

Deep Learning for Genomics

Jack Lanchantin

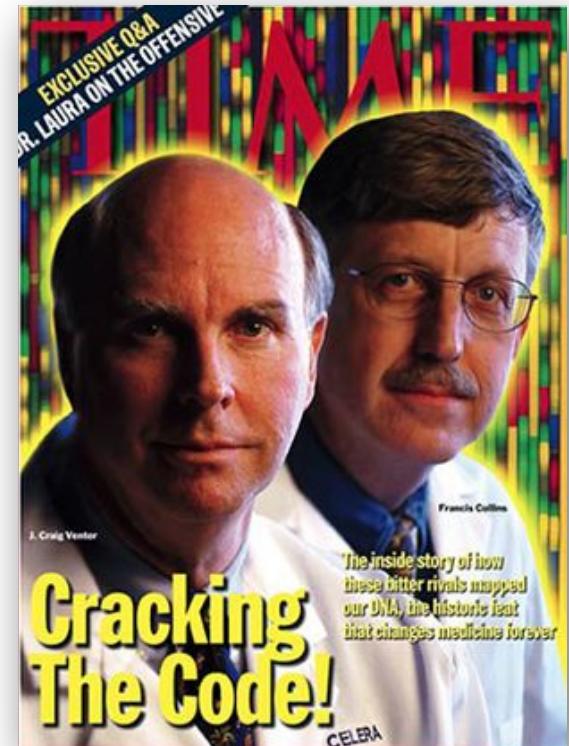
University of Virginia, Department of Computer Science



UNIVERSITY of VIRGINIA

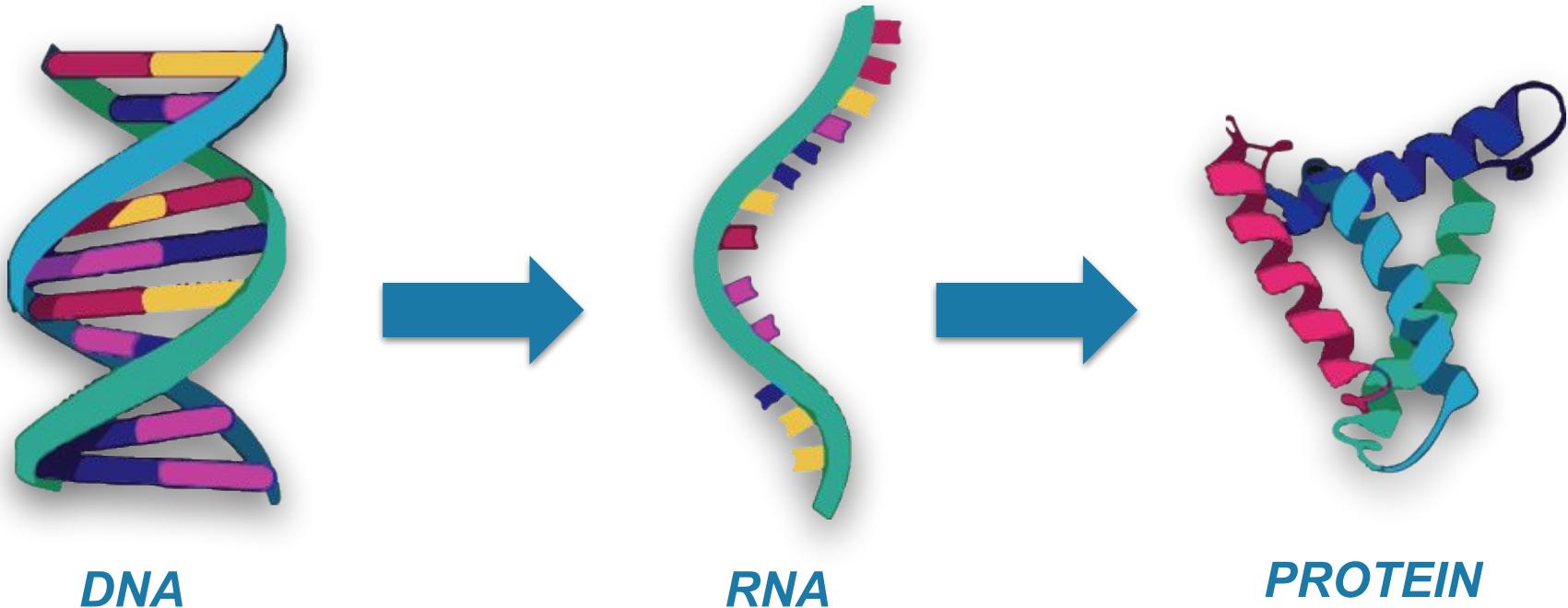
The Human Genome Project (1990-2003)

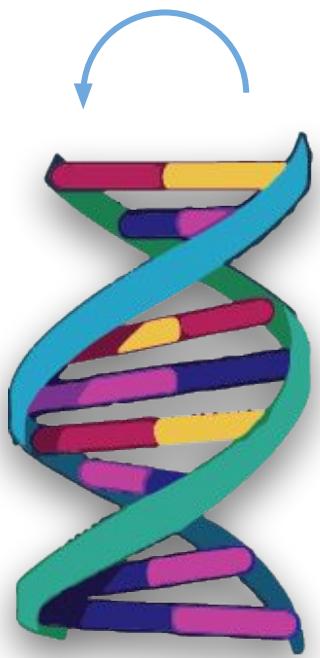
Mapped the human genome sequence and identified the genes



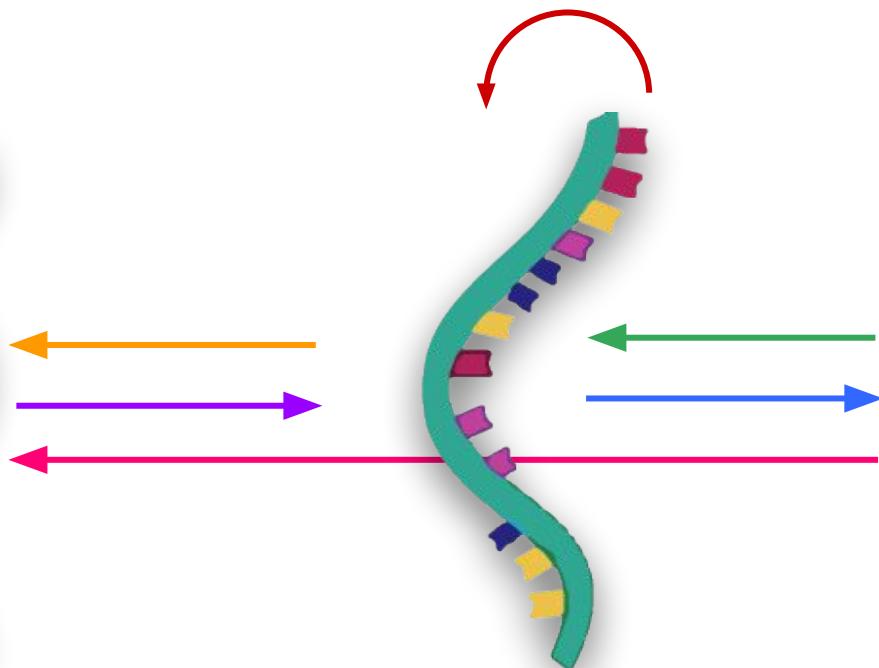
“Genome. Bought the book. Hard to read.”

-Eric Lander, Principal Leader of the Human Genome Project





DNA

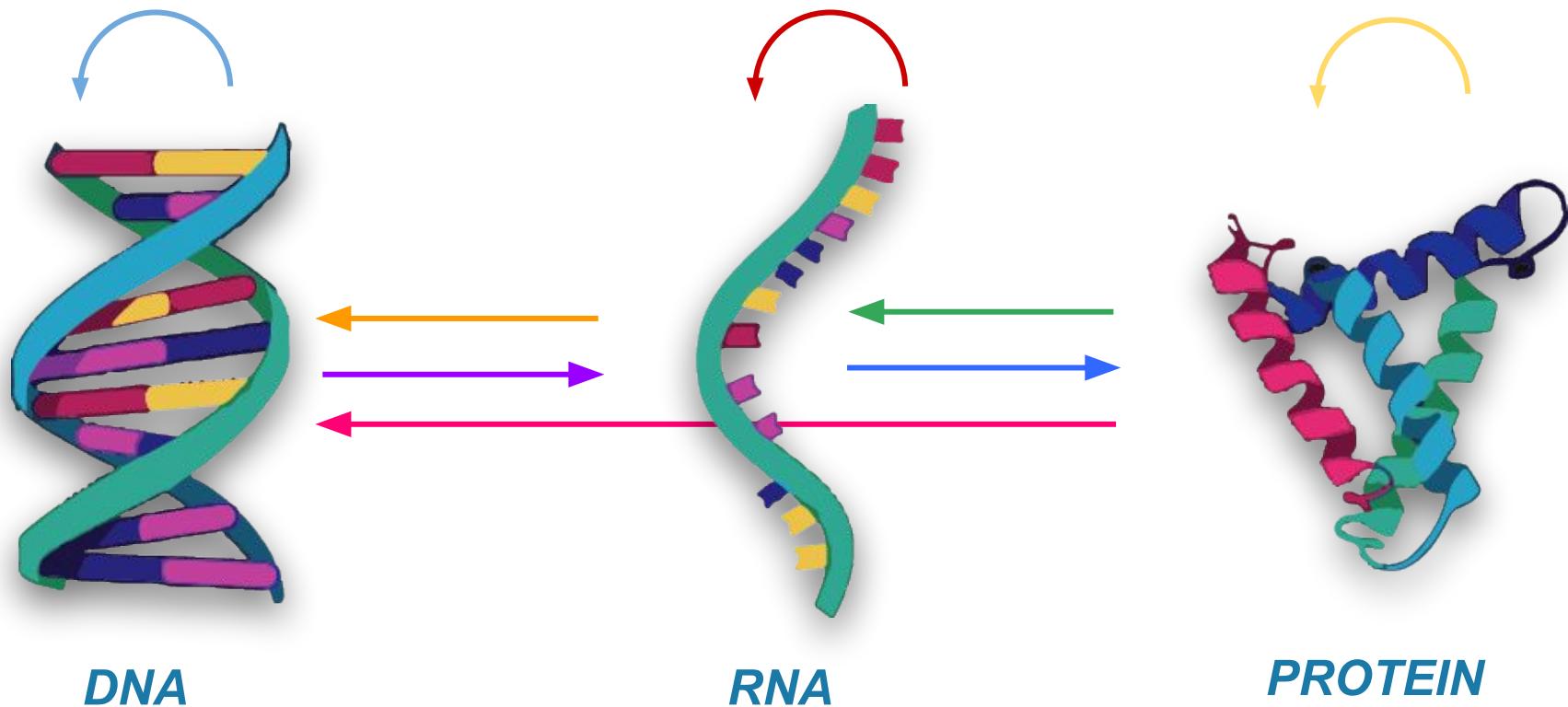


RNA



PROTEIN

This Talk: *Using Machine Learning to Read the Genome*





65%

LIFETIME RISK OF
GENETIC DISEASE



65%

LIFETIME RISK OF
GENETIC DISEASE



8M

BIRTHS PER YEAR
WITH GENETIC DEFECT



65%

LIFETIME RISK OF
GENETIC DISEASE



8M

BIRTHS PER YEAR
WITH GENETIC DEFECT



\$5M

LIFETIME COST
PER CHILD IN US

HEALTHCARE BUDGET
US & EUROPE



GLOBAL
IT MARKET



Gene

ATGCTCGATACTGAGACTACTGAGACTTGAGACTCTAGATCTGACTACTCACG



Gene Expressed

ATGCTCGATACTGAGACTACTGAGACTTGAGACTCTAGATCTGACTACTCACG

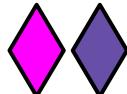


Gene Expressed

ATGCTCGATACTGAGACTACTGAGACTTGAGACTCTAGATCTGACTACTCACG

what causes a gene to be expressed?

1. Transcription
Factors



2. Histone
Modifications



Gene

ATGCTCGA**TACTGAGACTACTGAGACTTGAGACTCTAGATCTGACTACTCACG**

1. Predicting Transcription Factor Binding Sites from DNA
2. Predicting Gene Expression from Histone Modifications

Transcription Factors (TFs)

Proteins that bind to DNA **promoting** or **blocking** the recruitment of RNA polymerase

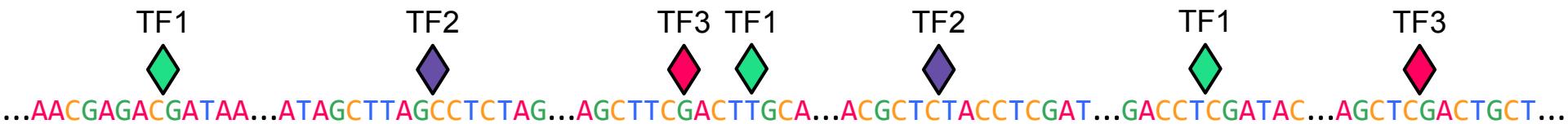


Transcription Factors (TFs)

Proteins that bind to DNA **promoting** or **blocking** the recruitment of RNA polymerase



Transcription Factor Binding Sites (TFBSs)



TFBS Datasets

GCGACGAATCG
CTCGA ◆ TCTCA
CGAT ◆ TGCTTC
AAGAAGCATTA
AA ◆ GAT ◆ A ◆ GCT
TGTCAAGCAAG
ATATC ◆ ATATA
AGCATAT ◆ CGA
CATATCATTTC
TA ◆ CAAGCT ◆ G
CGAATGCATAC
ACGA ◆ GATTAT

TFBS Classification Task



TFBS Classification Task

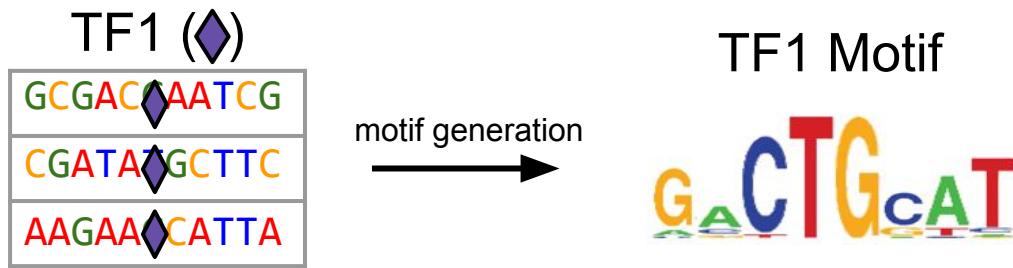


Why Study DNA → TF Binding?

- All gene regulation starts with TF binding
- Transcription Factors search for **specific sequence patterns** in DNA to bind.

Understanding the type of patterns will give us insights into the regulatory code of DNA, and help us understand alterations

Background: Motifs



Prediction with Motifs

?
G A C T G c A
...G T C T G C A ...A G T C G T C ...G A C T G C T ...T C G G C G A ...C C A G T T C ...

Prediction with Motifs

...**G**T**C**T**G**CA...**A**GTC**G**TC...**G**ACT**G**CT...**T**CG**G**CGA...**C**C**A**G**T**TC...



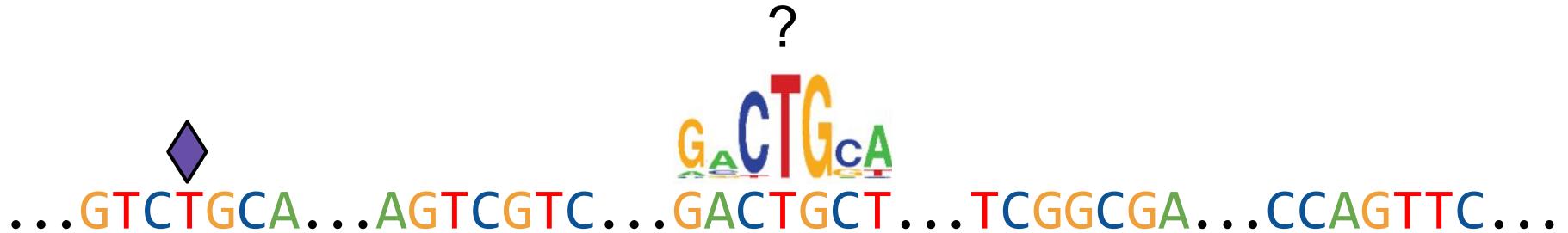
Prediction with Motifs

?

...**G**T**C**T**G**CA...**A**G**T**C**G**TC...**G**A**C**T**G**CT...**T**CG**G**CG**A**...**C**C**A****G**T**T**C...

Prediction with Motifs

?



...GTCTGCA...AGTCGTC...GACTGCT...TCGGCGA...CCAGTTC...

Prediction with Motifs

...**G**TCT**G**CA...**A**GTC**G**TC...**G**ACT**G**CT...**T**CGG**G**GA...**C**CAG**T**TC...



Prediction with Motifs

...**G**T**C**T**G**CA...**A**GTC**G**TC...**G**ACT**G**CT...**T**CG**G**CGA...**C**C**A**GT**T**C...



Prediction with Motifs

?

The diagram shows a sequence of DNA bases represented by colored diamonds: orange, red, green, blue, and yellow. Two specific motifs are highlighted with purple diamond markers above them. The first motif is 'GTCTGCA' and the second is 'GACTGCT'. To the right of the sequence, a question mark is positioned above a sequence of bases: 'G_AC_TG_cA'. Below this sequence, the bases are shown again with colored diamonds, indicating they are part of the sequence being analyzed.

...GTCTGCA...AGTCGTC...GACTGCT...TCGGCGA...CCAGTTC...

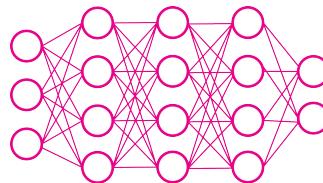
G_AC_TG_cA

A decorative graphic consisting of a repeating pattern of colored DNA base pairs (A, T, G, C) arranged in two rows. The top row contains 'G', 'A', 'C', 'T', 'G', 'C', 'A', 'T'. The bottom row contains 'A', 'T', 'G', 'C', 'A', 'T', 'G', 'C'. The letters are in various colors: 'G' is blue, 'A' is red, 'C' is green, 'T' is yellow, 'G' is orange, 'C' is green, 'A' is red, 'T' is yellow.

- ✖ Good motifs are hard to generate



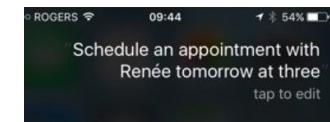
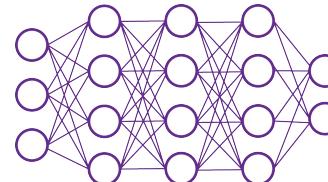
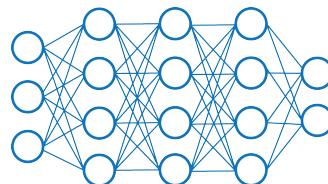
- ✖ Good motifs are hard to generate
- ✓ Deep neural networks are good at automatic feature extraction

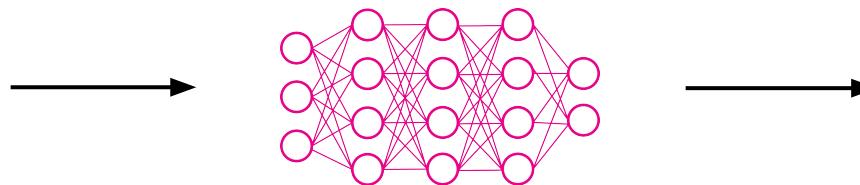


Dog

Can get overly sentimental at times, but Gus Van Sant's sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

[Full Review...](#) | May 25, 2006

A blue rectangular box containing a movie review. The text describes the film as being sentimental at times but praises Gus Van Sant's direction and the use of the city. It includes a link to a full review and the date May 25, 2006.



Dog

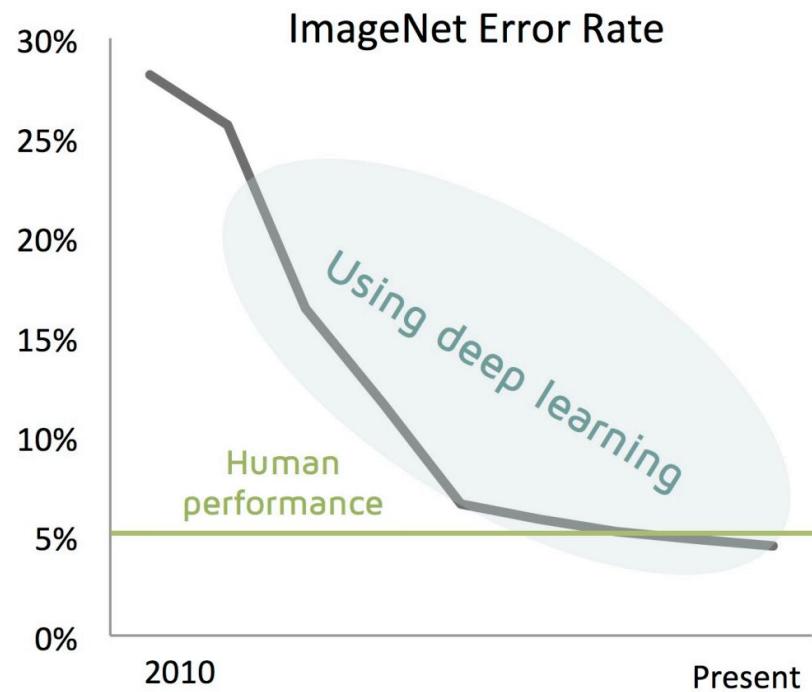


Image Recognition



Extracted Features



Dog

Image Recognition



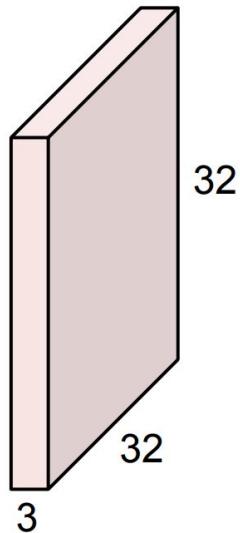
Extracted Features



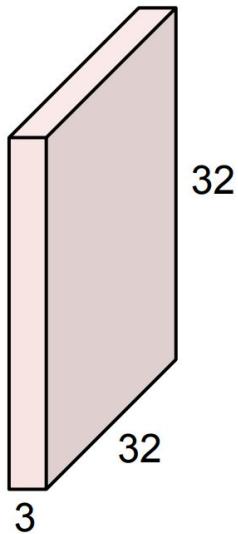
?

Dog

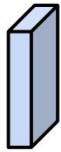
Key Operator: Convolutional Neural Networks



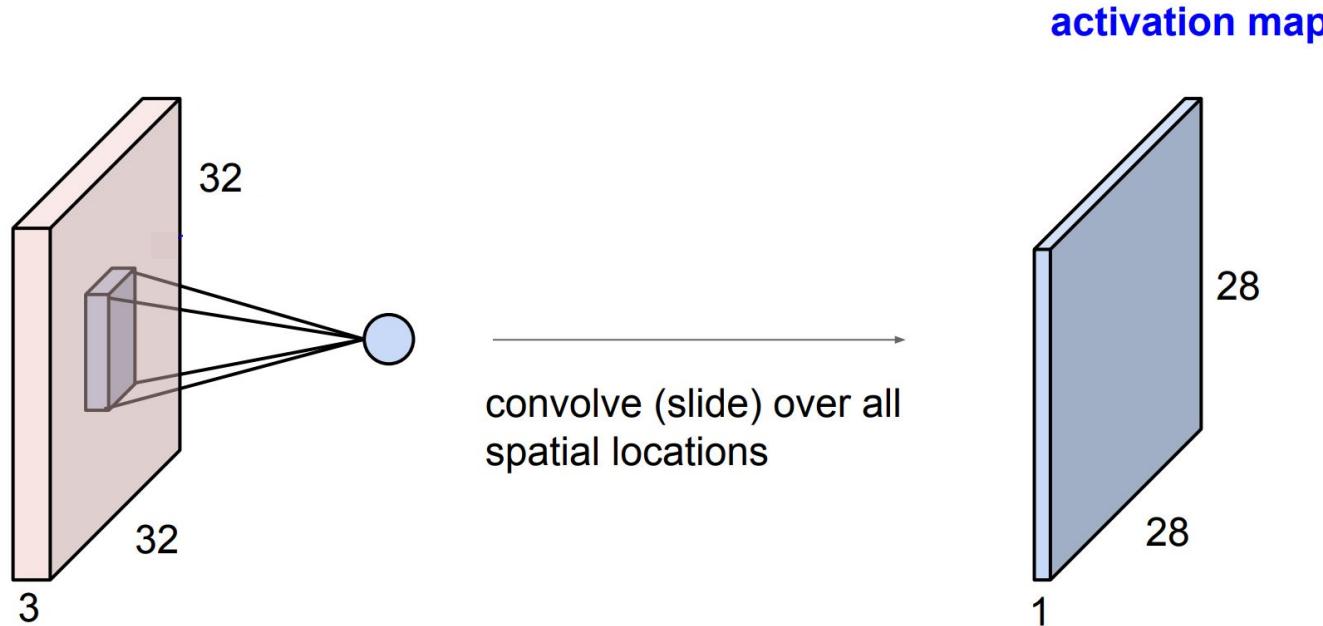
Key Operator: Convolutional Neural Networks



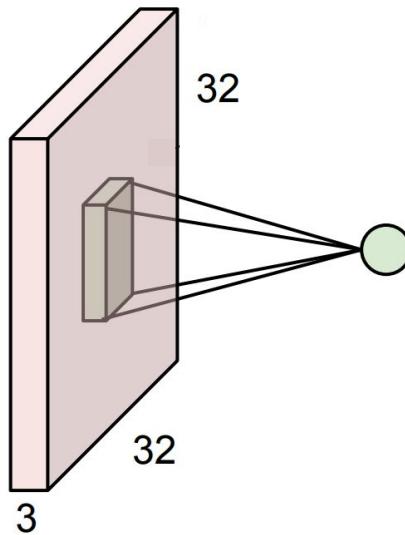
$5 \times 5 \times 3$ filter



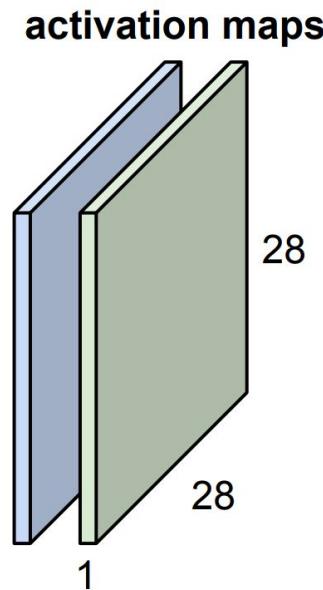
Key Operator: Convolutional Neural Networks



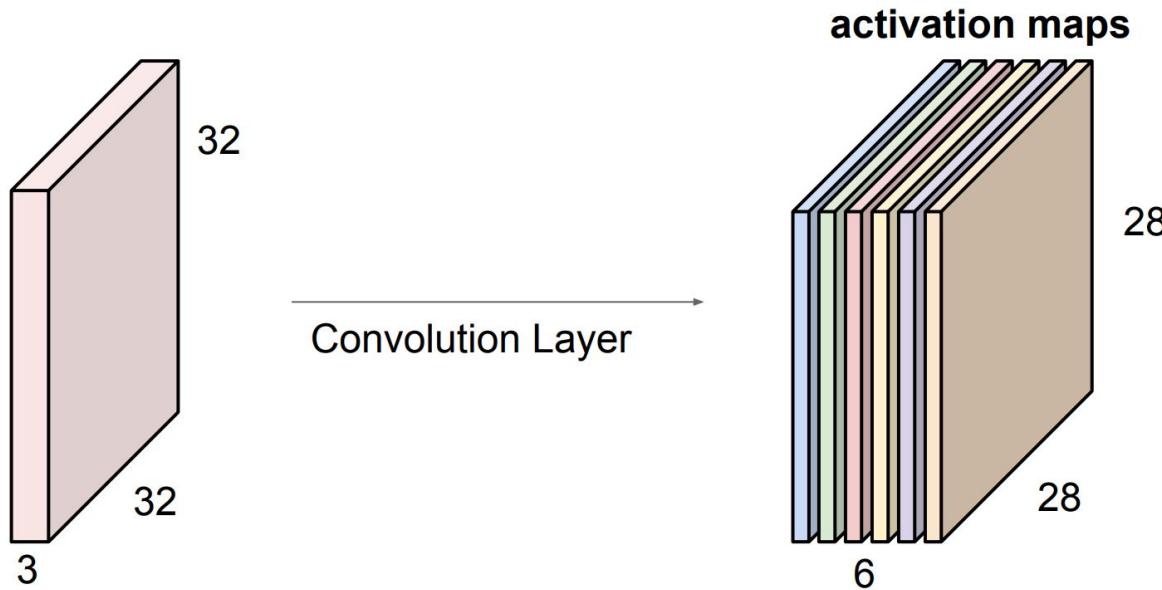
Key Operator: Convolutional Neural Networks



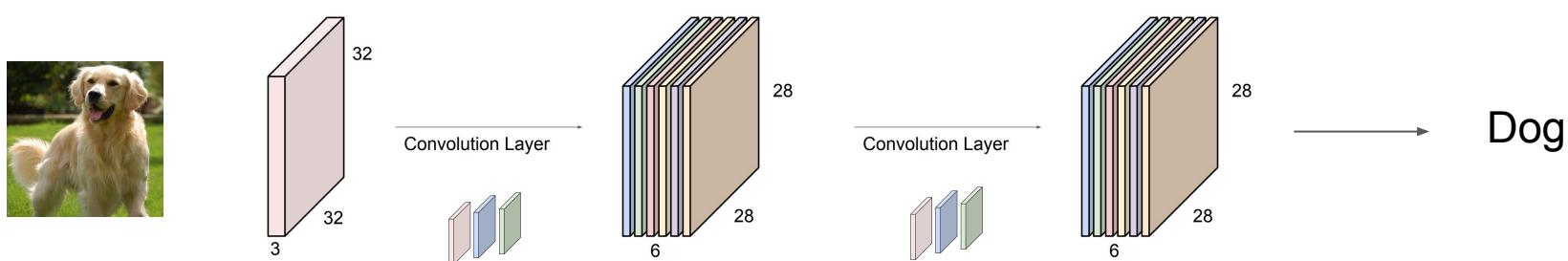
convolve (slide) over all
spatial locations



Key Operator: Convolutional Neural Networks



Key Operator: Convolutional Neural Networks

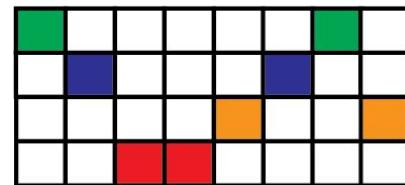


How does this relate to Genomics?

Convolutional Neural Network for DNA

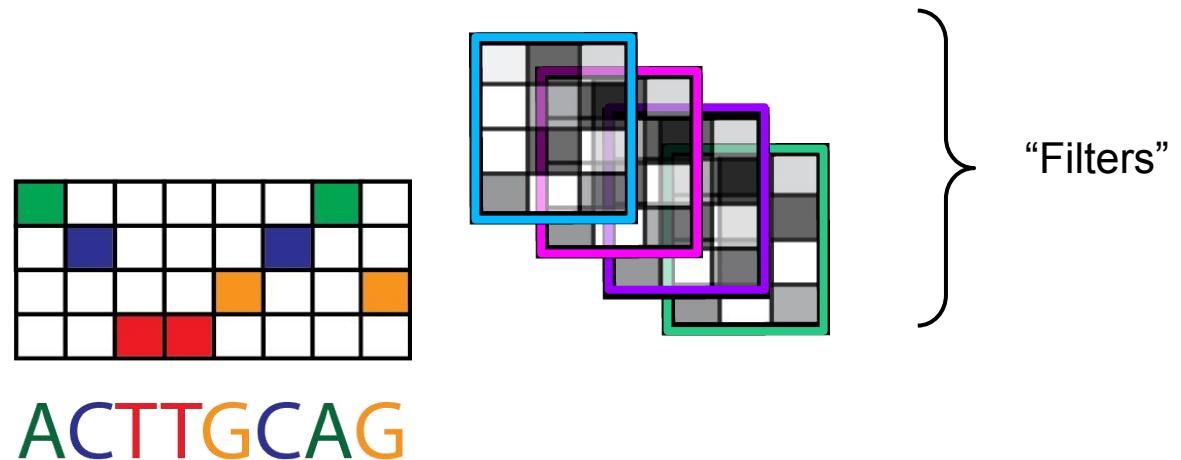
ACTTGCAG

Convolutional Neural Network for DNA

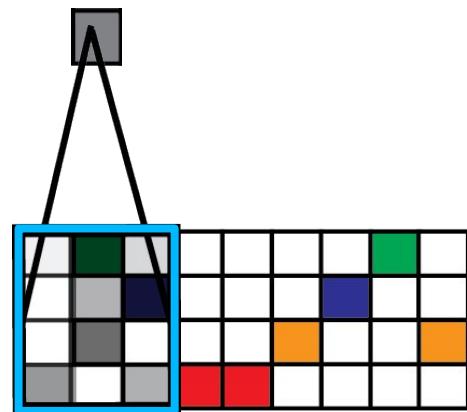


ACTTGCAG

Convolutional Neural Network for DNA

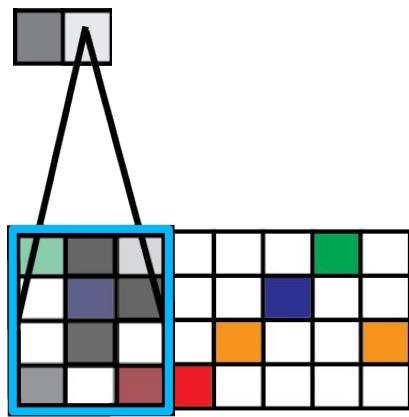


Convolutional Neural Network for DNA



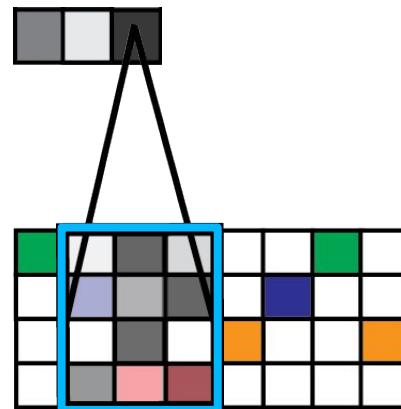
ACTTGCAG

Convolutional Neural Network for DNA



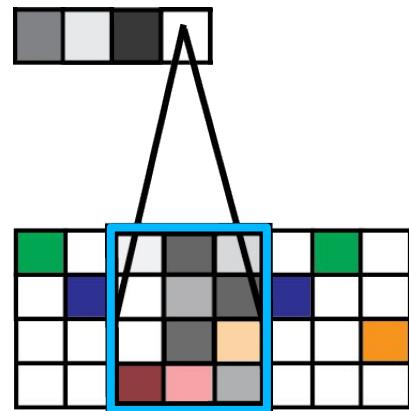
ACTTGCAG

Convolutional Neural Network for DNA



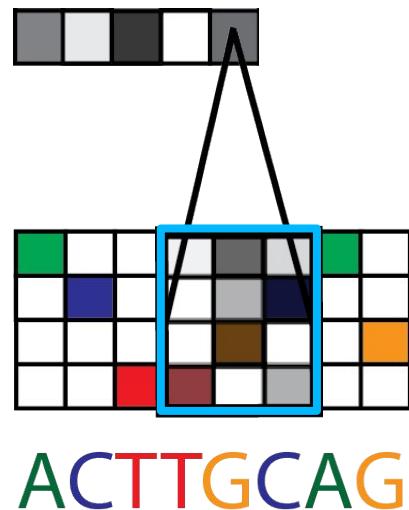
ACTTGCAG

Convolutional Neural Network for DNA

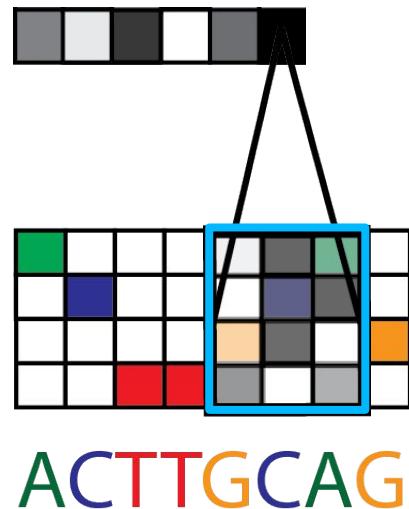


ACTTGCAG

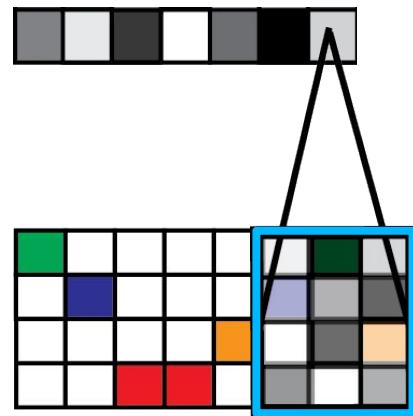
Convolutional Neural Network for DNA



Convolutional Neural Network for DNA

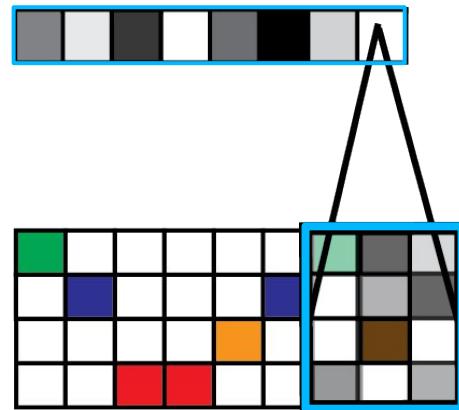


Convolutional Neural Network for DNA



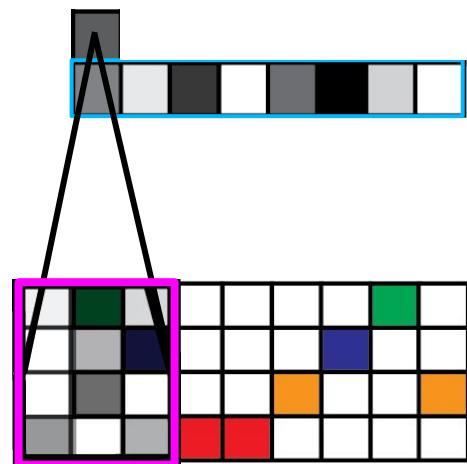
ACTTGCAG

Convolutional Neural Network for DNA



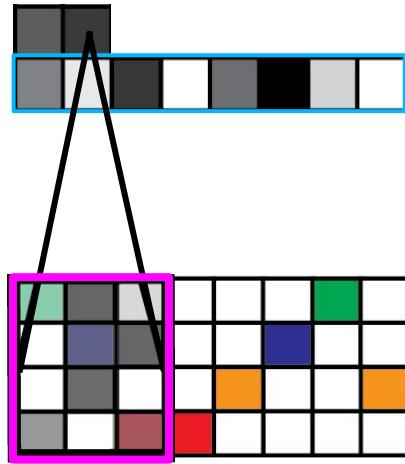
ACTTGCAG

Convolutional Neural Network for DNA



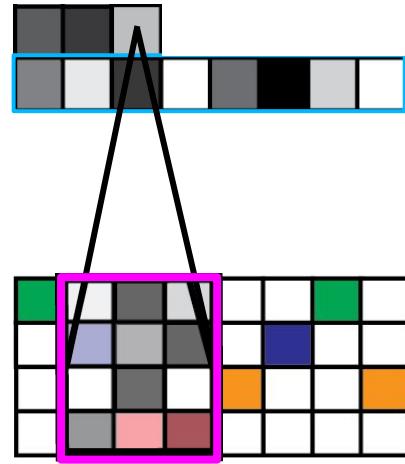
ACTTGCAG

Convolutional Neural Network for DNA



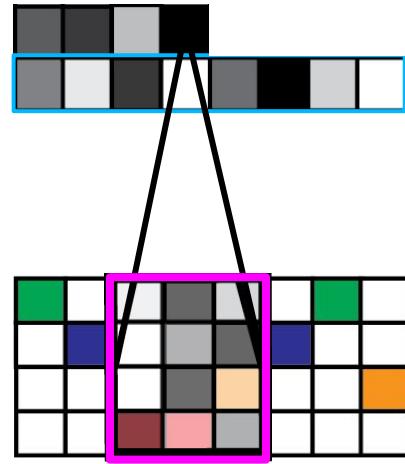
ACTTGCAG

Convolutional Neural Network for DNA



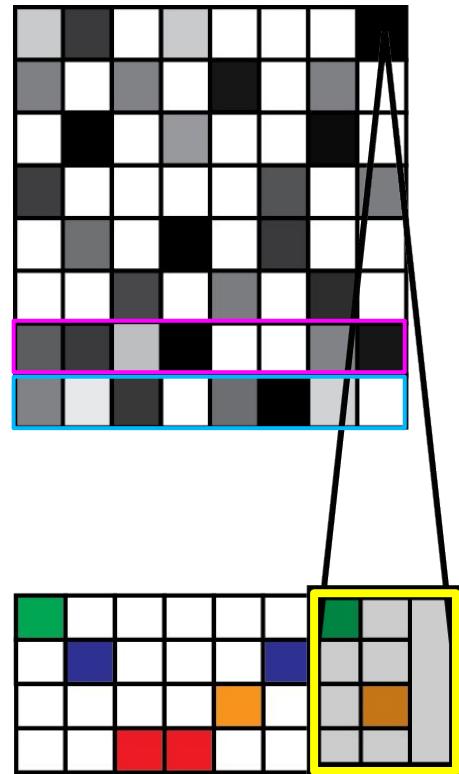
ACTTGCAG

Convolutional Neural Network for DNA



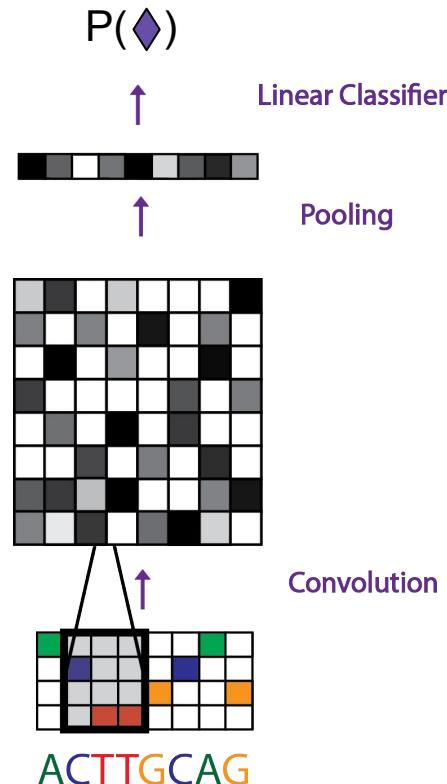
ACTTGCAG

Convolutional Neural Network for DNA



ACTTGCAG

Convolutional Neural Network for TFBS Classification



Why Convolution For Genomics?

Why Convolution For Genomics?

Dogs



Translational Invariant Features



Why Convolution For Genomics?

Dogs



Translational Invariant Features



AGT GAG ATCT CTT CA
GAAG CTCG ATGC ACG
CTCG ATT CAT GT CCT
ATGT CAAT CGAT CAC

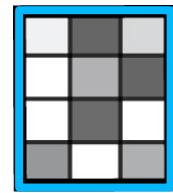


Translational Invariant Features



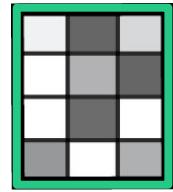
A CAG
CTT ATCT

Convolutional Filters learn motifs!



≈

C T T
T C

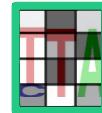


≈

A T G
T G

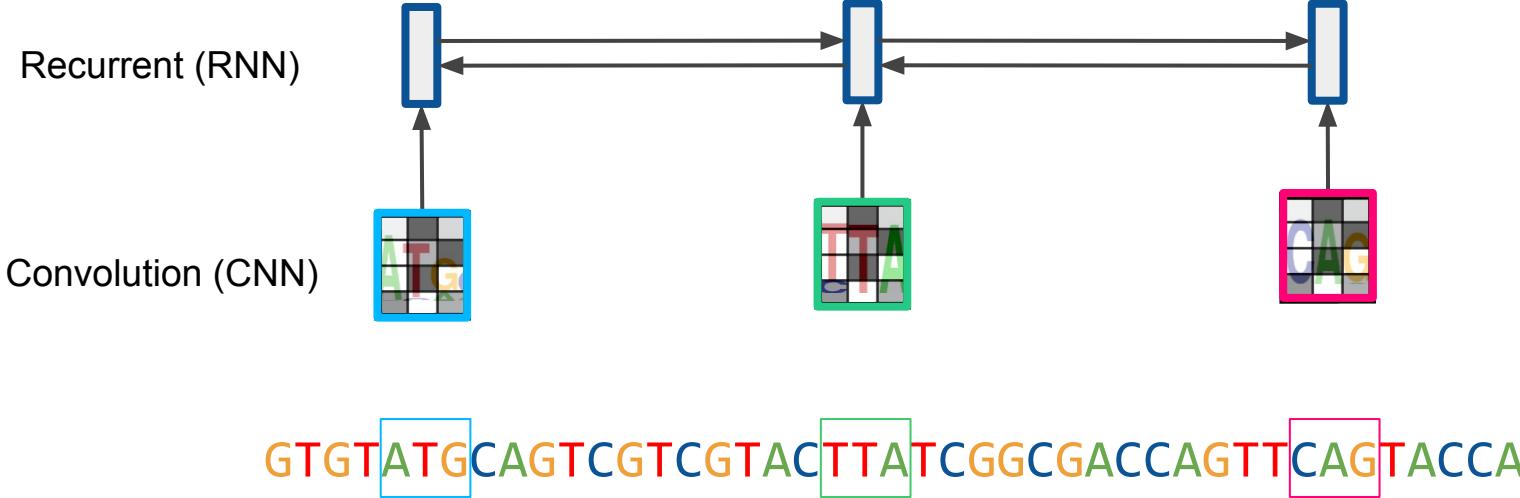
What about relationships among motifs?

Convolution (CNN)

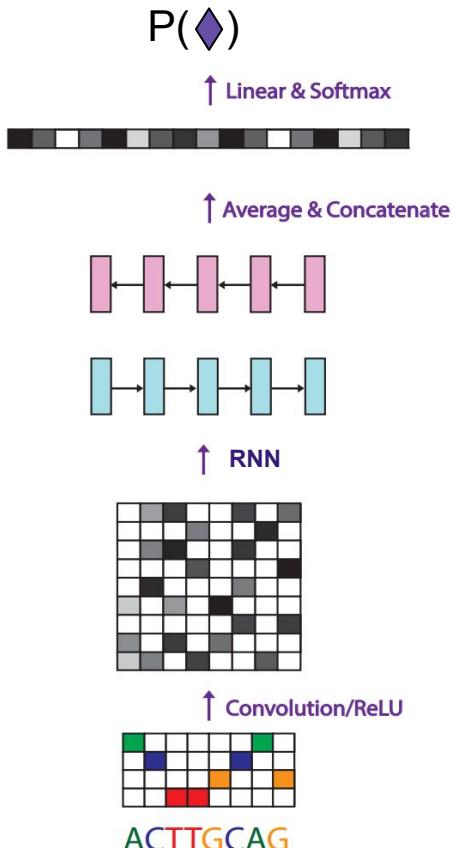


GTGTATGCAGTCGTAC

What about relationships among motifs?

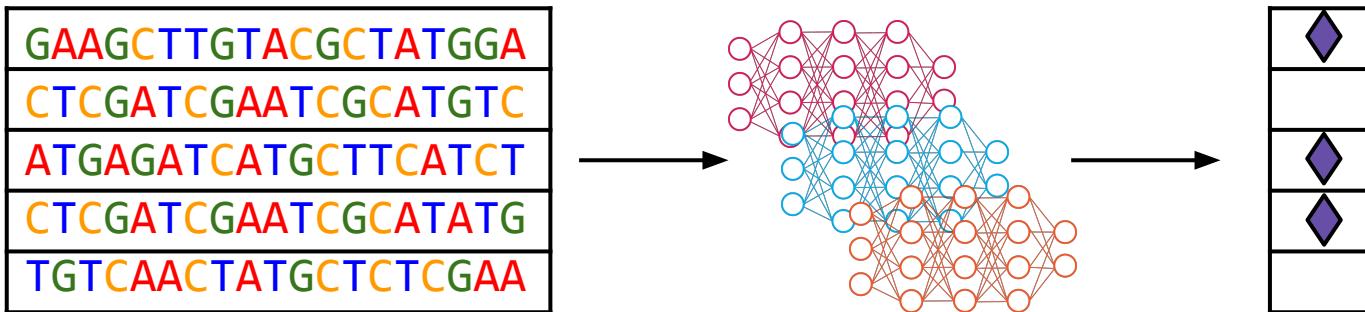


Convolution and Recurrent (CNN-RNN)



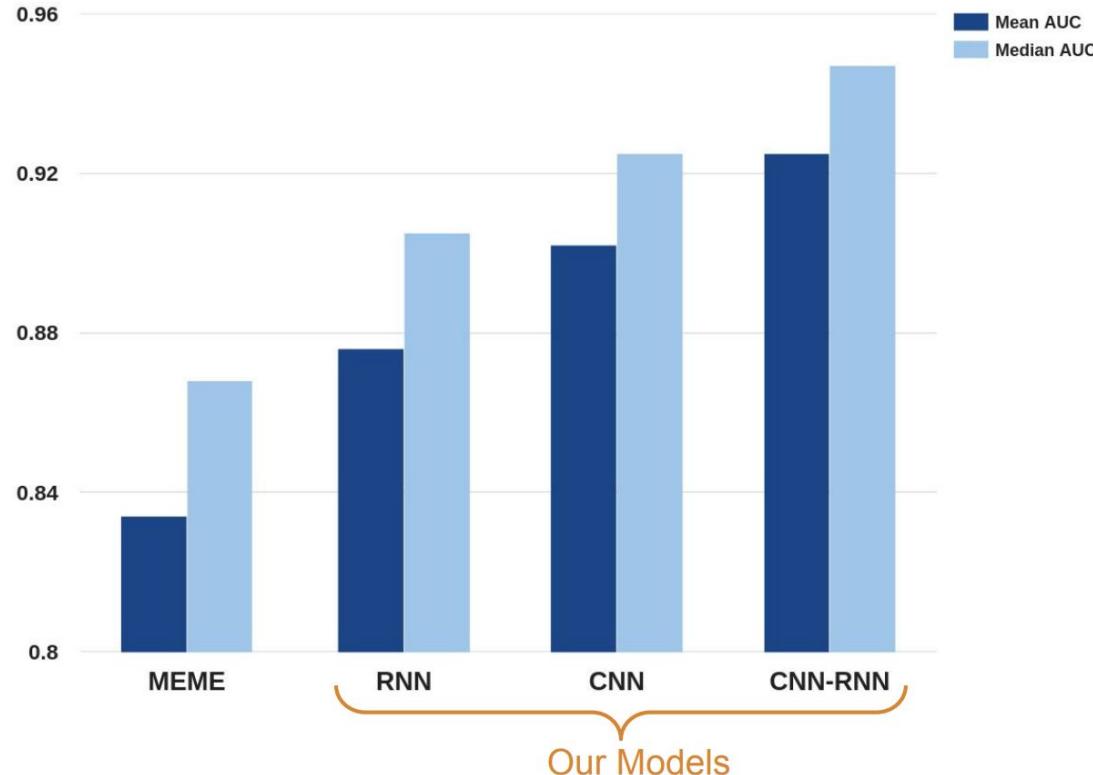
Deep Motif Dashboard

Lanchantin, Singh, Wang & Qi - ICLR Workshops 2016, PSB 2017



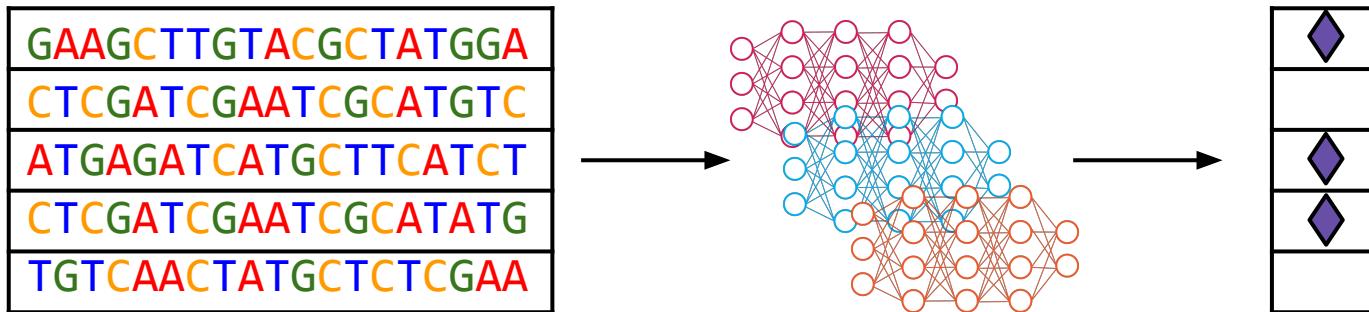
1. Convolutional (CNN)
2. Recurrent (RNN)
3. Convolutional-
Recurrent (CNN-RNN)

Model Accuracy on 108 Cancer Cell TFs



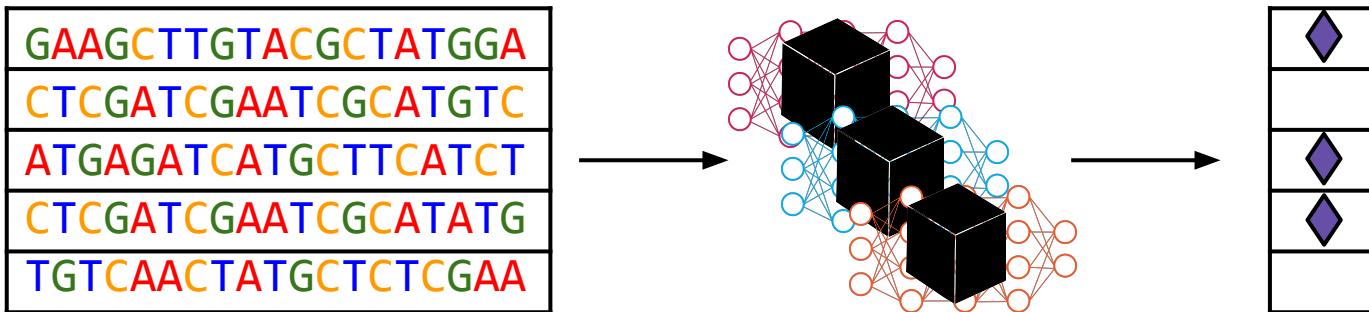
Deep Motif Dashboard

Lanchantin, Singh, Wang & Qi - ICLR Workshops 2016, PSB 2017



Deep Motif Dashboard

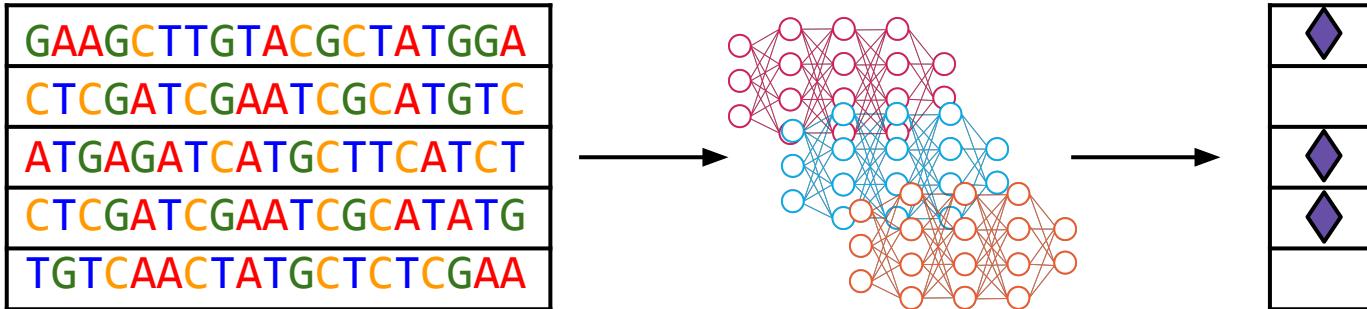
Lanchantin, Singh, Wang & Qi - ICLR Workshops 2016, PSB 2017



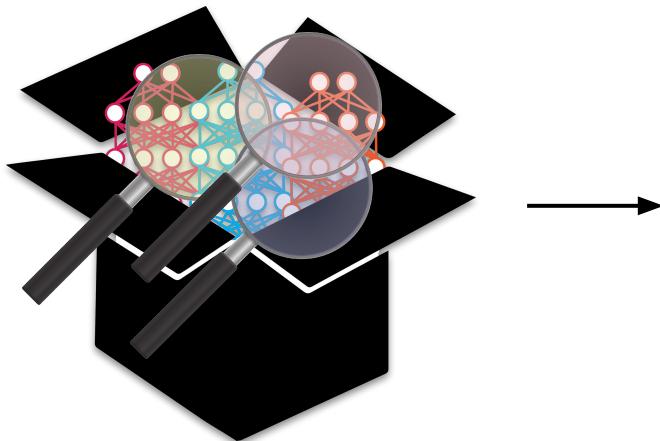
Deep Motif Dashboard

Lanchantin, Singh, Wang & Qi - ICLR Workshops 2016, PSB 2017

1.



2.



CNN Positive Class Maximization	
RNN Positive Class Maximization	
CNN-RNN Positive Class Maximization	
Positive Test Sequence	
CNN Saliency (0.90)	
RNN Saliency (0.96)	
CNN-RNN Saliency (0.99)	
Positive Test Sequence	
RNN Forward Temporal Outputs	
RNN Backward Temporal Outputs	
CNN-RNN Forward Temporal Outputs	
CNN-RNN Backward Temporal Outputs	

Visualization Methods

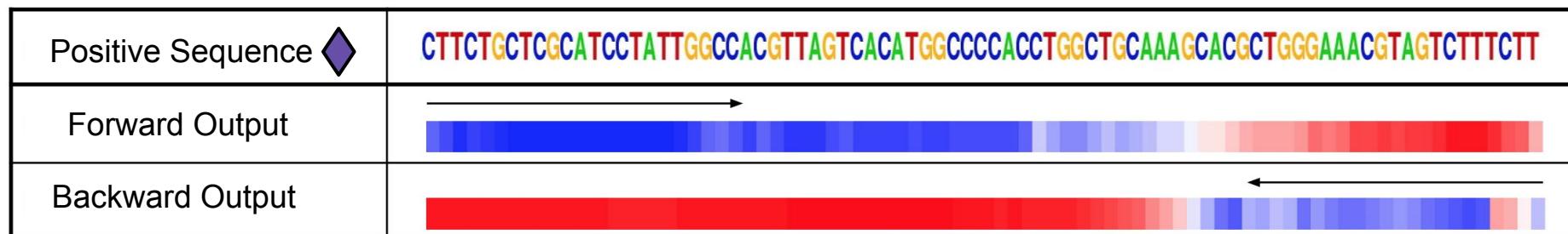
1. Saliency Maps: which nucleotides are most important for classification?

Positive Sequence	 TGCTCGCATCCTATTGGCCACGTTAGTCACATGGCCCCACCTGGCTGCAAAGCACGCTGGAAACGTAGTCTTCCTT
Saliency Map	

 = important nucleotide for prediction

Visualization Methods

2. Temporal Output Scores: what is the prediction at each timestep of the sequence?



█ = negative binding site prediction

█ = positive binding site prediction

Visualization Methods

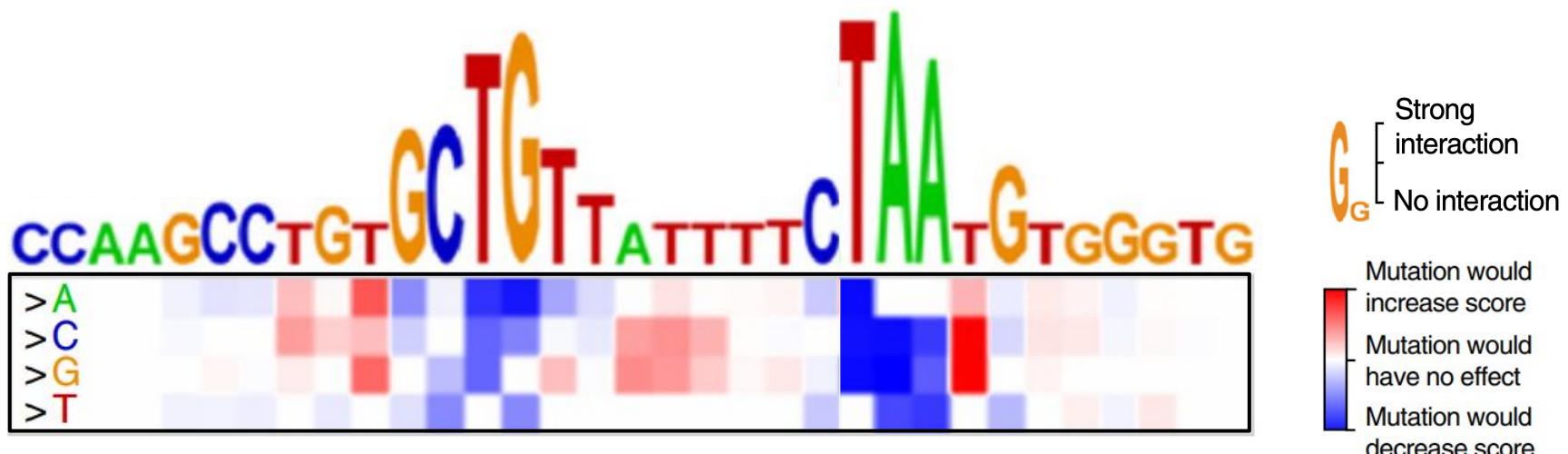
3. Class Optimization: for a particular TF, what does the optimal binding site look like?

Optimal binding site
for TF “CBX3” ♦

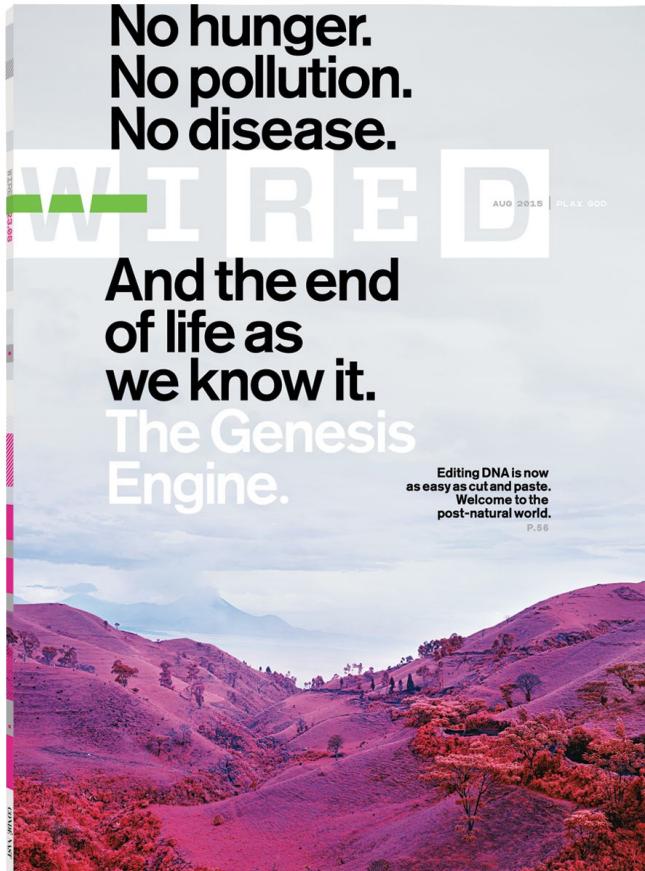


Visualization Methods

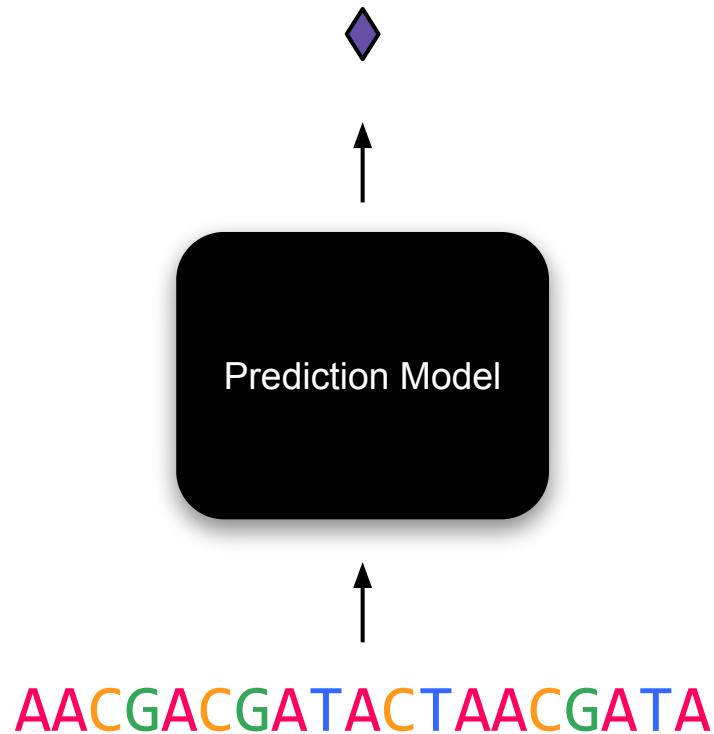
4. Mutation Simulations: How does the TFBS prediction change if mutations occur?



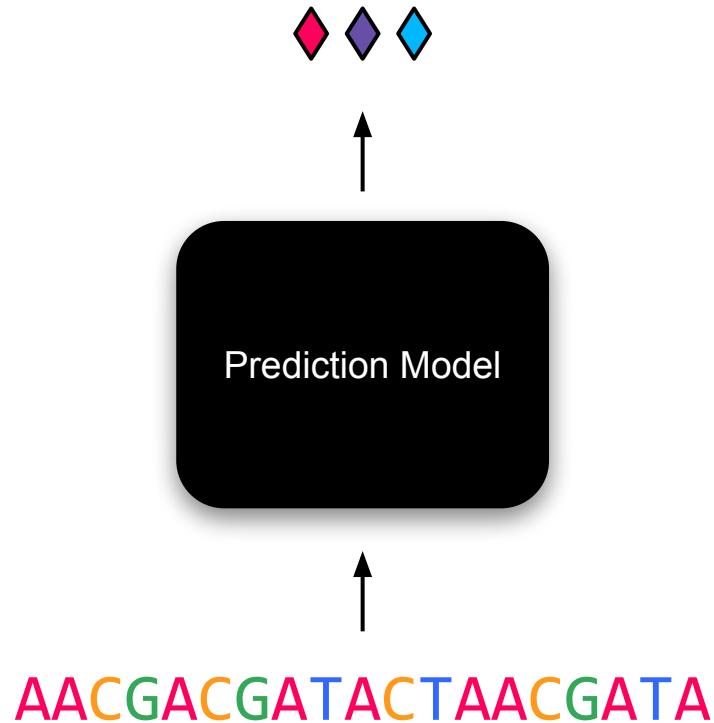
Fixing Mutations with CRISPR-Cas9?



Multi Label Classification for TFBS Prediction

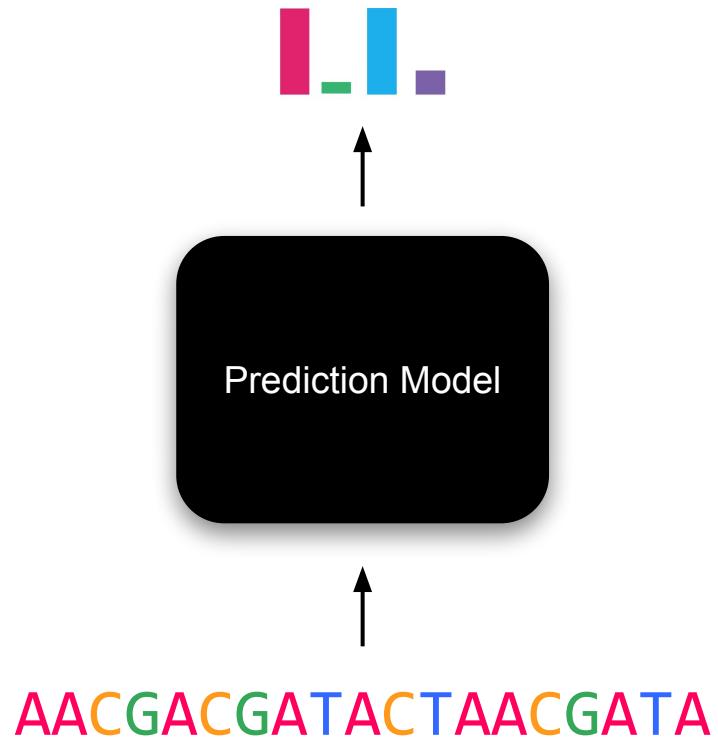


Multi Label Classification for TFBS Prediction



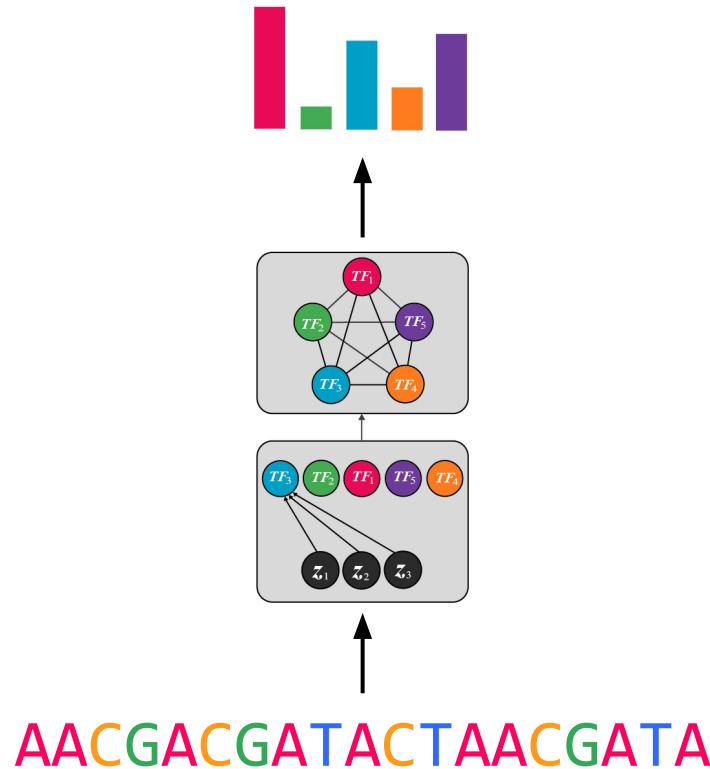
Multi Label Classification for TFBS Prediction

Lanchantin, Singh, & Qi - ICLR Workshops 2017

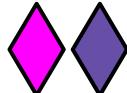


Neural Message Passing for MLC

Lanchantin, Sekhon, & Qi - Under Review 2019



1. Transcription
Factors



2. Histone
Modifications

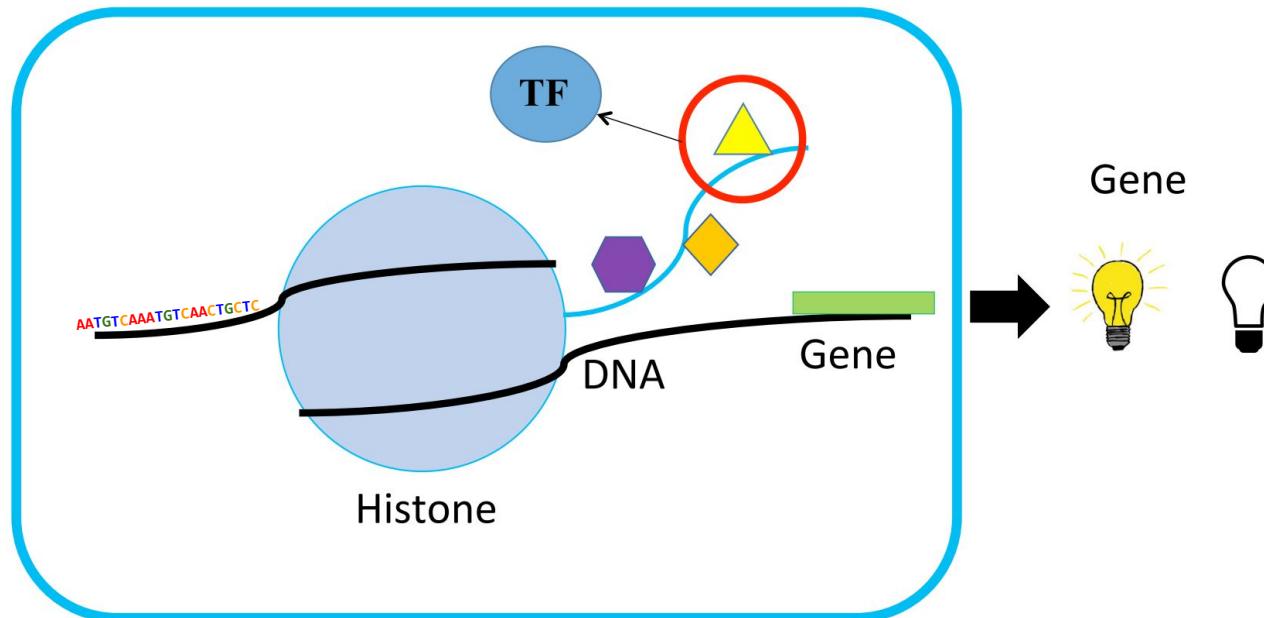


Gene

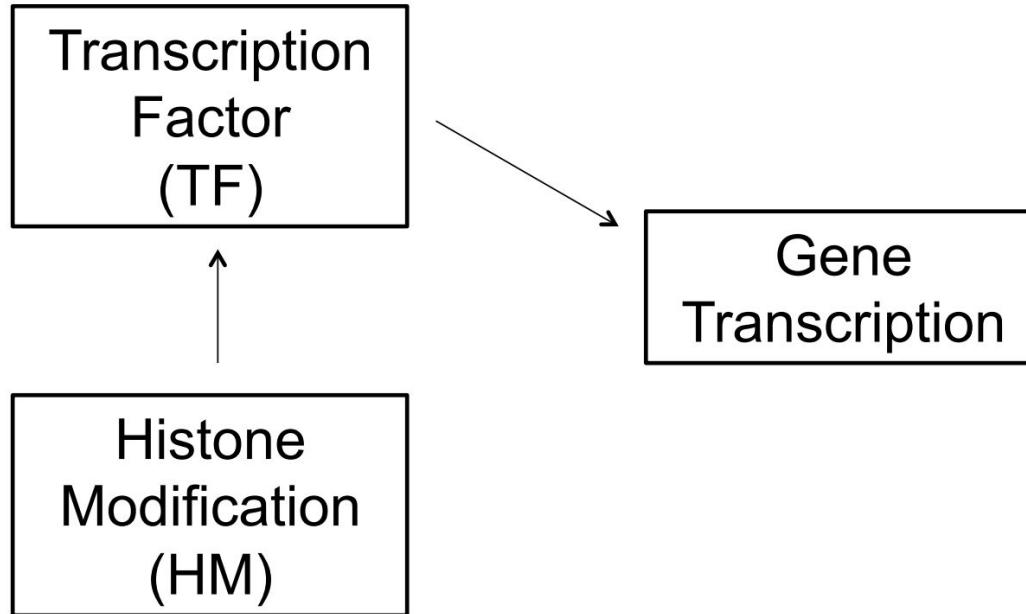
...AACGACG**A**TACTAACG**GATATGCTCGATACGATA**GCTCGACTATGCT...

1. Predicting Transcription Factor Binding Sites from DNA
2. Predicting Gene Expression from Histone Modifications

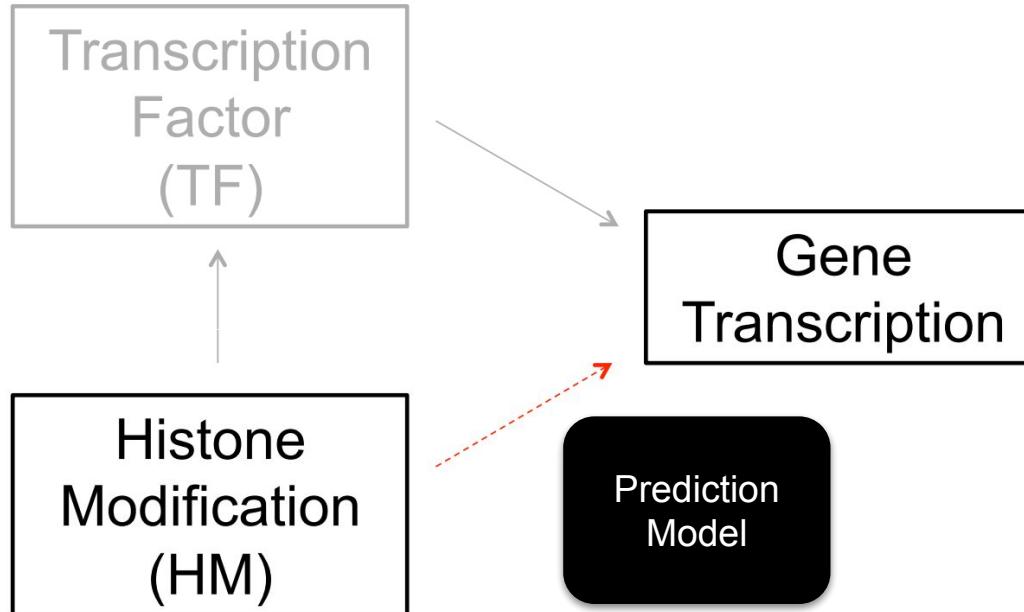
Histone Modifications (HMs)



Histone Modifications

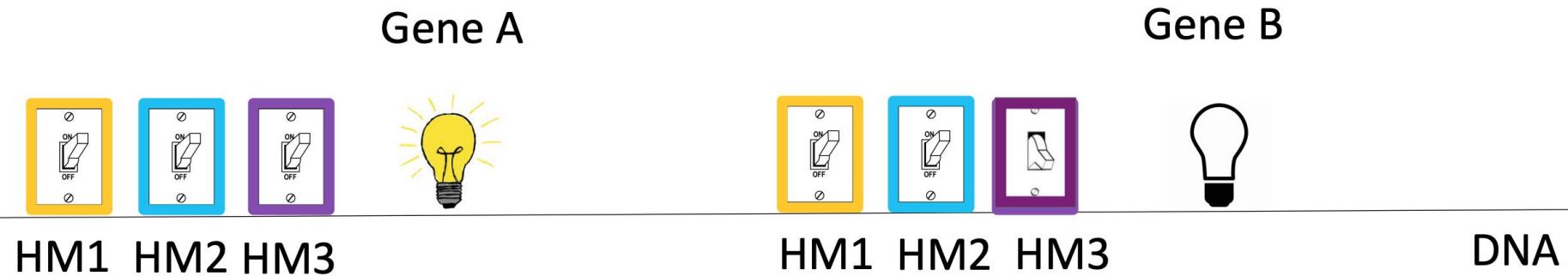


Histone Modifications



Can we predict gene expression from histone modification signals?

What HMs affect which genes in what cells?



Gene Transcription Prediction Task

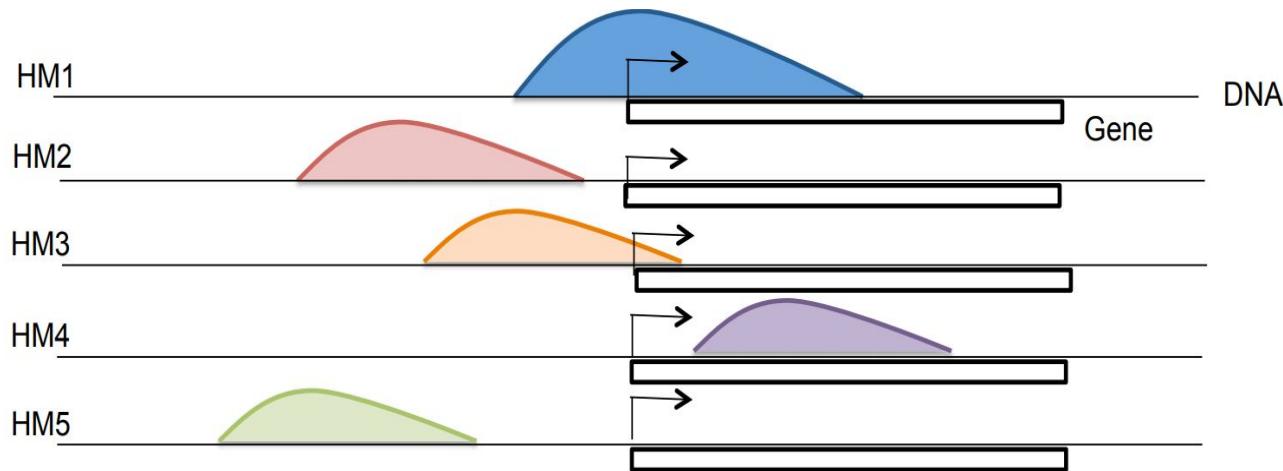


Why Study HM → Gene Expression?

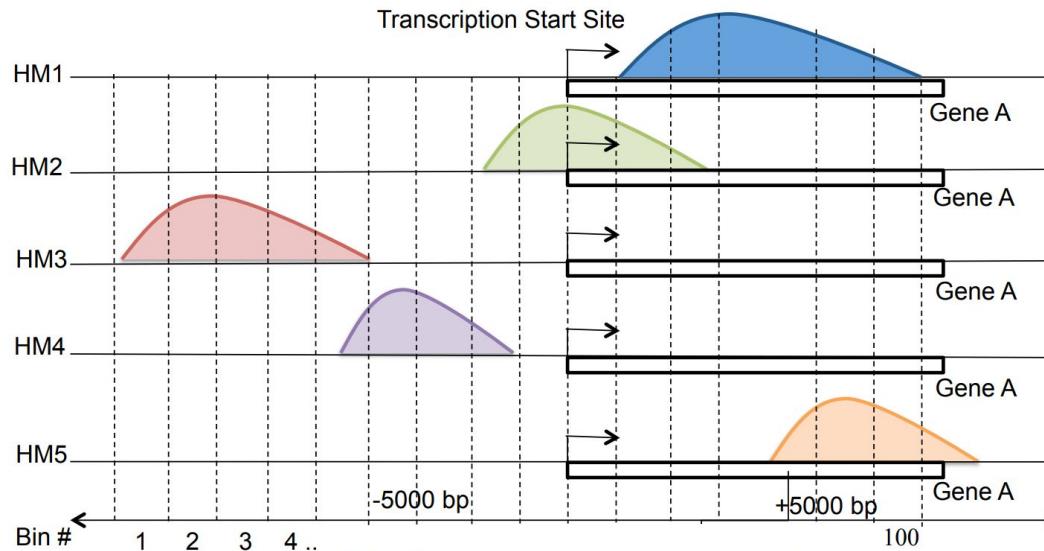
- **Epigenomics:** study of chemical changes in DNA and histones (without altering DNA sequence)
- **Epigenome is dynamic:** can be altered by environmental conditions.

Unlike genetic mutations, epigenomic changes such as histone modifications are potentially reversible → Epigenome drug for cancer cells?

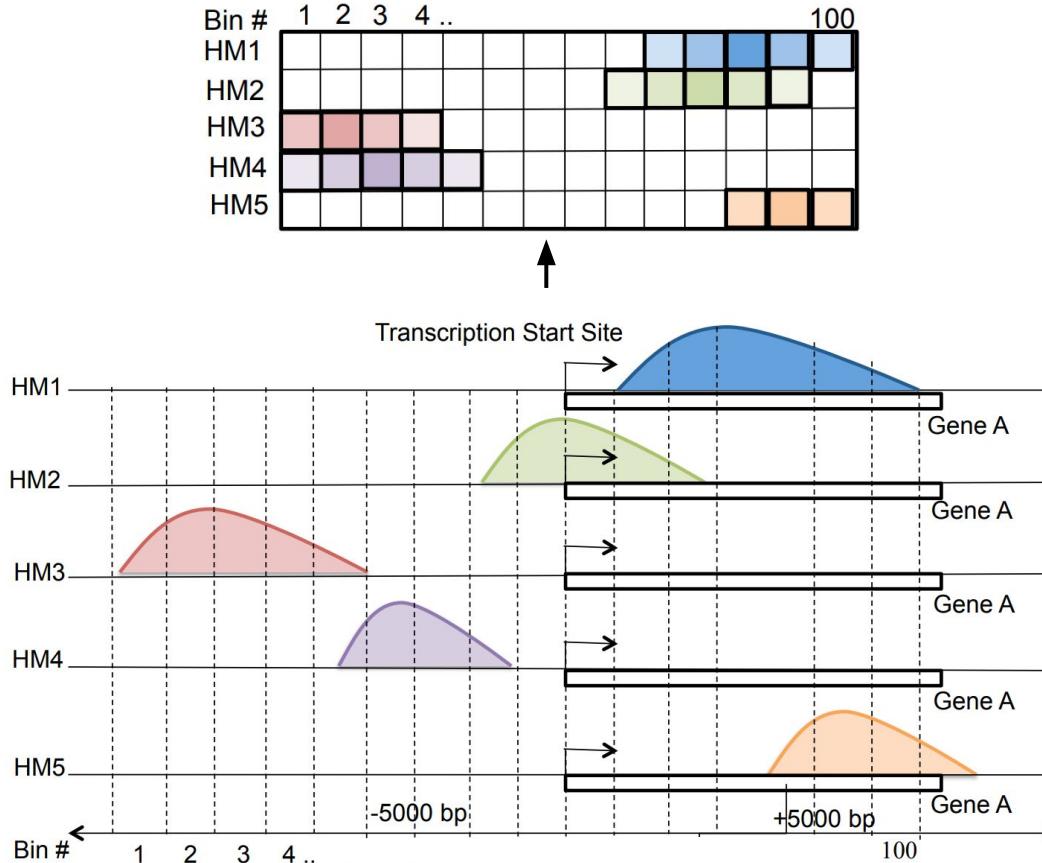
Histone Modification Input Data



Histone Modification Input Data

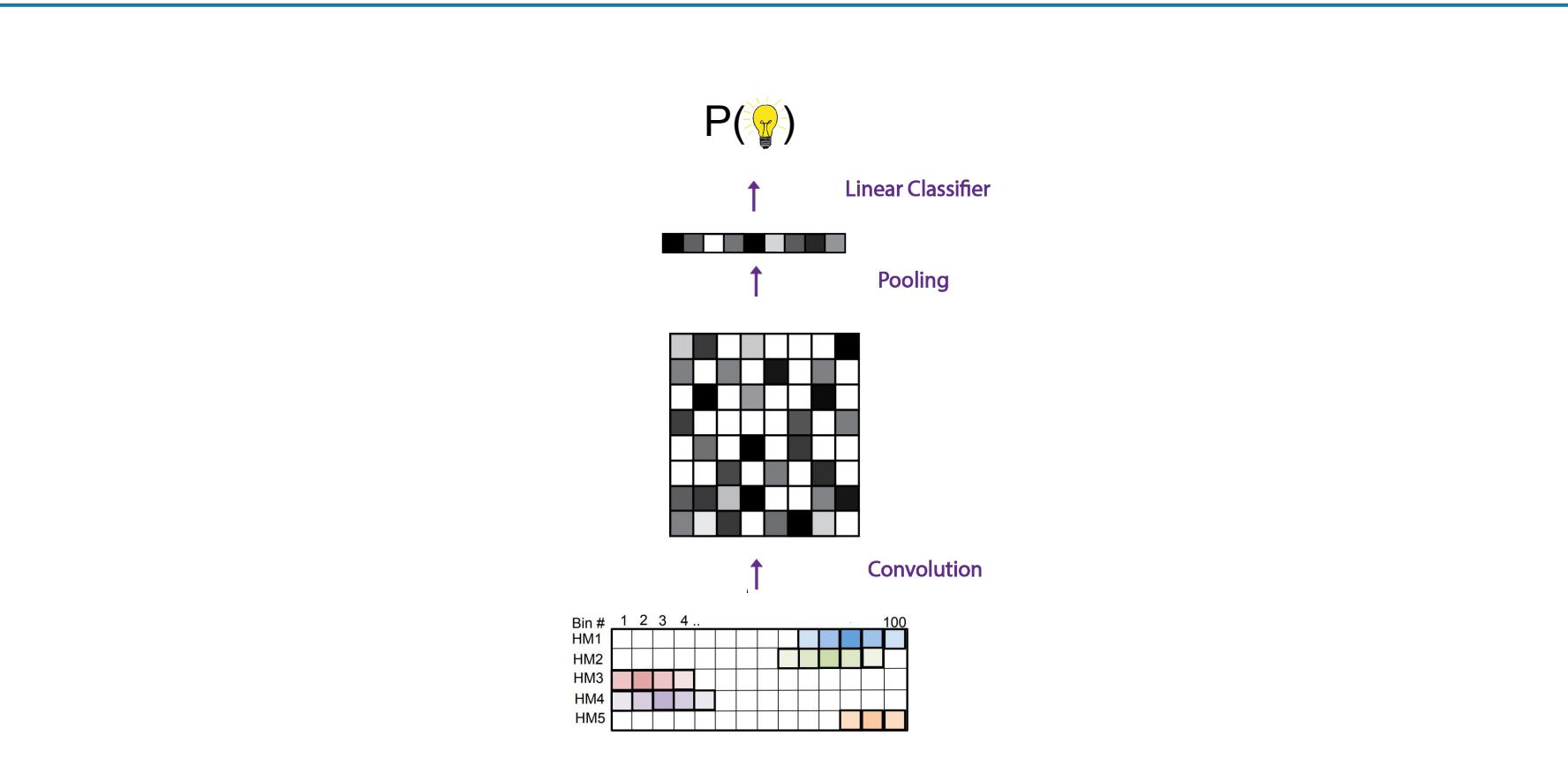


Histone Modification Input Data



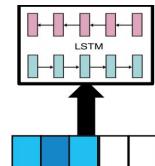
DeepChrome

Singh, Lanchantin, Robins & Qi - Bioinformatics 2016

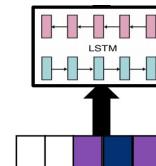


Attentive Chrome

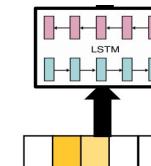
Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



HM1



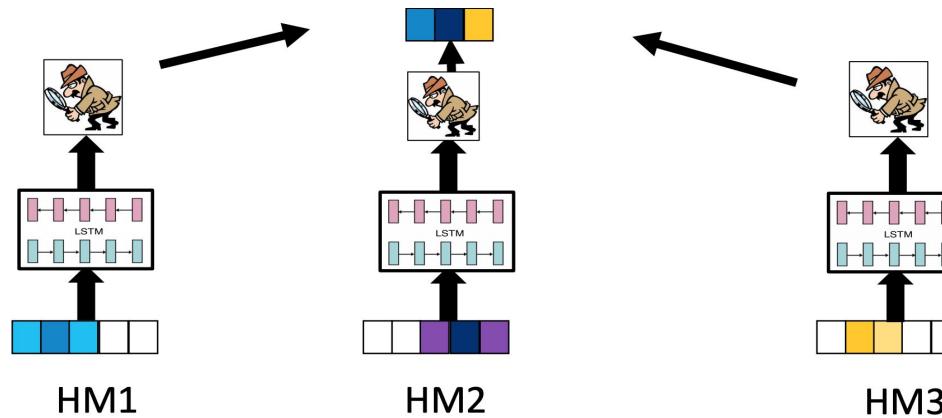
HM2



HM3

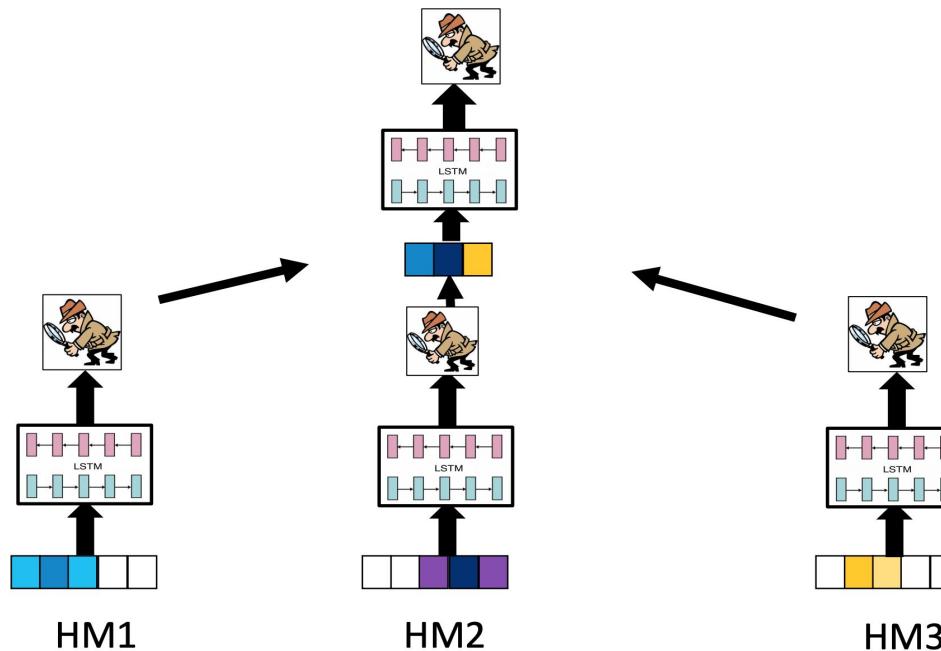
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



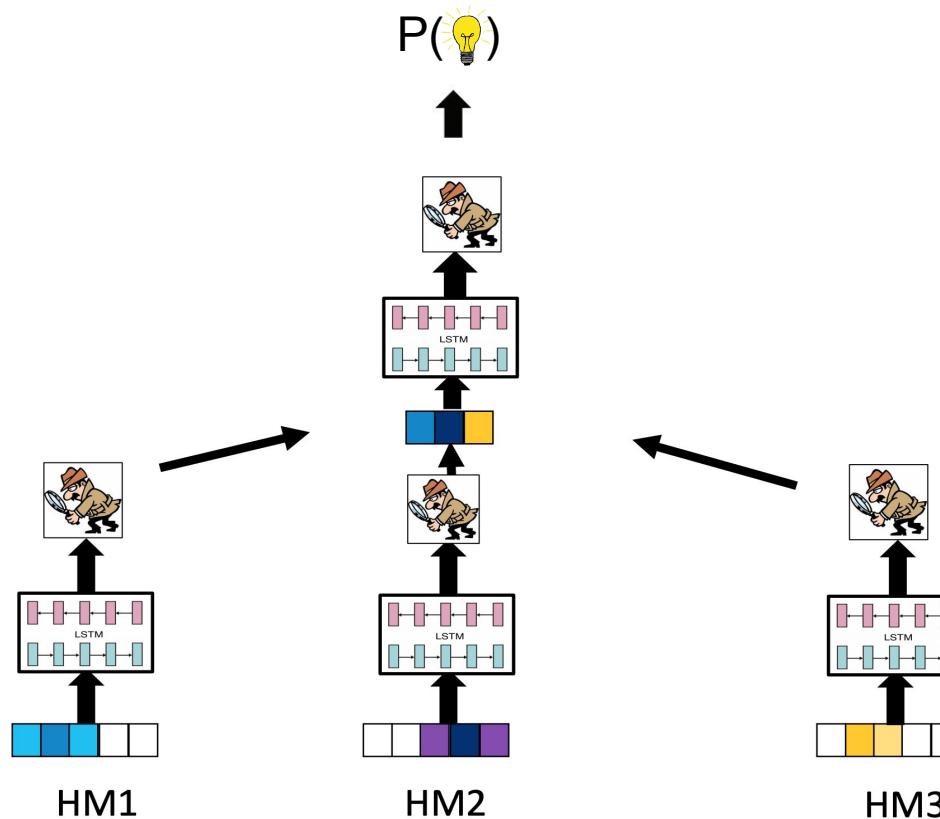
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



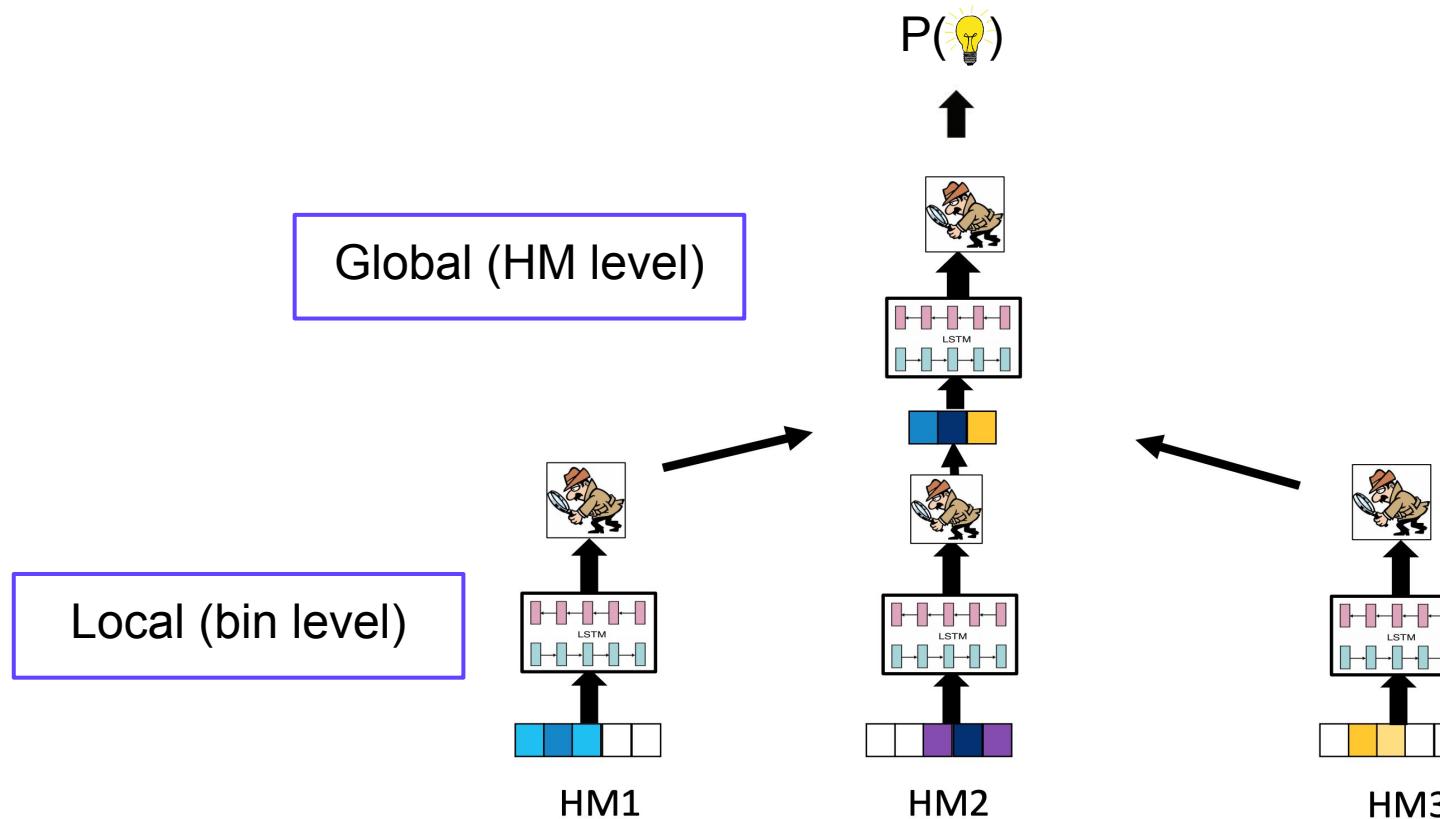
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



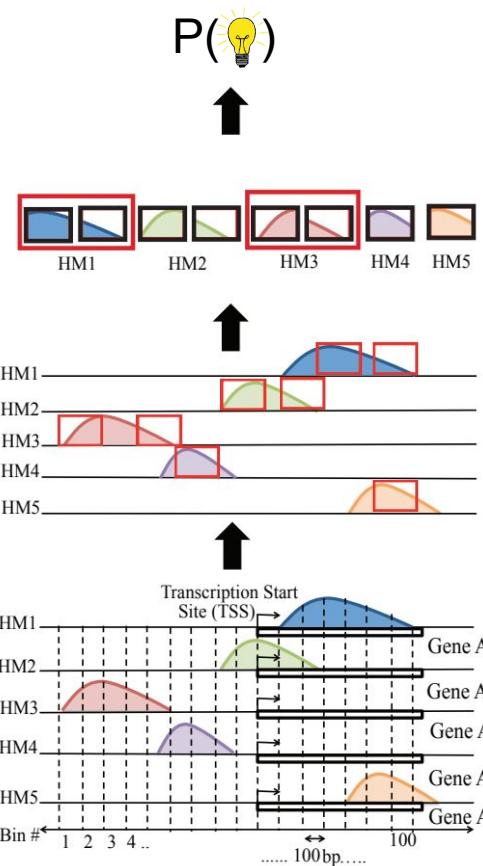
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017

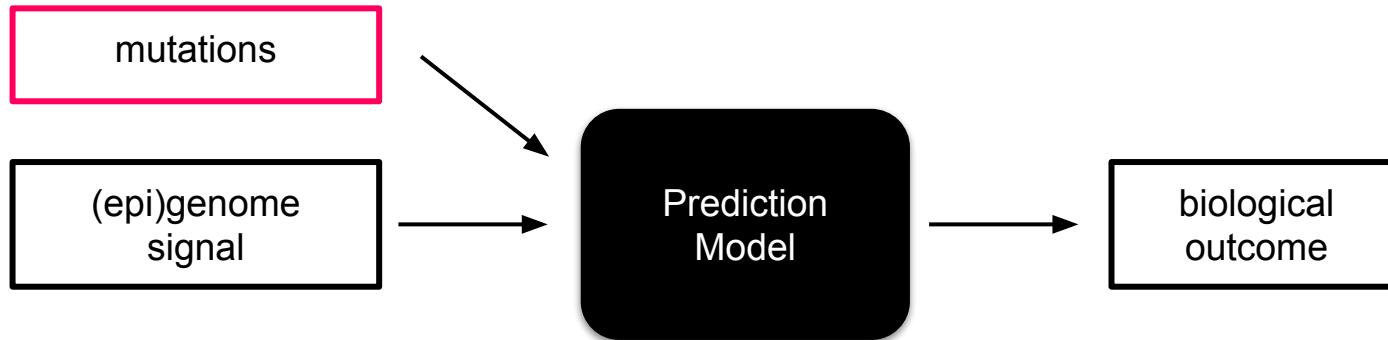


Summarizing Machine Learning for Genomics

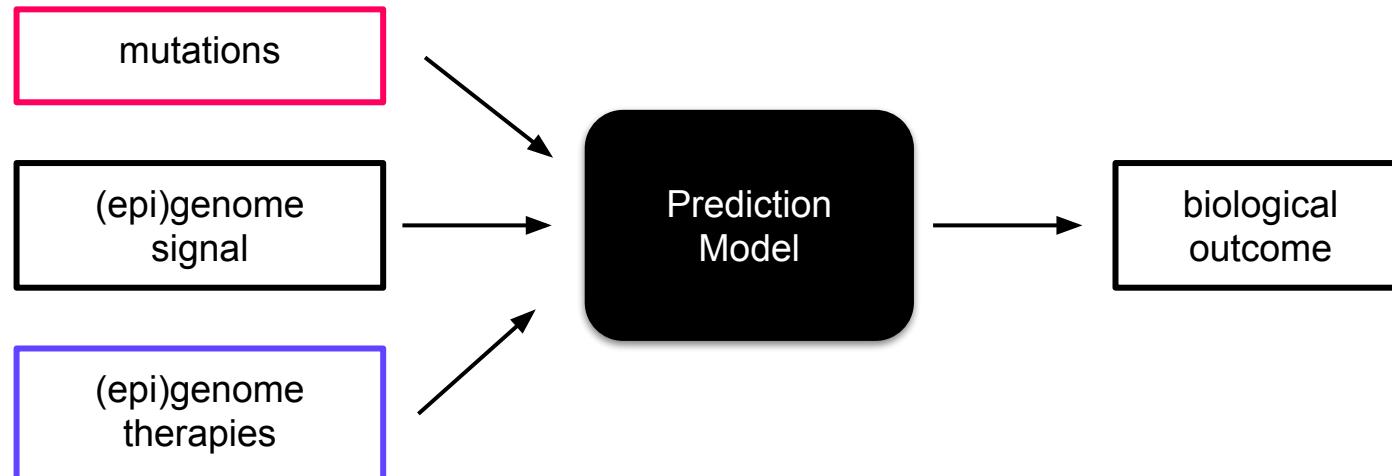
Summarizing Machine Learning for Genomics



Summarizing Machine Learning for Genomics

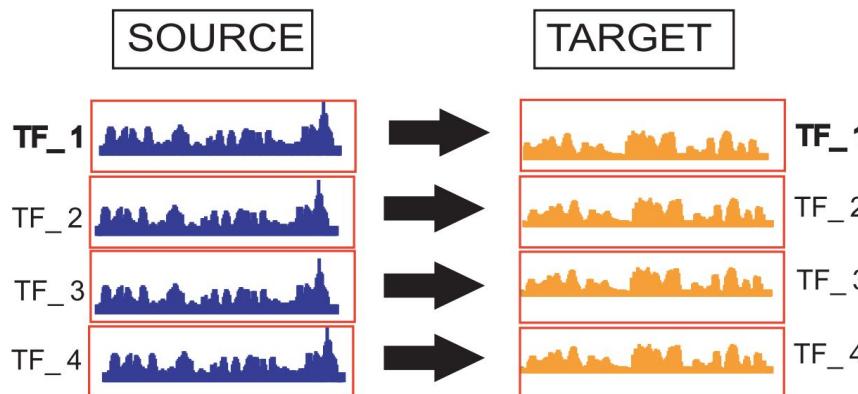
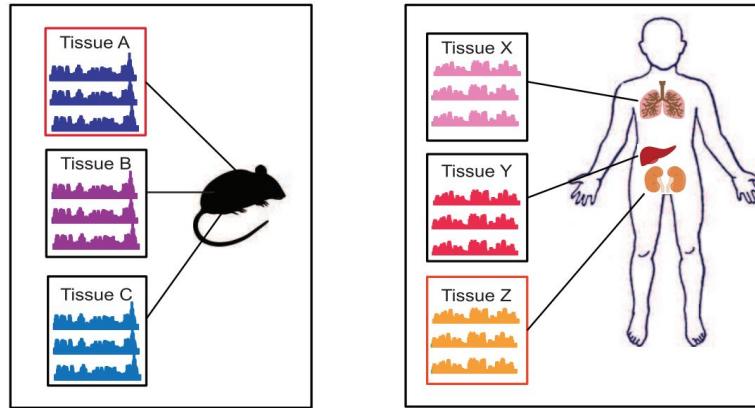


Summarizing Machine Learning for Genomics

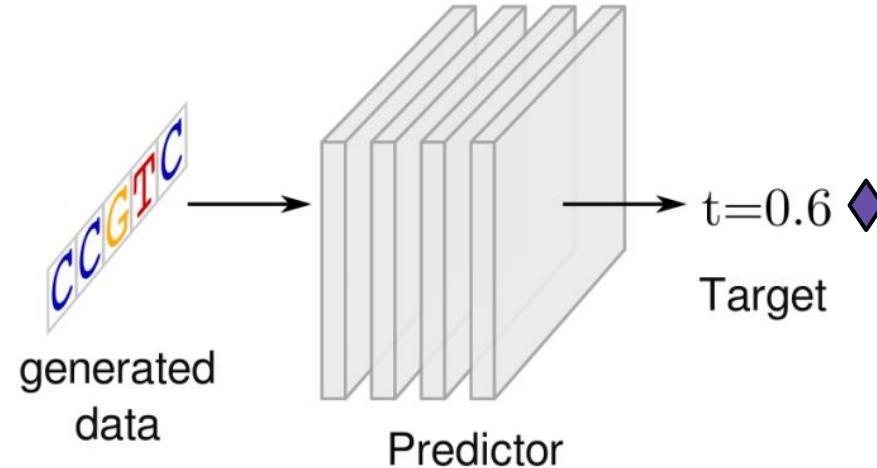


Open Problems

Transfer Learning



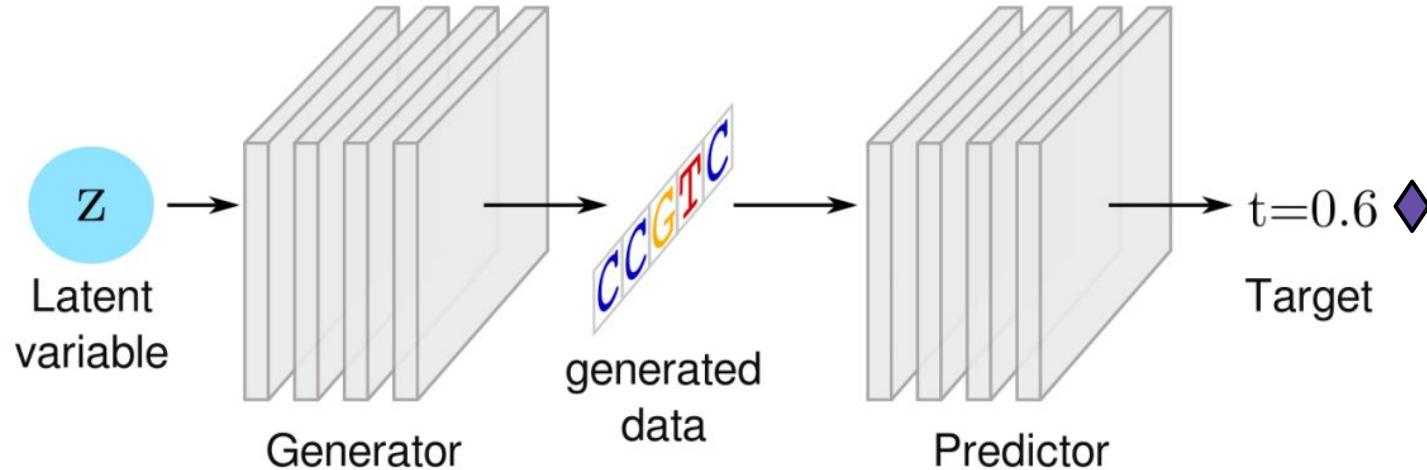
Generative Models



Optimal binding site
for TF "CBX3" ♦



Generative Models



Thank You!

code available at: github.com/qdata



Ritambhara Singh



Arshdeep Sekhon



Beilun Wang



Dr. Yanjun Qi



UNIVERSITY of VIRGINIA