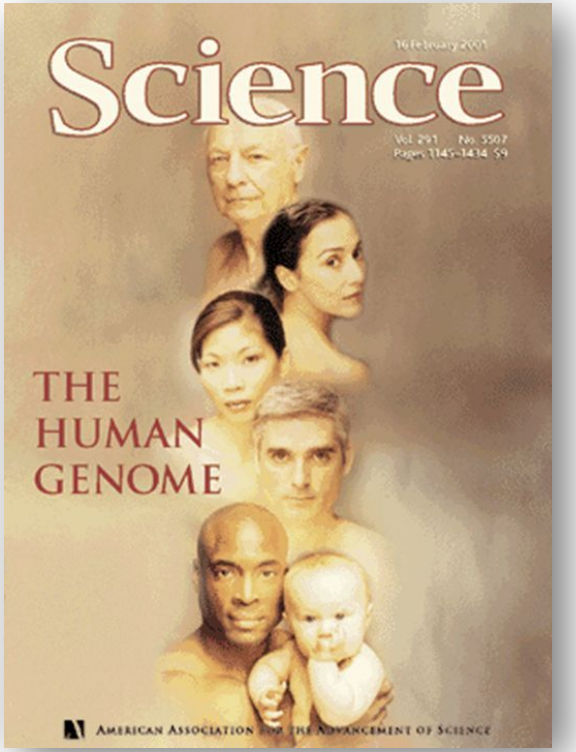
Deep Learning for Genomics

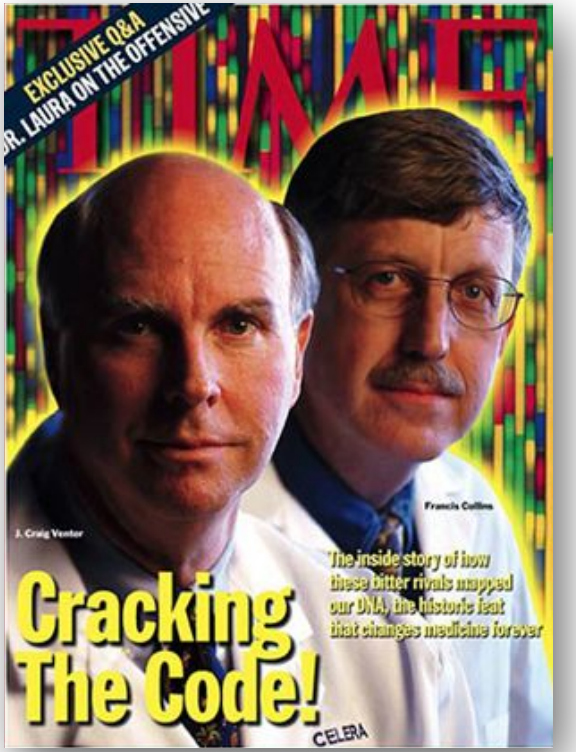
Jack Lanchantin
University of Virginia, Department of Computer Science

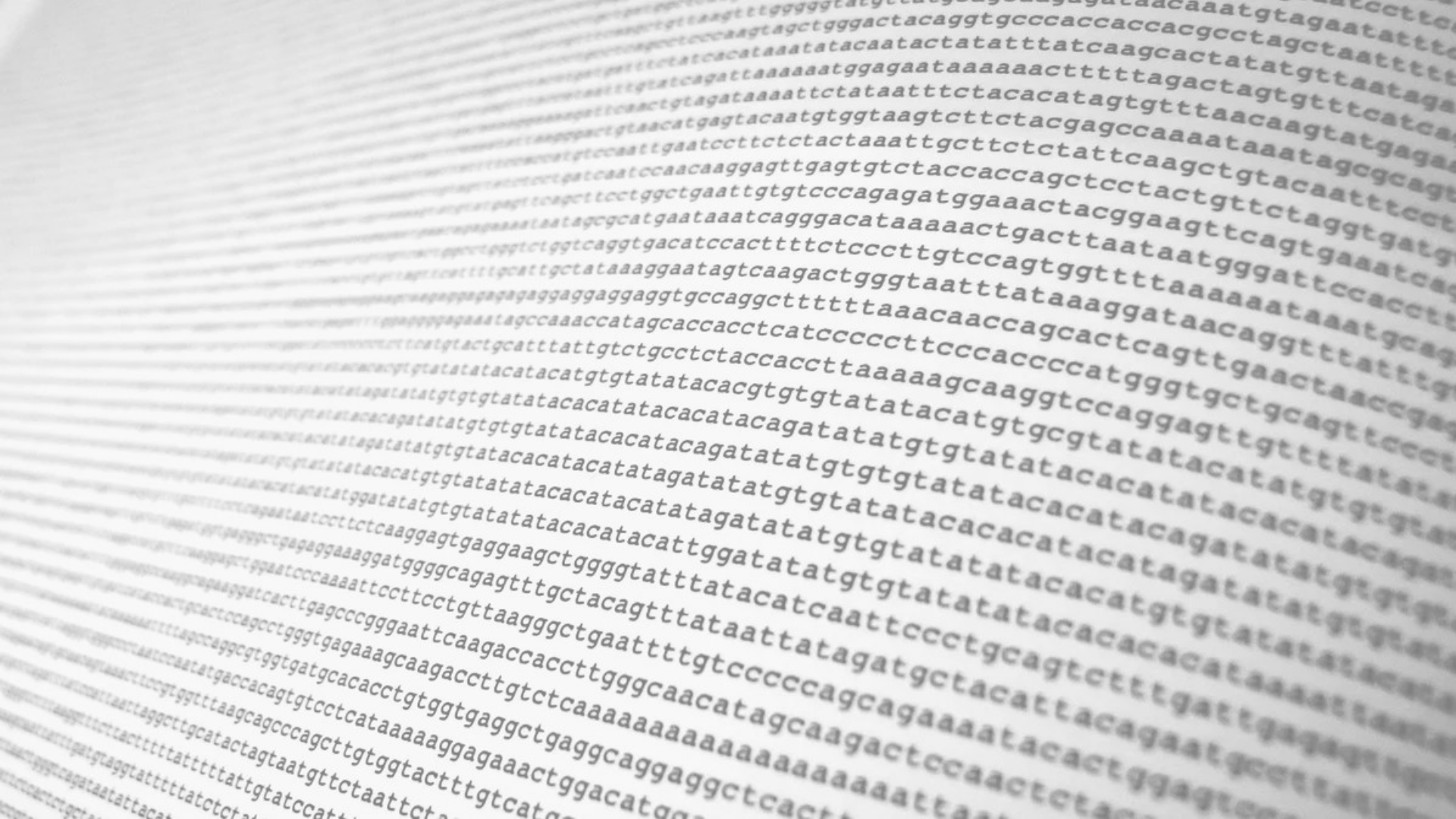
UNIVERSITY of VIRGINIA

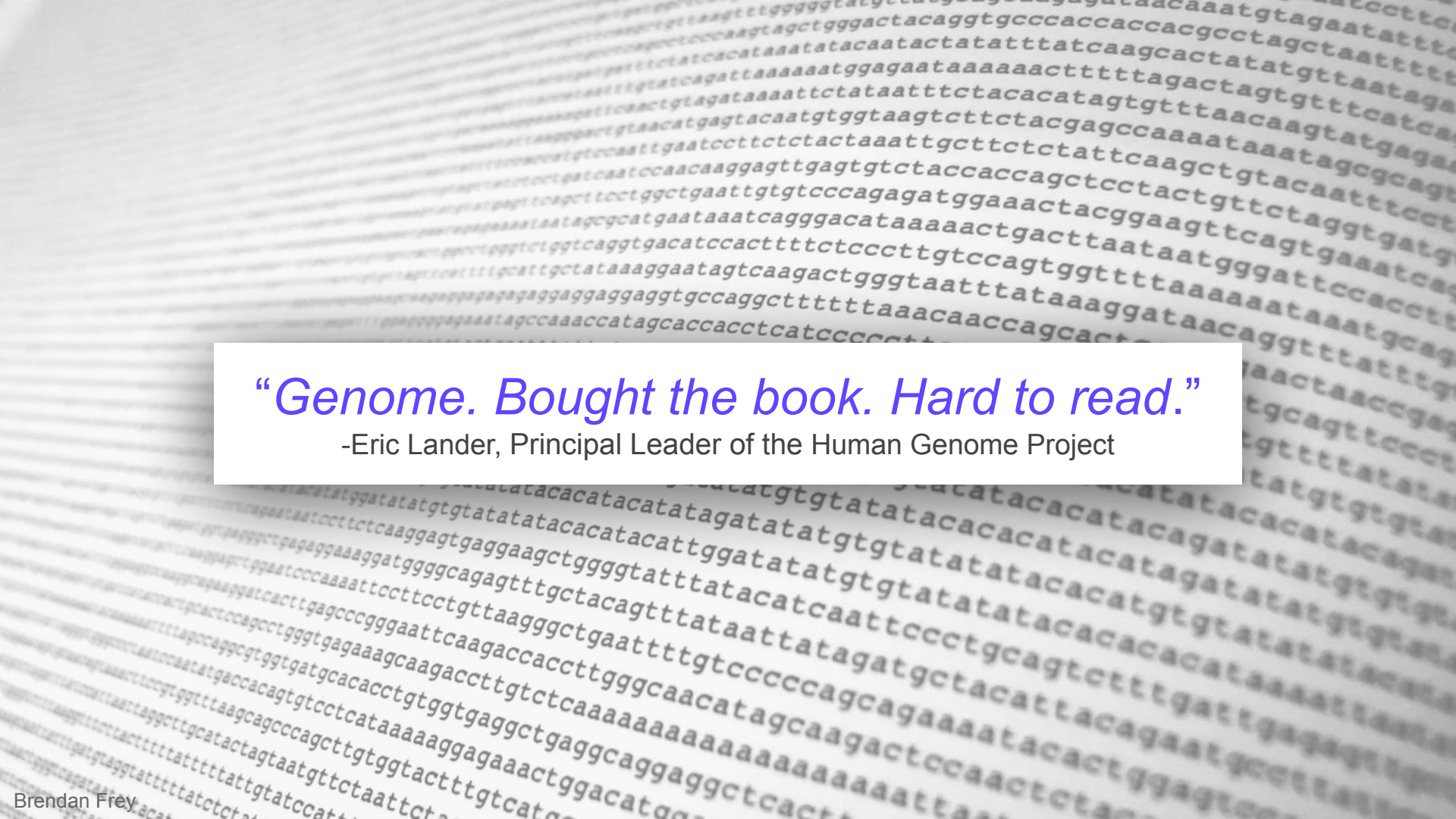
The Human Genome Project (1990-2003)

Mapped the human genome sequence and identified the genes

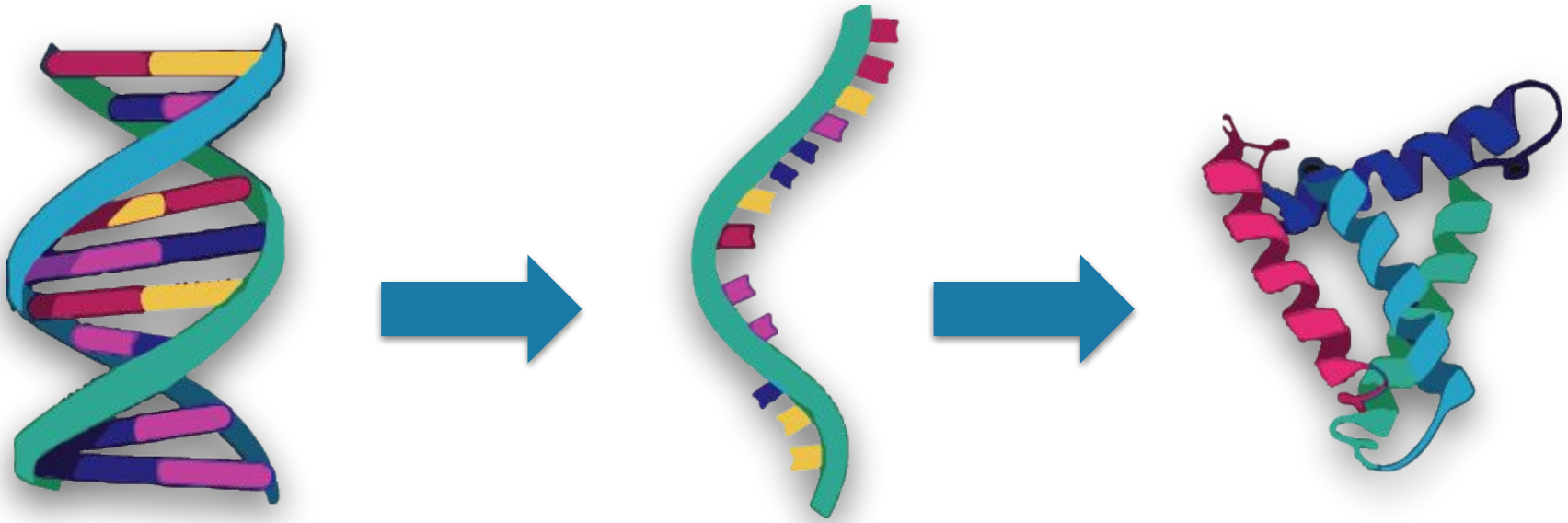


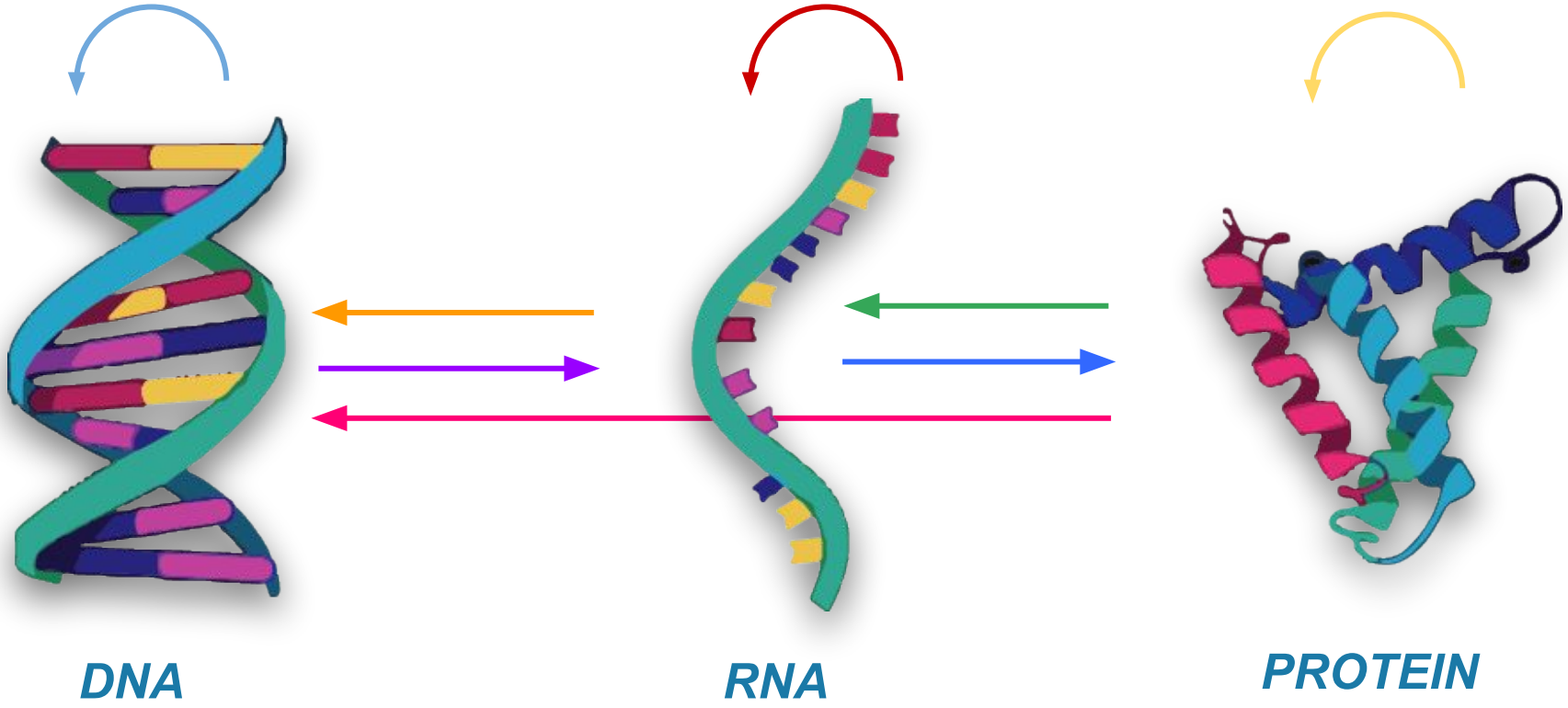




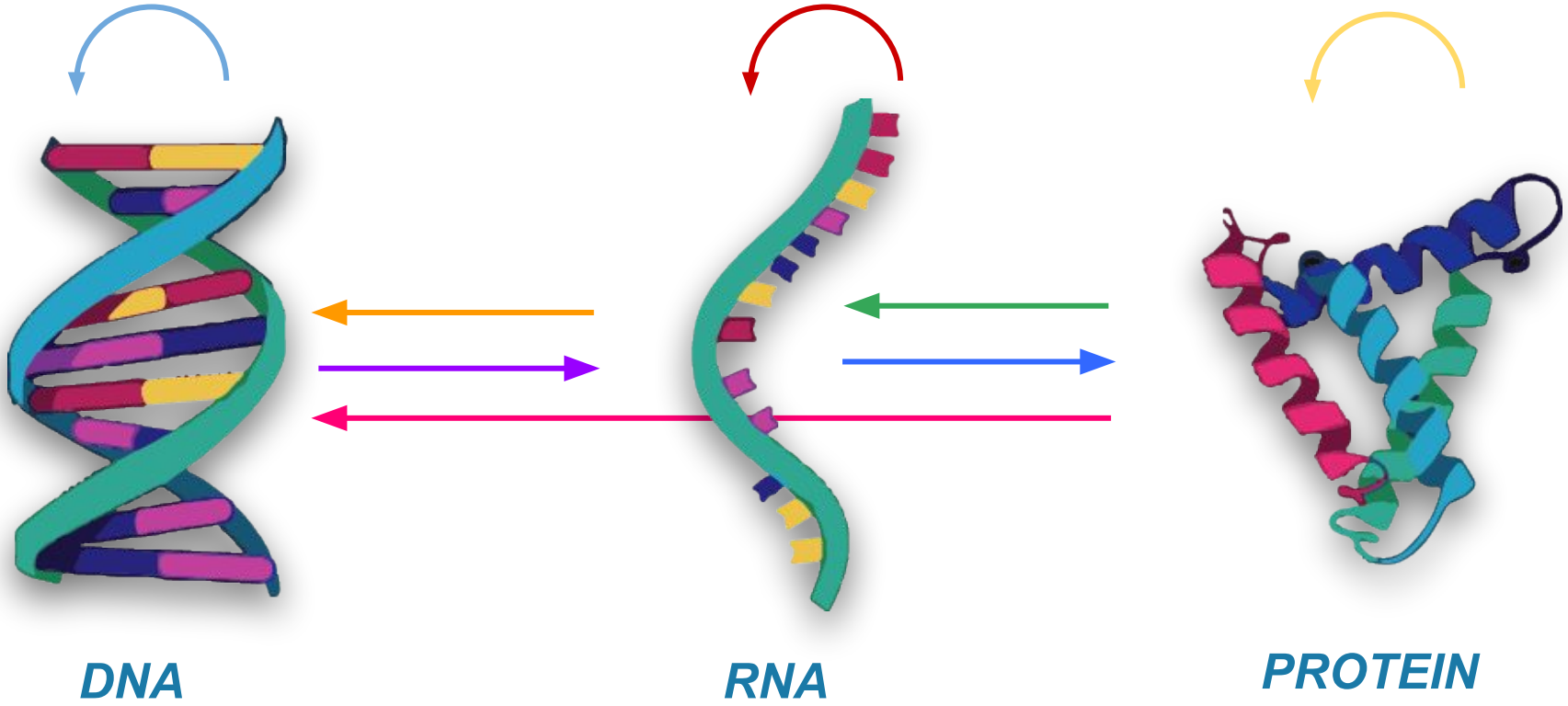


Brendan Frey


DNA
RNA
PROTEIN


DNA
RNA
PROTEIN

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



Brendan Frey

65%

LIFETIME RISK OF GENETIC DISEASE

8M

BIRTHS PER YEAR WITH GENETIC DEFECT

Brendan Frey

65%

LIFETIME RISK OF GENETIC DISEASE

8M

BIRTHS PER YEAR WITH GENETIC DEFECT

$5M

LIFETIME COST PER CHILD IN US

Brendan Frey

Brendan Frey

Gene

ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

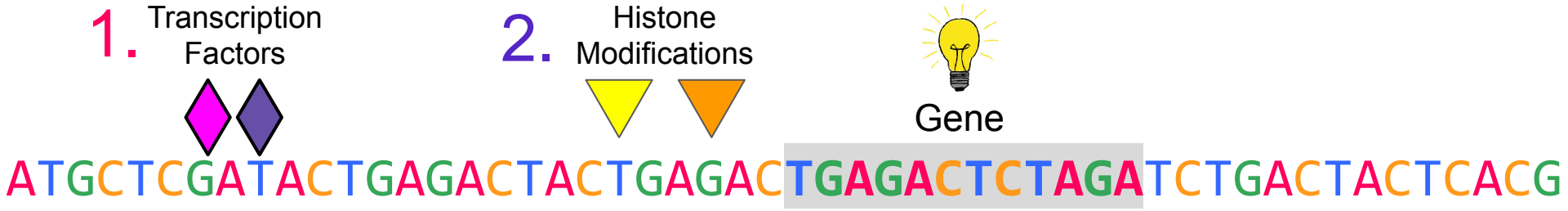
Gene Expressed

ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

Gene Expressed

ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

what causes a gene to be expressed?


1.
Transcription Factors
2.
Histone Modifications
Gene
ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

1. Predicting Transcription Factor Binding Sites from DNA

2. Predicting Gene Expression from Histone Modifications

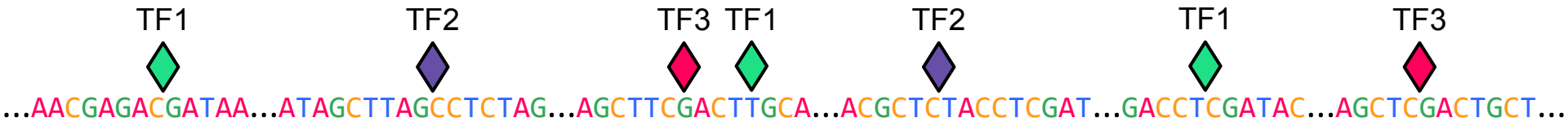
Transcription Factors (TFs)

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

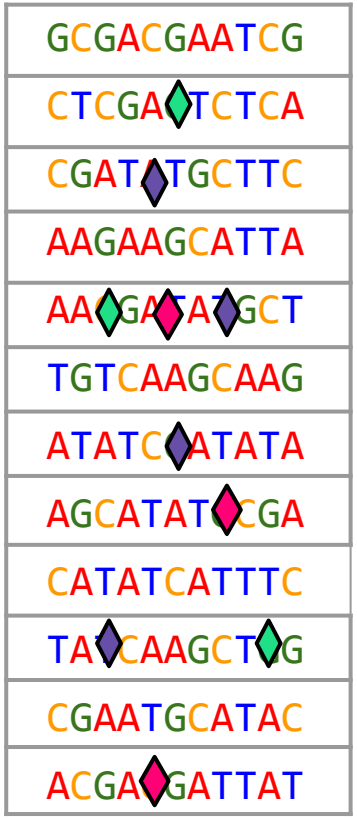
Transcription Factors (TFs)

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

Transcription Factor Binding Sites (TFBSs)



TFBS Datasets



TFBS Classification Task

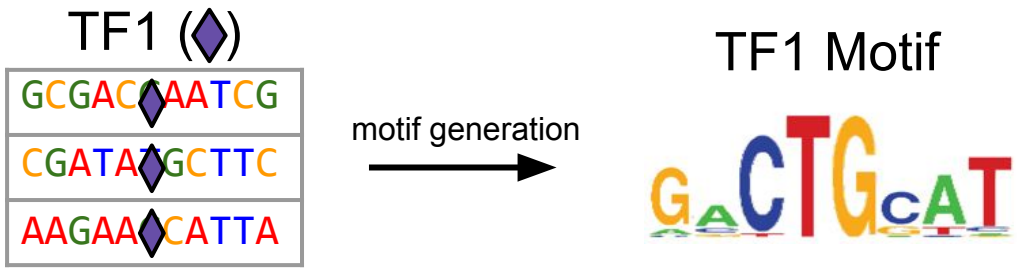
TFBS Classification Task

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



Bailey et. al. - Nucleic Acids Research 2006

Prediction with Motifs

?

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

Bailey et. al. - Nucleic Acids Research 2006

Prediction with Motifs

Bailey et. al. - Nucleic Acids Research 2006

Prediction with Motifs

?

Bailey et. al. - Nucleic Acids Research 2006

Prediction with Motifs

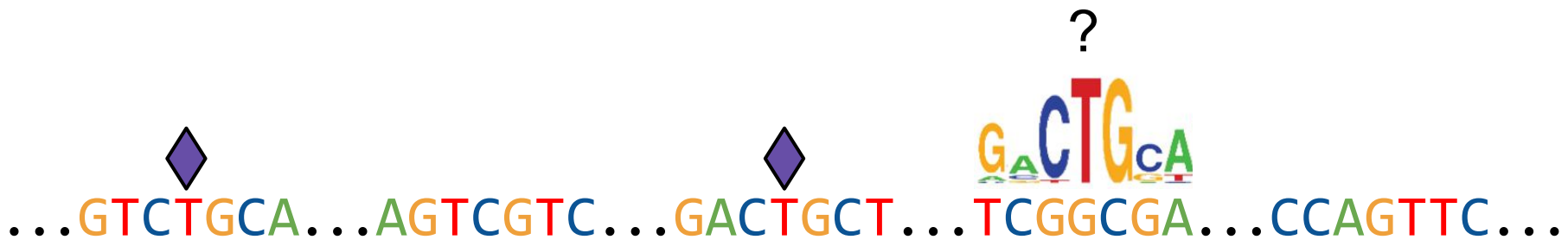
?

Bailey et. al. - Nucleic Acids Research 2006

Prediction with Motifs

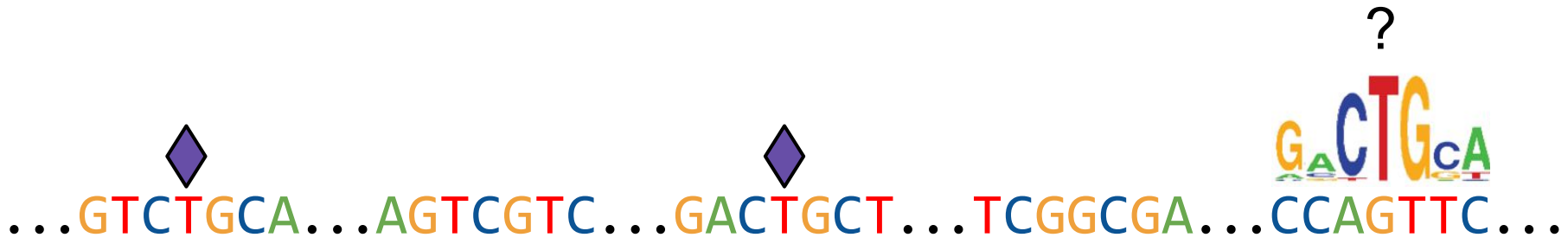
Bailey et. al. - Nucleic Acids Research 2006

Prediction with Motifs



Bailey et. al. - Nucleic Acids Research 2006

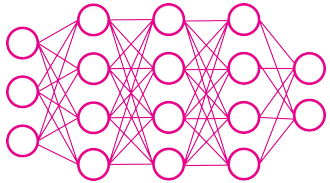
Prediction with Motifs



Bailey et. al. - Nucleic Acids Research 2006

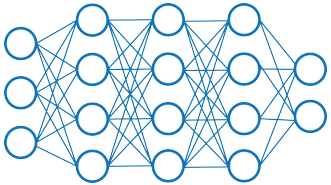
✗ Good motifs are hard to generate

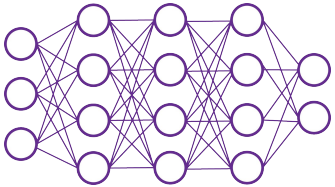
- Good motifs are hard to generate
- Deep neural networks are good at automatic feature extraction

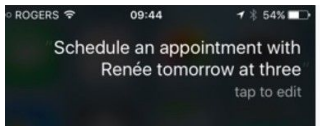


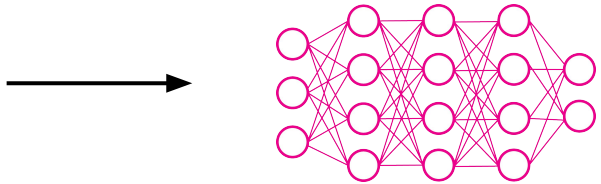
Dog


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






ROGERS
09:44
54%
Schedule an appointment with Renée tomorrow at three
tap to edit



Dog

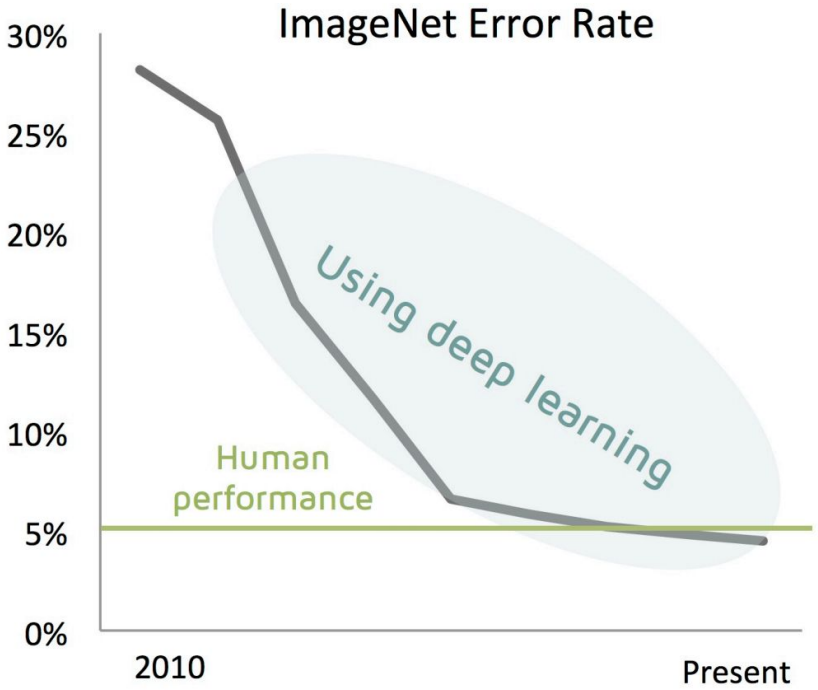

ImageNet Error Rate
30%
25%
20%
15%
10%
5%
0%
Using deep learning
Human performance
2010
Present

Image Recognition

Extracted Features

Dog

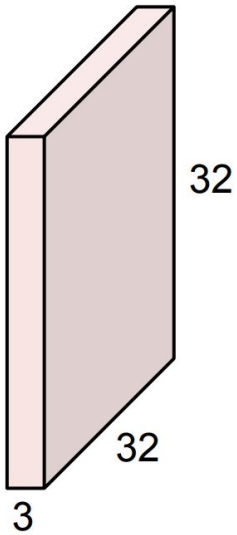
Image Recognition

Extracted Features

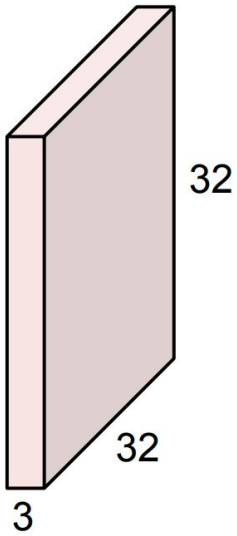
?

Dog

Key Operator: Convolutional Neural Networks

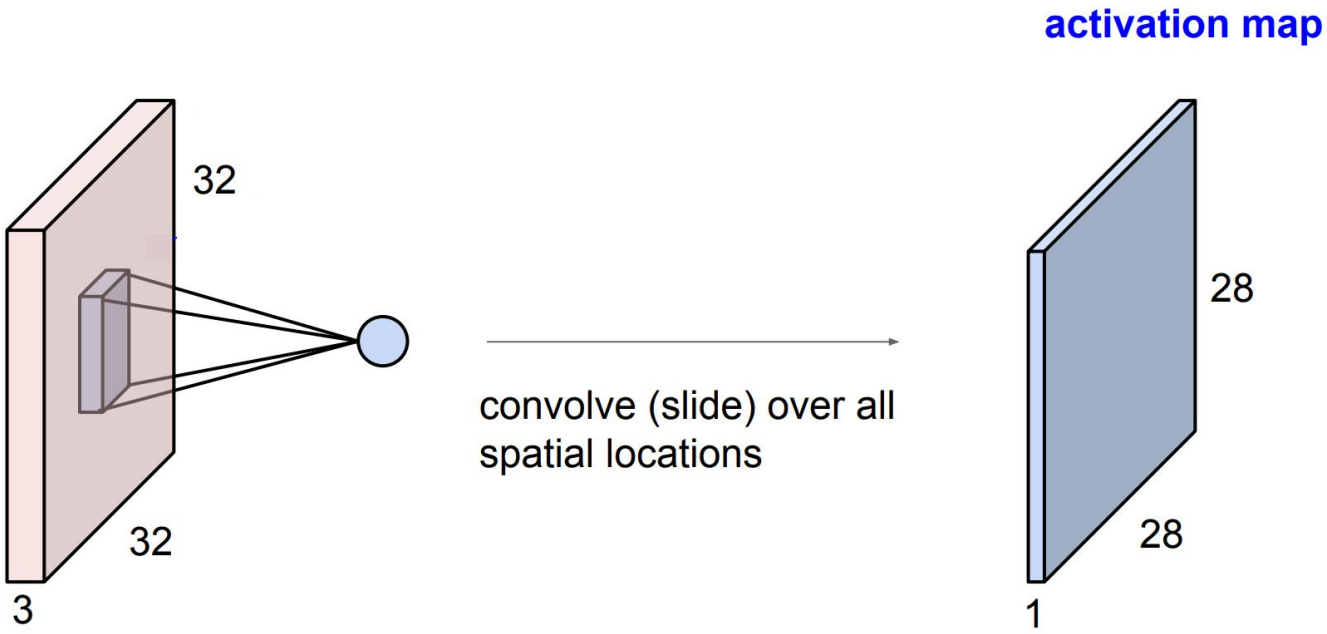


Key Operator: Convolutional Neural Networks

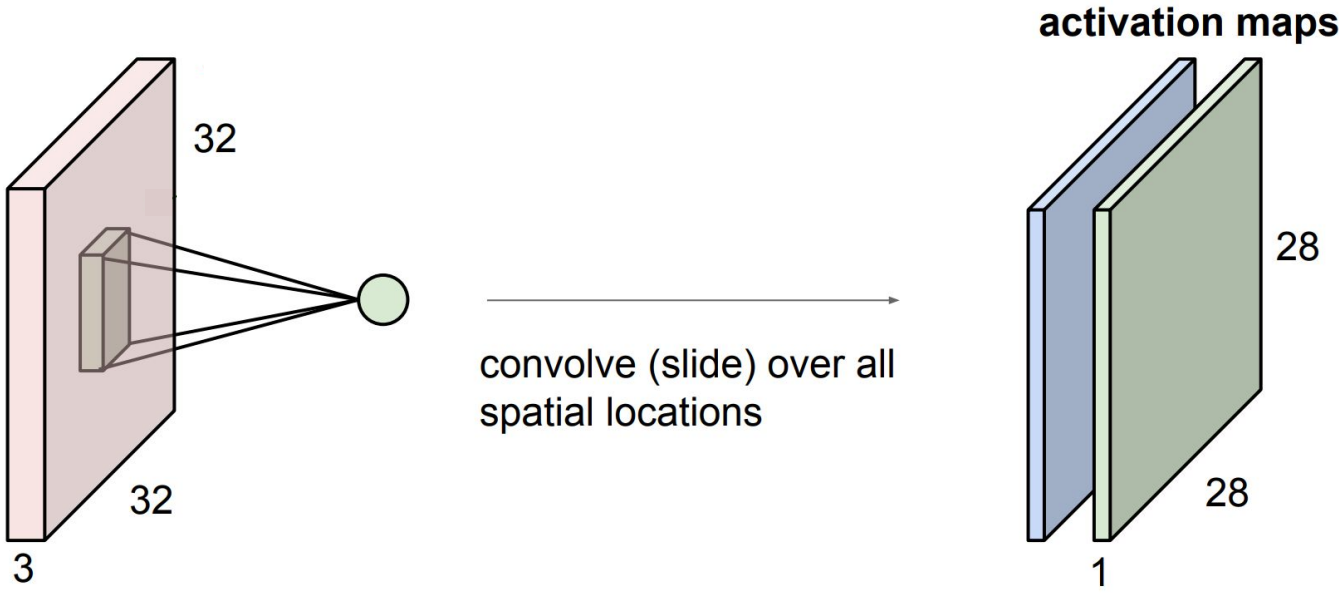


5x5x3 filter

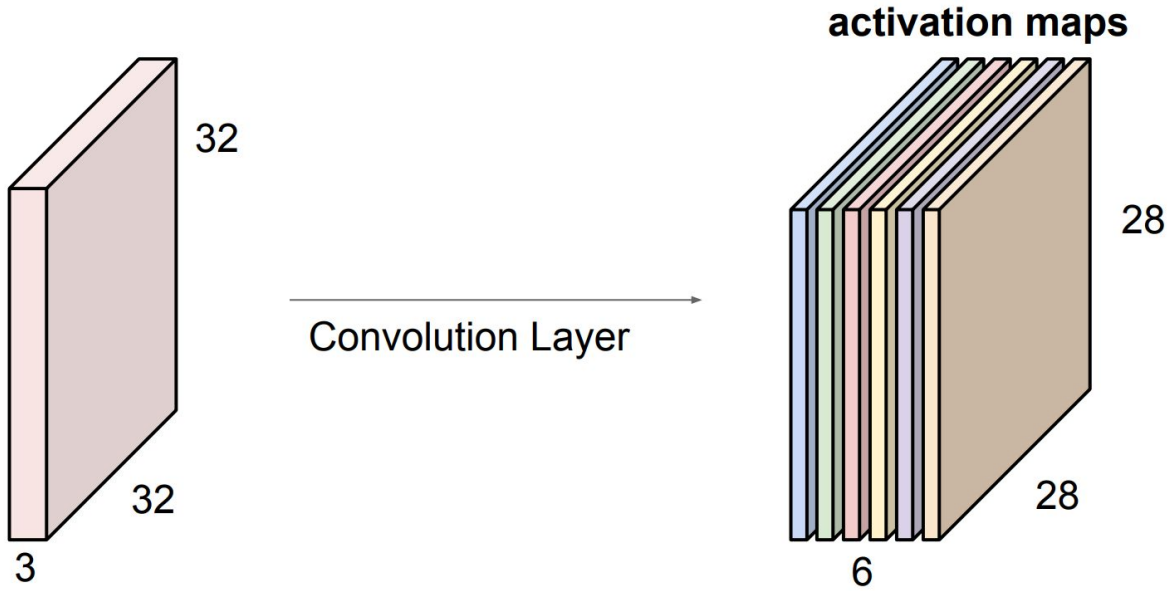
Key Operator: Convolutional Neural Networks



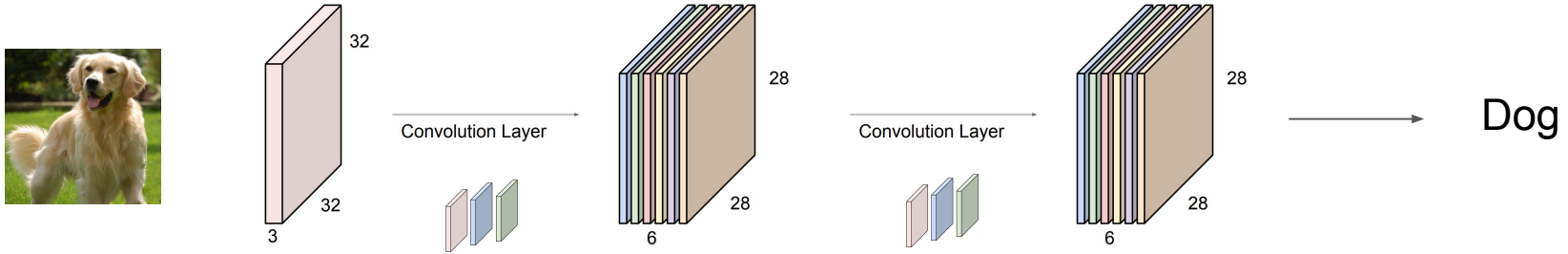
Key Operator: Convolutional Neural Networks



Key Operator: Convolutional Neural Networks



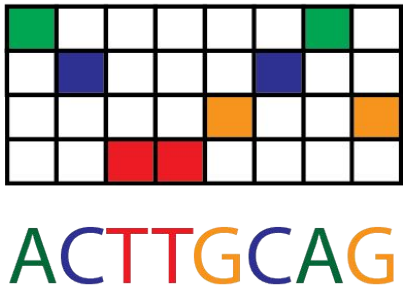
Key Operator: Convolutional Neural Networks



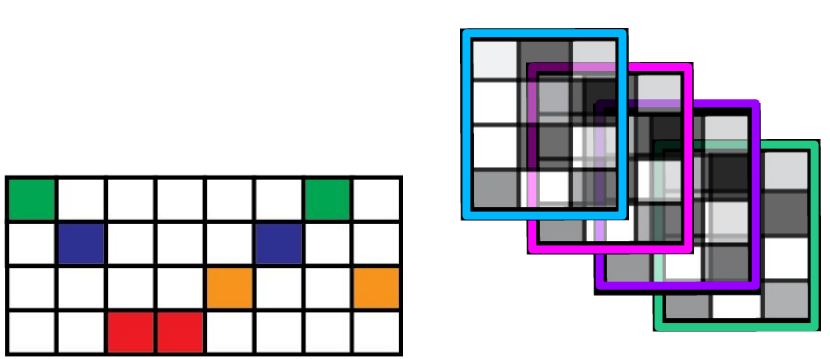
How does this relate to Genomics?

Convolutional Neural Network for DNA

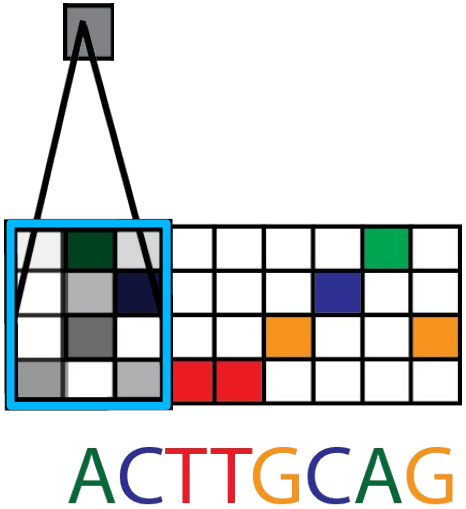
Convolutional Neural Network for DNA



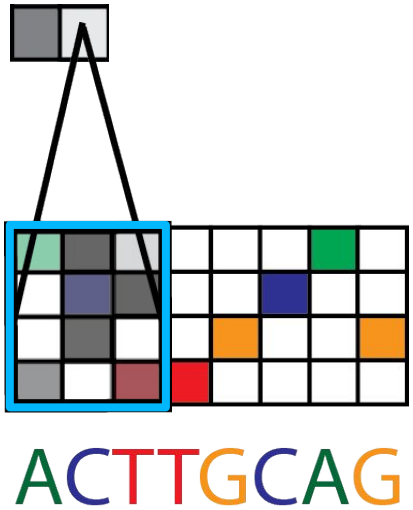
Convolutional Neural Network for DNA



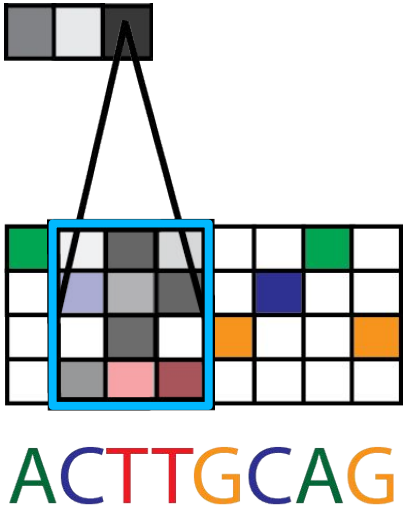
Convolutional Neural Network for DNA



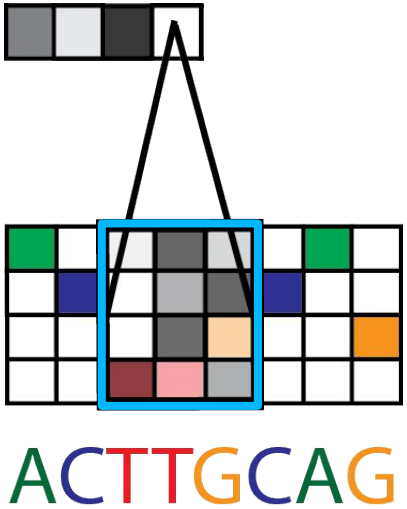
Convolutional Neural Network for DNA



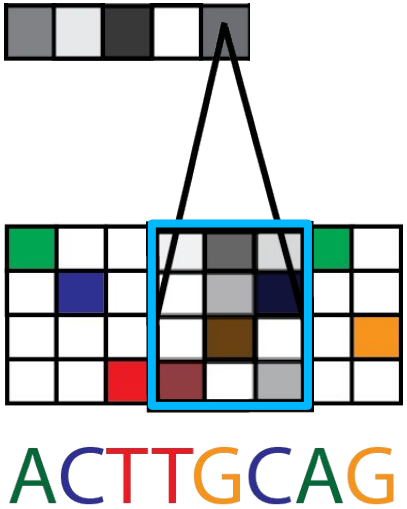
Convolutional Neural Network for DNA



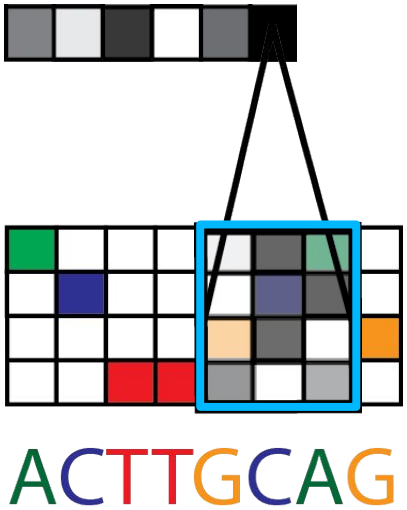
Convolutional Neural Network for DNA



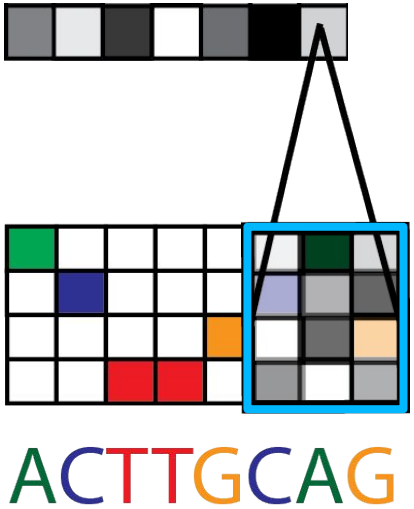
Convolutional Neural Network for DNA



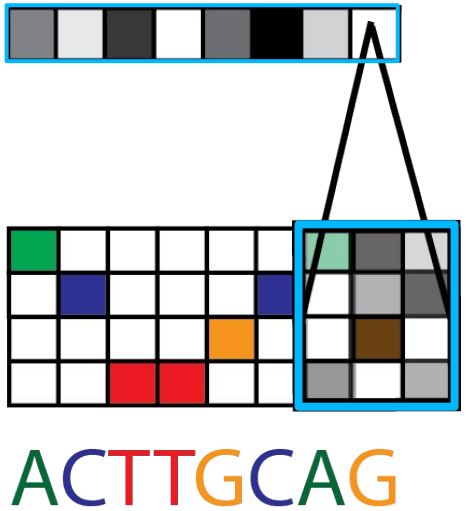
Convolutional Neural Network for DNA



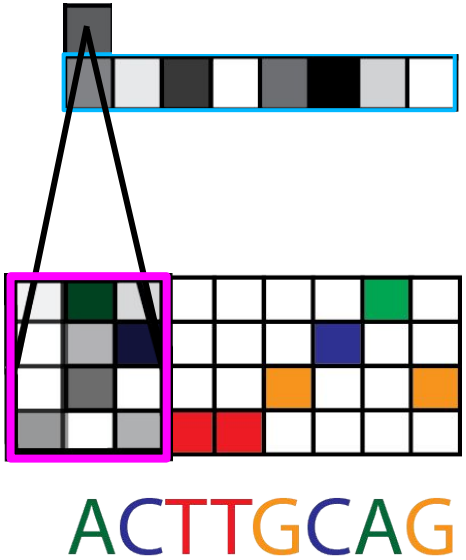
Convolutional Neural Network for DNA



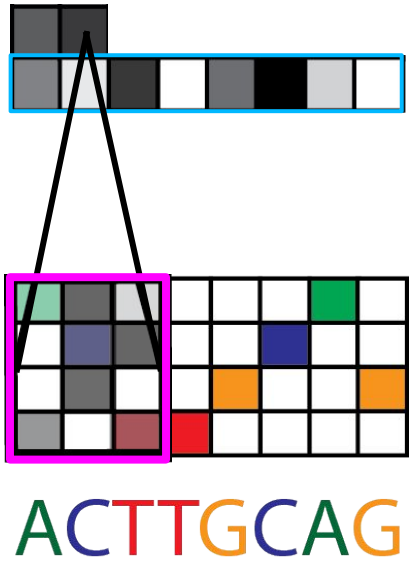
Convolutional Neural Network for DNA



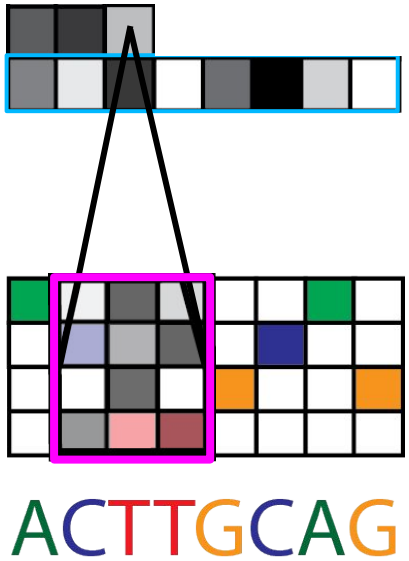
Convolutional Neural Network for DNA



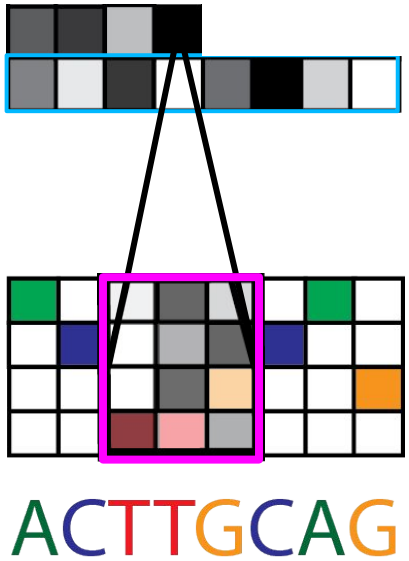
Convolutional Neural Network for DNA



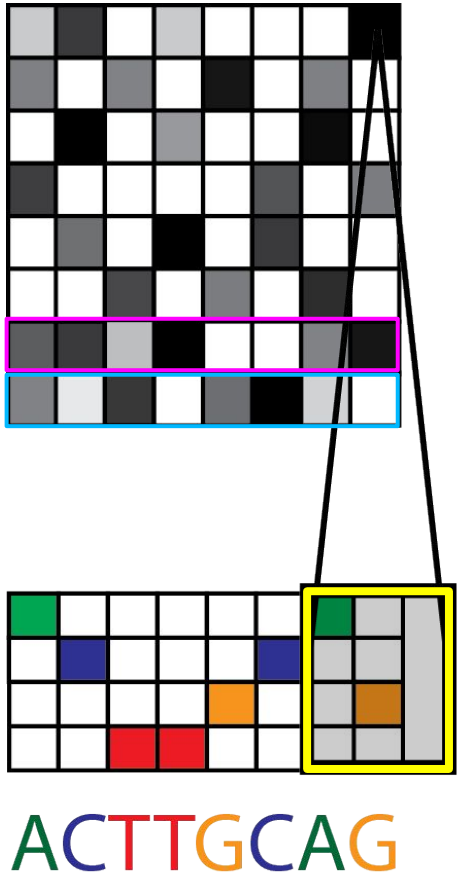
Convolutional Neural Network for DNA



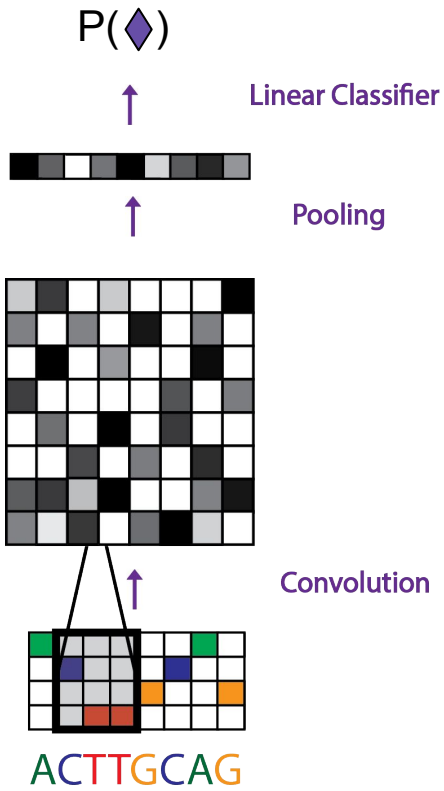
Convolutional Neural Network for DNA



Convolutional Neural Network for DNA



Convolutional Neural Network for TFBS Classification



Why Convolution For Genomics?

Why Convolution For Genomics?

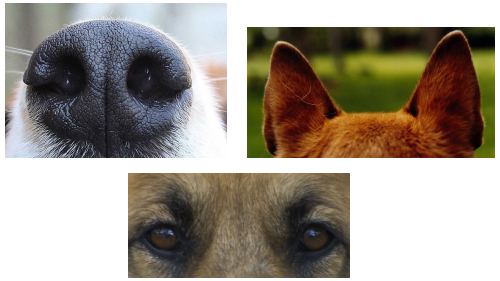
Dogs

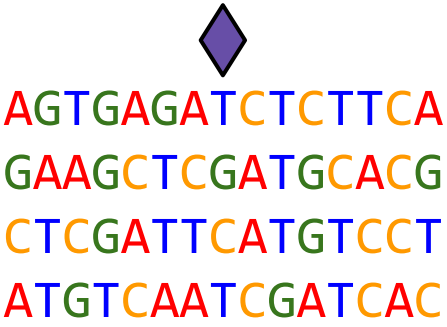
Translational Invariant Features

Why Convolution For Genomics?

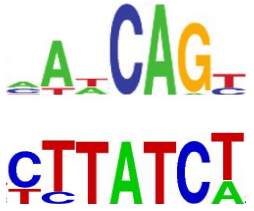
Dogs

Translational Invariant Features

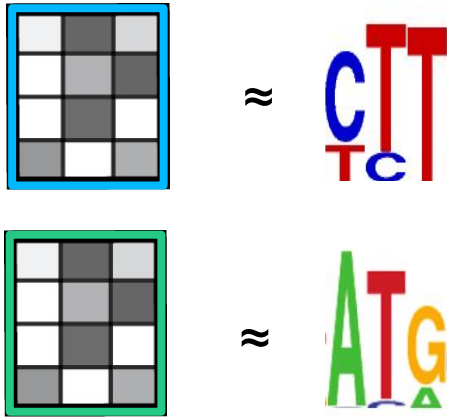




Translational Invariant Features

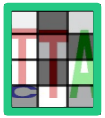


Convolutional Filters learn motifs!



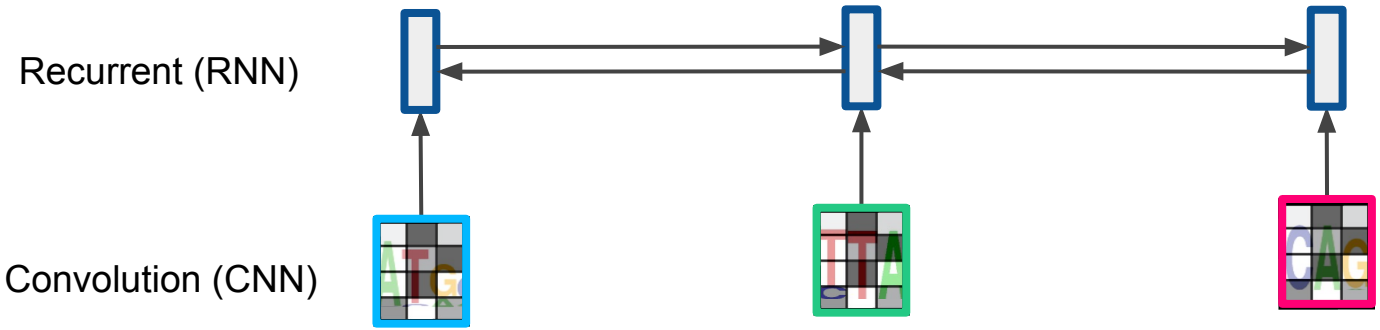
What about relationships among motifs?

Convolution (CNN)

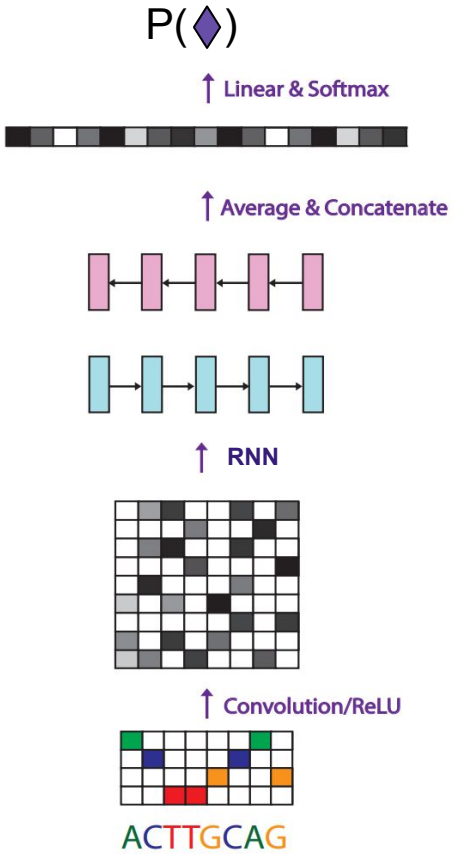


GTGTATGCAGTCGTCGTACTTATCGGCGACCAGTTCAGTACCA

What about relationships among motifs?

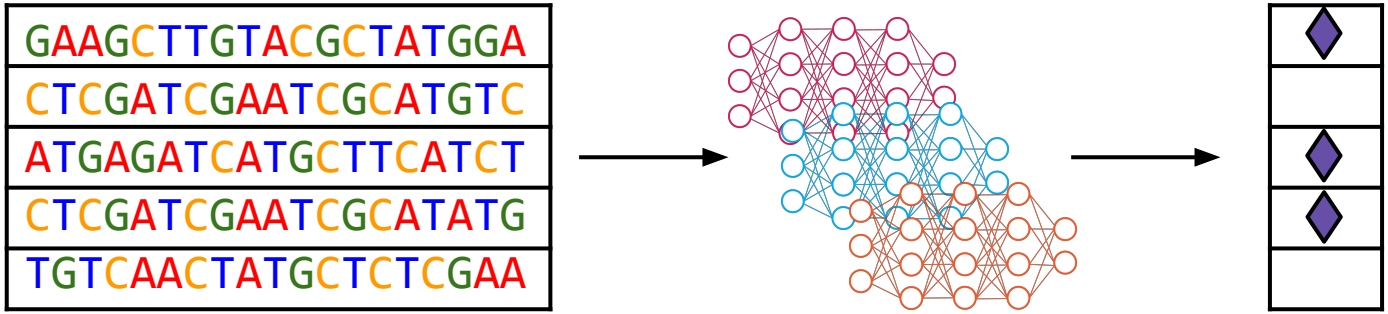


Convolution and Recurrent (CNN-RNN)



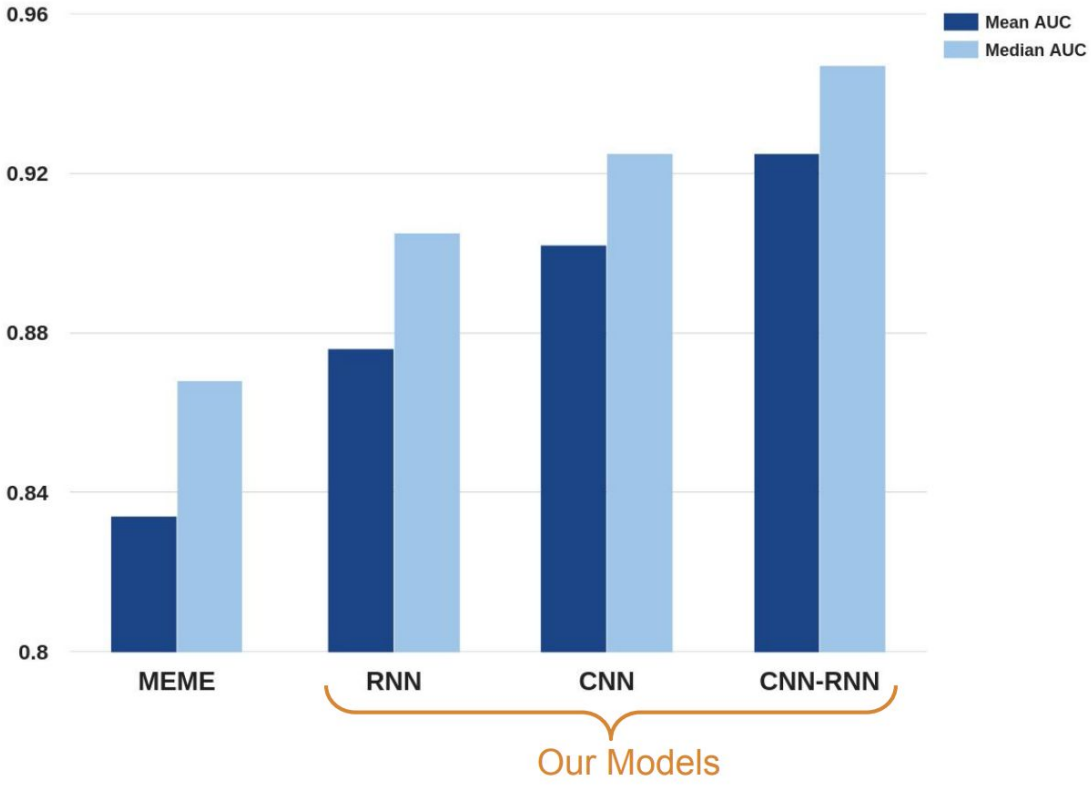
Deep Motif Dashboard

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



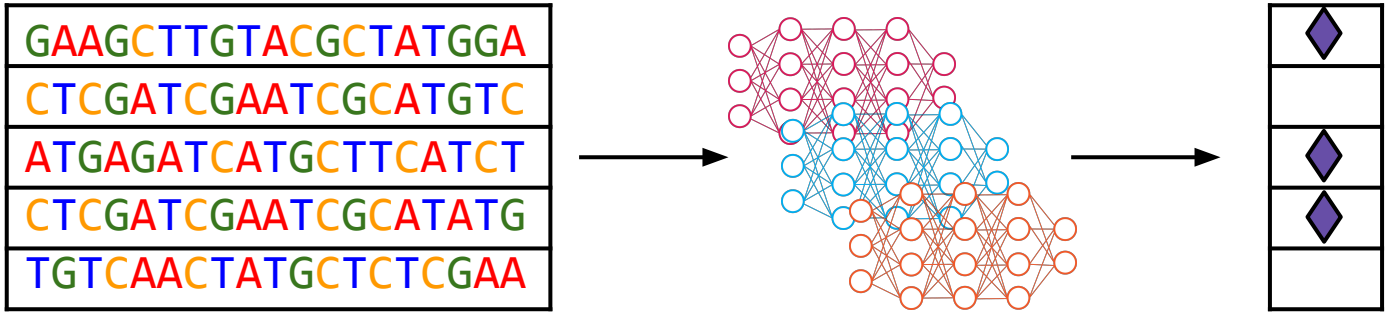
1. Convolutional (CNN)
2. Recurrent (RNN)
3. Convolutional-Recurrent (CNN-RNN)

Model Accuracy on 108 Cancer Cell TFs



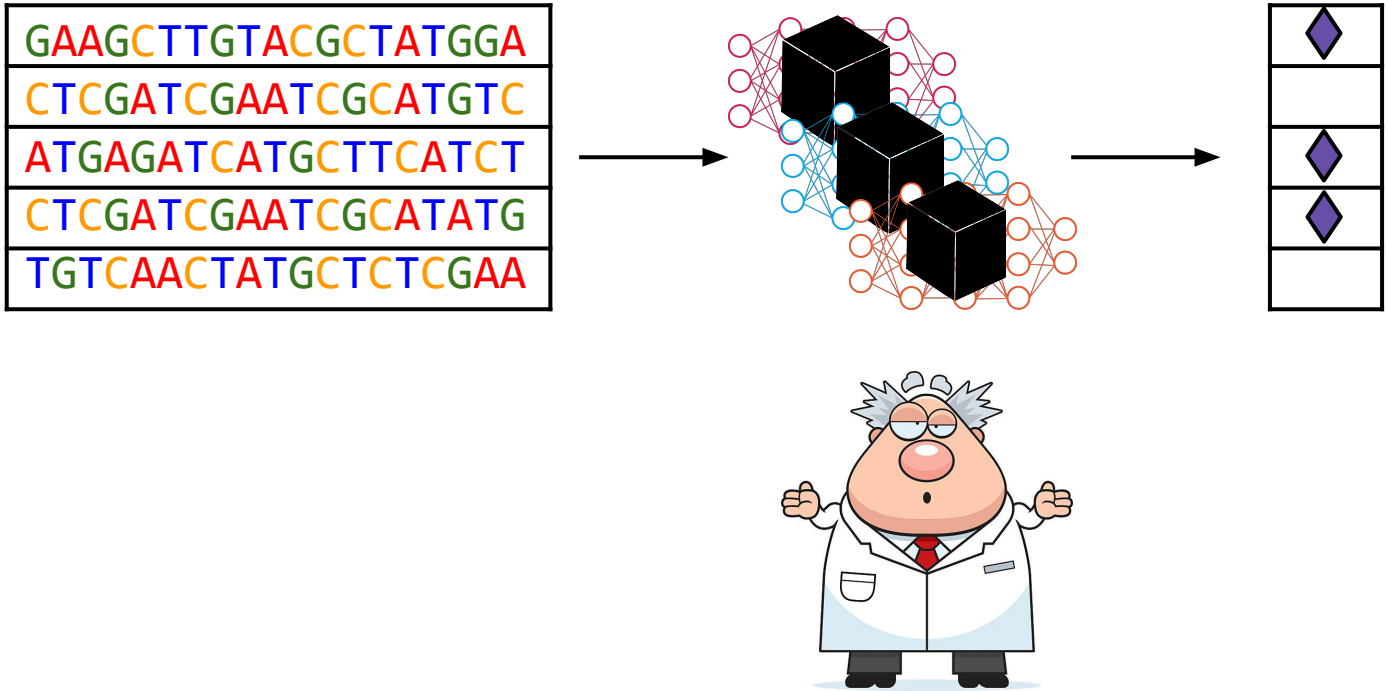
Deep Motif Dashboard

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



Deep **Mo**tif Dashboard

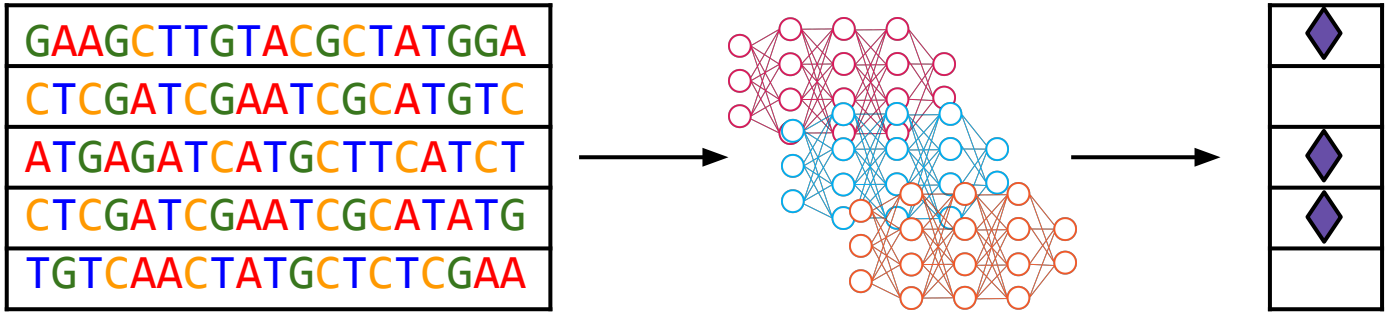
Lanchantin, Singh, Wang & Qi - ICLR Workshops 2016, PSB 2017



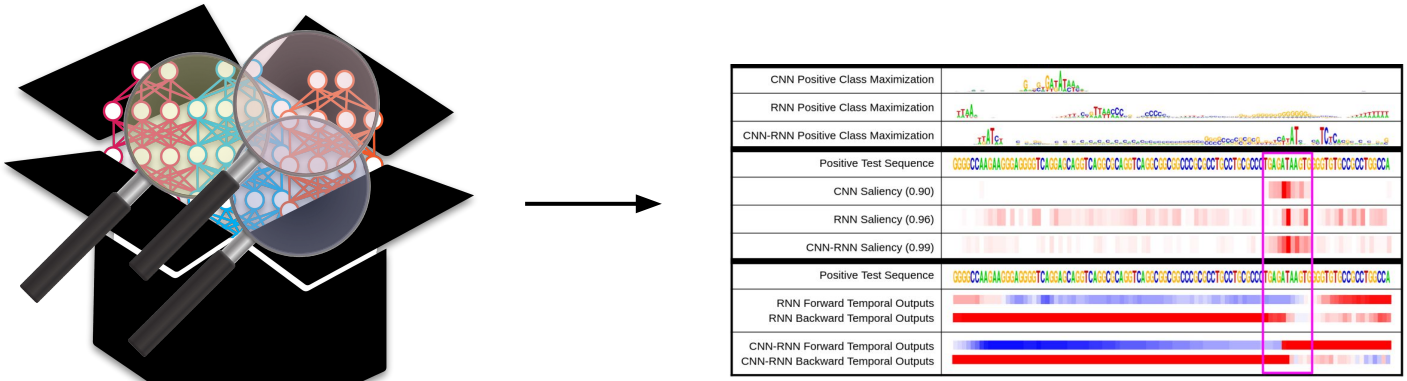
Deep Motif Dashboard

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

1.



2.



Visualization Methods

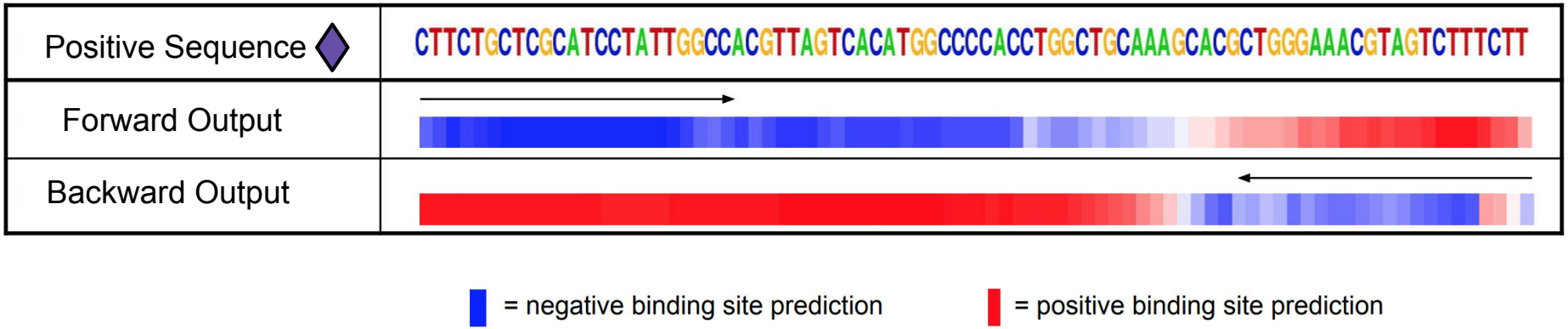
1. Saliency Maps: which nucleotides are most important for classification?

Positive Sequence 	TGCTCGCATCCTATTGGCCACGTTAGTCACATGGCCCCACCTGGCTGCAAAGCACGCTGGGAAACGTAGTCTTTCTT
Saliency Map	

= important nucleotide for prediction

Visualization Methods

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



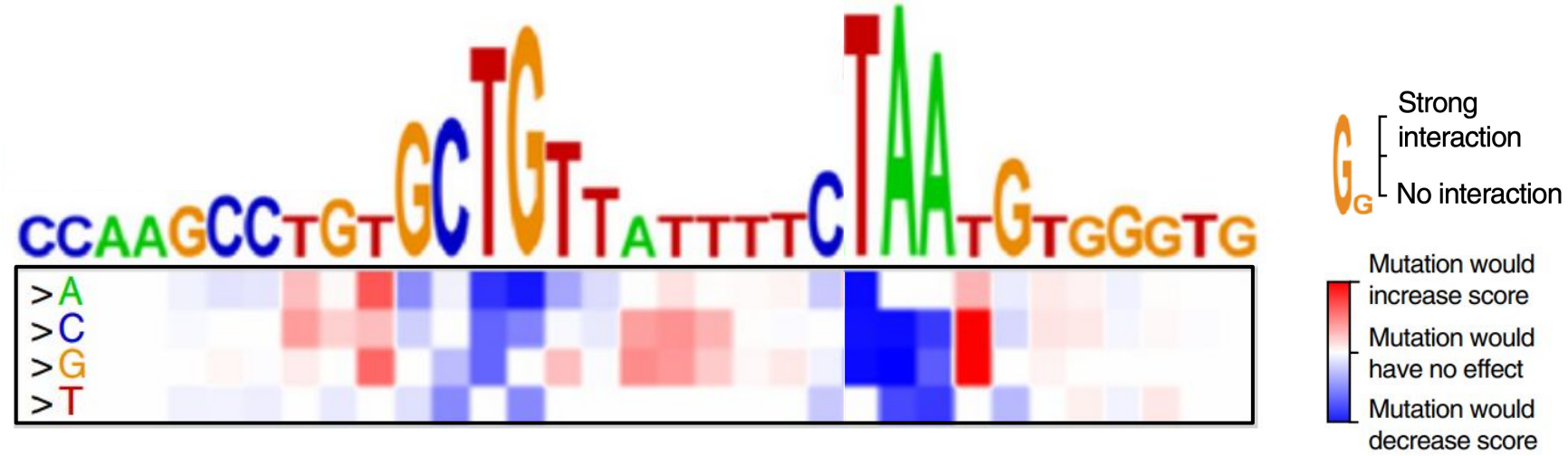
Visualization Methods

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

Optimal binding site for TF “CBX3”	

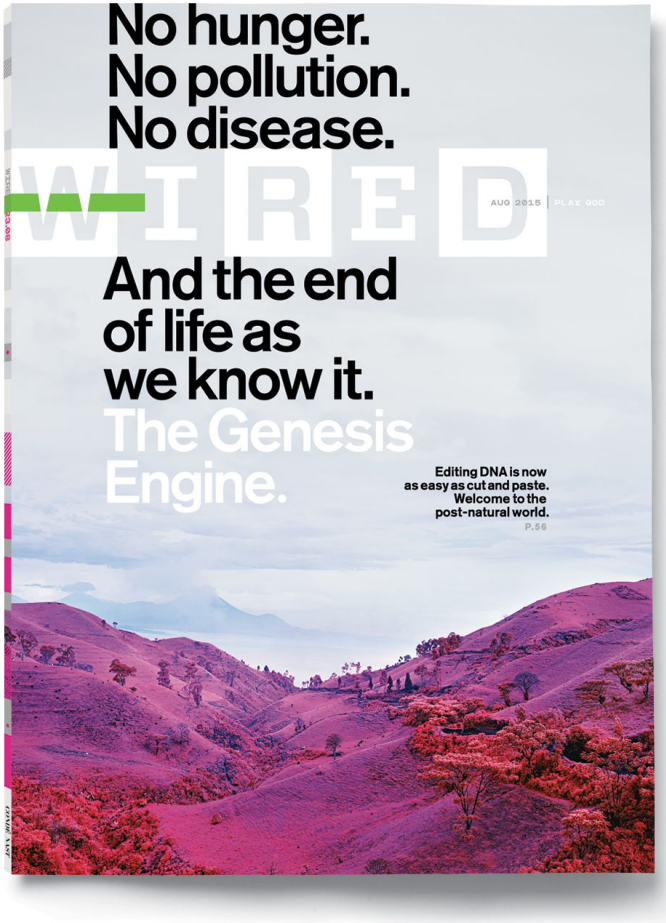
Visualization Methods

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

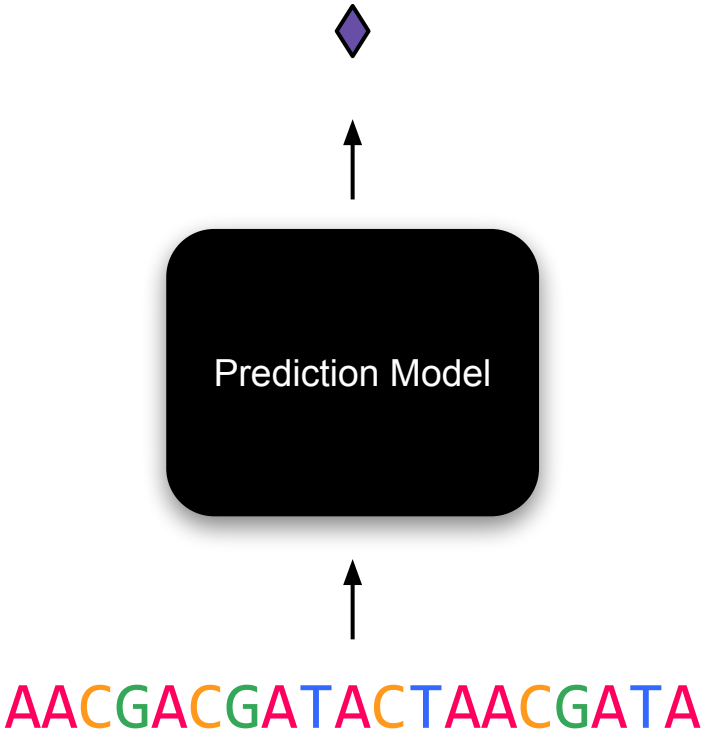


Alipanahi et. al. - Nature Biotech 2015

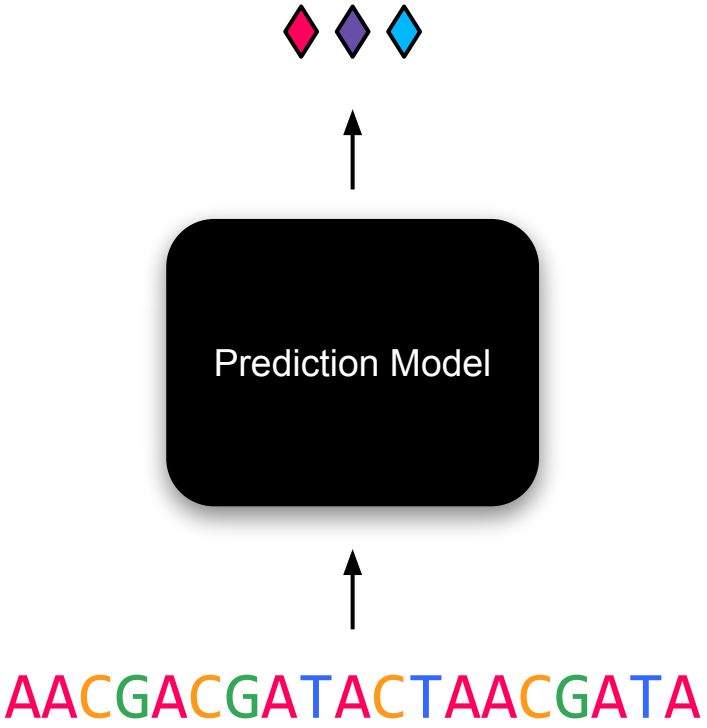
Fixing Mutations with CRISPR-Cas9?



Multi Label Classification for TFBS Prediction

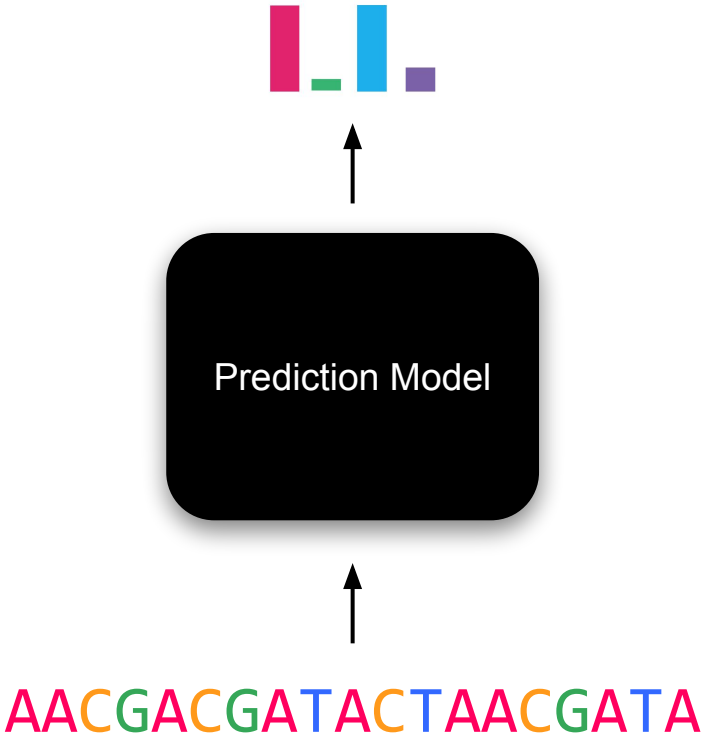


Multi Label Classification for TFBS Prediction



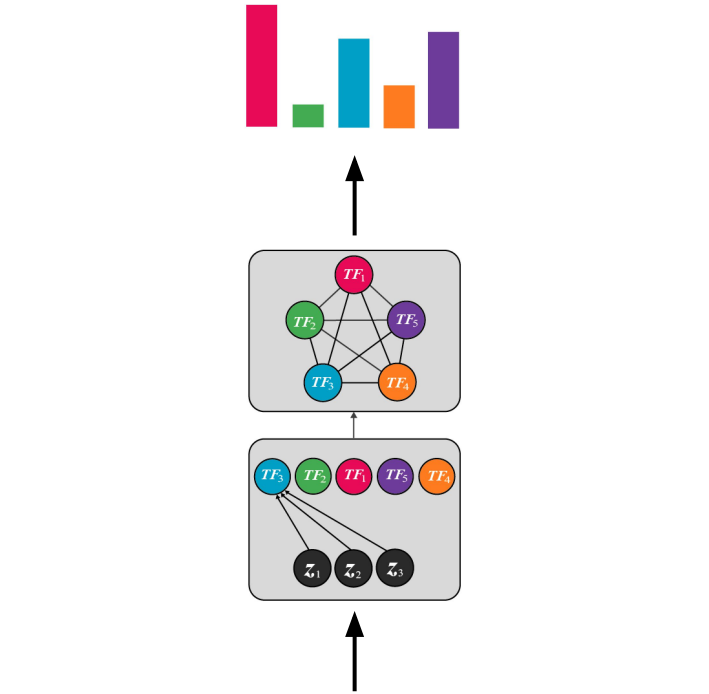
Multi Label Classification for TFBS Prediction

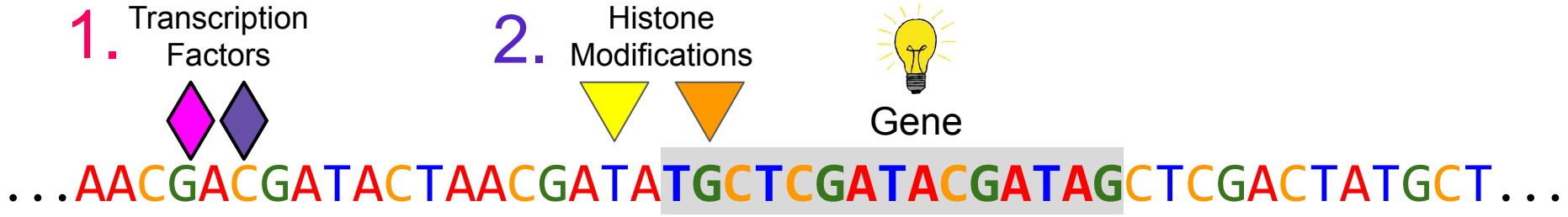
Lanchantin, Singh, & Qi - ICLR Workshops 2017



Neural Message Passing for MLC

Lanchantin, Sekhon, & Qi - Under Review 2019

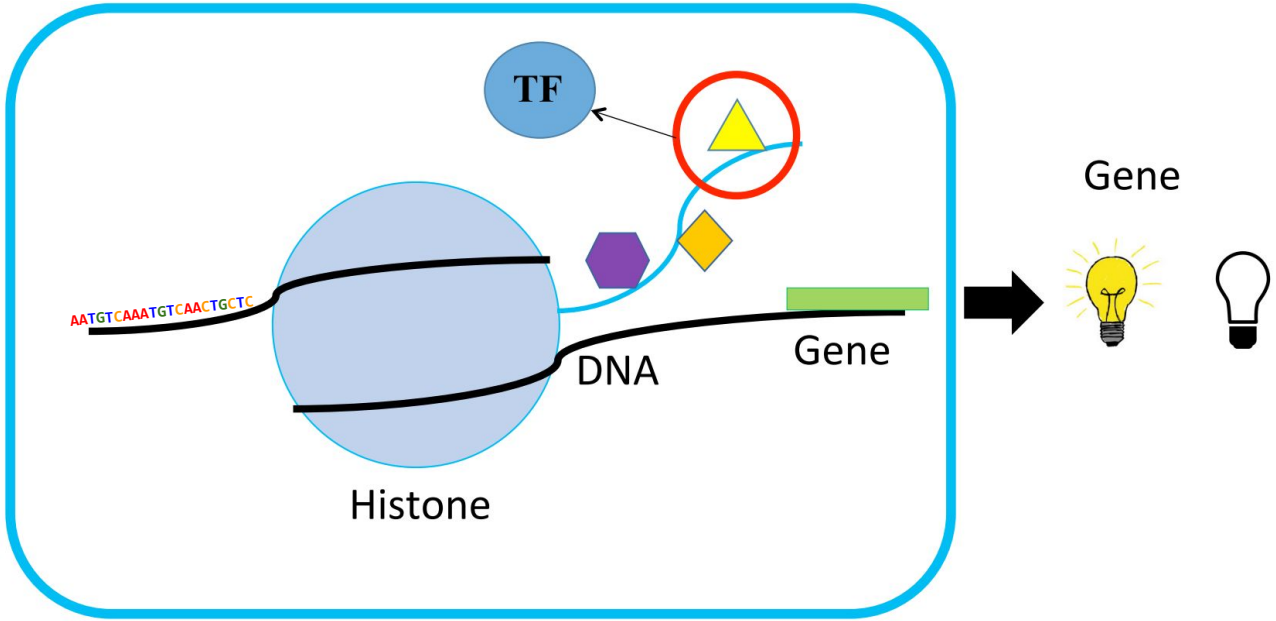



1.
Transcription Factors
2.
Histone Modifications
Gene
...AACGACGATACTAACGATATGCTCGATACGATAGCTCGACTATGCT...

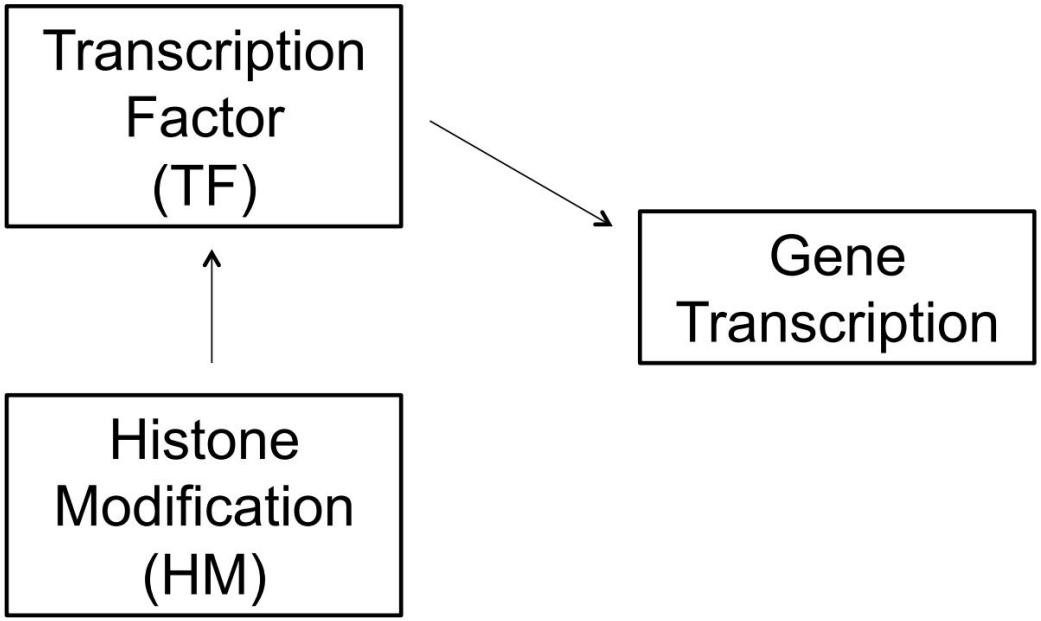
1. Predicting Transcription Factor Binding Sites from DNA

2. Predicting Gene Expression from Histone Modifications

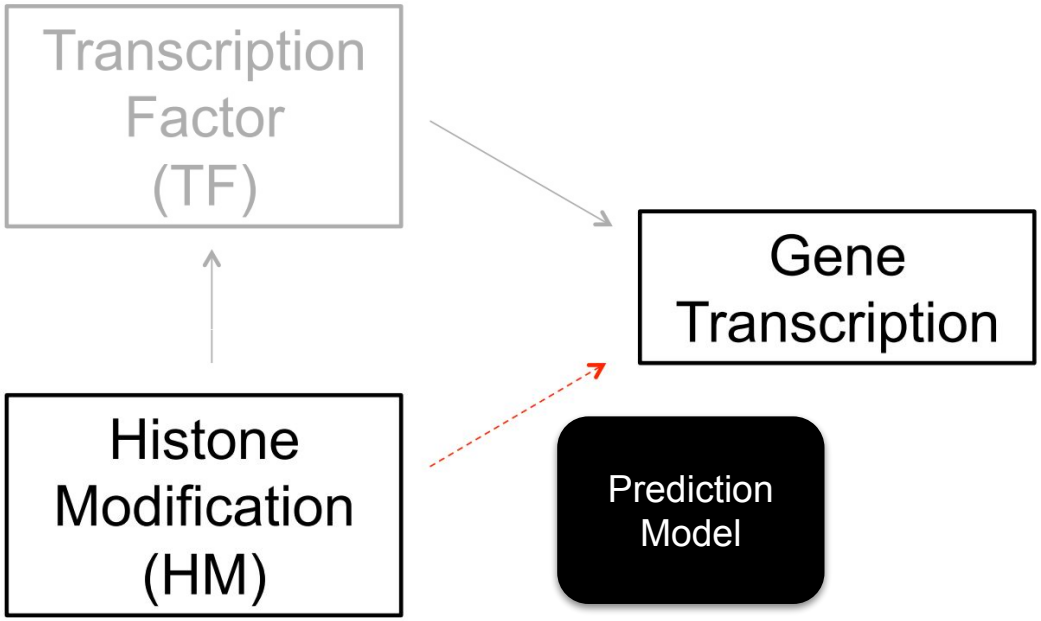
Histone Modifications (HMs)



Histone Modifications

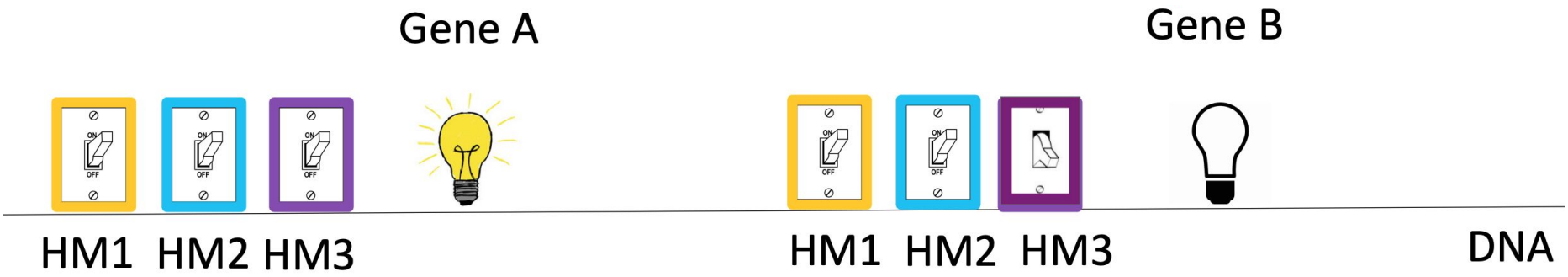


Histone Modifications



Can we predict gene expression from histone modification signals?

What HMs affect which genes in what cells?



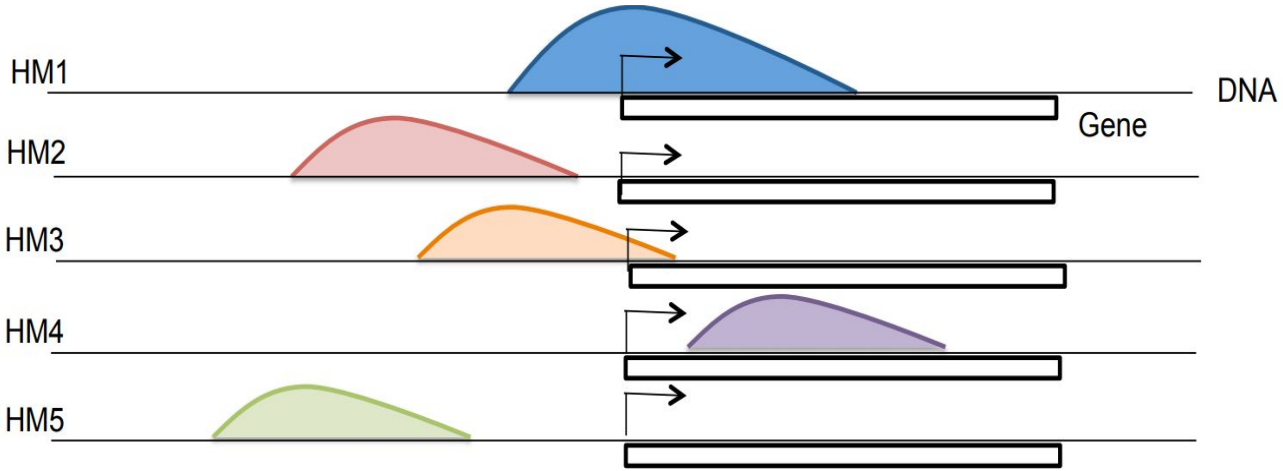
Gene Transcription Prediction Task

Why Study HM → Gene Expression?

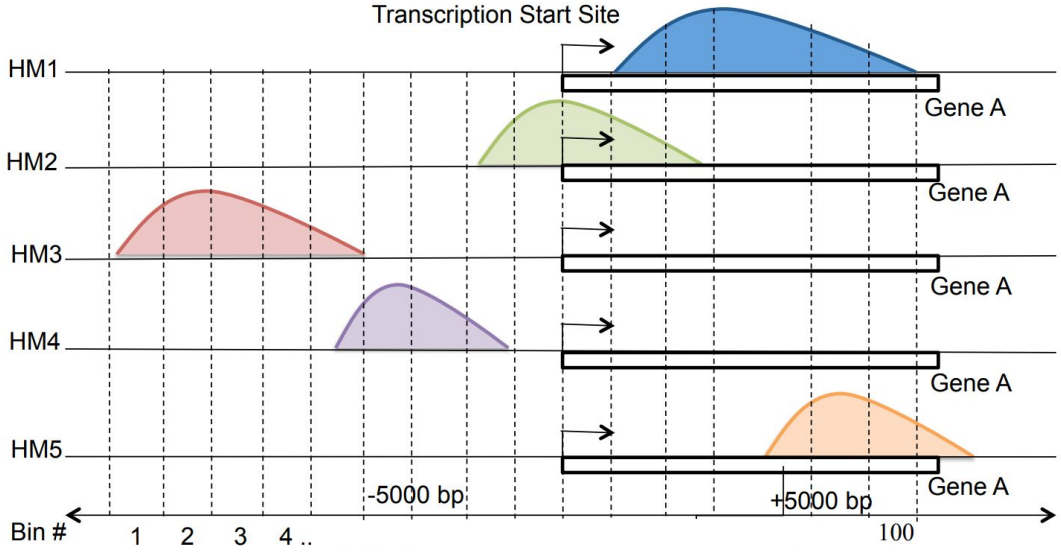
- **Epigenomics:** study of chemical changes in DNA and histones (without altering DNA sequence)
- **Epigenome is dynamic:** can be altered by environmental conditions.

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

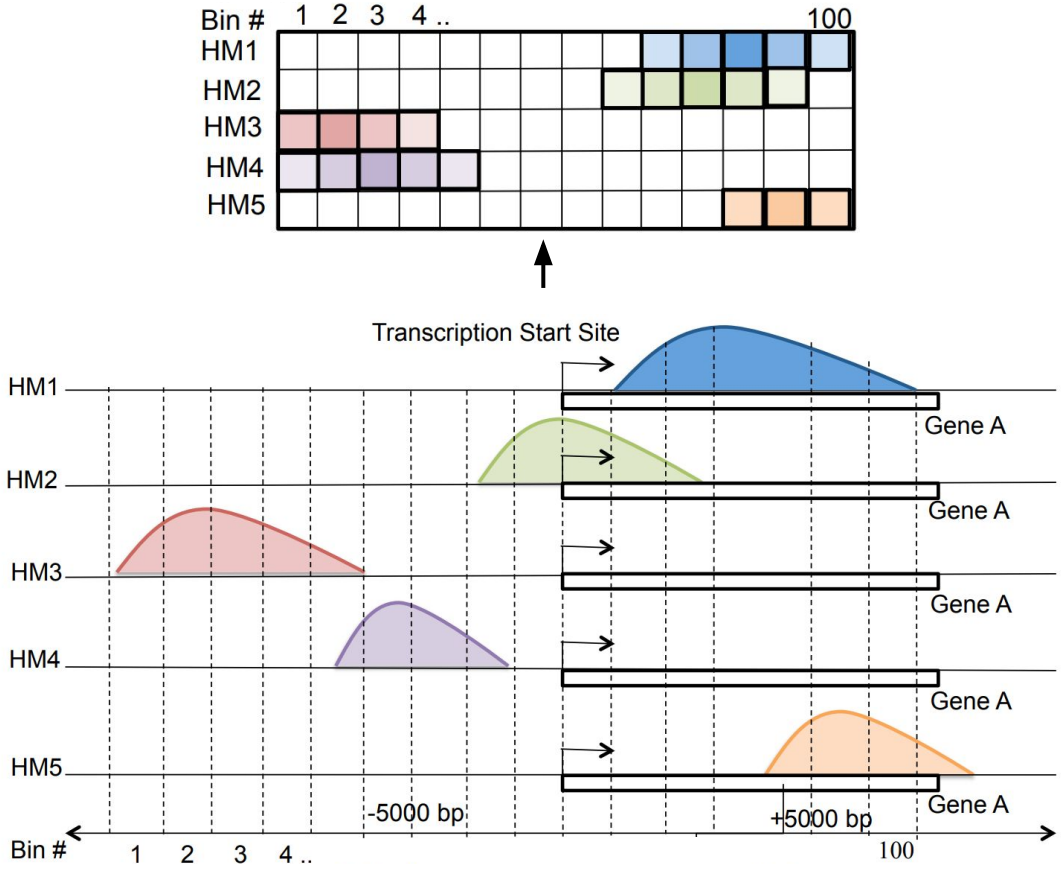
Histone Modification Input Data



Histone Modification Input Data

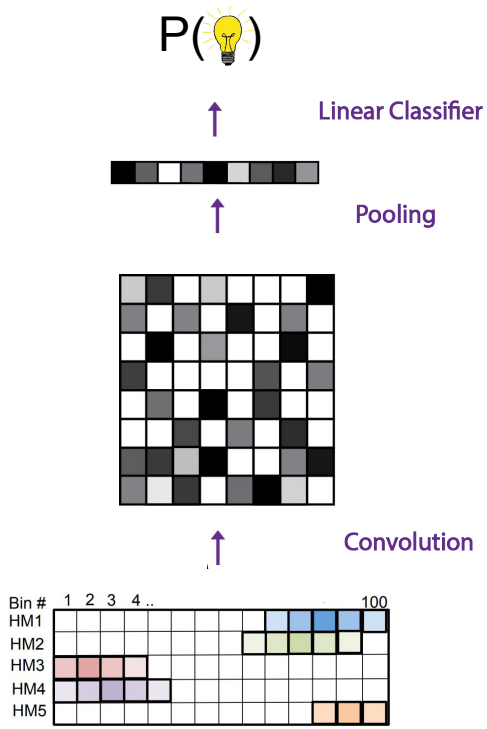


Histone Modification Input Data



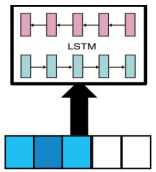
DeepChrome

Singh, Lanchantin, Robins & Qi - Bioinformatics 2016

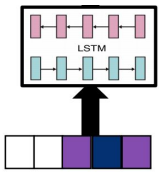


Attentive Chrome

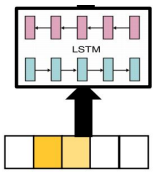
Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



HM1



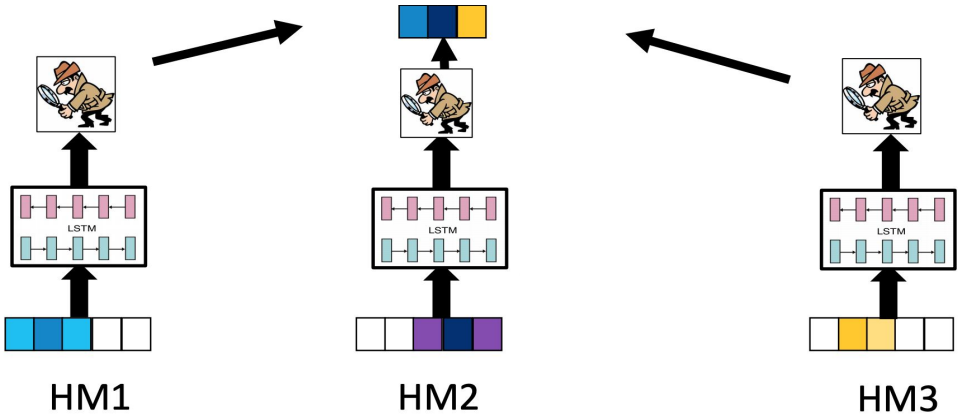
HM2



HM3

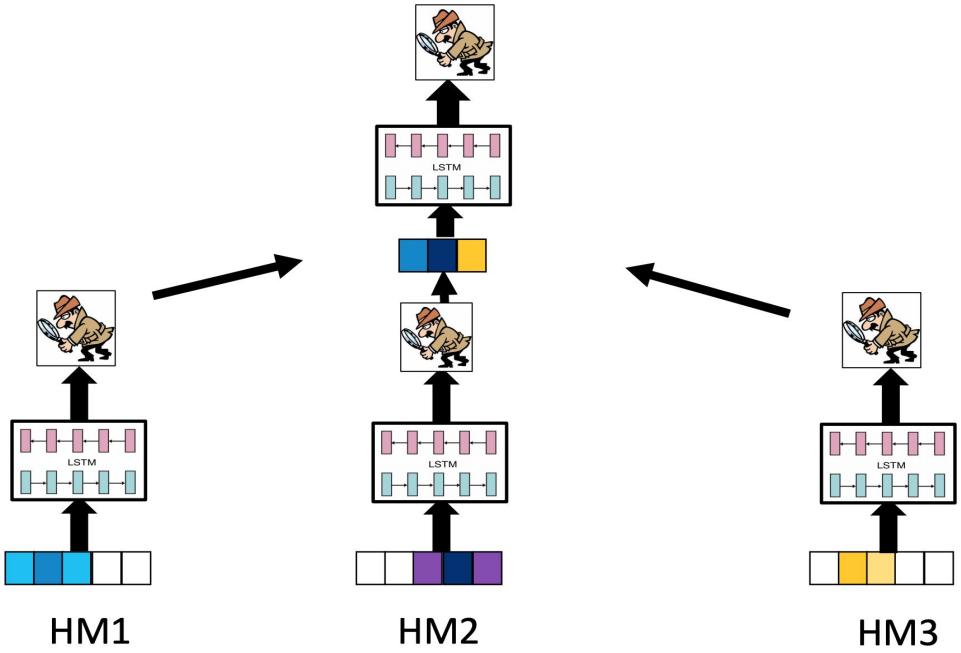
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



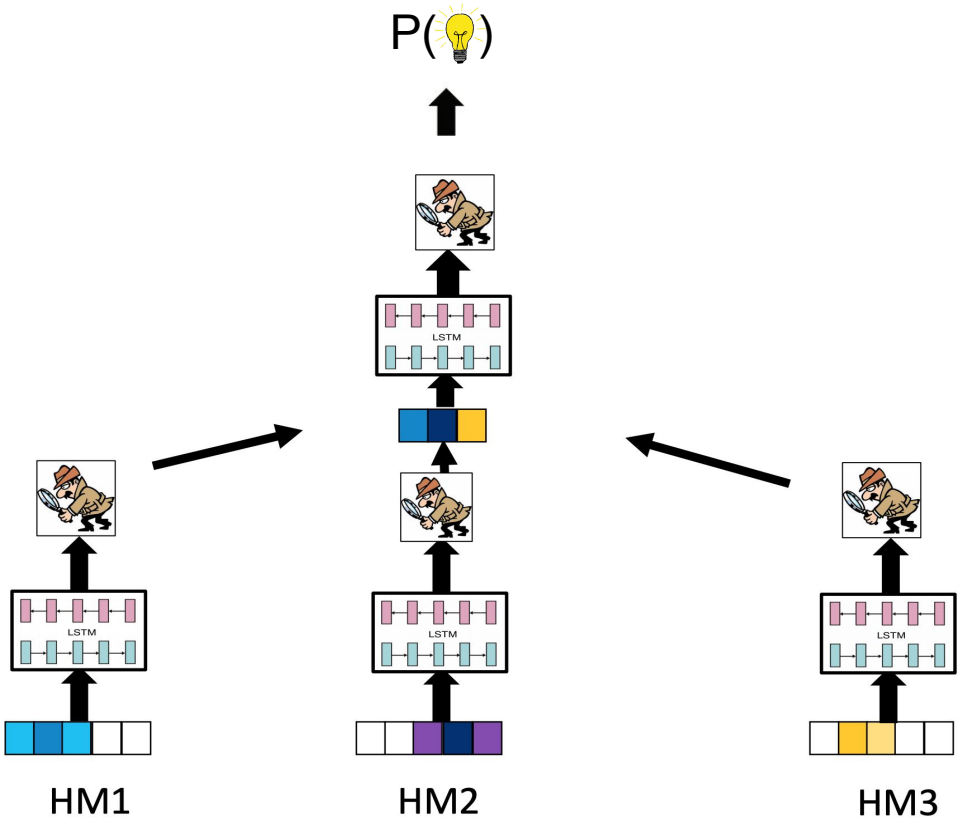
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



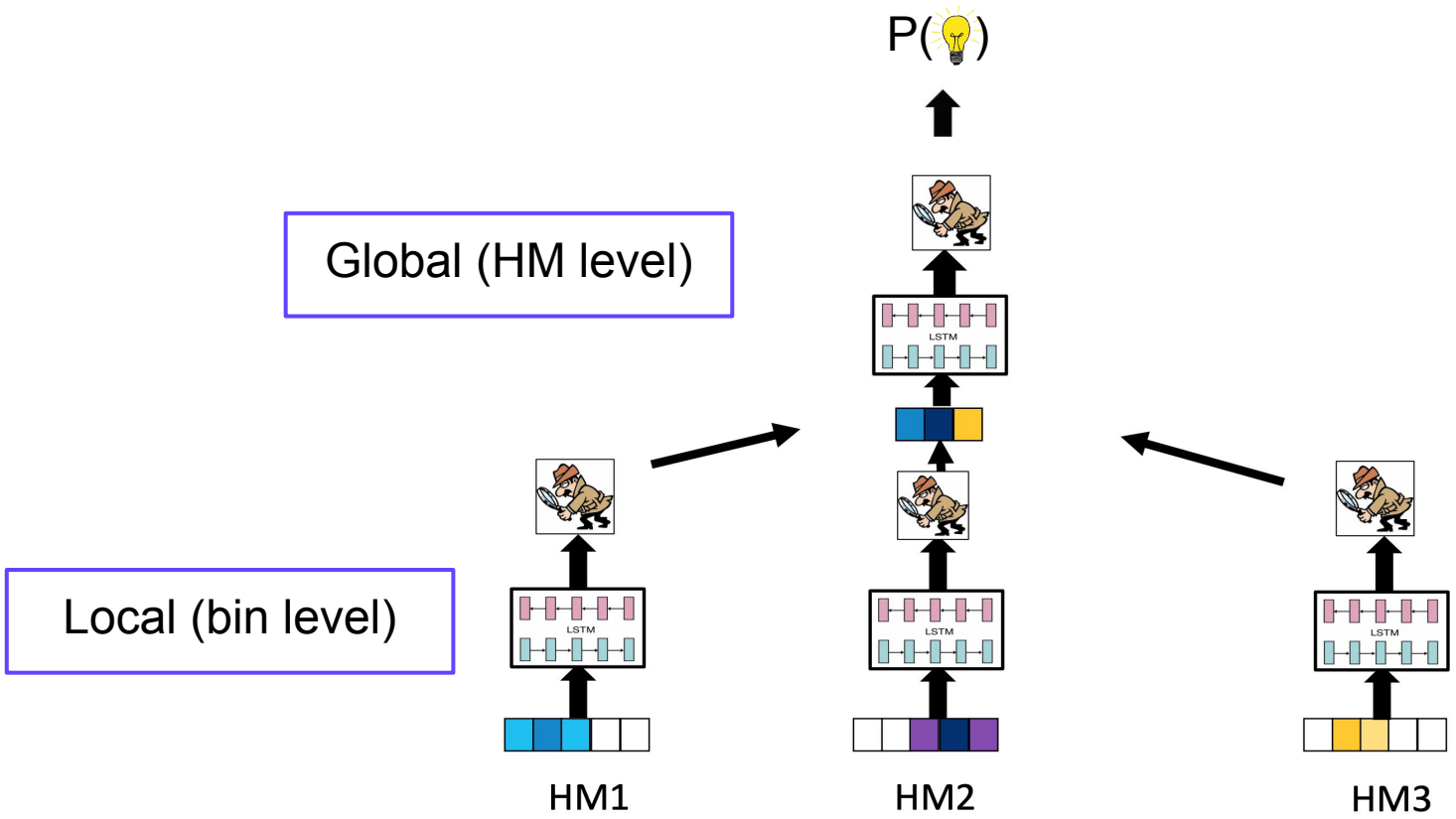
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



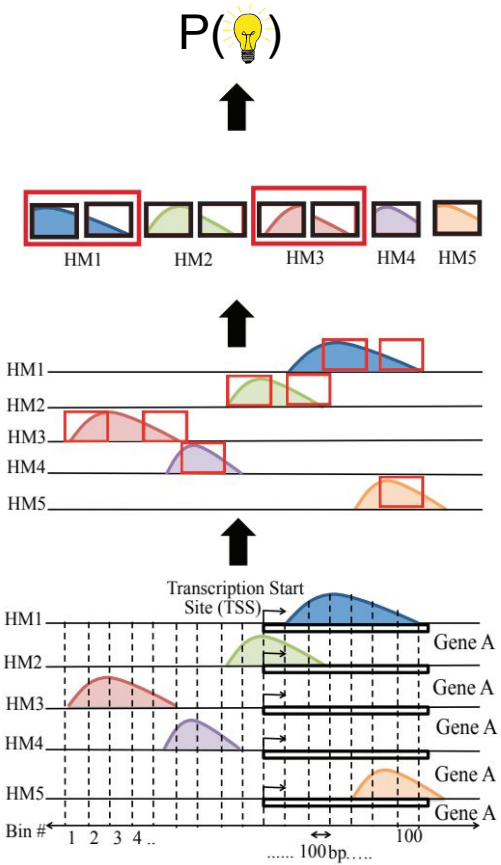
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



Attentive Chrome

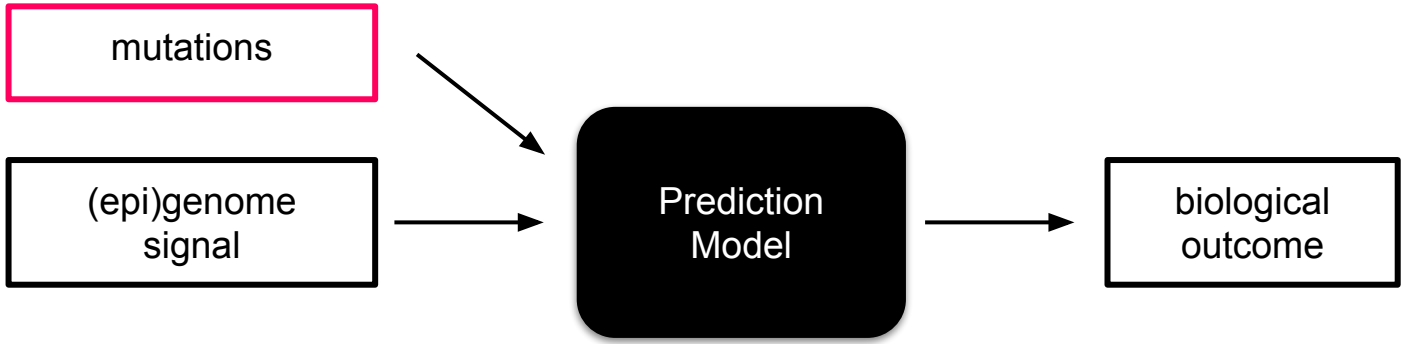
Singh, Lanchantin, Sekhon, & Qi - NIPS 2017



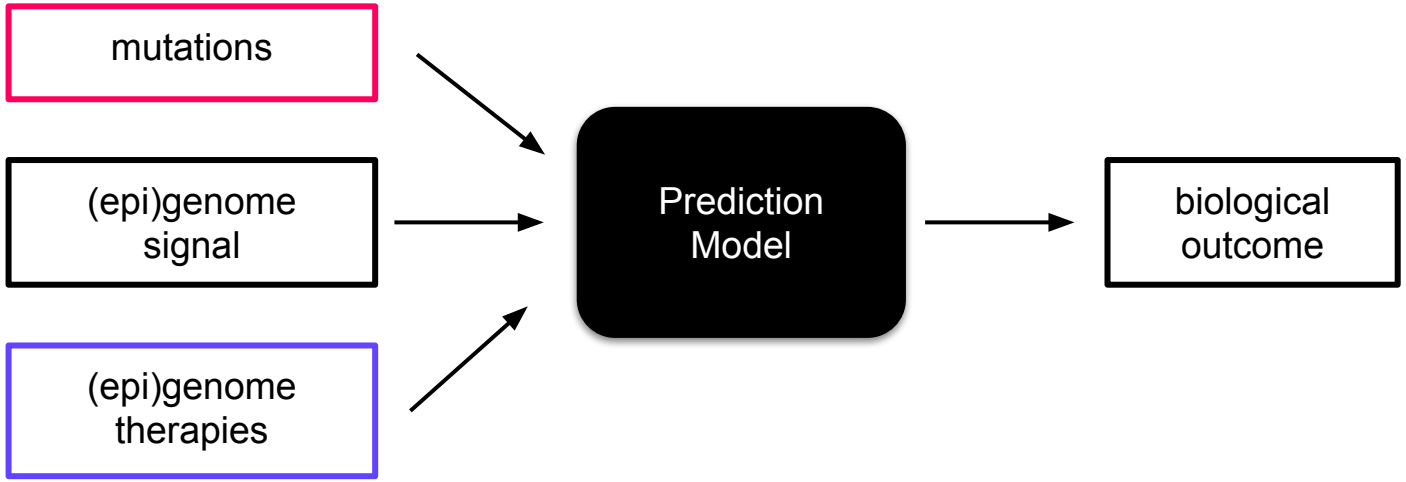
Summarizing Machine Learning for Genomics

Summarizing Machine Learning for Genomics

Summarizing Machine Learning for Genomics

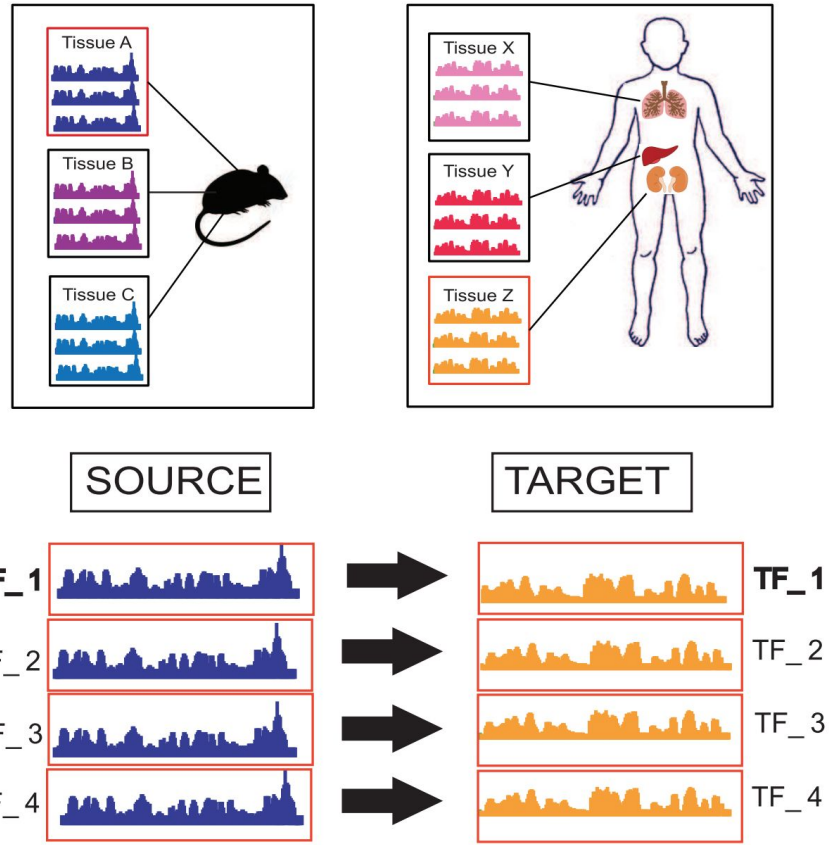


Summarizing Machine Learning for Genomics



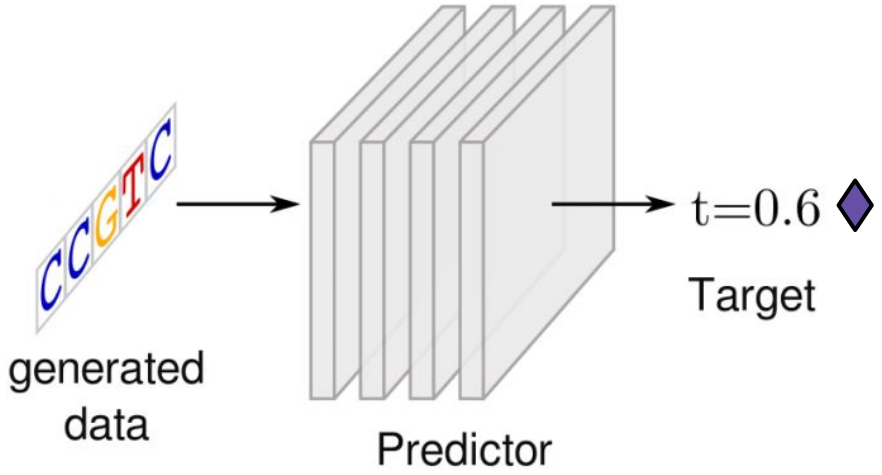
Open Problems

Transfer Learning



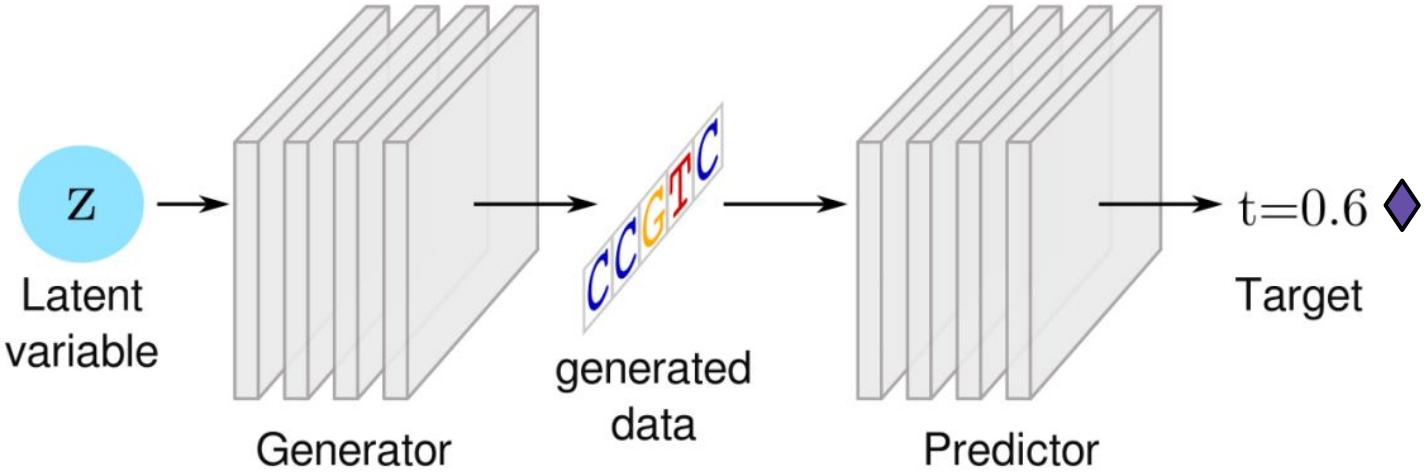
Singh, Lanchantin, Robins, & Qi - TCBB 2016

Generative Models



Optimal binding site for TF “CBX3”	

Generative Models



“Generating and designing DNA with deep generative models” Killoran et. al. 2017

Thank You!

code available at: **github.com/qdata**

Ritambhara Singh

Arshdeep Sekhon

Beilun Wang

Dr. Yanjun Qi

UNIVERSITY of VIRGINIA