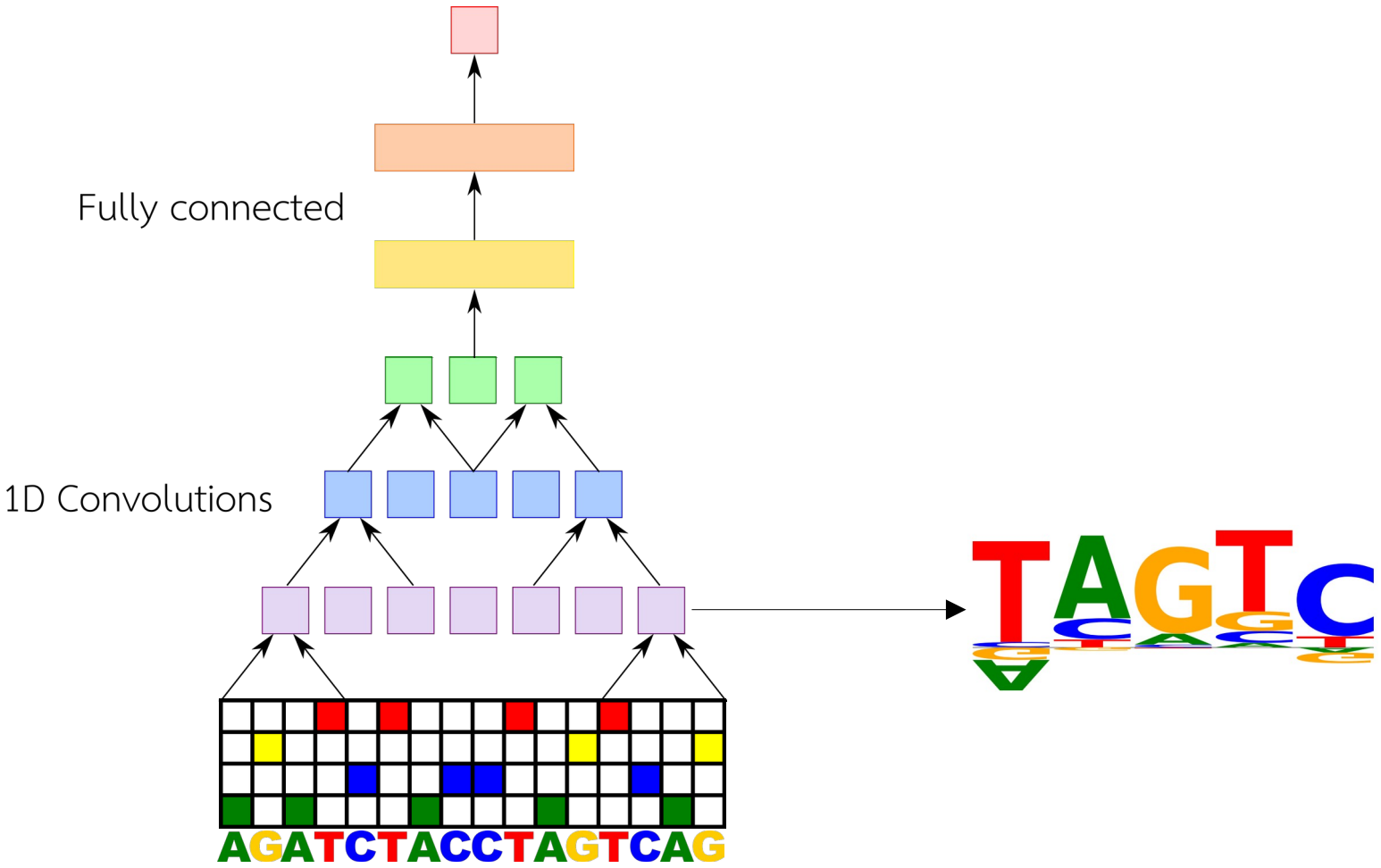
Deep learning in regulatory genomics

Neural networks and sequence-to-activity models

Deep neural network (DNN) models have emerged as the leading type of predictive model in regulatory genomics[1,4,6]. For this Review, we focus on sequence-to-activity models based on neural networks. These models take a putative regulatory DNA sequence (usually 100–10,000 bp) as input and aim to predict some dynamic property (that is, cell or context specificity) of the sequence's activity. For example, a model may predict whether a given TF binds to that sequence in a given cell type as measured by a chromatin immunoprecipitation followed by sequencing (ChIP–seq) experiment[7,8,9,10]. Other common prediction targets include chromatin accessibility[11,12], RNA binding[13], gene expression[14,15,16,17], splicing[18,19] or aspects of chromatin 3D organization[20].

Novakovsky, *et. al.*, 2023; Illustrations by Audrey Tseng

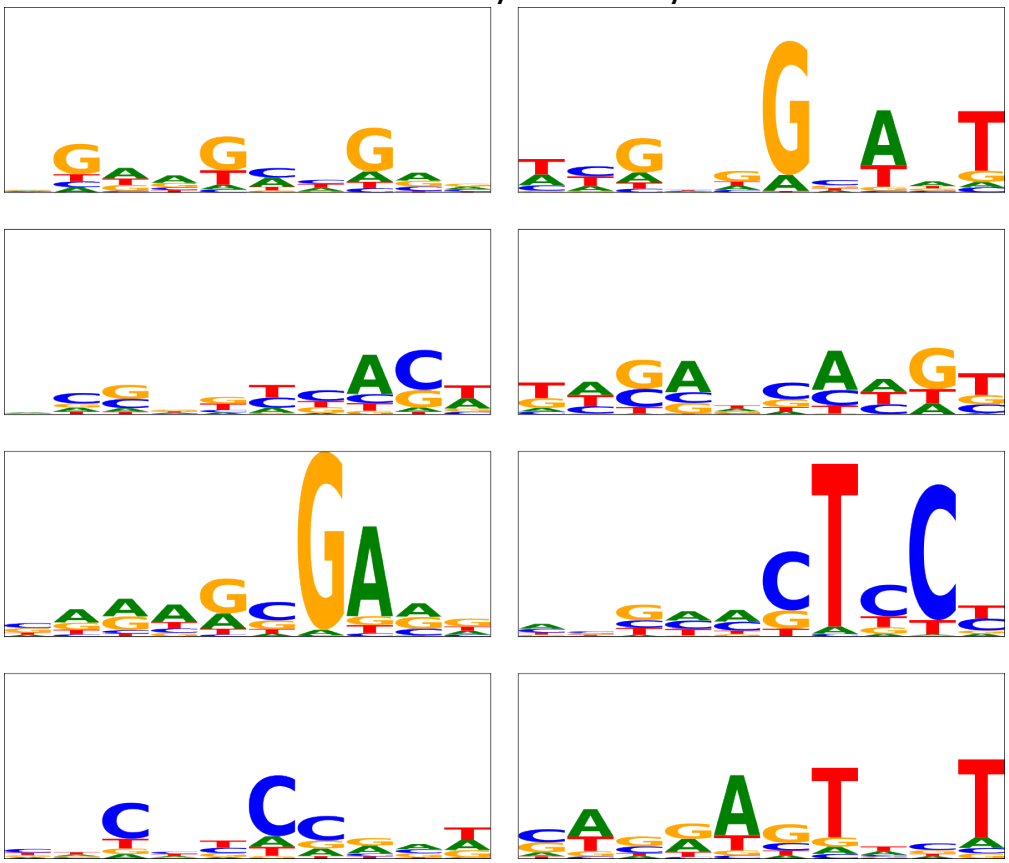
Identifying motifs from a NN by interpreting filters



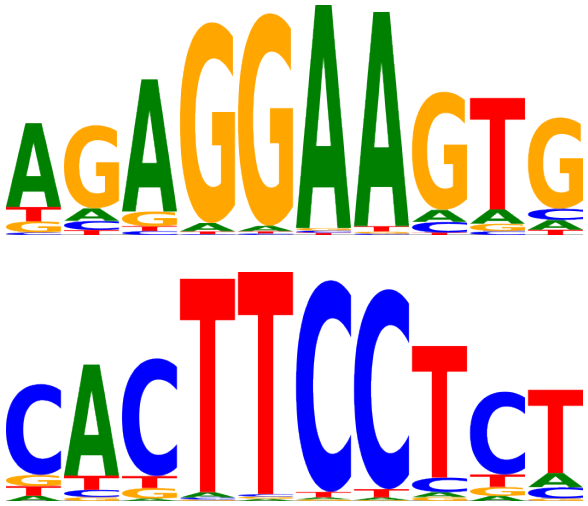
Alipanahi *et. al.,* 2015; Kelley *et. al.,*, 2016

Problem: a single filter's weights don't always encode motifs

Motifs encoded by first-layer filters



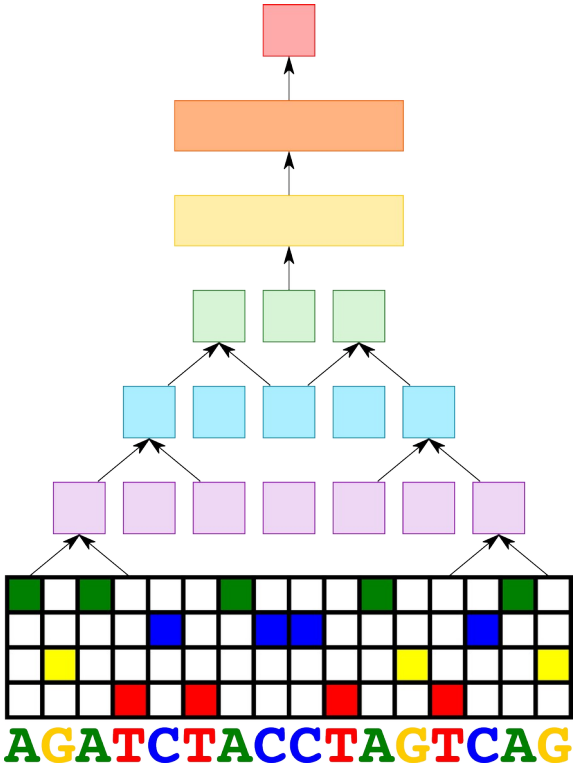
True motifs



State-of-the-art accuracy, **dreadful** interpretability

Motifs are *distributed* across filters

Identifying motifs from a NN by integrating through the whole NN



Train model

↓

Compute importance scores

↓

Segment into putative motif instances

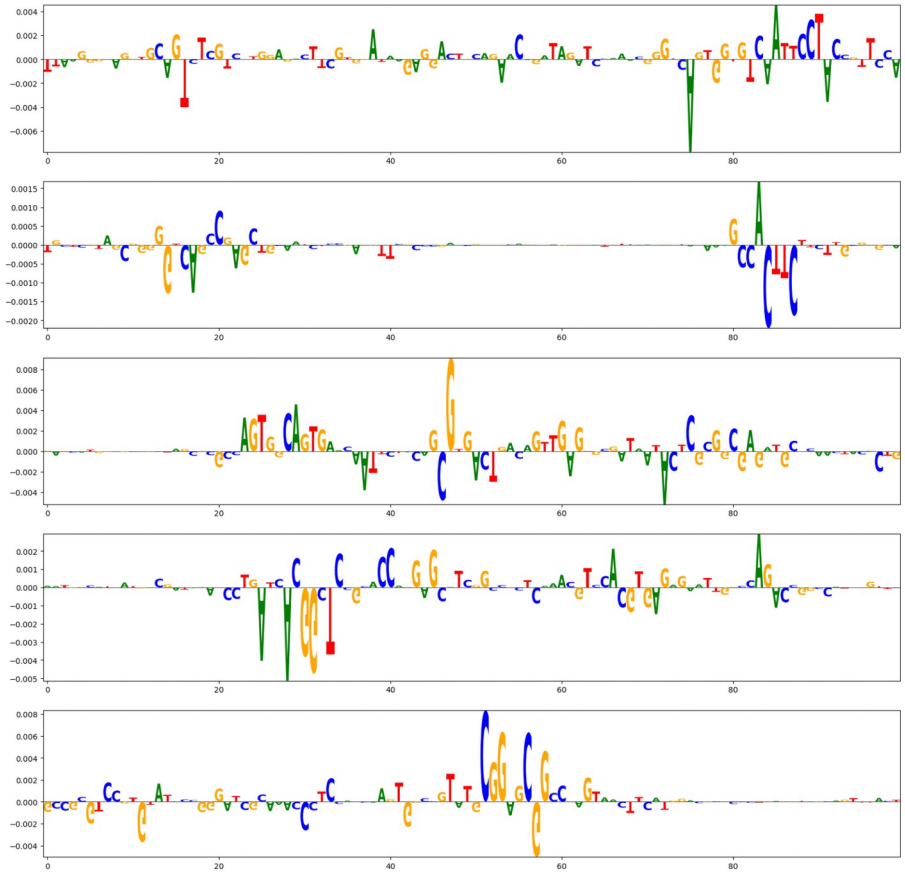
↓

Cluster into recurring motifs

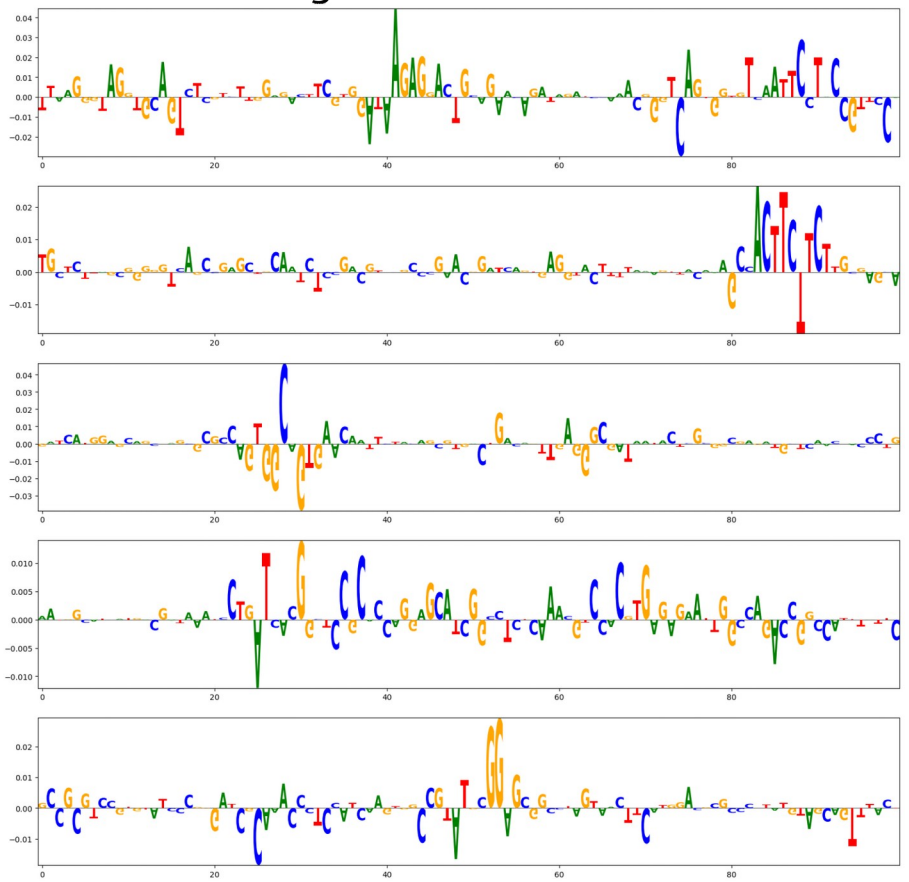
Shrikumar *et. al., 2020*

Problem: importance scores are noisy and unstable

DeepLiftSHAP



Integrated Gradients

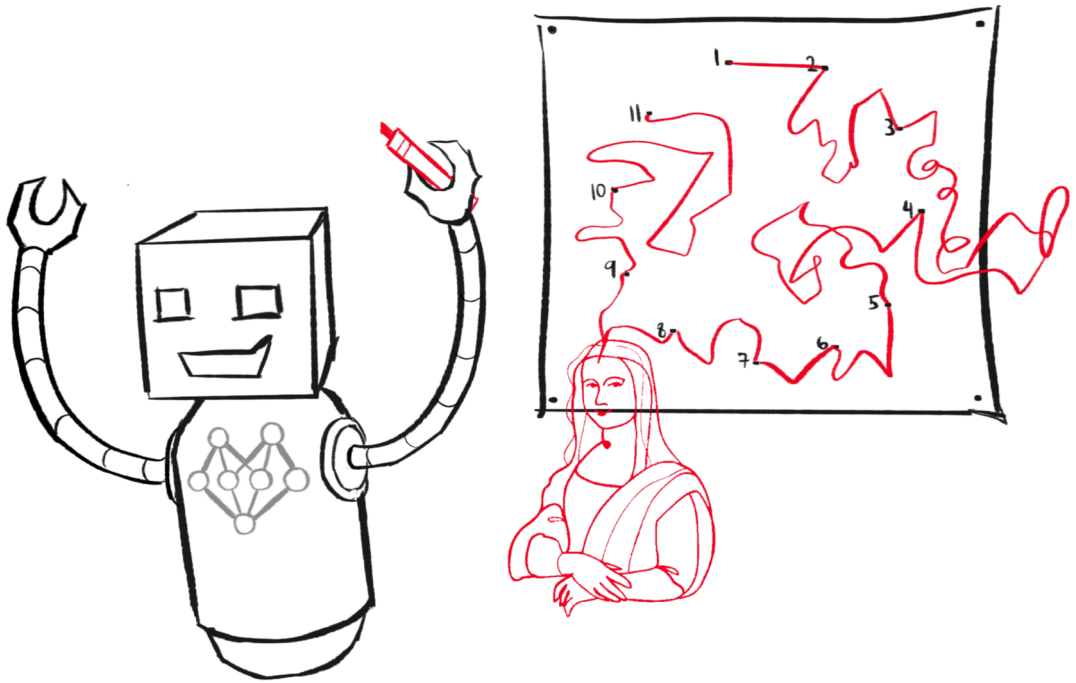


Bad: hard to pick out what the motifs are

Worse: hard to pick out *where* the motifs are

Even worse: the two methods disagree even on the *sign* of importance

The central limitation: expressivity vs interpretability



- Filters do not learn biologically interpretable units
- Importance scores are extremely suboptimal and require enormous *post hoc* processing

Illustration by Audrey Tseng

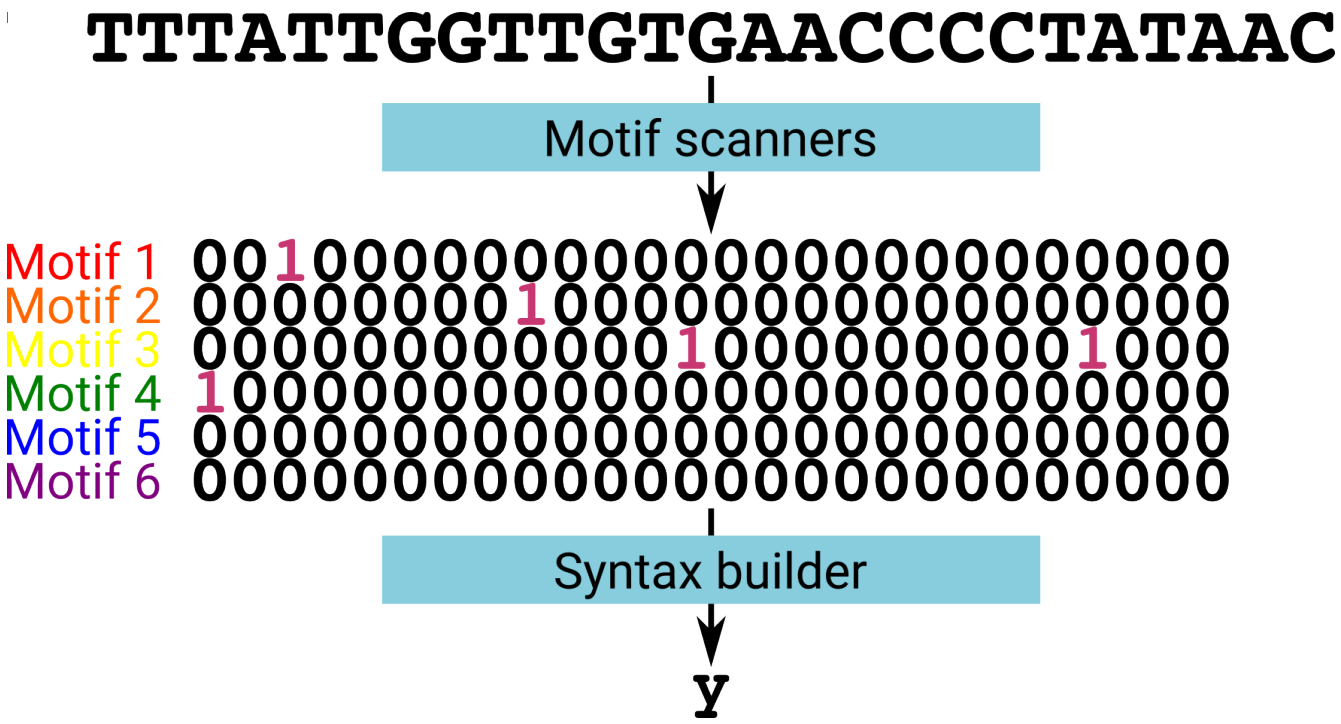
An alternative solution: mechanistic interpretability

- Carefully limit the expressivity of the architecture so the *only* solutions are directly interpretable (while retaining performance)

- In a *mechanistically interpretable* architecture, learned decisions are directly encoded in the weights and activations

Analysis of Regulatory Genomics with a Mechanistically Interpretable Neural Network (ARGMINN)

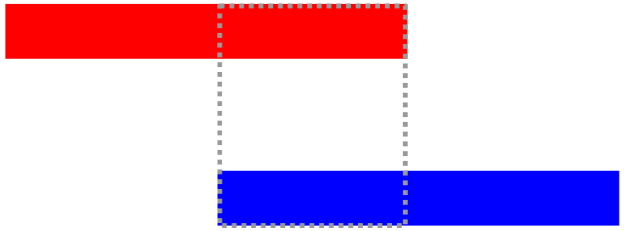
A mechanistically interpretable architecture for regulatory DNA



Motif-scanning module

TTTATTGGTTGTGAACCCCTATAAC

Filters

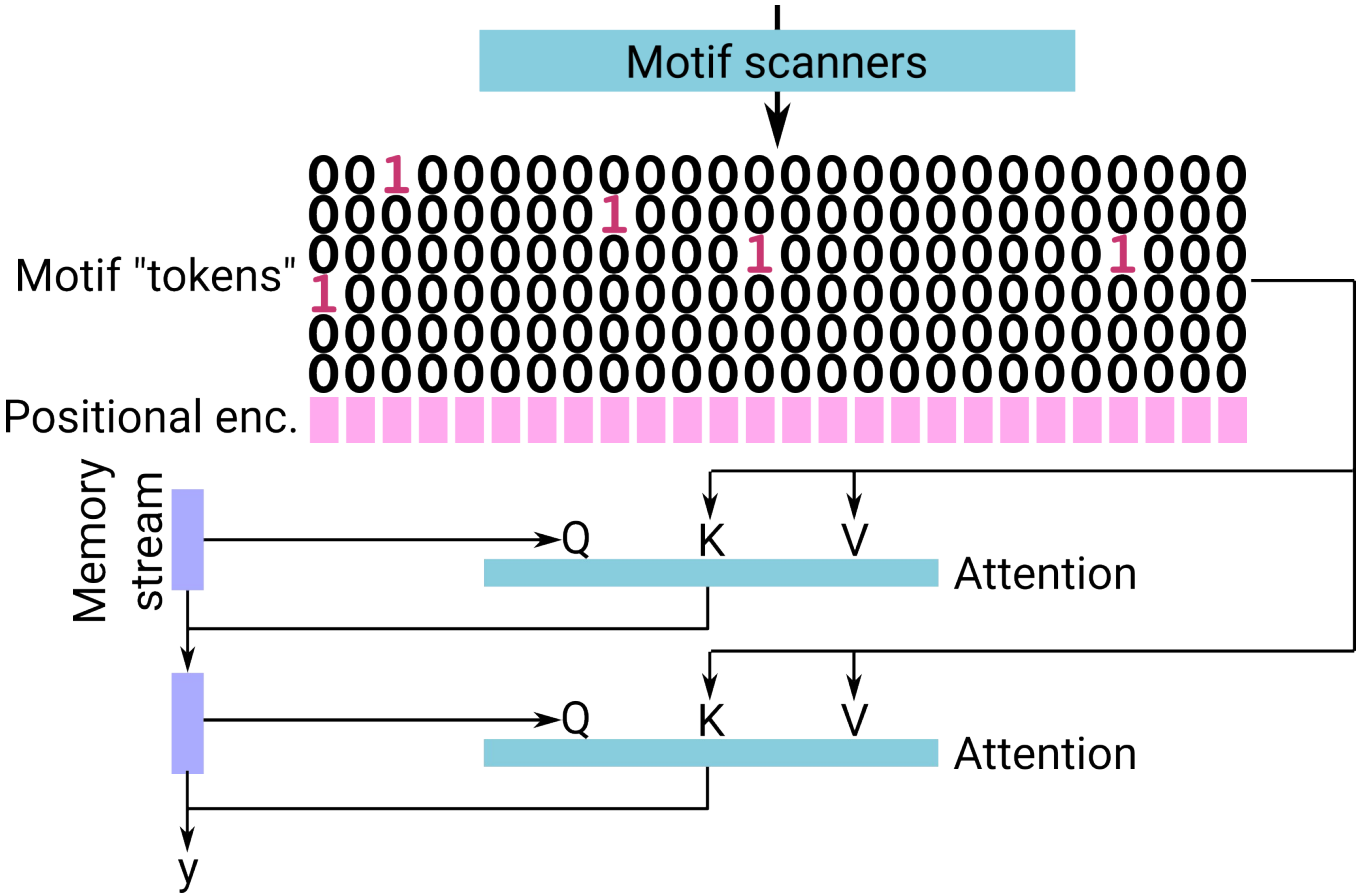


If both neurons fire, the boxed weights should not both be non-zero

A single convolutional layer learns all motifs.

Unique regularization to ensure a one-to-one mapping between motifs and filters.

Syntax-building module



Attention with *single* query from *explicit* memory stream (modified each layer).

Key/value vectors always directly derived from the original motif-scanner activations.

Additional attention layers learn 2nd-order, 3rd-order (etc.) interactions.

Extracting motifs from ARGMINN

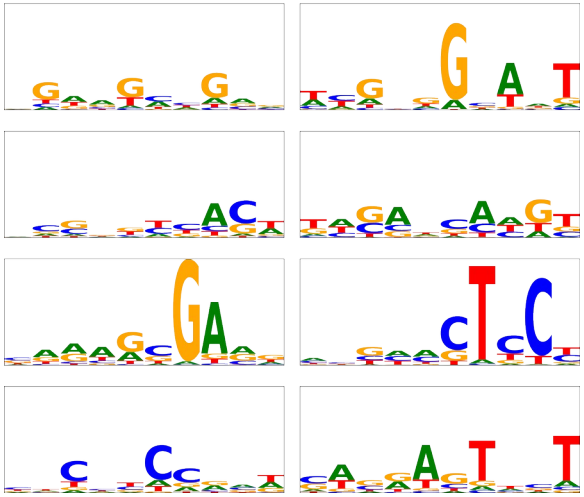
TTTATTGGTTGTGAACCCCTATAAC

Filters

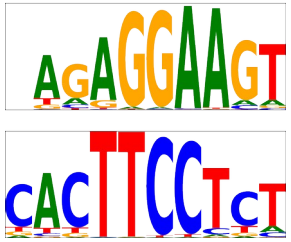
One-to-one mapping between filters and relevant motifs

Discovered motifs

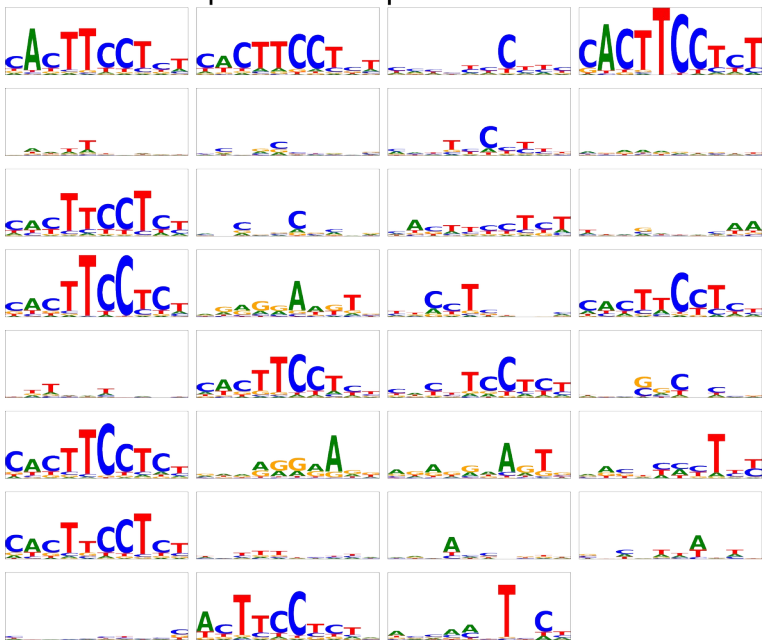
Traditional CNN



ARGMINN



DeepLIFTShap + MoDISco

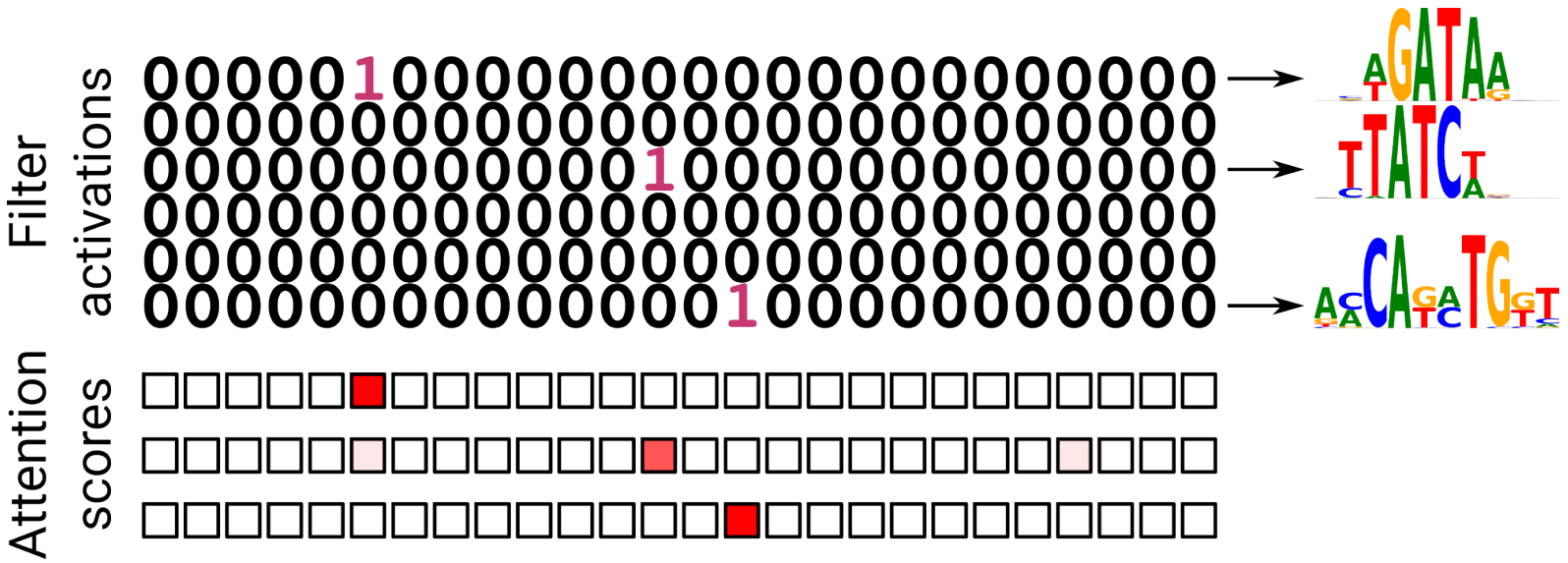


Interpreting filters directly in a traditional model leads to non-motif patterns:
Information is distributed across the model

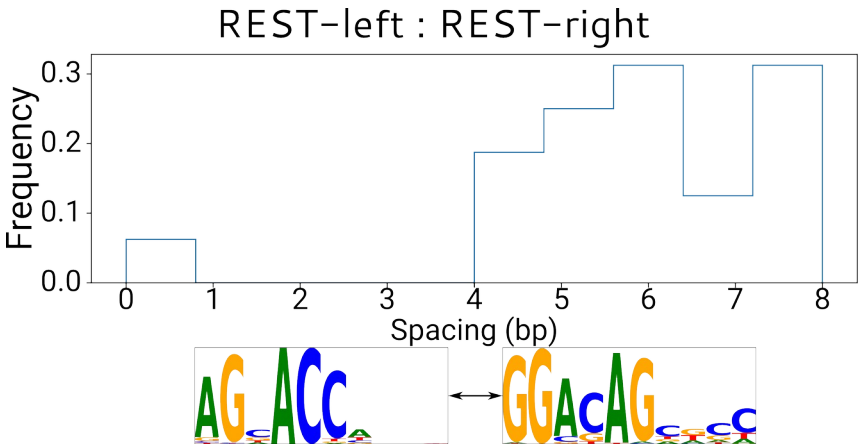
ARGMINN finds *high-quality*, *relevant*, *non-redundant* motifs

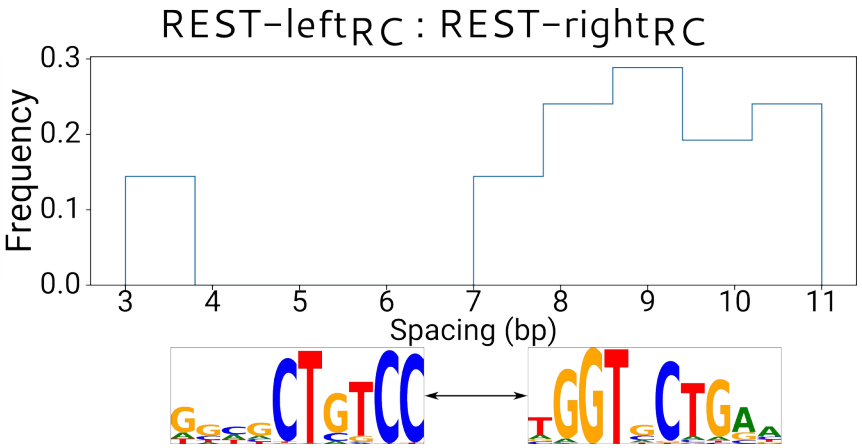
Clustering importance scores leads to redundancy and non-motif patterns:
Fundamental limitations in importance scores and in clustering

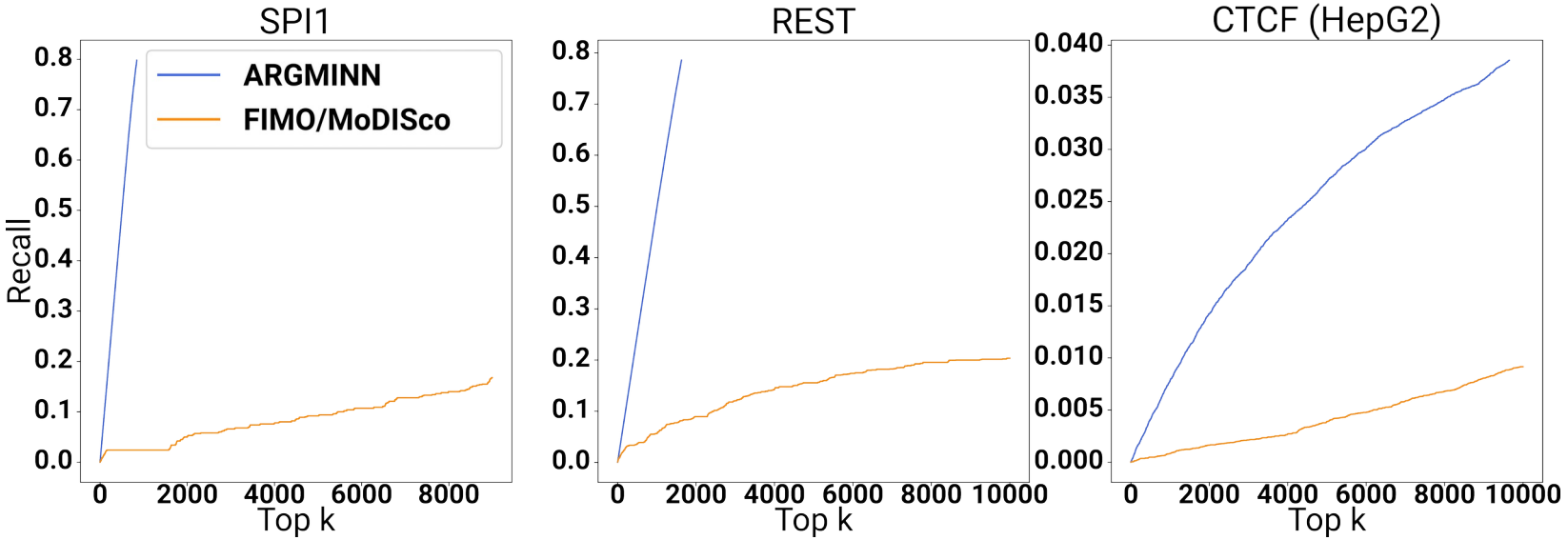
Extracting motif instances and syntax with ARGMINN



Motif instances and syntax are revealed with a single forward pass







Summary

- ARGMINN: a mechanistically interpretable neural network for regulatory genomics
- ARGMINN encodes biologically relevant motifs in its filters in a one-to-one fashion
- ARGMINN reveals motif instances and syntax with a single forward pass
- Through improved interpretability, the quality of motifs and their instances surpasses current state-of-the-art methods

Acknowledgments




Thanks for listening!

tseng.alex@gene.com

Biology Research | AI Development

Genentech
A Member of the Roche Group