STANFORD SCHOOL of MEDICINE | Genetics

STANFORD COMPUTER SCIENCE

Deep learning frameworks for regulatory genomics and epigenomics

Chuan Sheng Foo

Avanti Shrikumar

Nicholas Sinnott-Armstrong

ANSHUL KUNDAJE

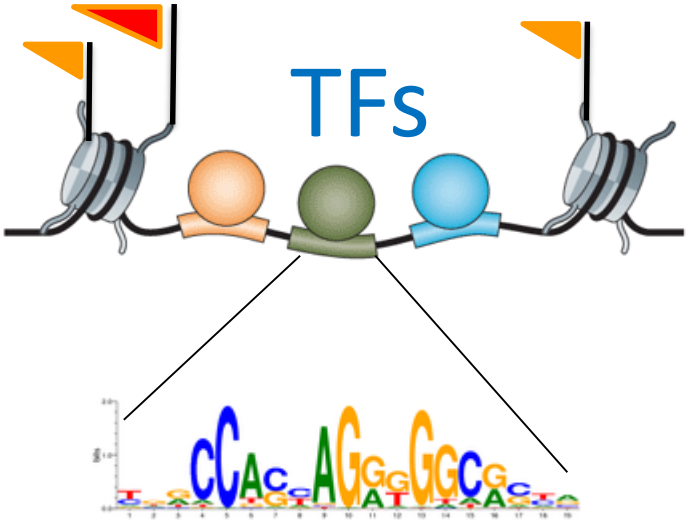
Genetics, Computer science

Stanford University

Johnny Israeli

Local **chromatin architecture** of regulatory elements

histone marks

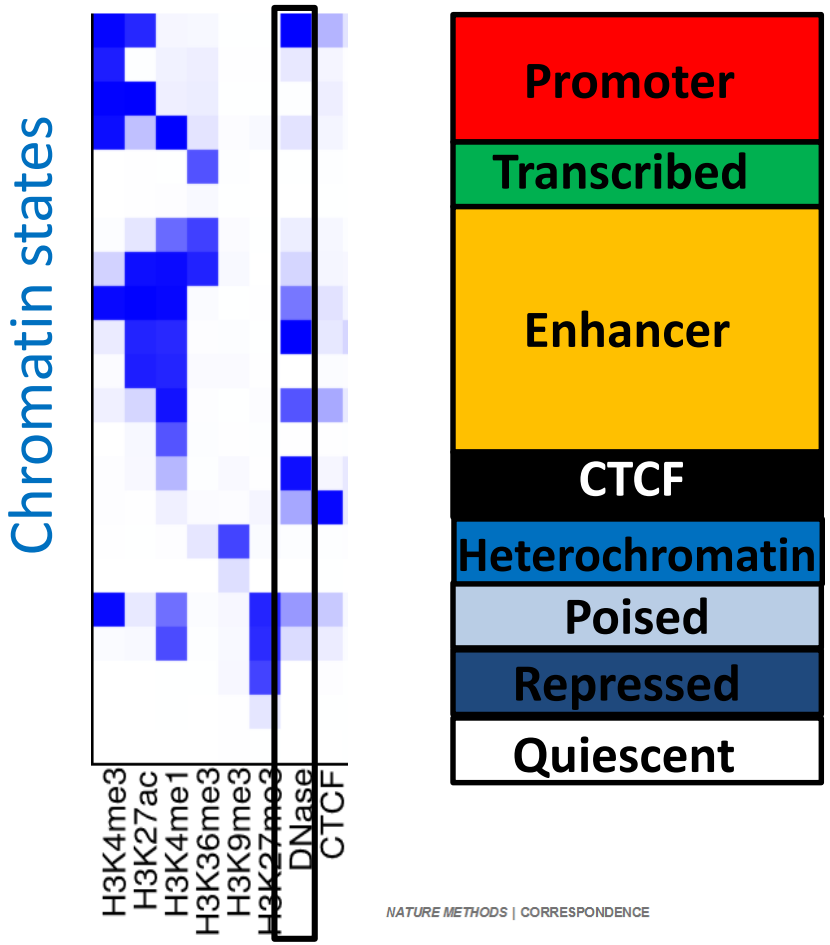


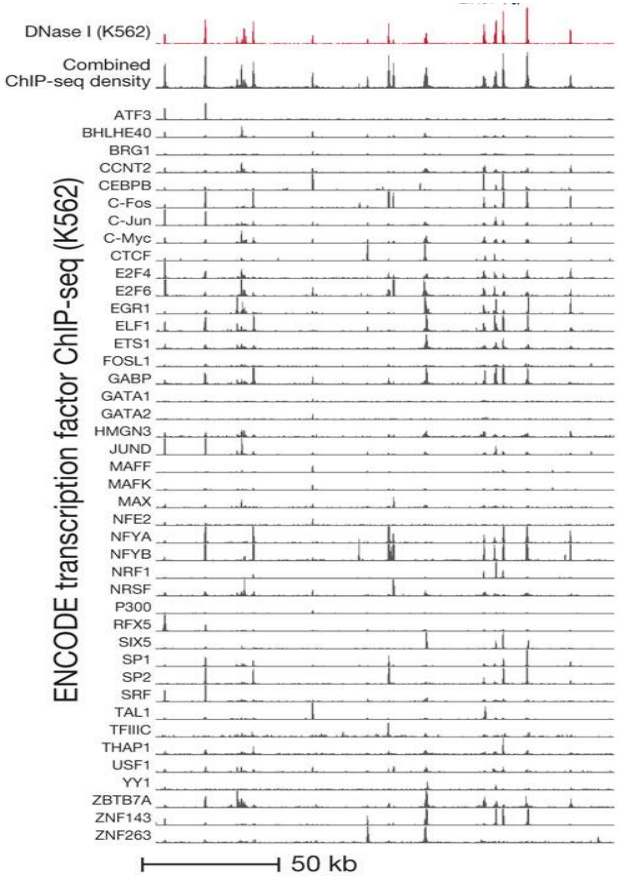
nucleosomes

sequence motifs

Adapted from Shlyueva et al. (2014) Nature Reviews Genetics.

Combinatorial **chromatin states** define broad classes of elements





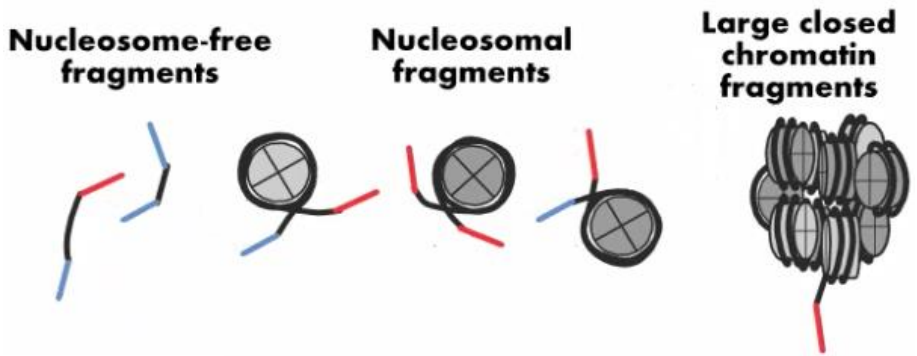
Thurman et al. (2012) Nature

NATURE METHODS | CORRESPONDENCE

ChromHMM: automating chromatin-state discovery and characterization

Jason Ernst & Manolis Kellis

ATAC-seq: genome-wide **chromatin accessibility** from low input material

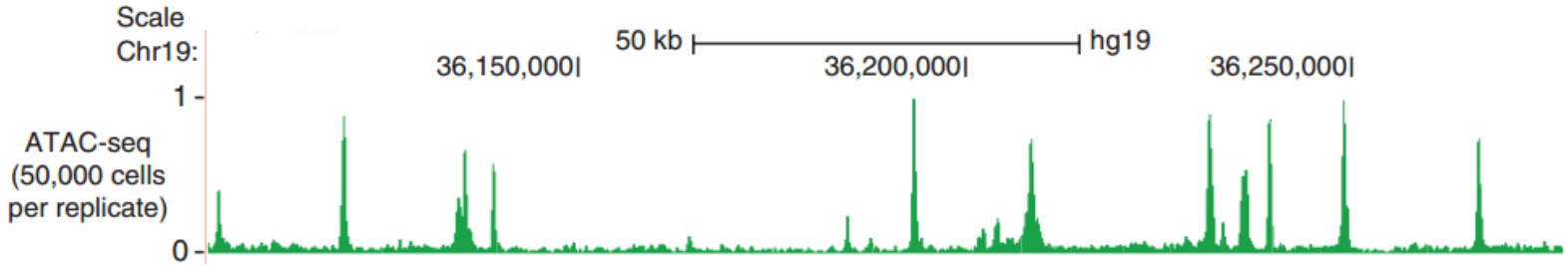


NATURE METHODS | ARTICLE

Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position

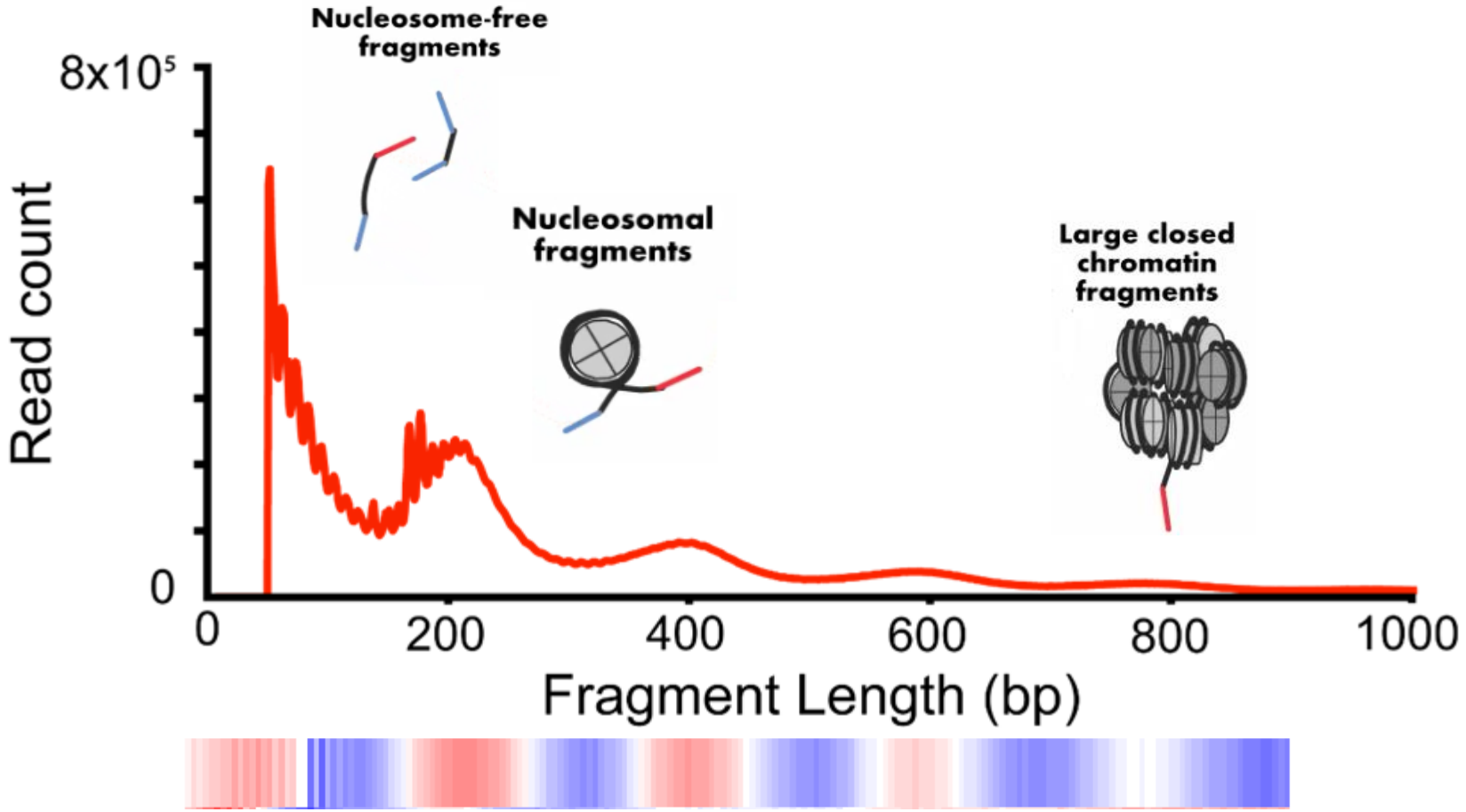
Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang & William J Greenleaf

Paired-end sequencing



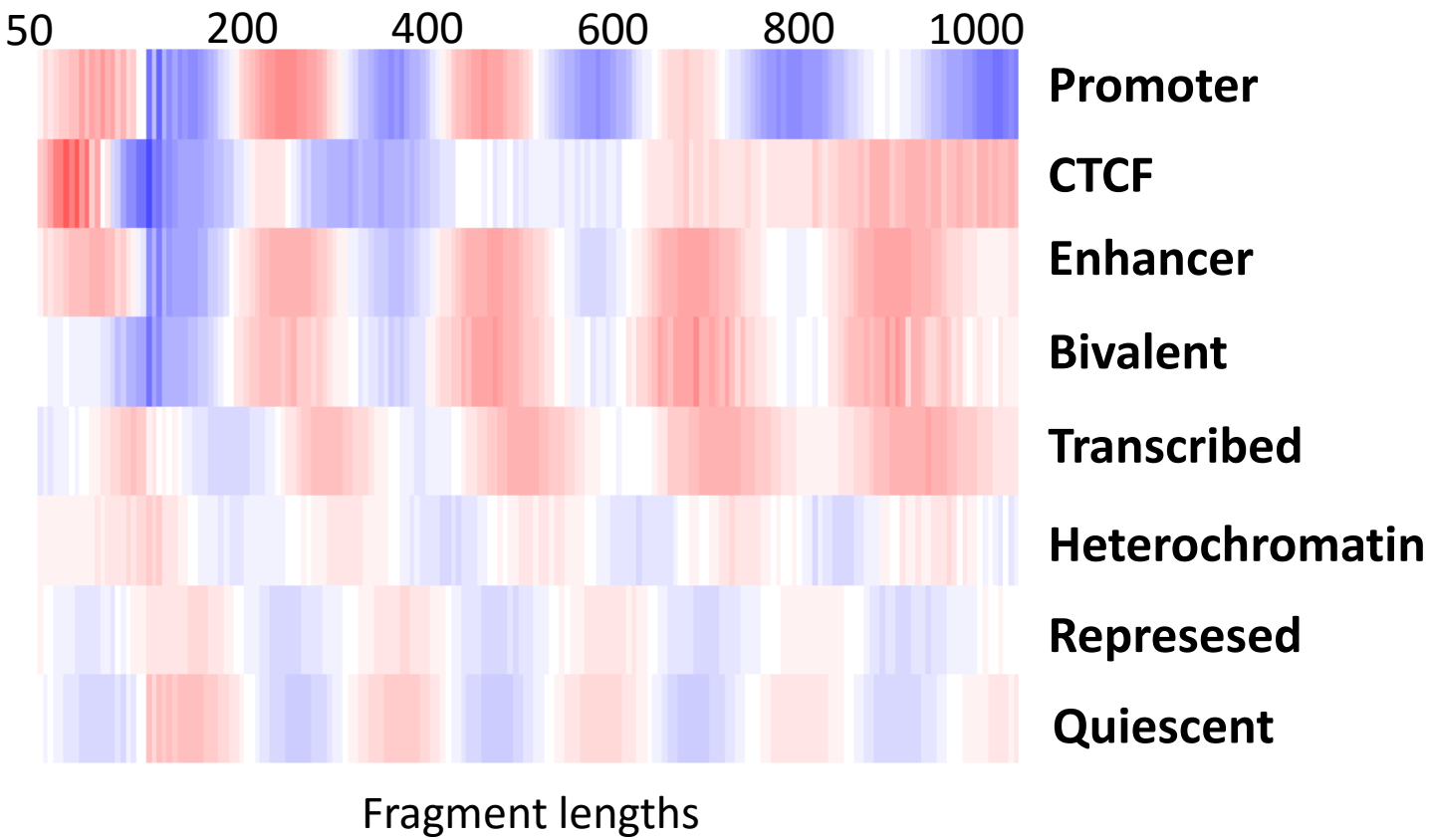
ATAC-seq peaks identify chromatin accessible regulatory elements

ATAC-seq reveals chromatin architecture in genome-wide **fragment length distributions**

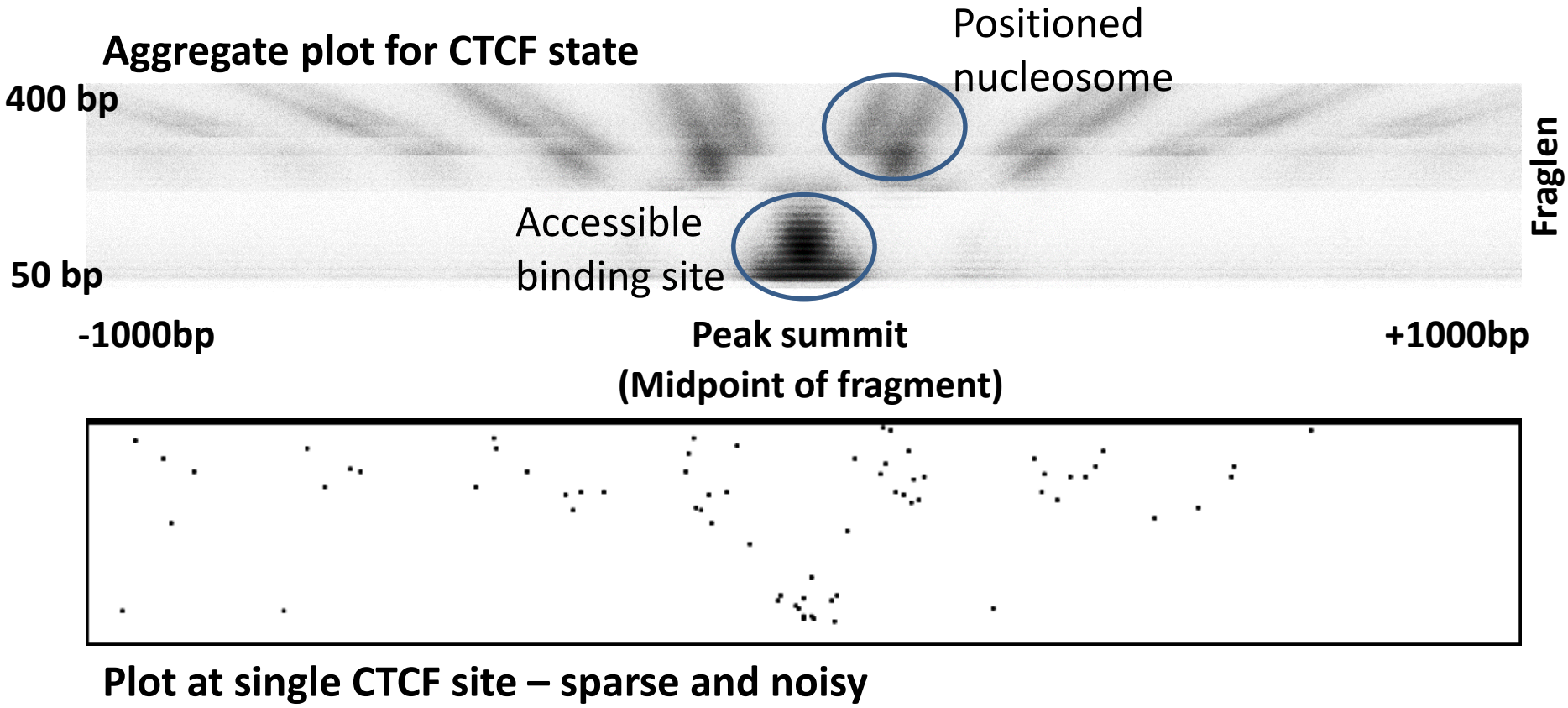


Buenrostro et al. (2013) Nature Methods.

Chromatin architecture reflects chromatin state

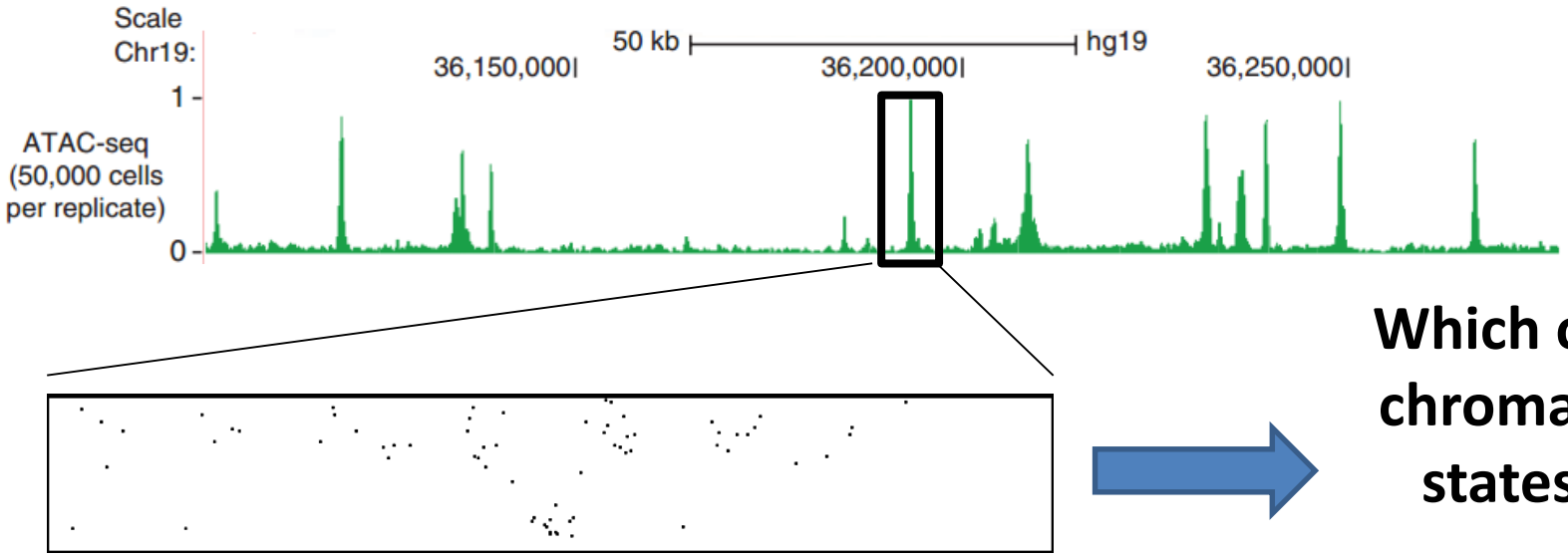


Position-aware 2D fragment length distributions (V-plots)



V-plots were first introduced by Henikoff et al. 2011, PNAS

Can we predict chromatin states/histone marks at ATAC-peaks?



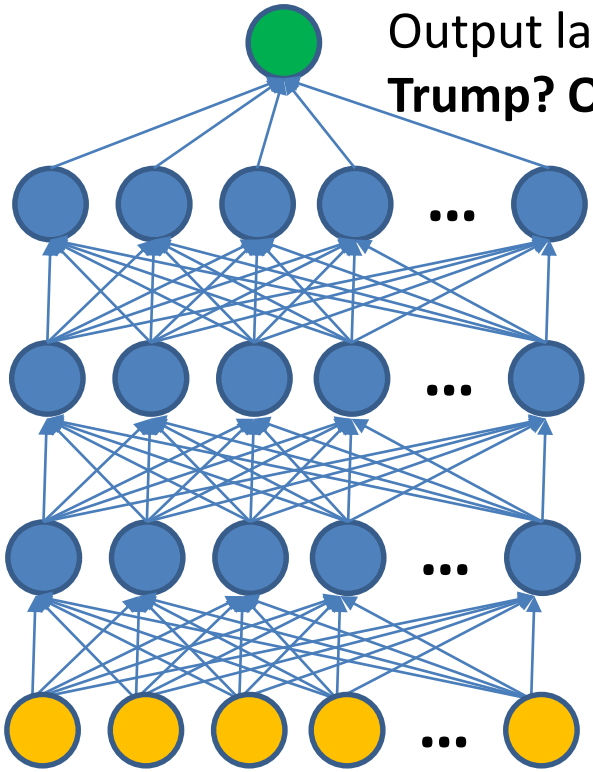
2kb around ATAC-peak

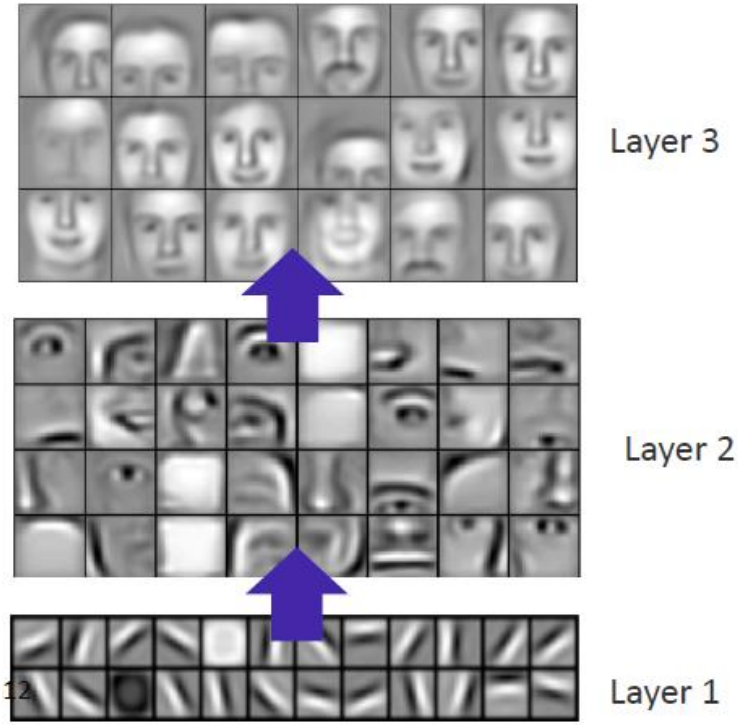
Which of 8 chromatin states?

Which histone mark is present?

Image classification task!

Deep neural networks (DNNs) for image classification

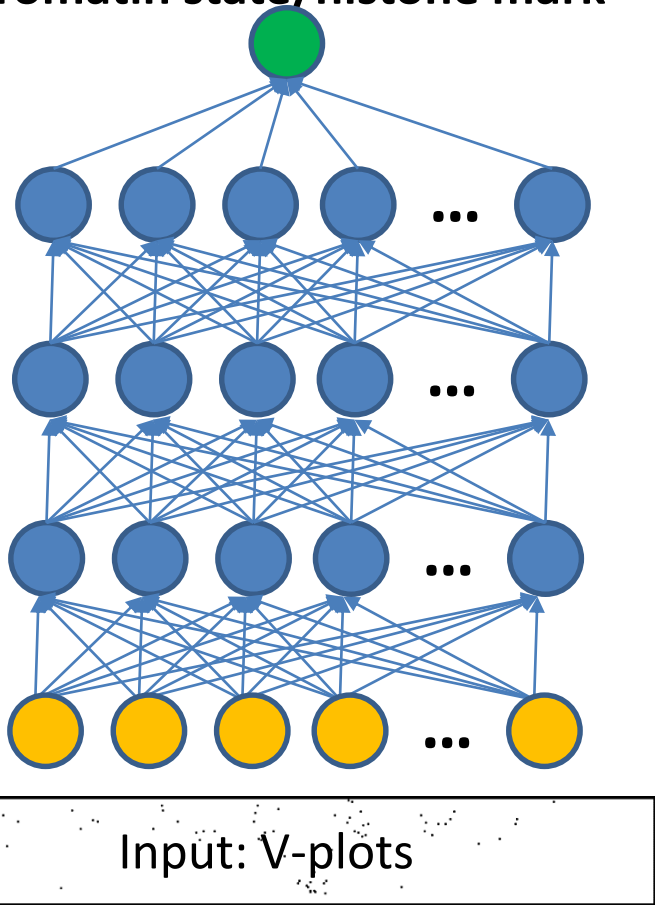




Lee et. al. (2009), ICML

Deep neural networks (DNNs) for V-plot classification

Output label:
Chromatin state/histone mark

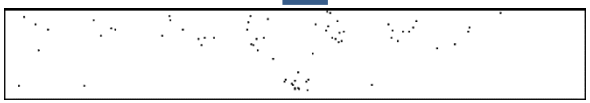


Chromatin state/histone mark

Nucleosome arrays,
Large regions of
open chromatin

+1 / -1 nucleosome,
Segments of open
chromatin

Nucleosomes,
Open chromatin

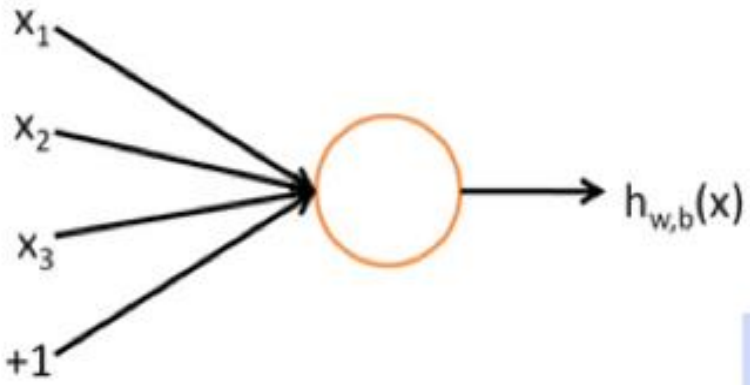


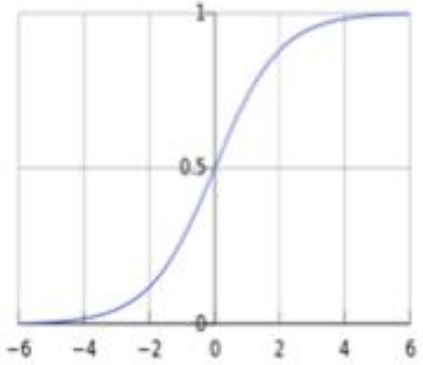
An artificial neuron

$$h_{w,b}(x) = f(w^T x + b)$$

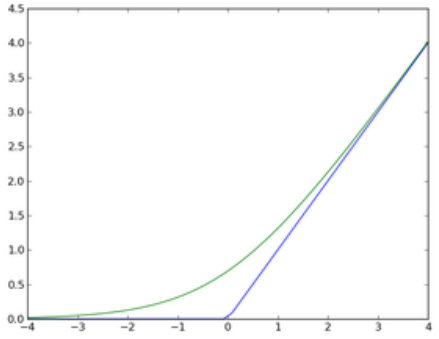
b: We can have an "always on" feature, which gives a class prior, or separate it out, as a bias term

$$f(z) = \frac{1}{1+e^{-z}}$$





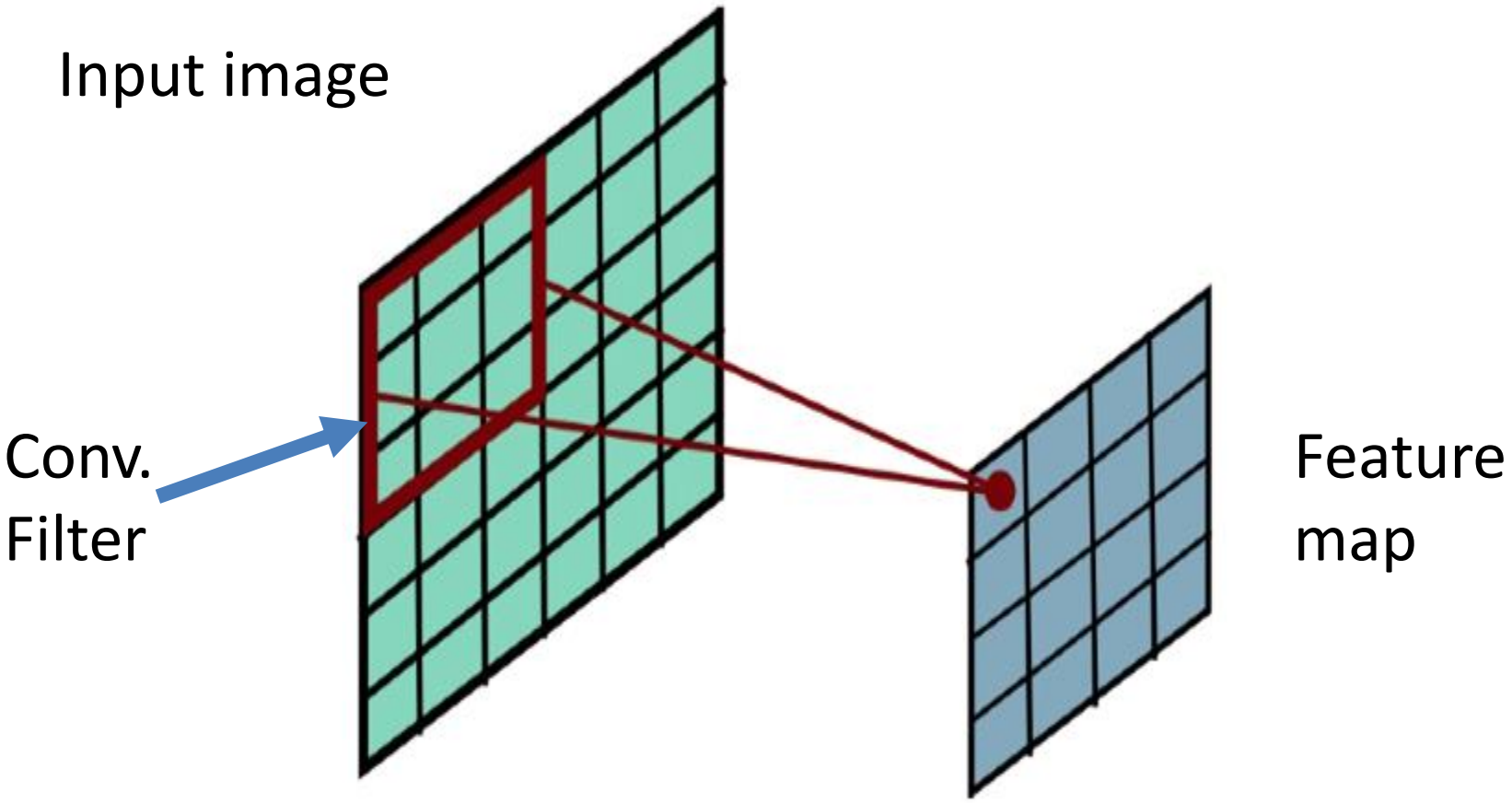
Sigmoid



ReLu (Rectified Linear Unit)

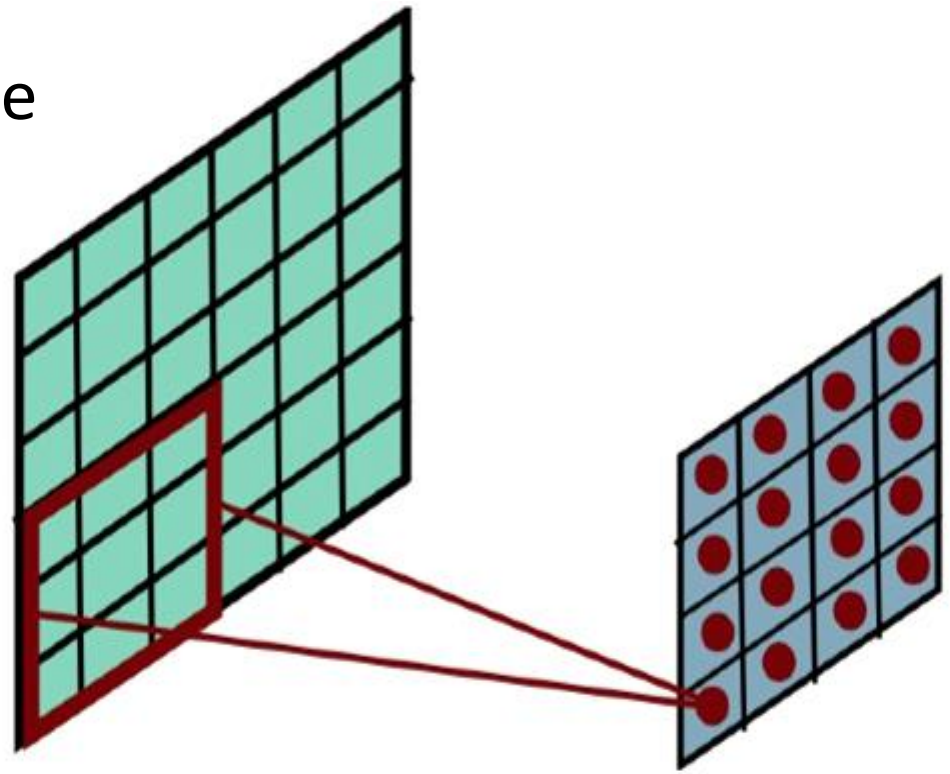
w, *b* are the parameters of this neuron i.e., this logistic regression model

Convolutional filters



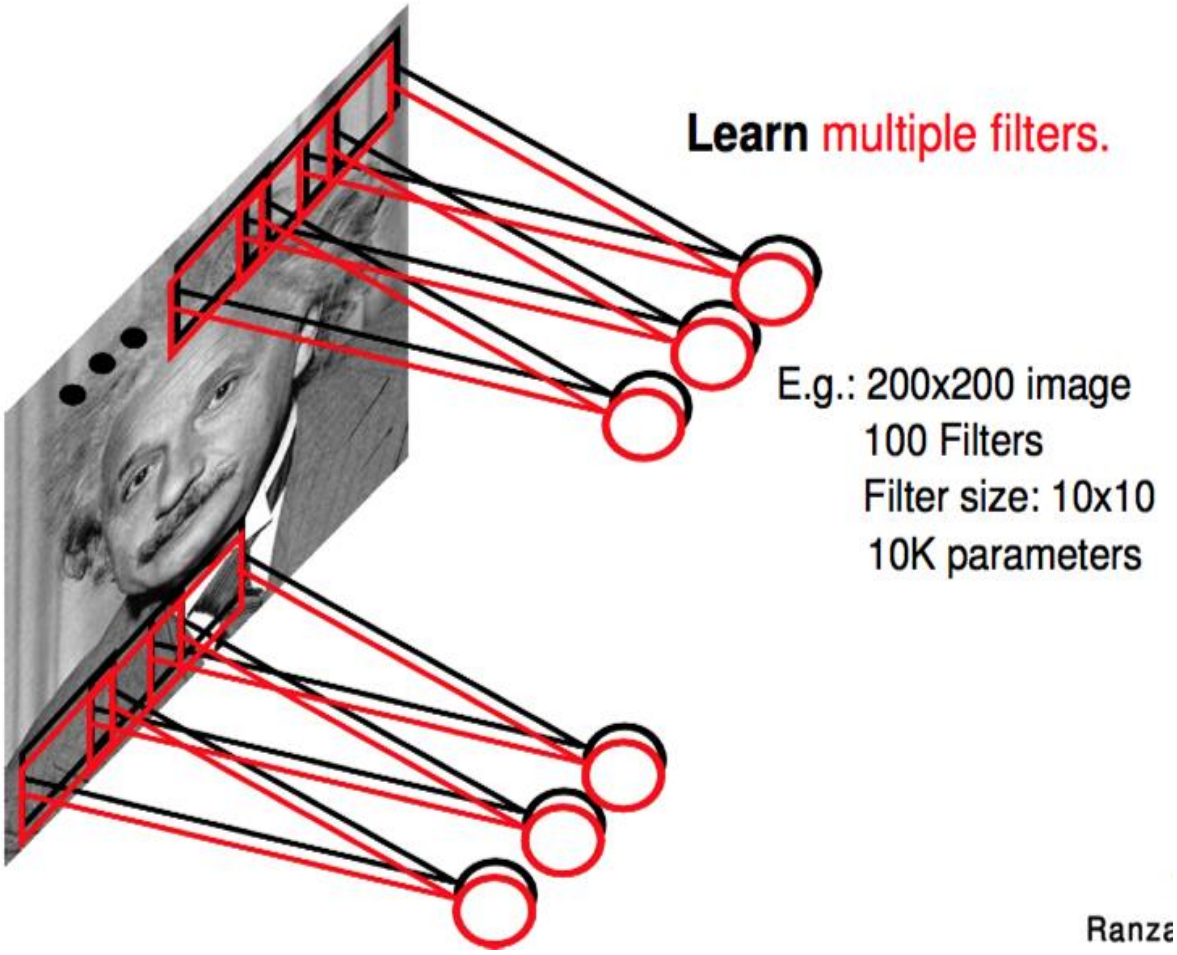
Convolutional filters

Input image

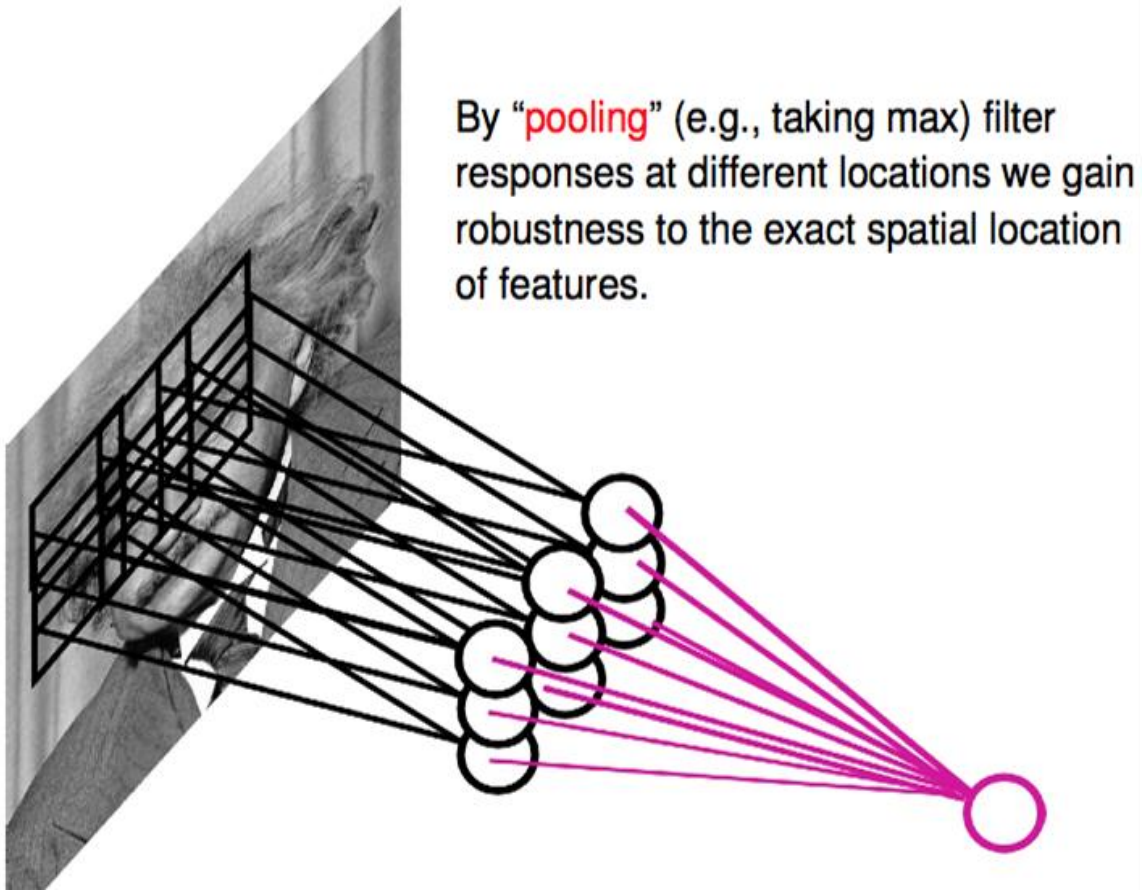


Feature map

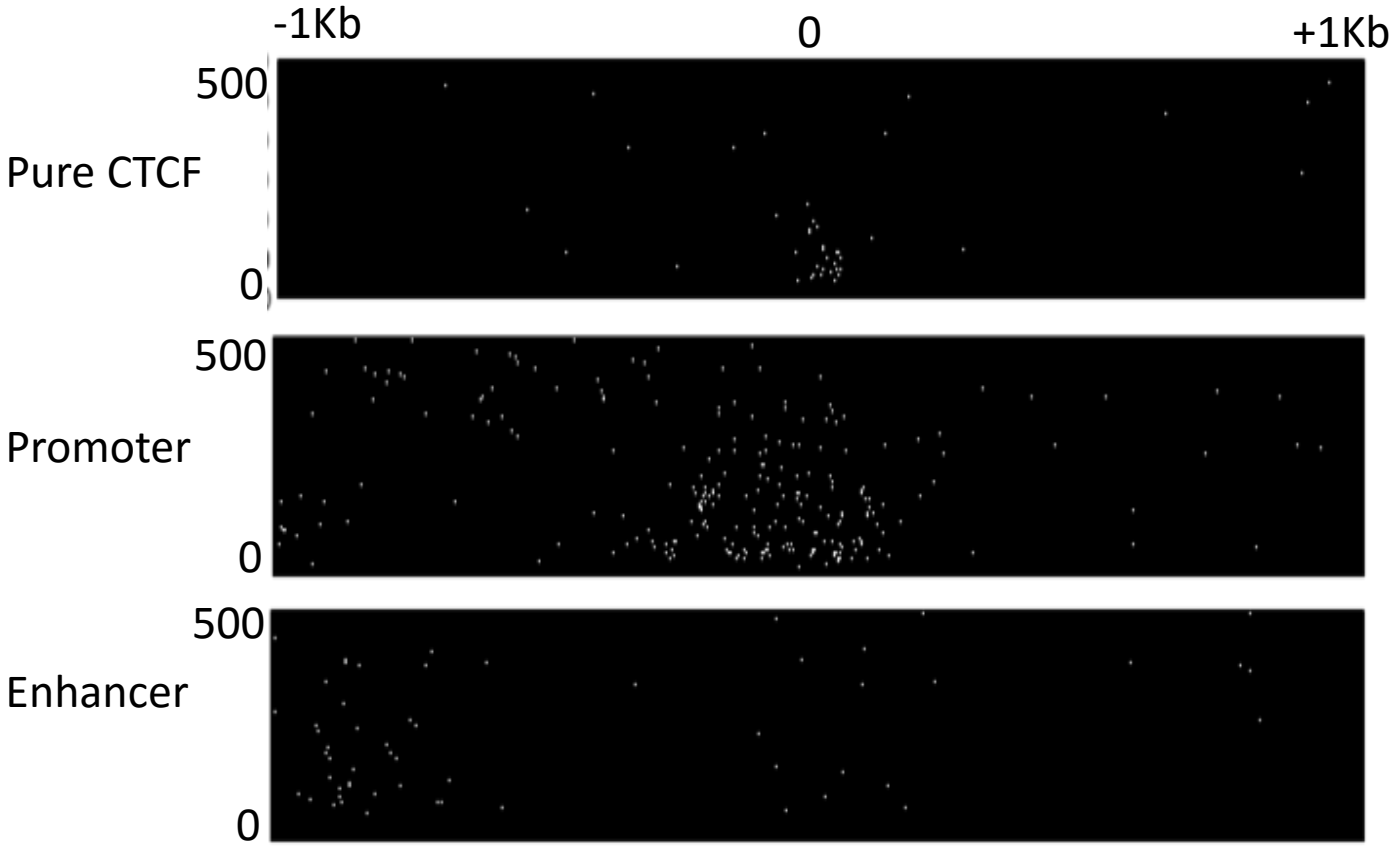
Convolutional layer: multiple filters learn distinct features

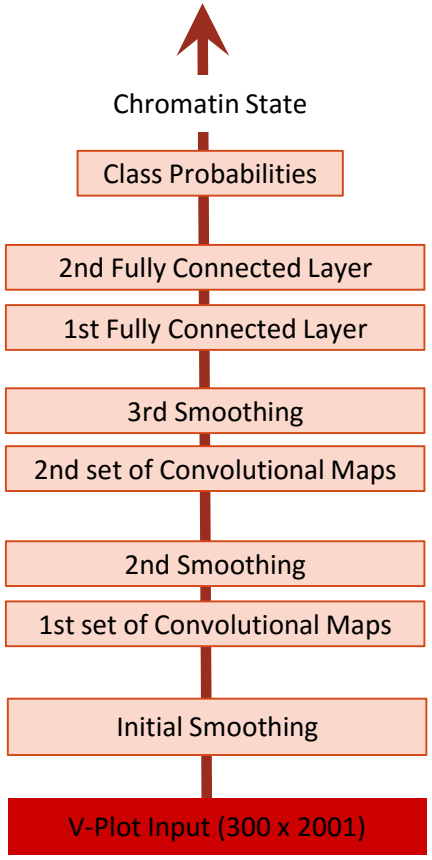


Pooling layers: locally smooth signal



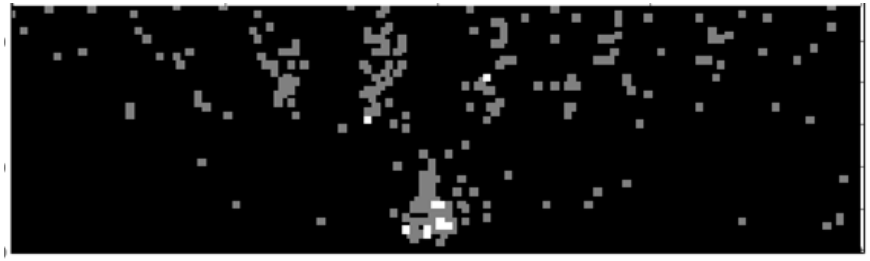
How does a deep conv. neural network transform the raw V-plot input at each layer



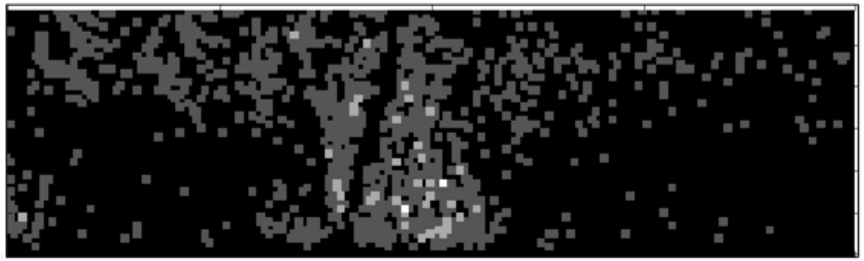


After initial pooling (smoothing)

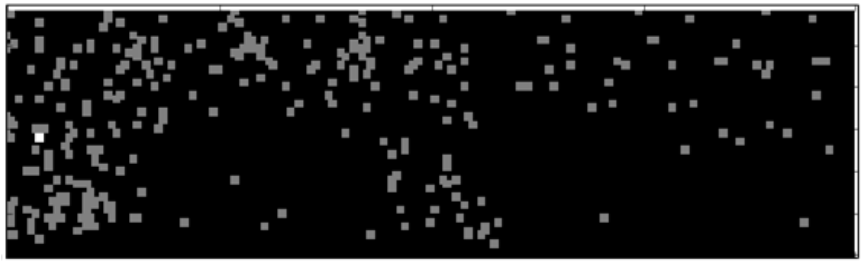
Pure CTCF

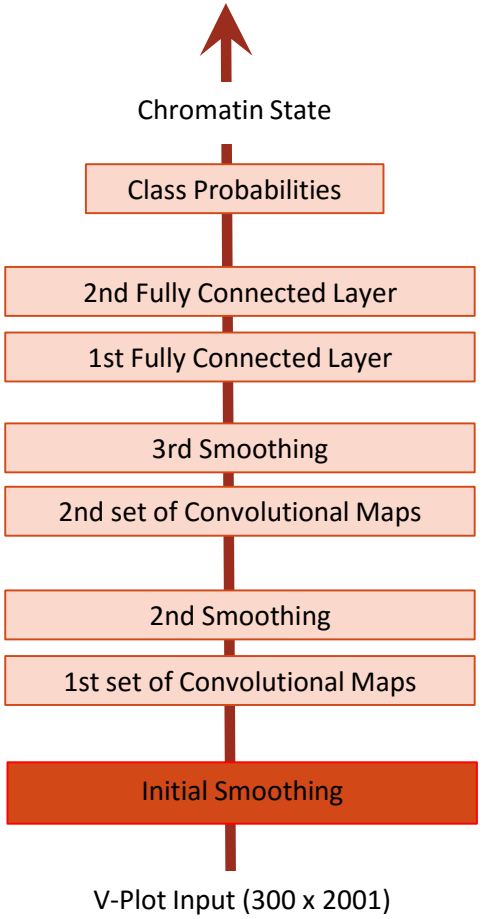


Promoter



Enhancer



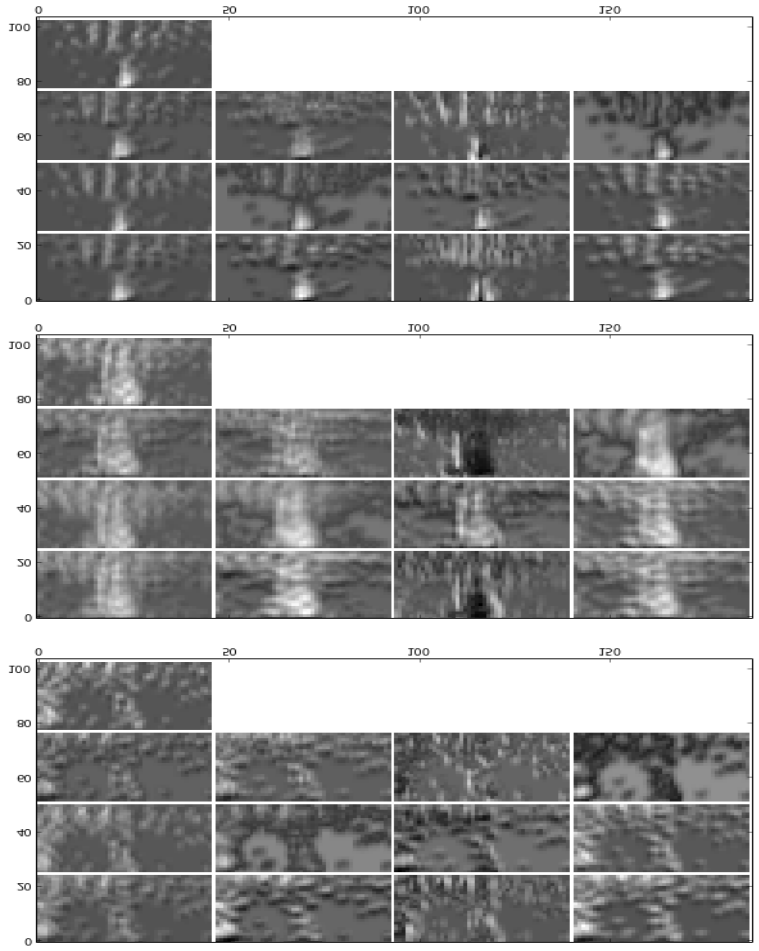


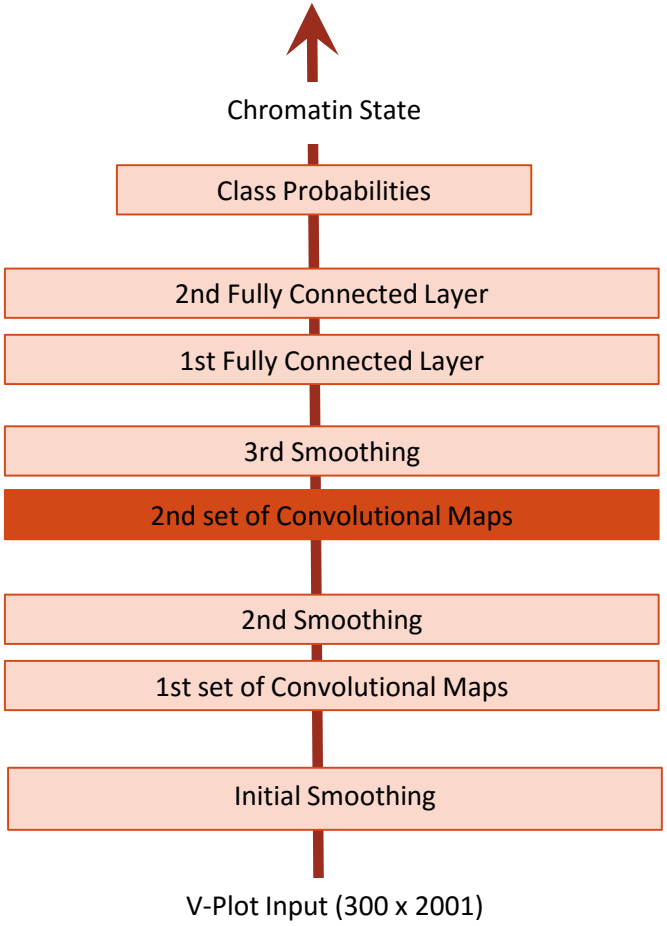
Second set of convolutional maps

Pure CTCF

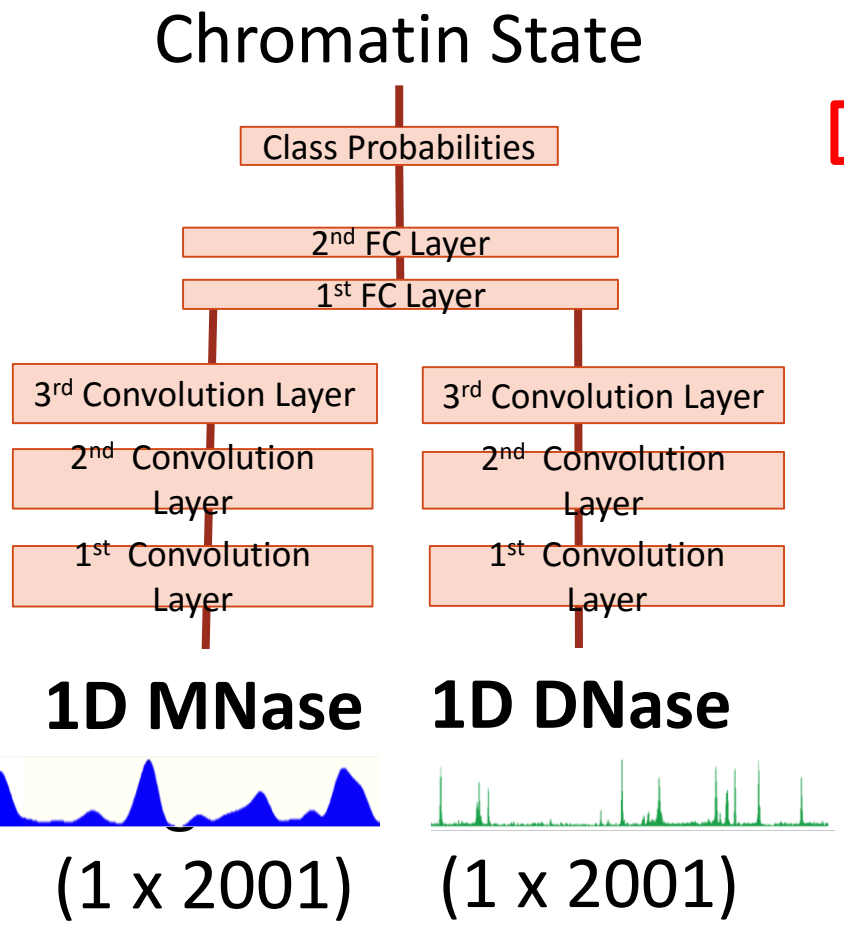
Promoter

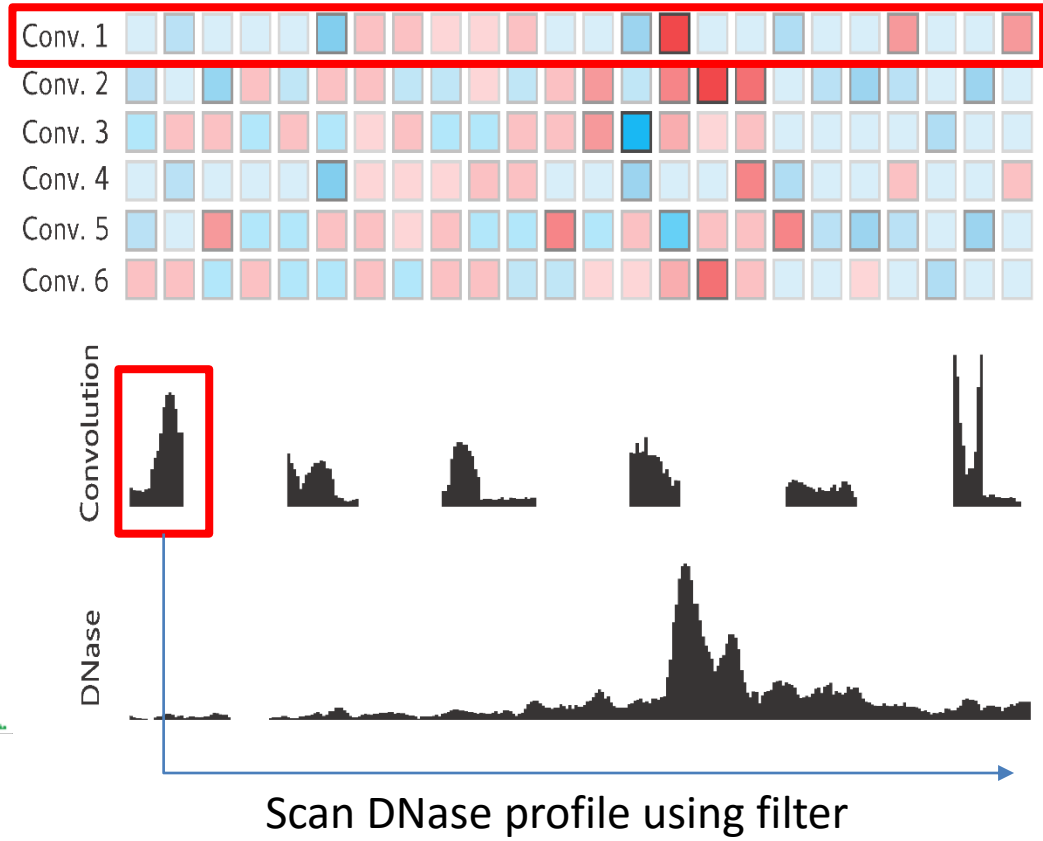
Enhancer



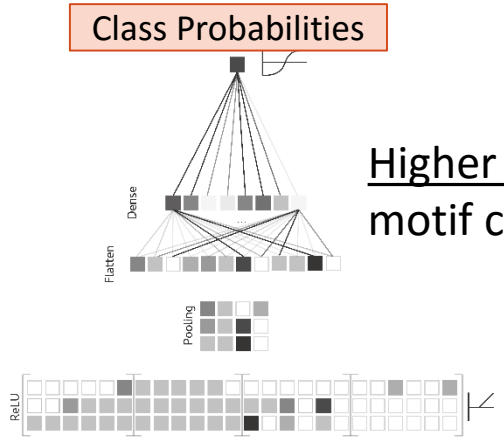


Learning from **multiple 1D functional data** (e.g. DNase, MNase)





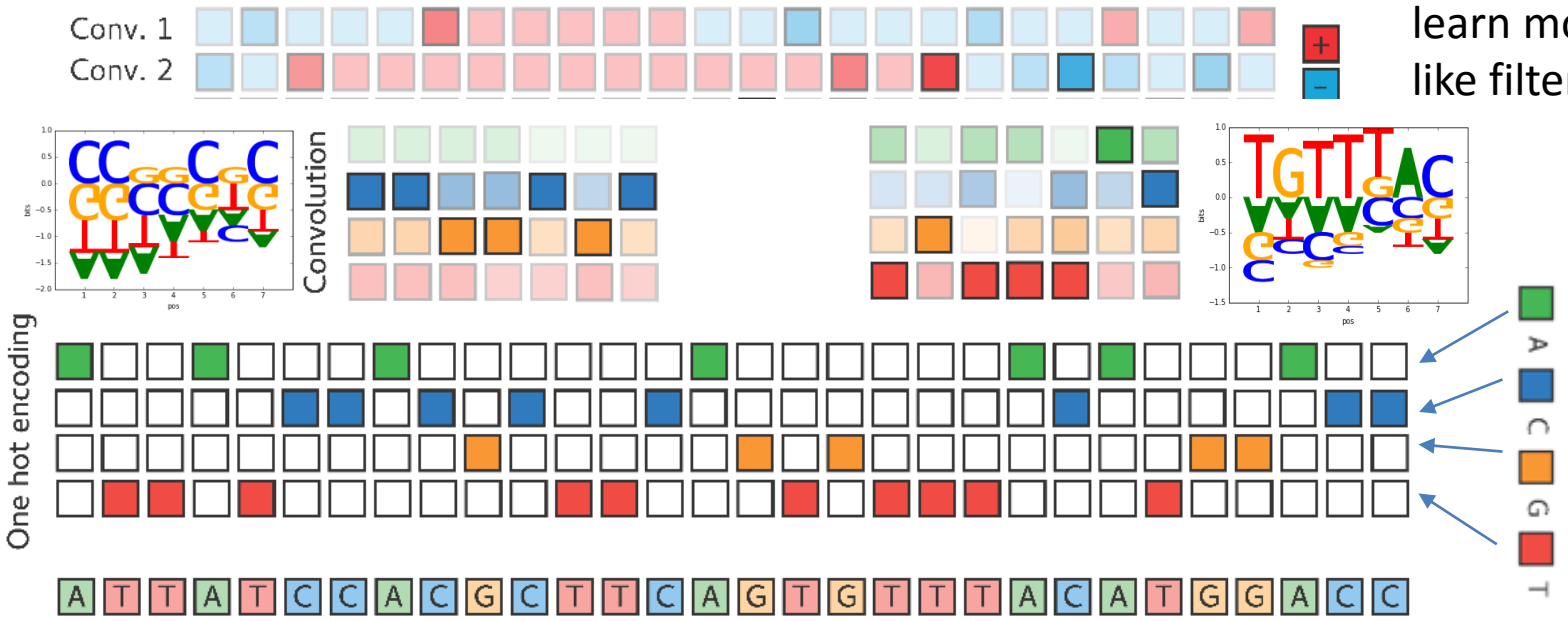
Learning from raw DNA sequence



Higher layers learn motif combinations

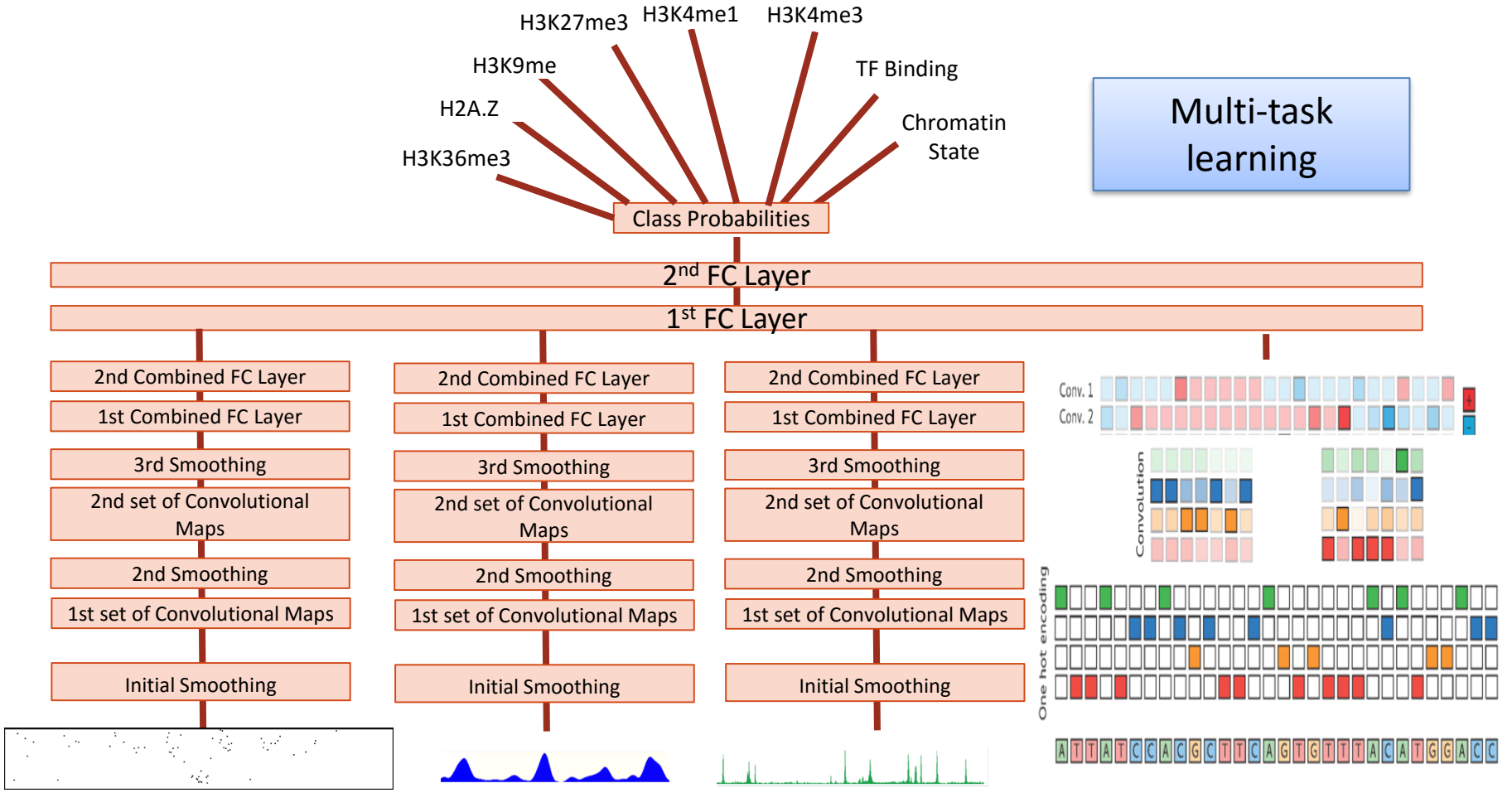
Score sequence using filters

Convolutional layers learn motif (PWM) like filters



THE CHROMPUTER

Integrating multiple inputs (1D, 2D signals, sequence) to simulatenously **predict multiple outputs**



Chromatin architecture can predict **chromatin state** in held out chromosome (same cell type)

Model + Input data types	8-class chromatin state accuracy (%)
Majority class (baseline)	**42%**
Gene proximity	59%
Random Forest: **ATAC-seq** (150M reads)	**61%**
Chromputer: **DNase** (60M reads)	68.1%
Chromputer: **Mnase** (1.5B reads)	69.3%
Chromputer: **ATAC-seq** (150M reads)	**75.9%**
Chromputer: **DNase** + **MNase**	**81.6%**
Chromputer: **ATAC-seq** + **sequence**	83.5%
Chromputer: **DNase** + **MNase** + **sequence**	**86.2%**
Label accuracy across replicates (upper bound)	**88%**

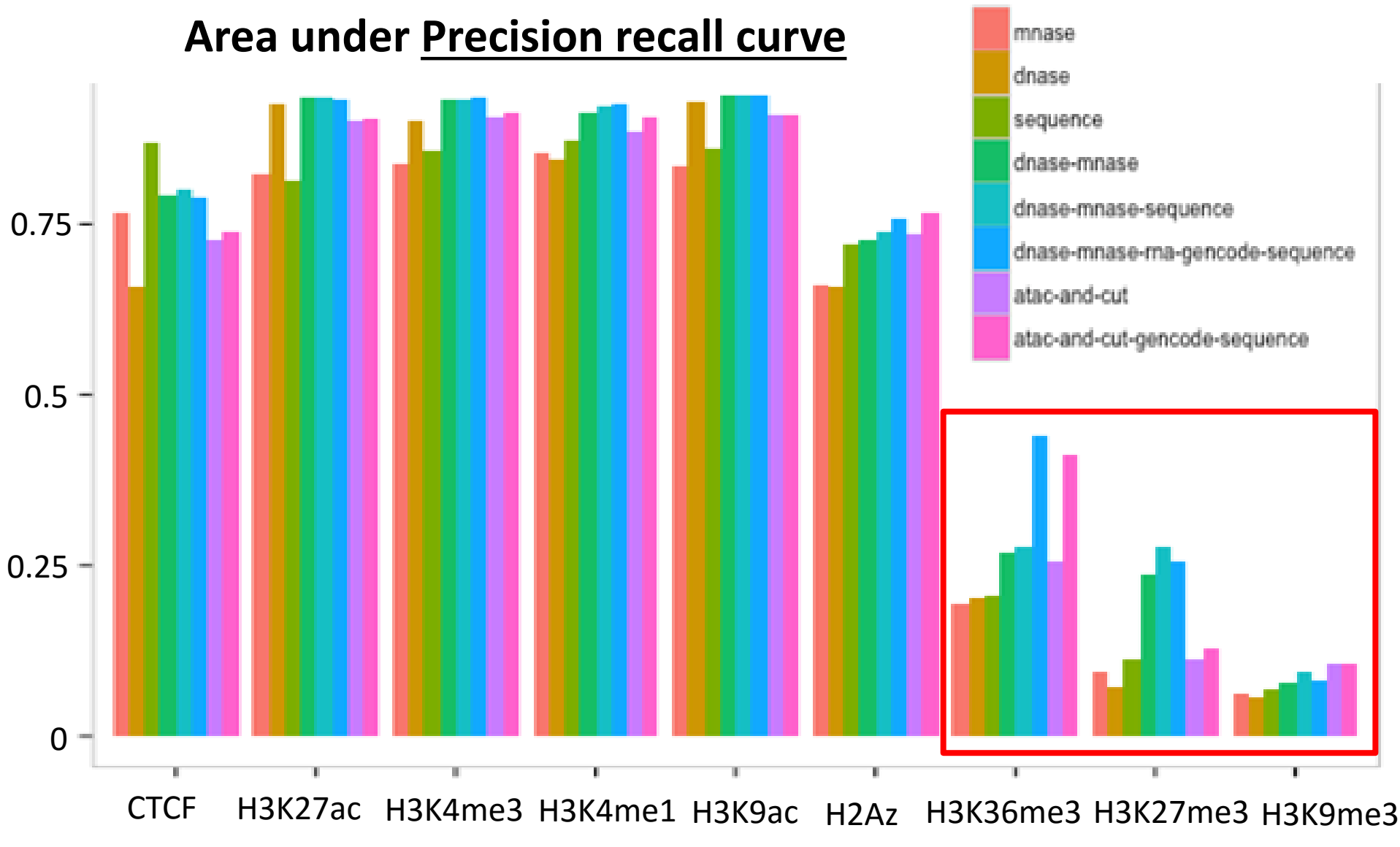
High cross cell-type chromatin state prediction

- Learn model on **DNase and MNase only**
- **Learn on GM12878, predict on K562 (and vice versa)**
- **Requires local normalization** to make signal comparable

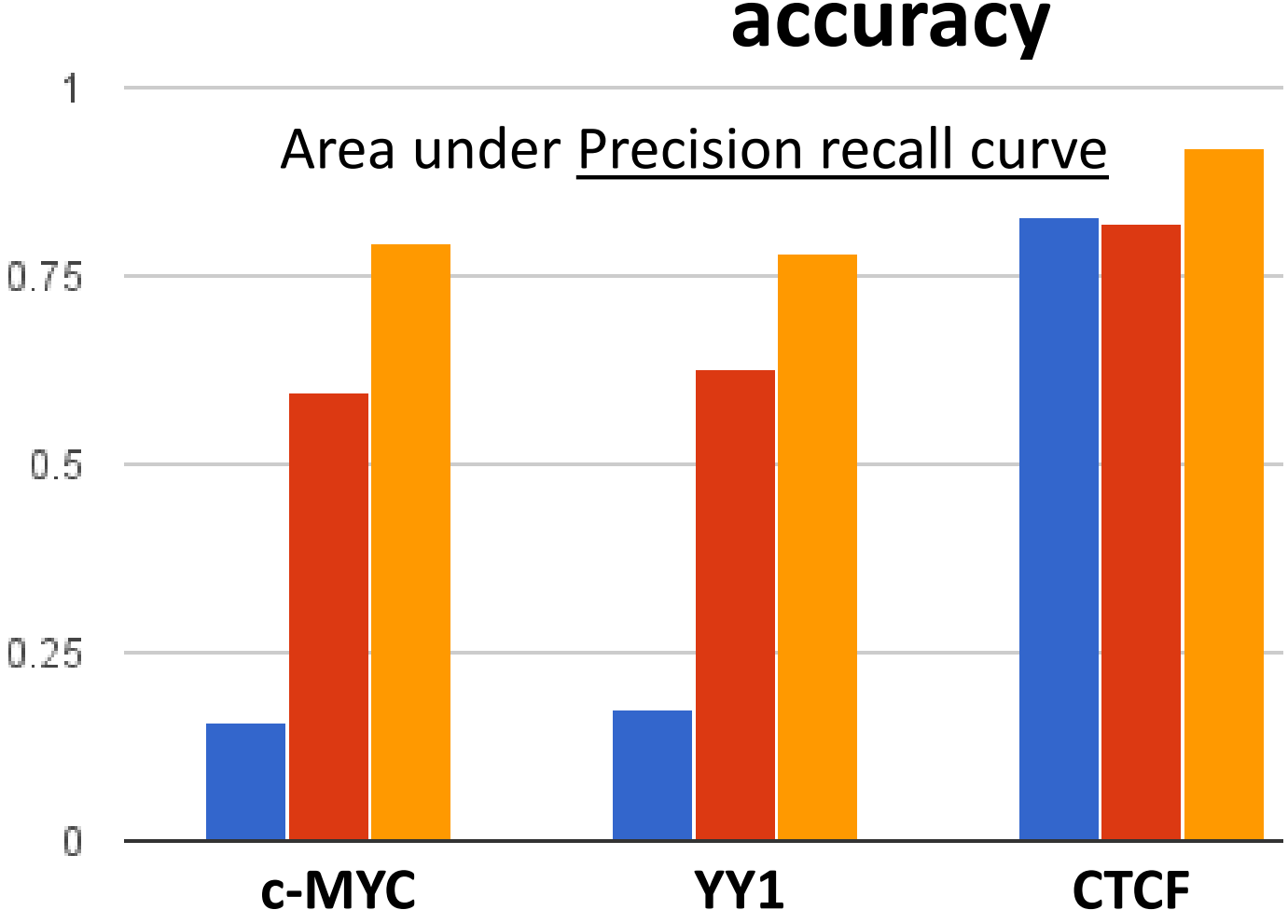
8 class chromatin state accuracy		
Train ↓ / Test →	GM12878	K562
GM12878	0.816	**0.818**
K562	**0.769**	0.844

Predicting individual histone marks

from ATAC/DNase/MNase/Sequence



Chromputer trained on TF ChIP-seq predicts cross cell-type in-vivo TF binding with high accuracy



DeepBind
DeepSEA
Chromputer

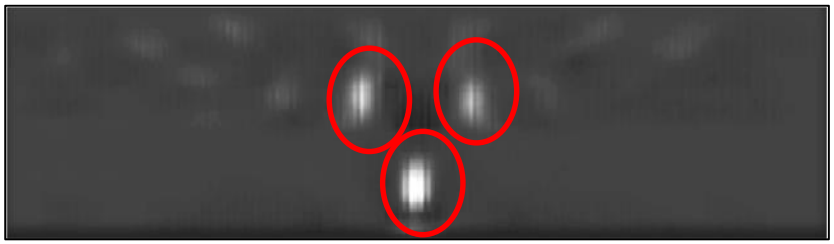
Inputs: Seq + DNA shape + DNase profile
Positives: Reproducible ChIP-seq peaks
Negatives: All other DNase peaks + flanks + matched random sites

Test sets: Held out chromosomes in held out cell types

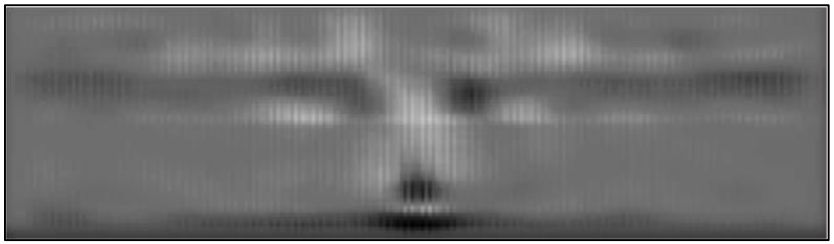
DeepLIFT: Scoring predictive power of features in Deep Neural Networks

- LIFT: Linear Importance Feature Tracker, or LIFTing the top off the black box.
- Provides a **predictive 'importance score'** for
  - any raw input feature (e.g. pixels in V-Plot images, each nucleotide in sequence)
  - intermediate learned features (e.g. convolutional filters)
- Linear breakdown of contribution of each input to immediate outputs
  - Recursively apply to get contribution of any input to any output
  - Can be **computed efficiently** with a single backpropagation (unlike in-silico mutagenesis)
  - **Less susceptible to buffering effects** than in-silico mutagenesis
- Technical details:
  - ReLU networks: equivalent to Taylor approximation of change in softmax/sigmoid logit if input eliminated.
  - i.e. gradient (w.r.t logit) * input

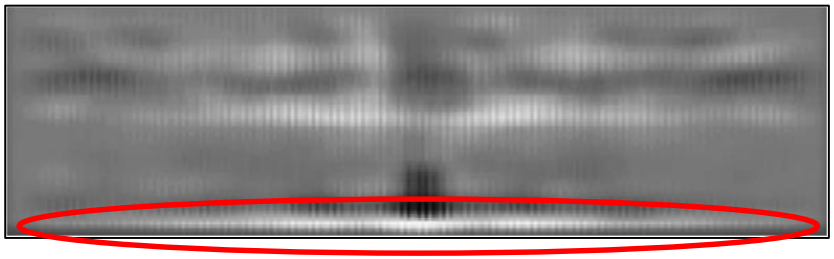
What architecture properties of the ATAC-seq V-plots predict different chromatin states?



CTCF state: centered binding, symmetric phased nucleosomes



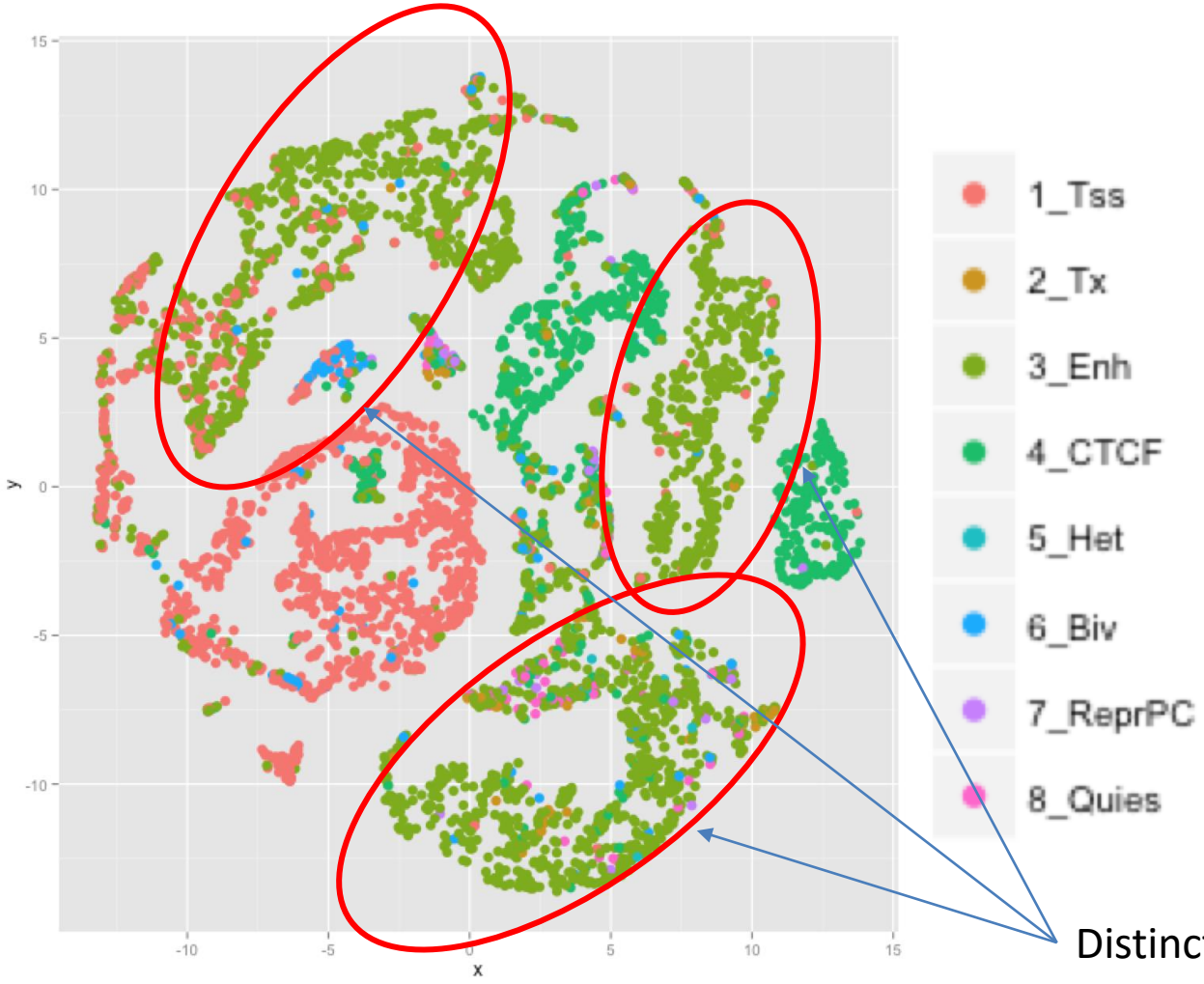
Enhancer state: localized signal, heterogeneity



Promoter state: broad regions of accessible chromatin

what is the change in classification probability relative to an unbiased classifier if we ***only*** consider the contributions from each pixel

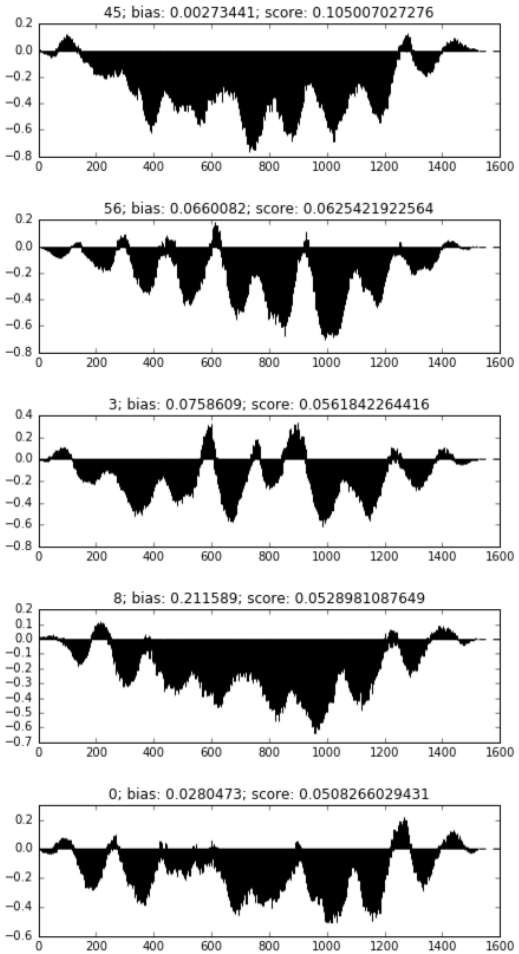
Architectural heterogeneity of accessible elements in different chromatin states

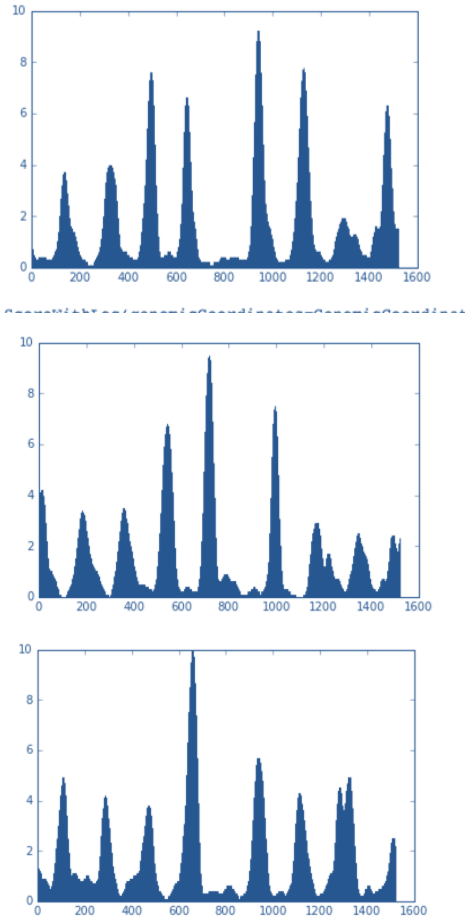


Low dimensional t-SNE embedding (like PCA) of accessible sites (colored by chromatin state)

Distinct classes of enhancers

Top scoring MNase filters and activating input patterns for CTCF state

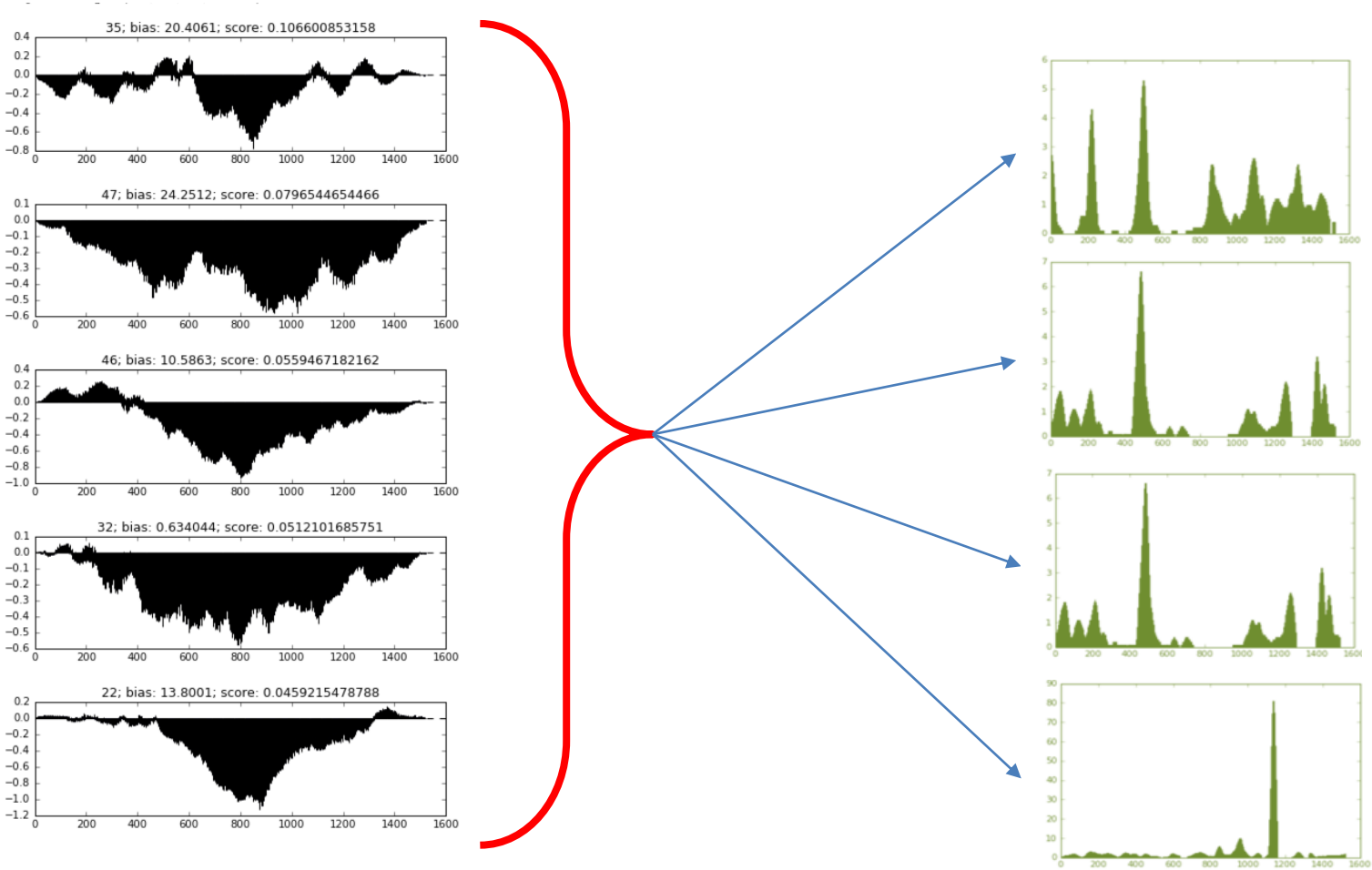




Top scoring MNase filters

Maximally activating input MNase profiles

Top scoring MNase filters and activating input patterns for promoter state



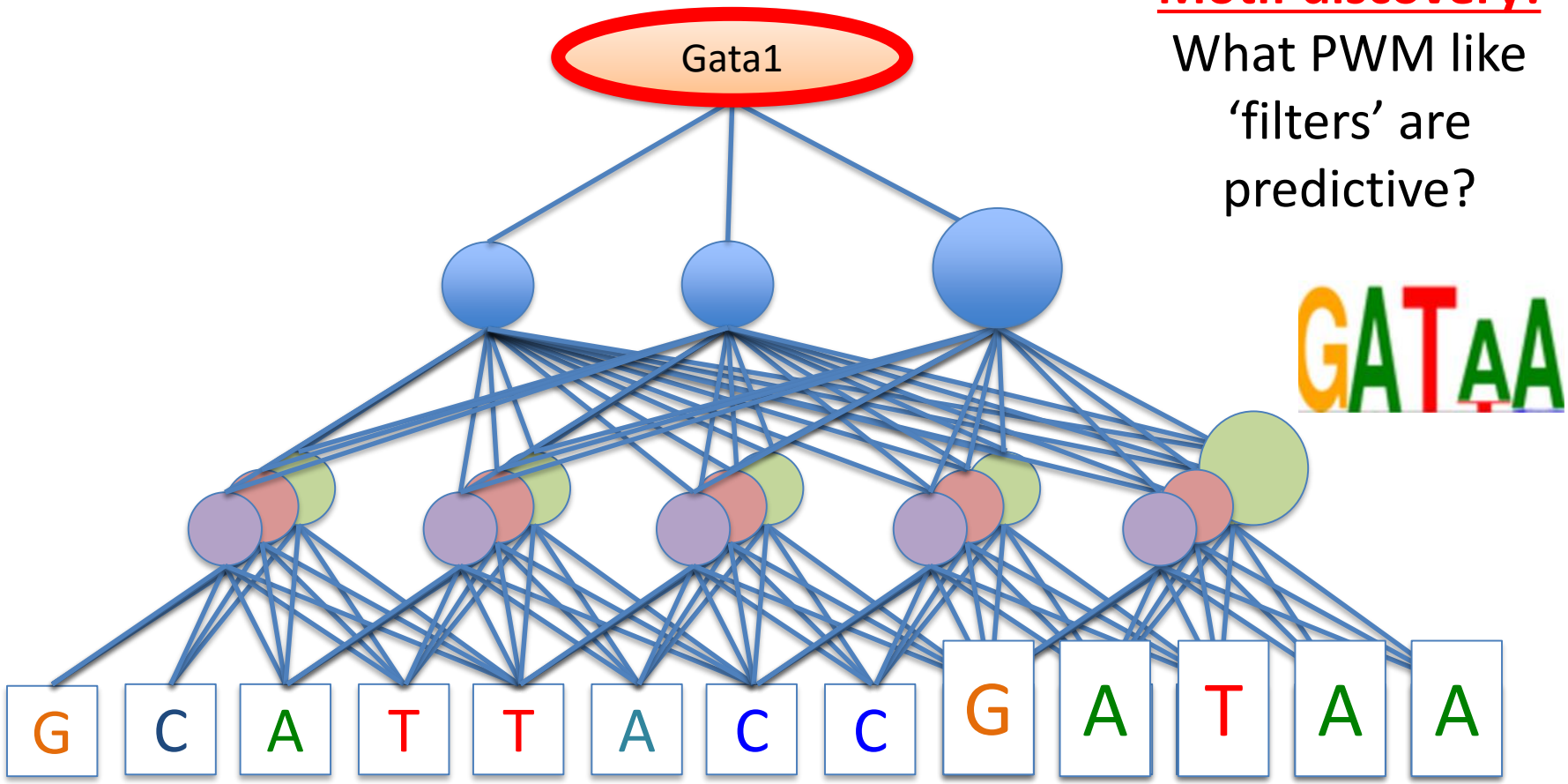
Top scoring MNase filters

Maximally activating input MNase profiles

What useful patterns can we extract from raw DNA sequence models?

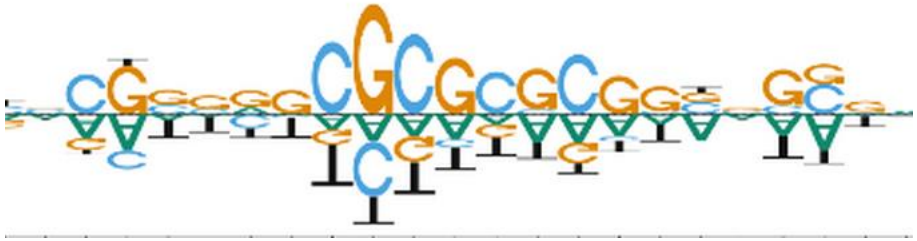
Motif discovery!

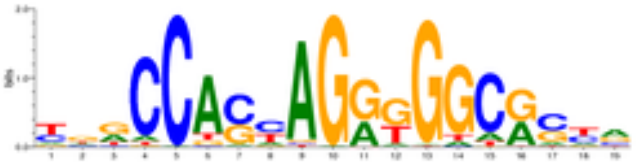
What PWM like 'filters' are predictive?



Which nucleotides in input sequence are contributing to binding!

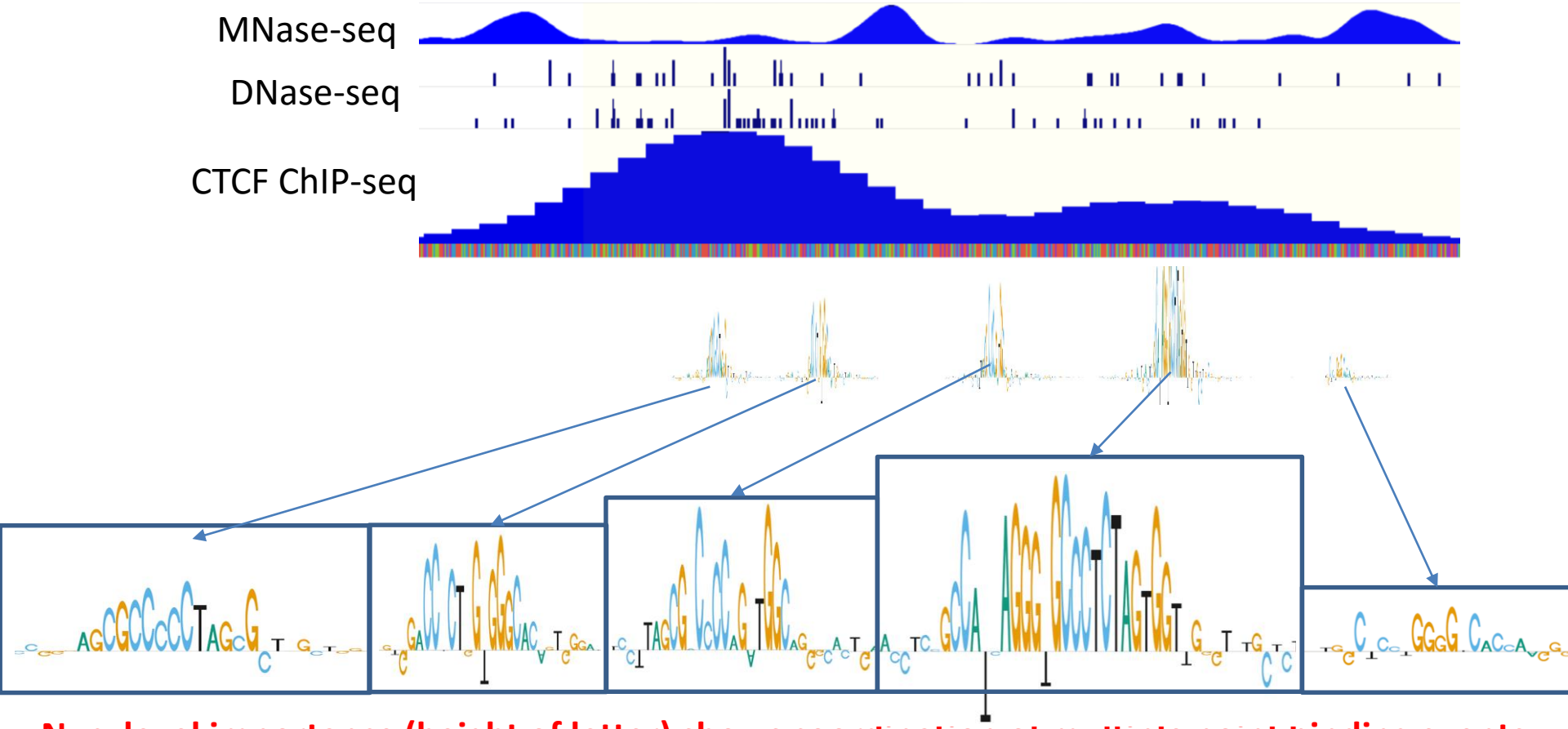
Top sequence filters for CTCF state





Canonical motif

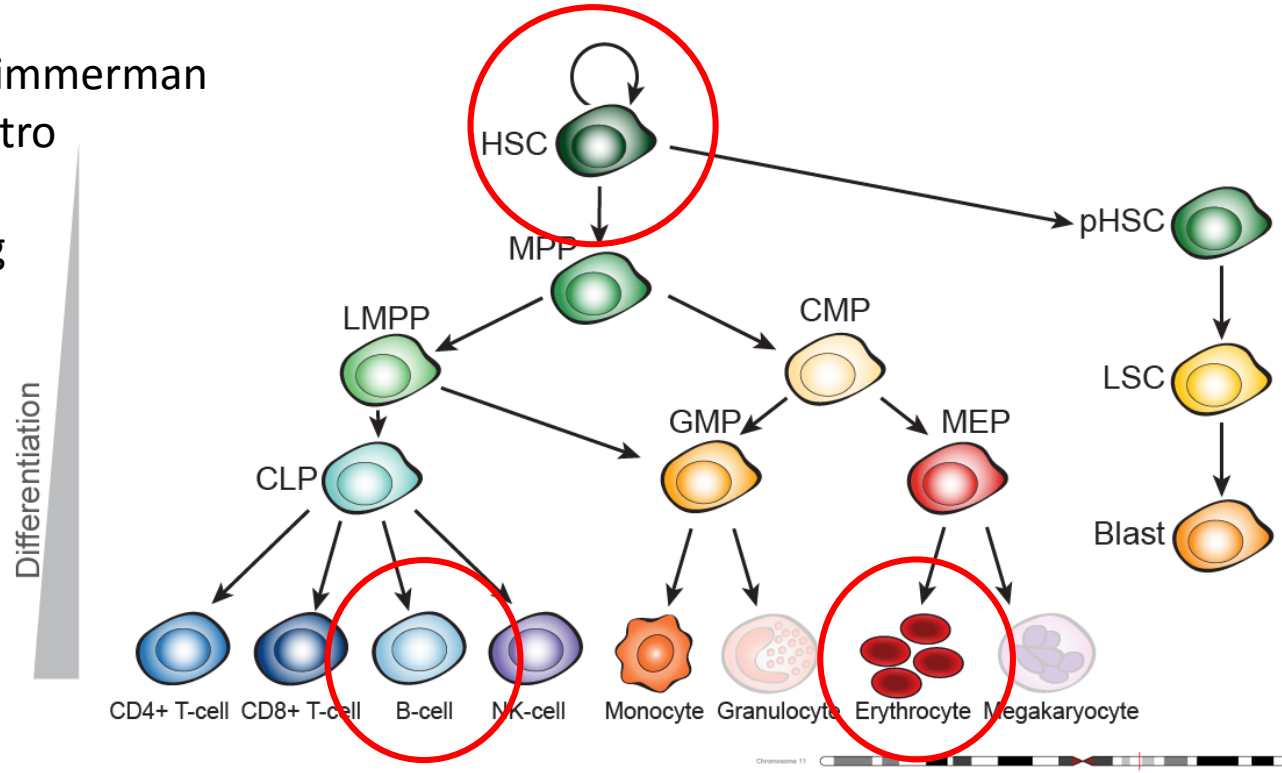
High resolution point binding events and sequence grammars at CTCF peaks

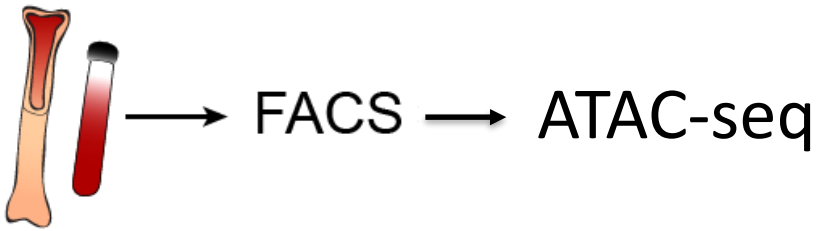


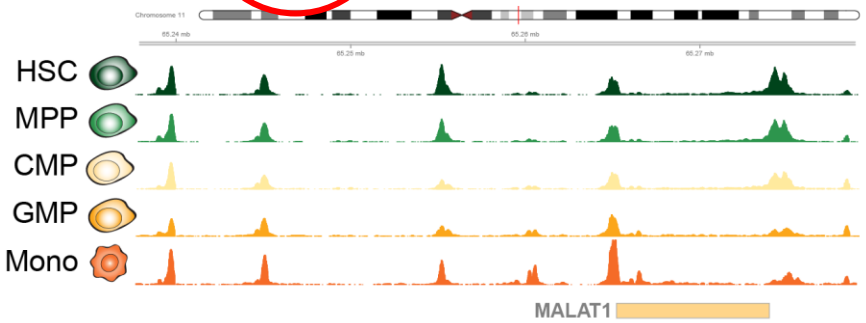
Nuc. level importance (height of letter) shows coordination of multiple point binding events

Context-specific reuse of regulatory sequence in chromatin accessibility changes during hematopoiesis

Ryan Corces-Zimmerman
Jason Buenrostro
Will Greenleaf
Howard Chang
Ravi Majeti

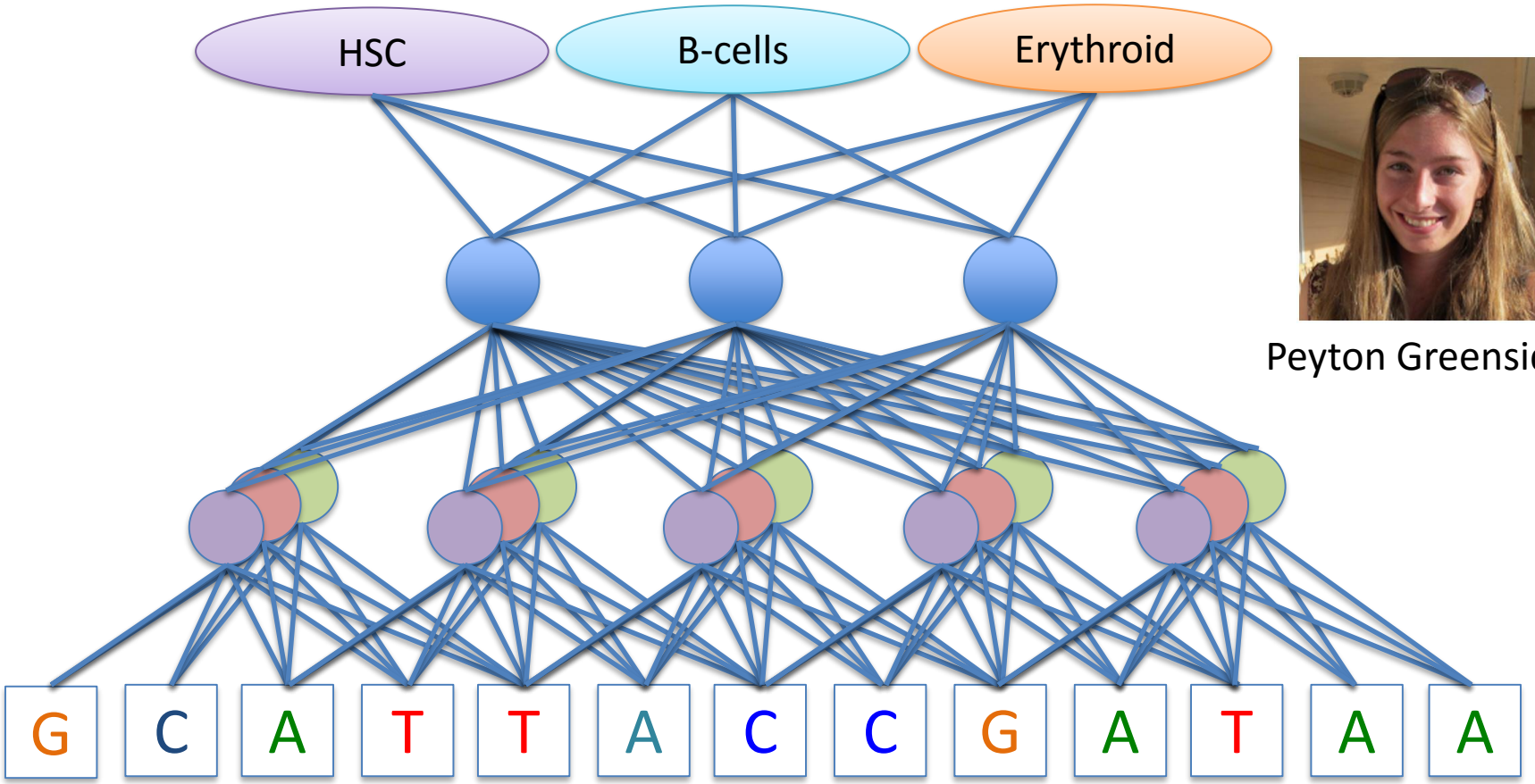






Deep learning sequence determinants of chromatin accessibility

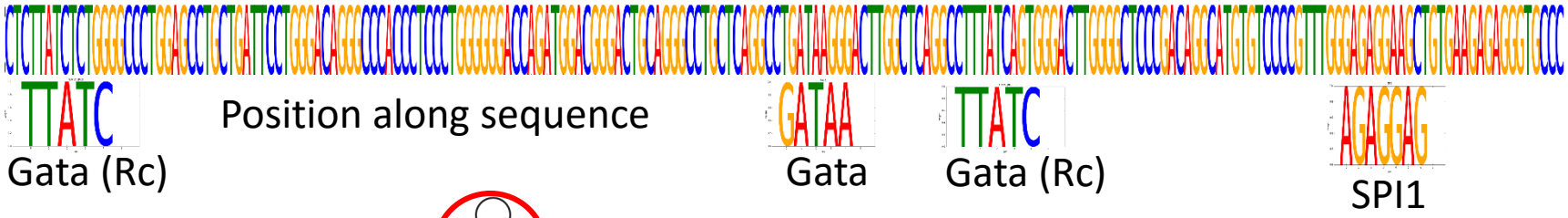
Output: Accessible (+1) vs. not accessible (0)

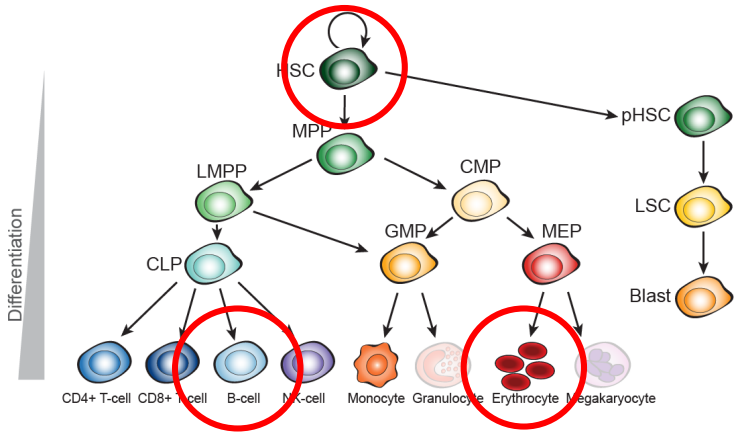


Peyton Greenside

Input: Raw DNA sequence

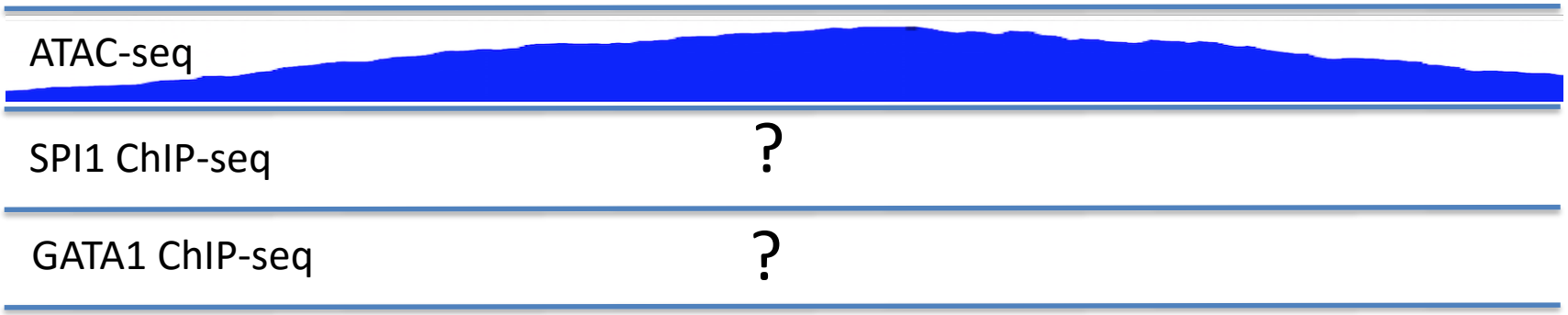
Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts



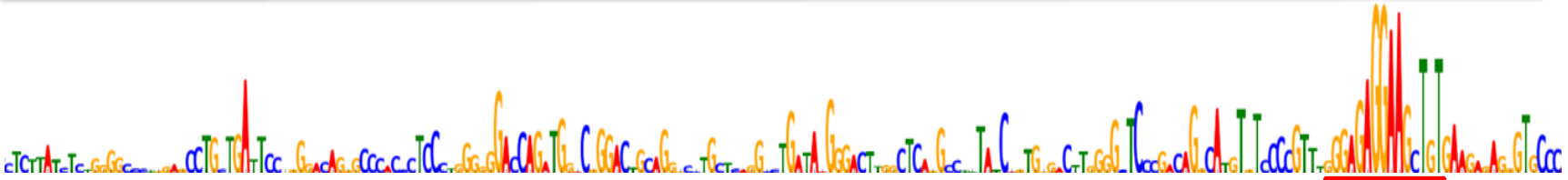


Peyton Greenside

Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts

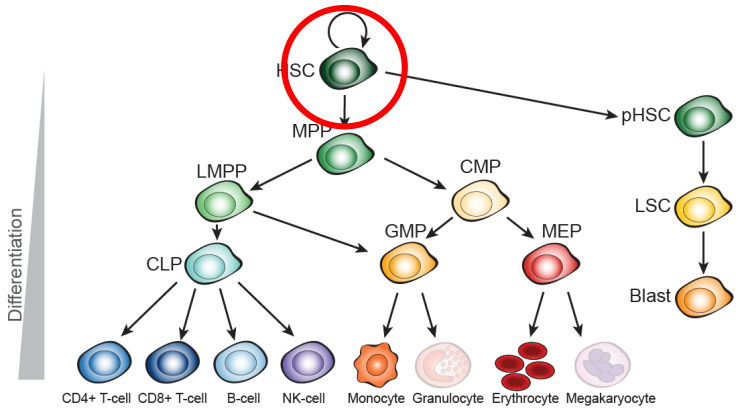


Importance in **HSC's**



Position along sequence

SPI1



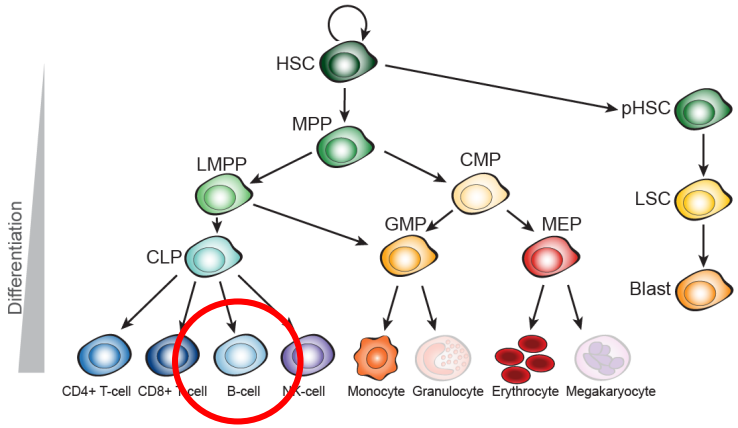
Peyton Greenside

Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts

ATAC-seq	No peak
SPI1 ChIP-seq	No peak
GATA1 ChIP-seq	Not expressed

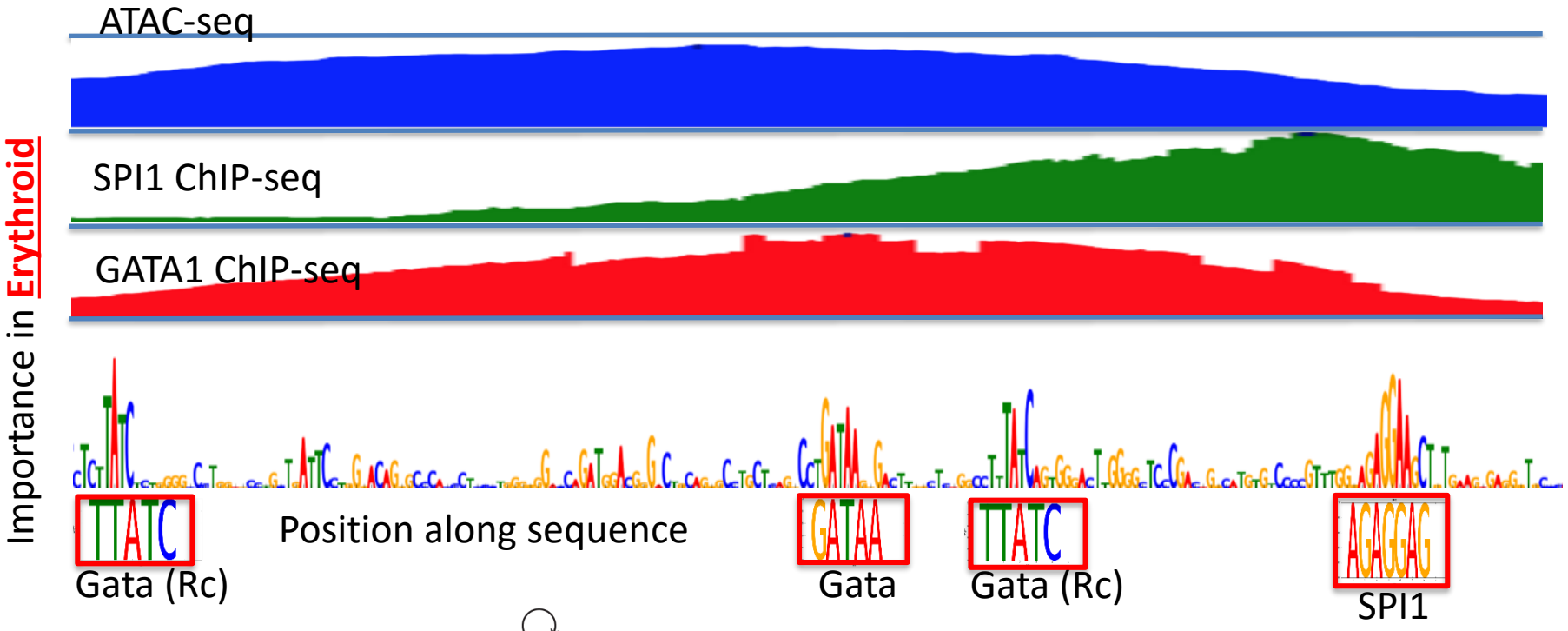
Importance in **B-cells**

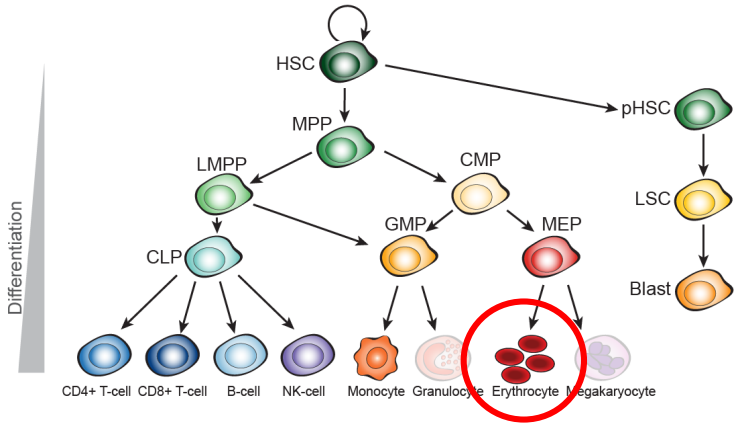
Position along sequence



Peyton Greenside

Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts

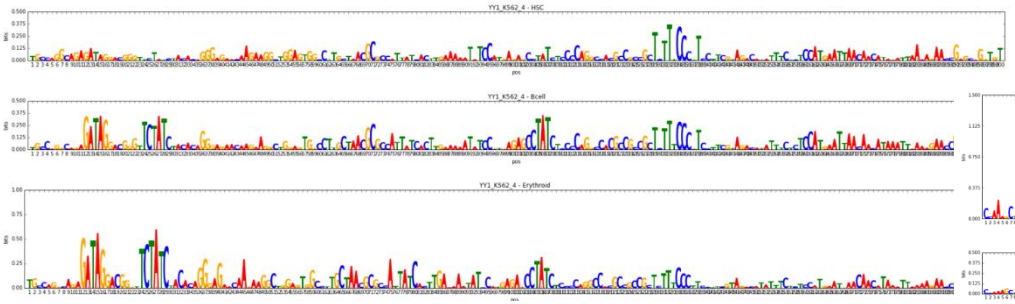




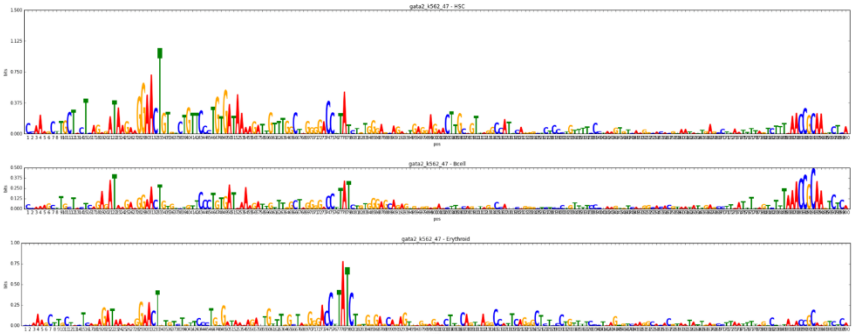
Peyton Greenside

…and much, much more

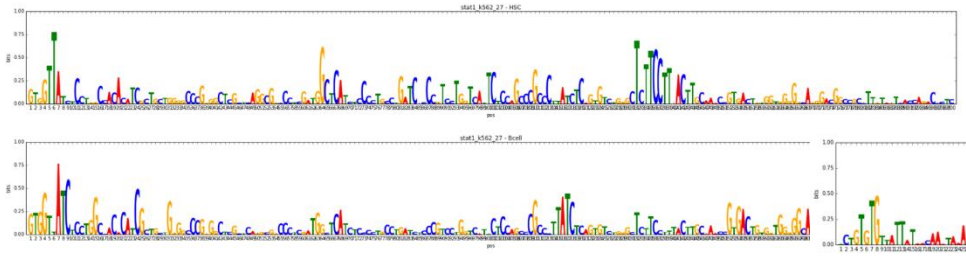
YY1 & GATA



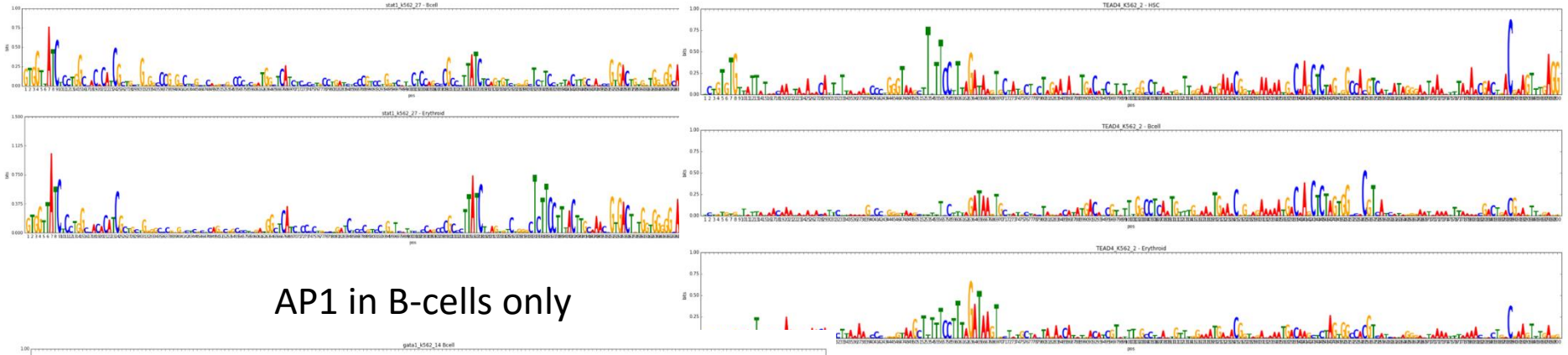
GATA, SPI1, RUNX2



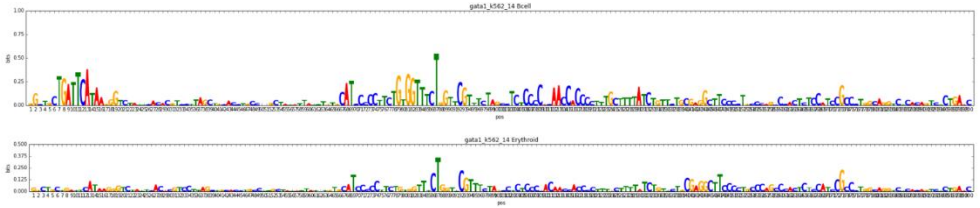
STAT1 & GATA



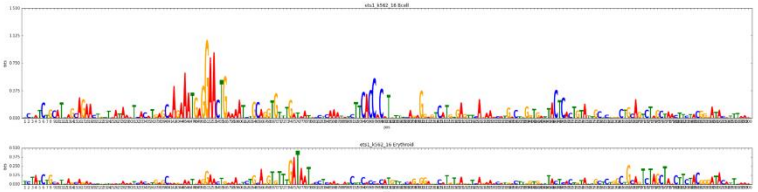
TEAD4 & GATA



AP1 in B-cells only



ETS & GATA



Summary and ongoing work

- **New predictive deep learning framework** (Chromputer) for integrative genomics
- **New interpretation engine** for deep learning models. We can extract predictive features (motifs, grammars, footprints, architecture features) from the deep neural networks
- **Local chromatin architecture is predictive of chromatin state** and histone marks within and across cell types
- We can predict **in-vivo binding profiles of TFs** in new cell types from sequence + shape + DNase/ATAC-seq with high accuracy
- Context-specific reuse of sequence grammars in accessible sites
- **Extensions:** From binary to continuous signal prediction
- **Extensions:** Functional variant (QTL, GWAS, rare variant) prediction from raw sequence models

Acknowledgements

Kundaje Lab members

Chuan Sheng Foo

Avanti Shrikumar

Nicholas Sinnott-Armstrong

Johnny Israeli

Rahul Mohan

Peyton Greenside

Nathan Boley

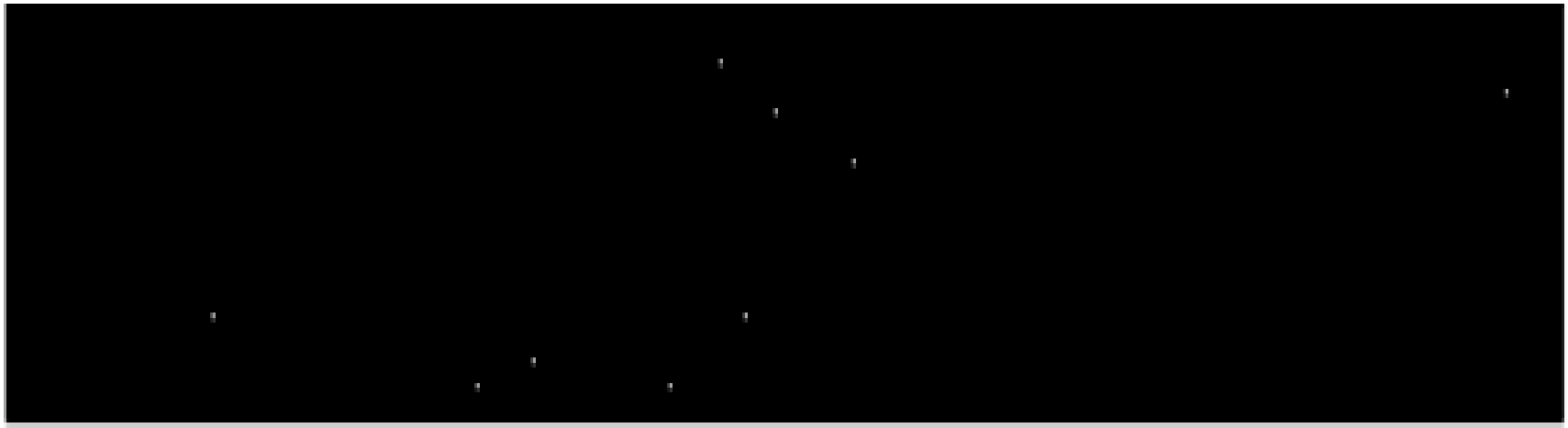
Will Greenleaf

Funding

U01HG007919-02 (GGR)
U41-HG007000-04S1
R01ES02500902

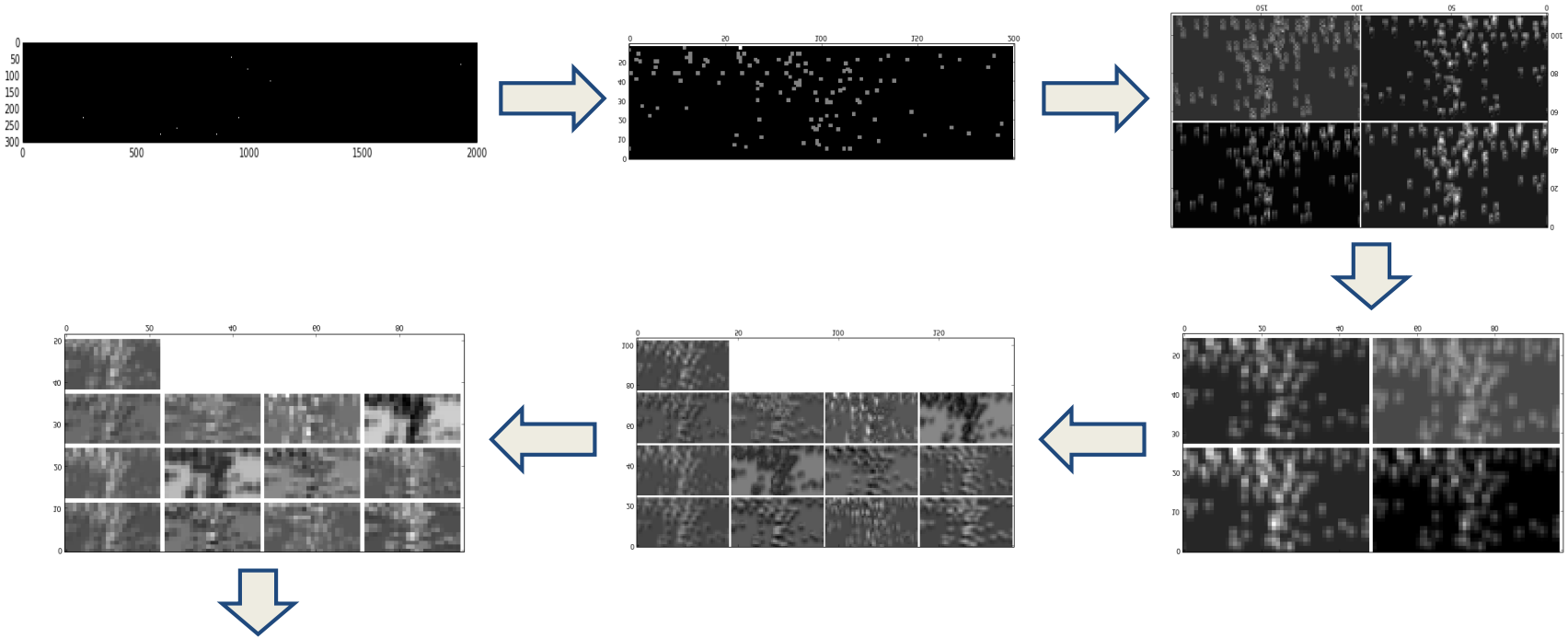
Conflict of Interest: Deep Genomics (SAB), Epinomics (SAB)

Guess the element from the V-plot
AI vs. human



What is this regulatory element?
Pure CTCF, Promoter, or Enhancer?

Its an enhancer!



1st Fully Connected Layer → 2nd Fully Connected Layer → Enhancer (correct)