http://deepchrome.net

DeepChrome:

Interpretable Deep Learning for Sequential Data Analysis in Biomedicine

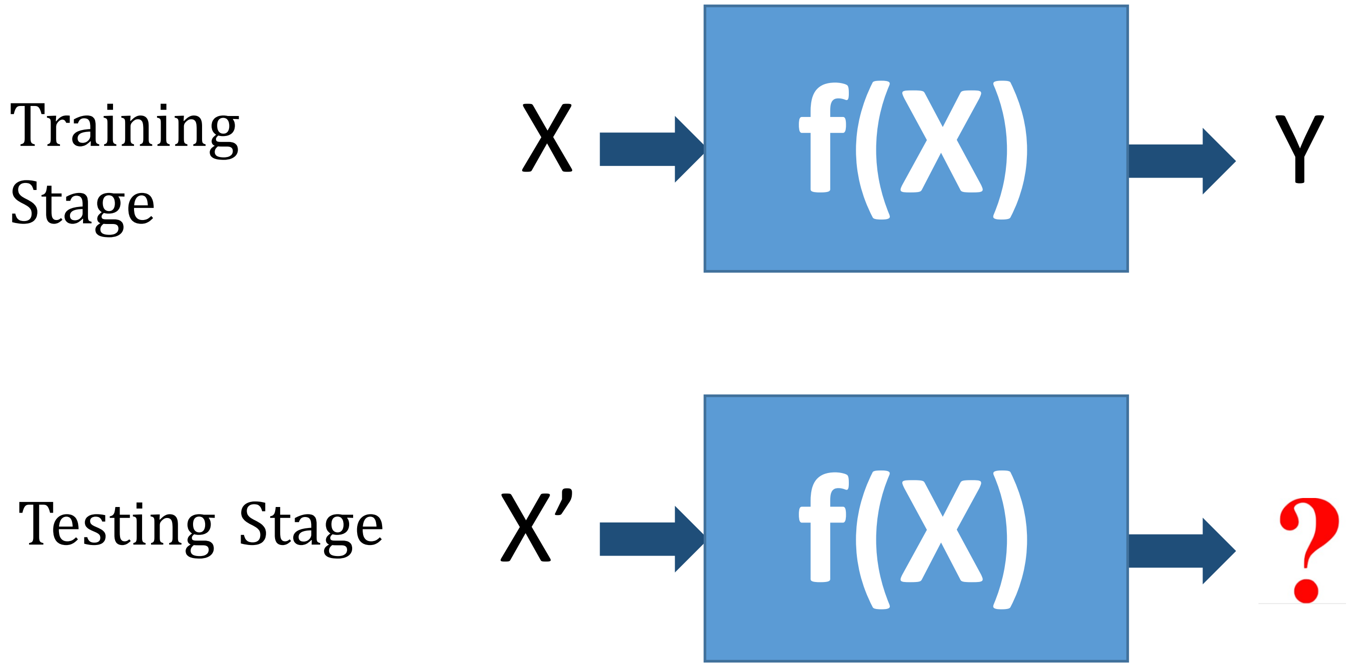
Dr. Yanjun Qi

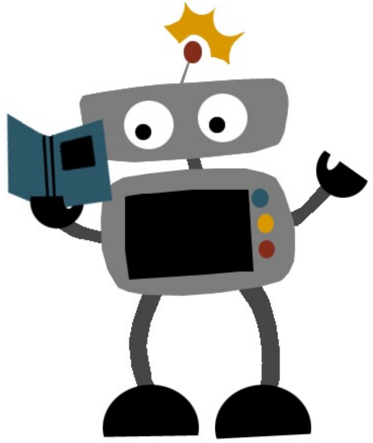
DATA Scholar 2021 @ NIA

Associate Professor, Department of Computer Science @

University of Virginia

Basics of Machine Learning



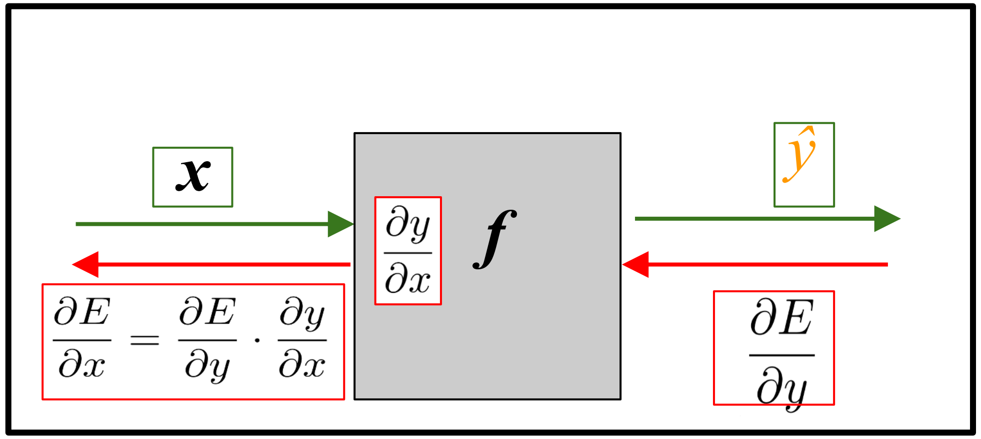


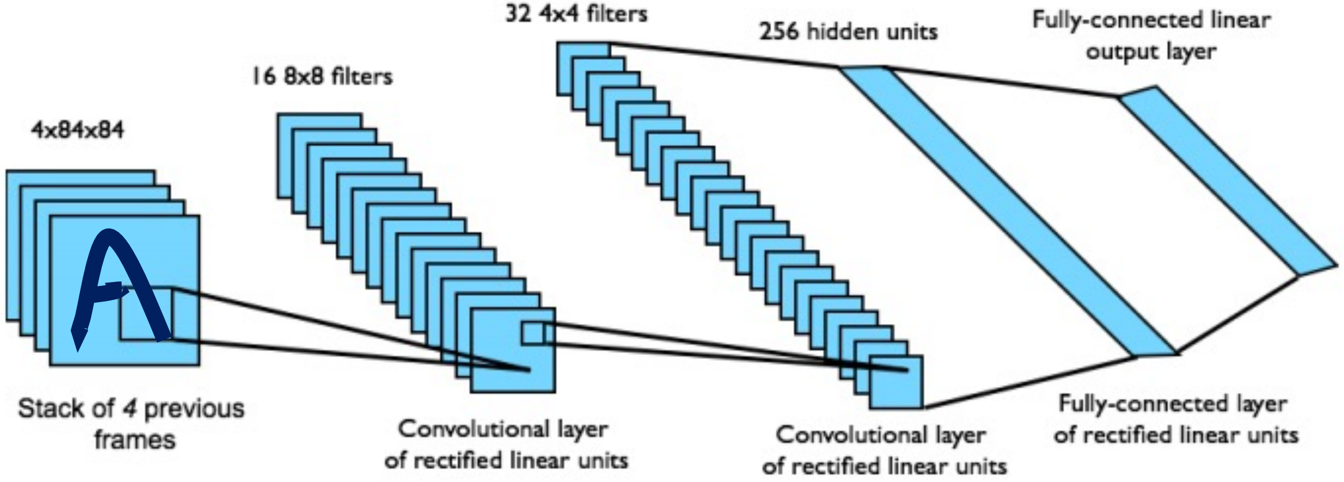
Learn f(x) to Generalize to Unseen X'

Supervised Learning

Generalisation: learn model f(x) from past data in order to "explain", "predict", "model" or "control" new data examples

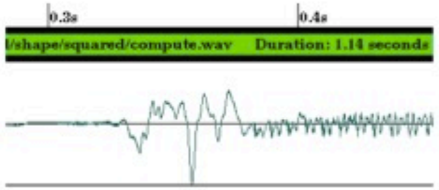
Building Deep Neural Nets


$\boldsymbol{x}$
$\hat{y}$
$\boldsymbol{f}$
$\frac{\partial y}{\partial x}$
$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial y} \cdot \frac{\partial y}{\partial x}$
$\frac{\partial E}{\partial y}$

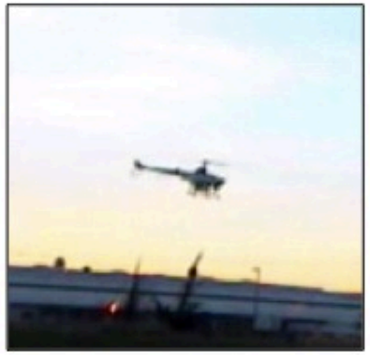

4x84x84
16 8x8 filters
32 4x4 filters
256 hidden units
Fully-connected linear output layer
Stack of 4 previous frames
Convolutional layer of rectified linear units
Convolutional layer of rectified linear units
Fully-connected layer of rectified linear units

Deep Learning is Changing the World

How may I help you, human?

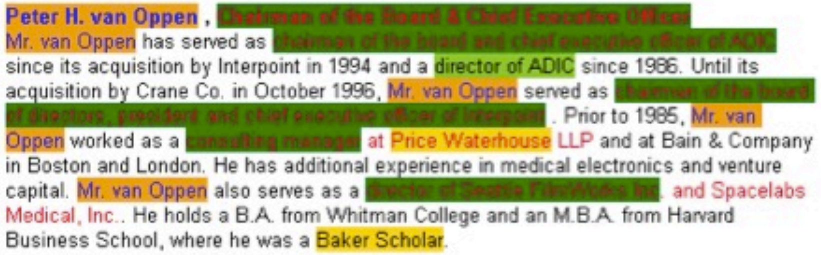


Speech Recognition



Control learning

Text analysis

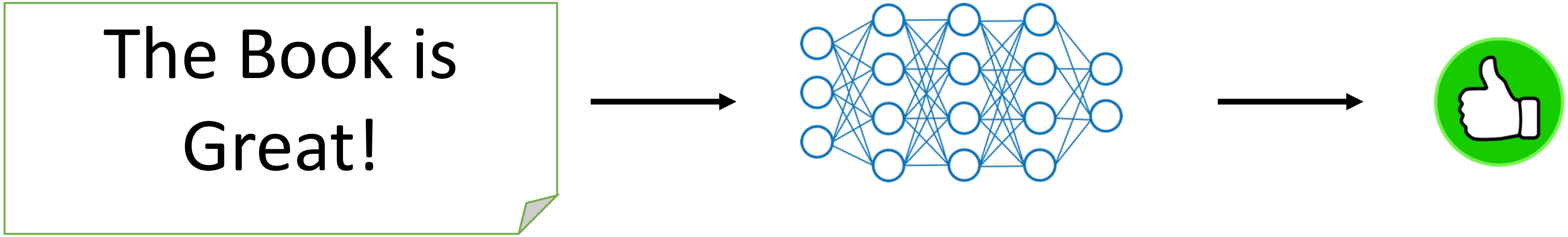


Peter H. van Oppen , Chairman of the Board & Chief Executive Officer
Mr. van Oppen has served as chairman of the board and chief executive officer of ADIC since its acquisition by Interpoint in 1994 and a director of ADIC since 1986. Until its acquisition by Crane Co. in October 1996, Mr. van Oppen served as chairman of the board of directors, president and chief executive officer of Interpoint . Prior to 1985, Mr. van Oppen worked as a consulting manager at Price Waterhouse LLP and at Bain & Company in Boston and London. He has additional experience in medical electronics and venture capital. Mr. van Oppen also serves as a director of Seattle FilmWorks Inc. and Spacelabs Medical, Inc.. He holds a B.A. from Whitman College and an M.B.A. from Harvard Business School, where he was a Baker Scholar.

Object recognition

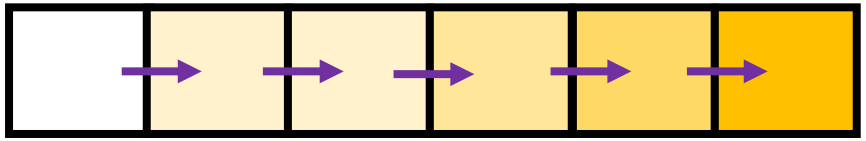
Many more !

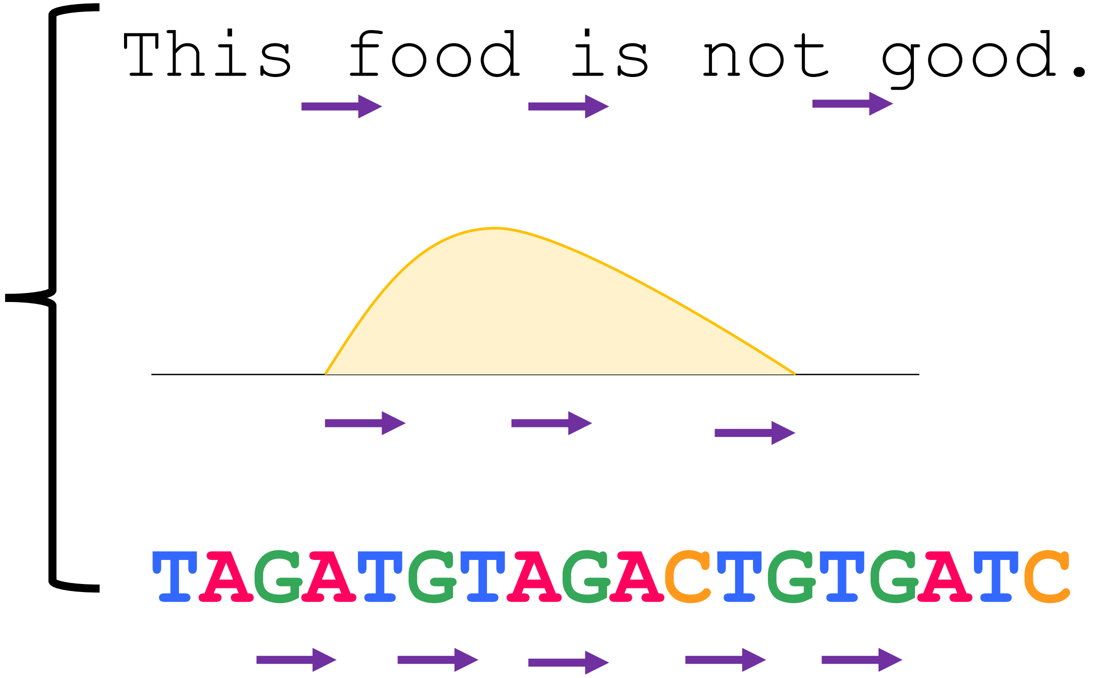
Deep Learning Excellence on Sequential Data



The Book is Great!

Sequential Data



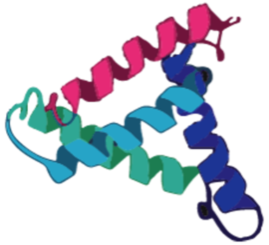


Strings, signals etc.

This food is not good.

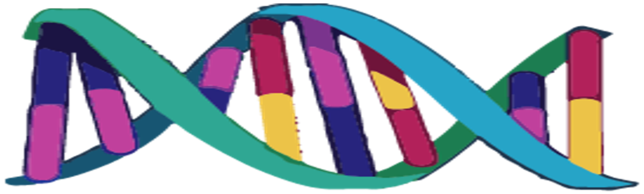
TAGATGTAGACTGTGATC

PROTEIN



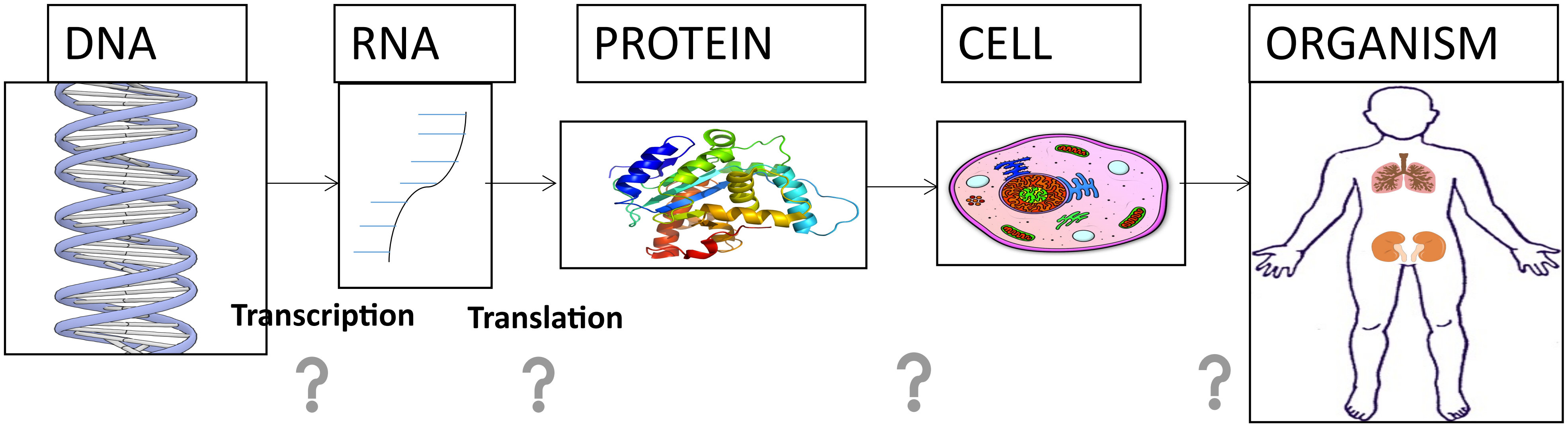
RNA

DNA



PROTEIN	TGKHQFTVKE
RNA	UAGACUGUAGACUGUGAC
DNA	TAGATGTAGACTGTGATC

Biological Modules



DNA

RNA

PROTEIN

CELL

ORGANISM

Transcription

Translation

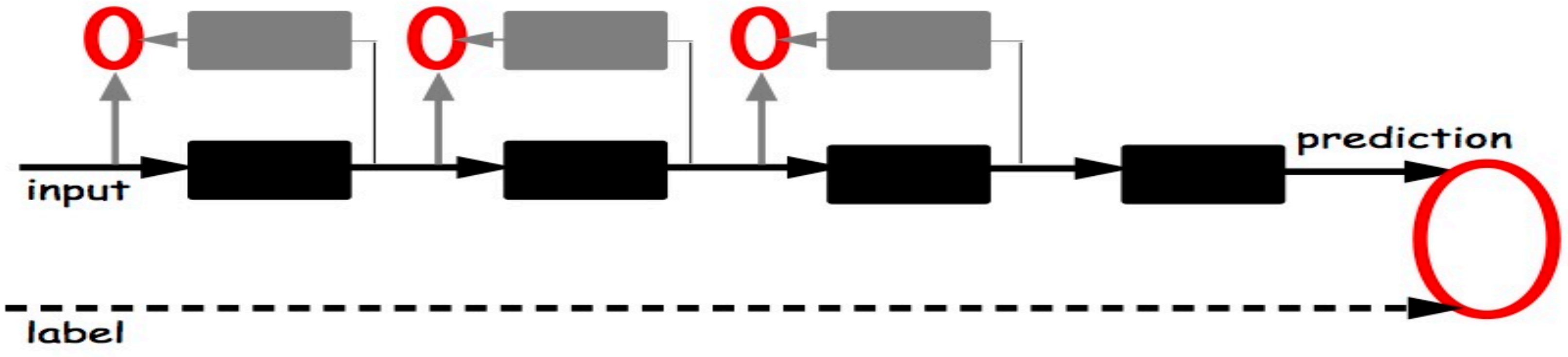
? ? ? ?

CATGACTG
CATGCCTG

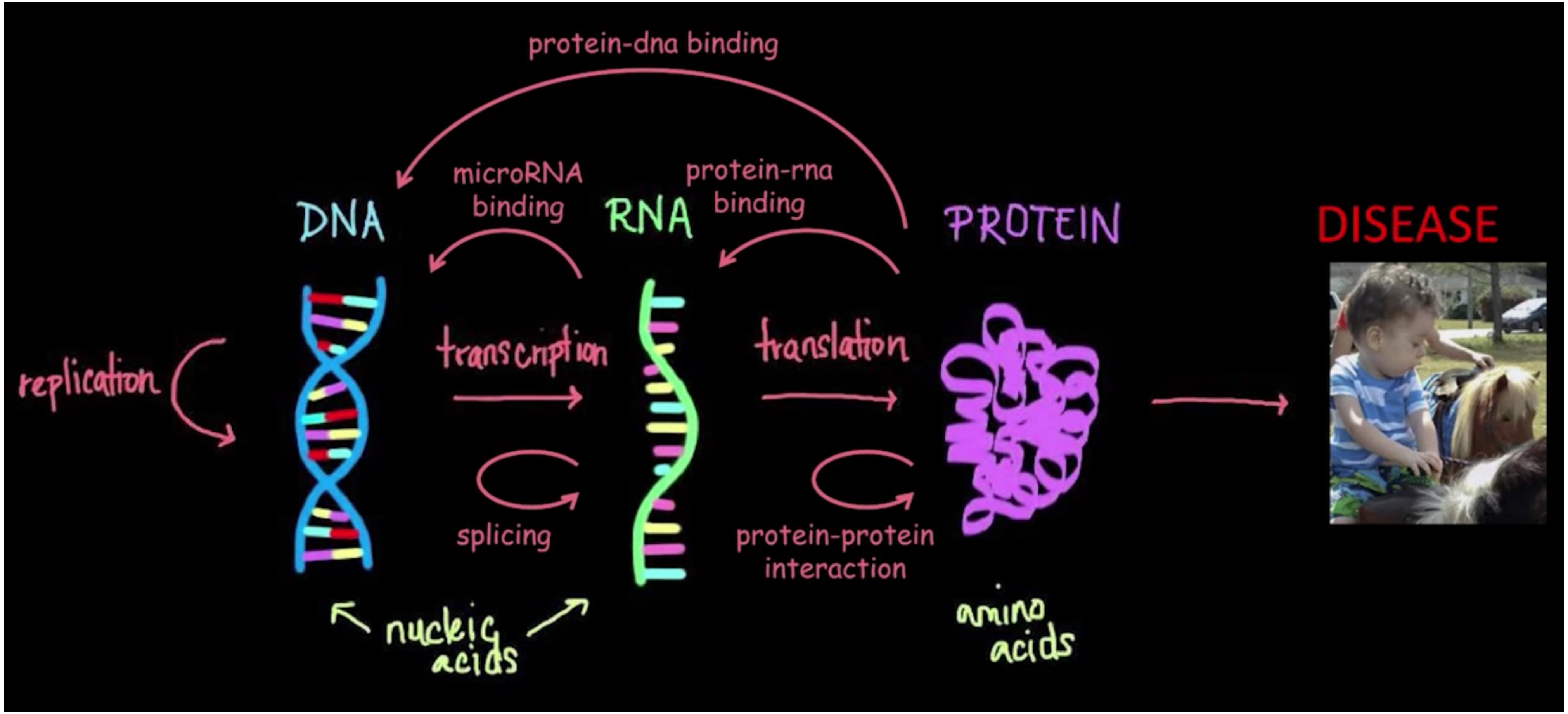
Genetic Variant

Disease

Deep Learning Modules (composable)



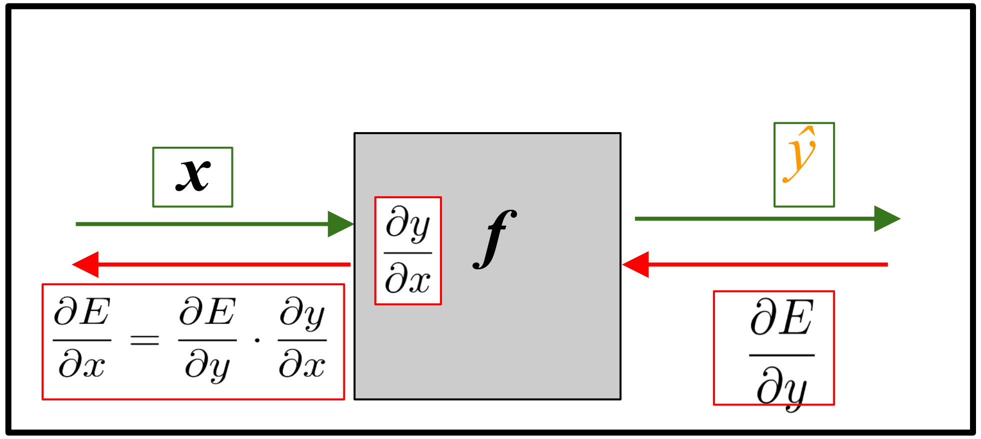
Biology is super complex

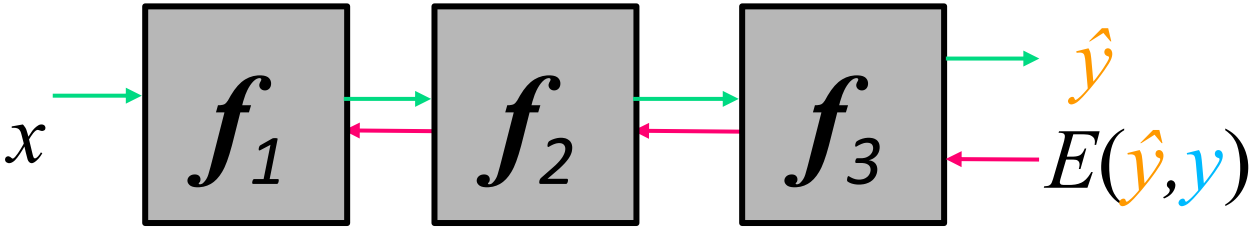


alternative splicing, reverse transcriptase, introns, junk DNA, epigenetics, RNA viruses, trans-splicing, transposons, prions, epigenetics, gene rearrangements and many more ……

Image Credit: Brendan Frey

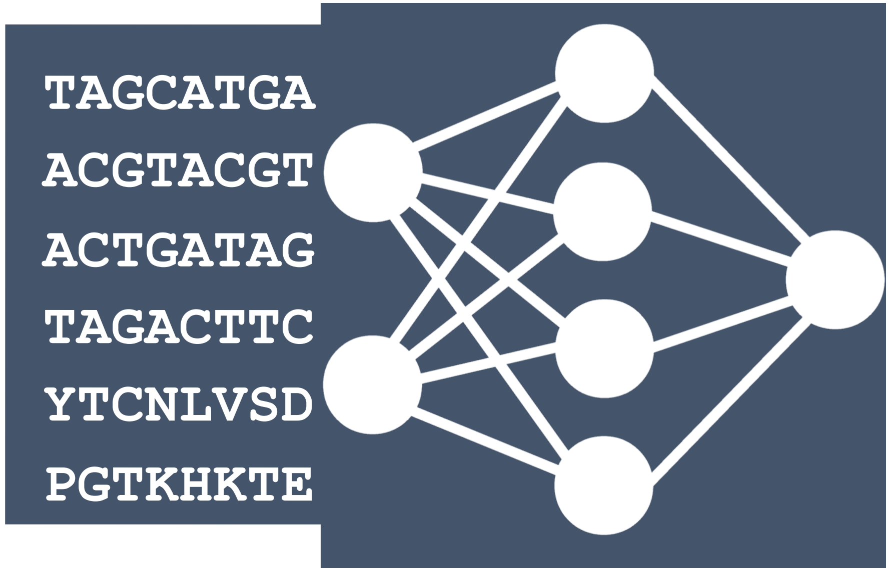
Building Deep Neural Nets


$\boldsymbol{x}$
$\hat{y}$
$\frac{\partial y}{\partial x}$
$\boldsymbol{f}$
$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial y} \cdot \frac{\partial y}{\partial x}$
$\frac{\partial E}{\partial y}$



http://cs231n.stanford.edu/slides/winter1516_lecture5.pdf

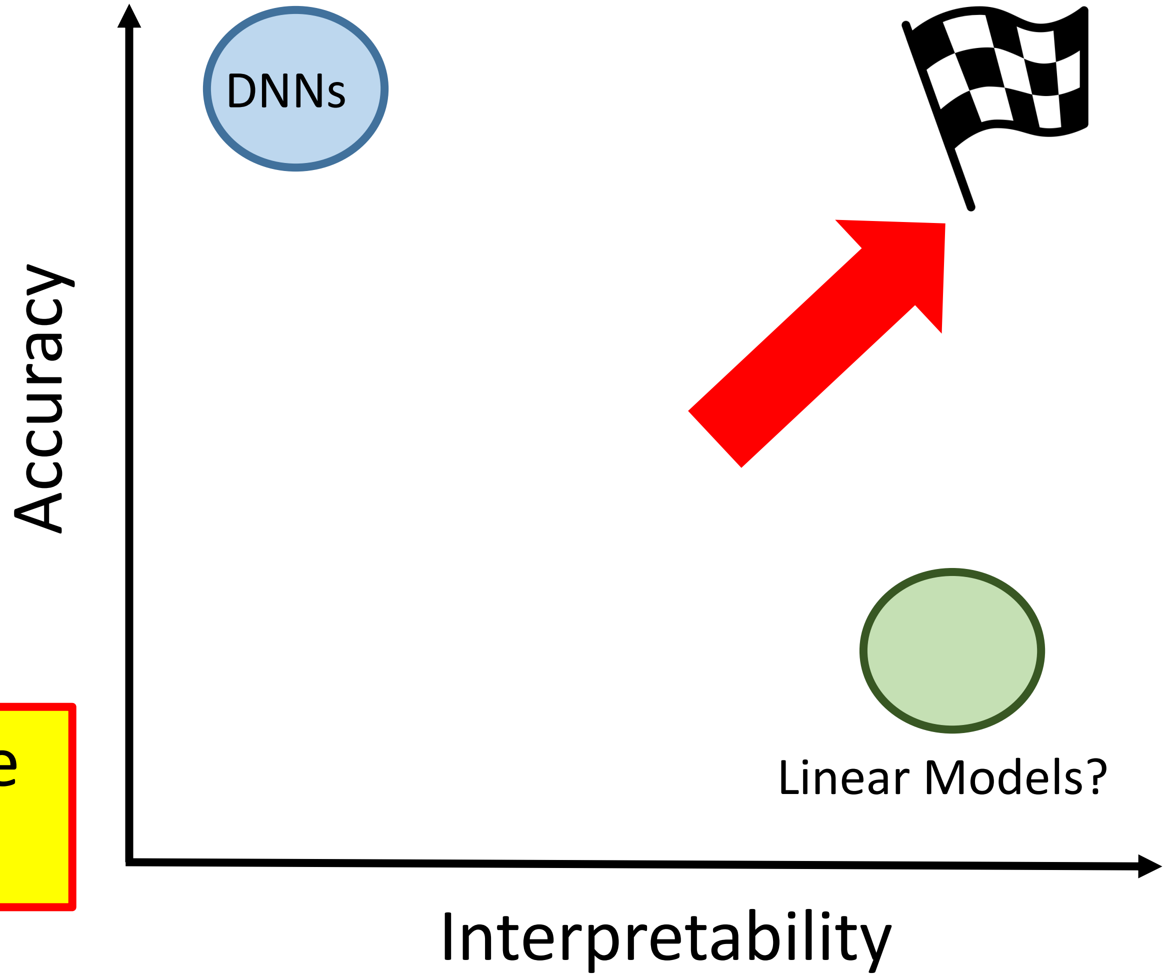
This Talk: Using Deep Representation Learning to Read and Understand the Human Genome and Proteome



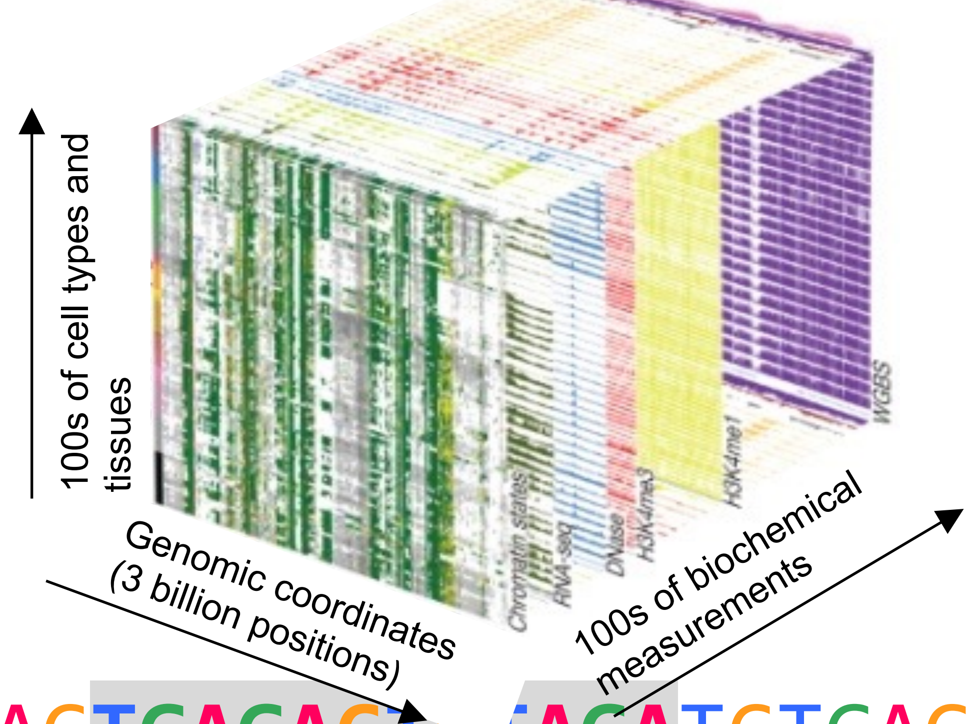
1. Predict

2. Interpret

Our Goal: Interpretable Deep Learning Models



Challenge : DNNs are hard to Interpret



100s of cell types and tissues

Genomic coordinates (3 billion positions)

100s of biochemical measurements

ATGCTCGATACTGAGACTACTGAGAC**TGAGACTCTAGA**TCTGACTACTCACG

Gene Expressed

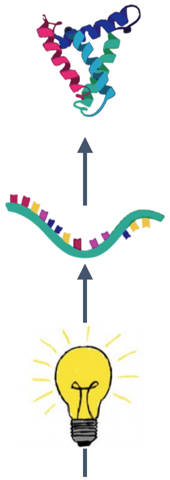
ATGCTCGATACTGAGACTACTGAGAC**TGAGACTCTAGA**TCTGACTACTCACG

Gene Expressed

ATGCTCGATACTGAGACTACTGAGAC**TGAGACTCTAGA**TCTGACTACTCACG

what causes a gene to be expressed?

To understand gene regulation



gene expressed

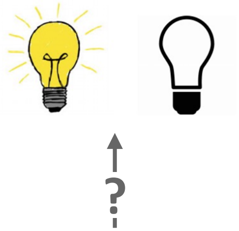
ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

To understand gene regulation

gene repressed

ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

What controls Gene Regulation? How?



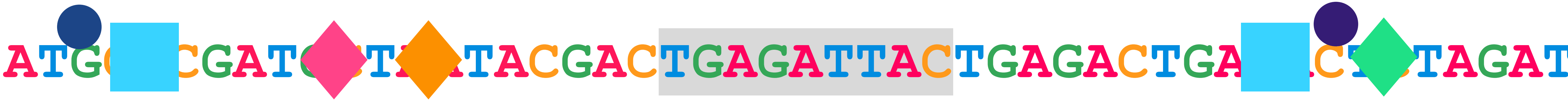
ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

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

-Eric Lander, Principal Leader of the Human Genome Project

Credit: Brendan Frey

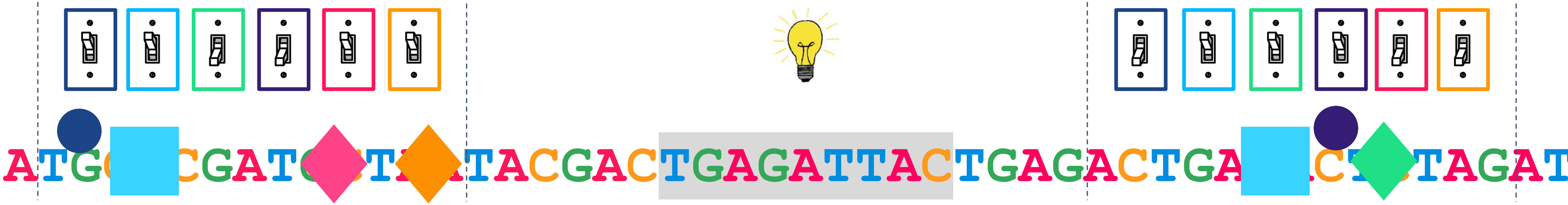
Chromatin Profile



Chromatin Profile Attributes

◆◆◆ Transcription Factors | ●● Histone Modifications | ■ DNA Accessibility

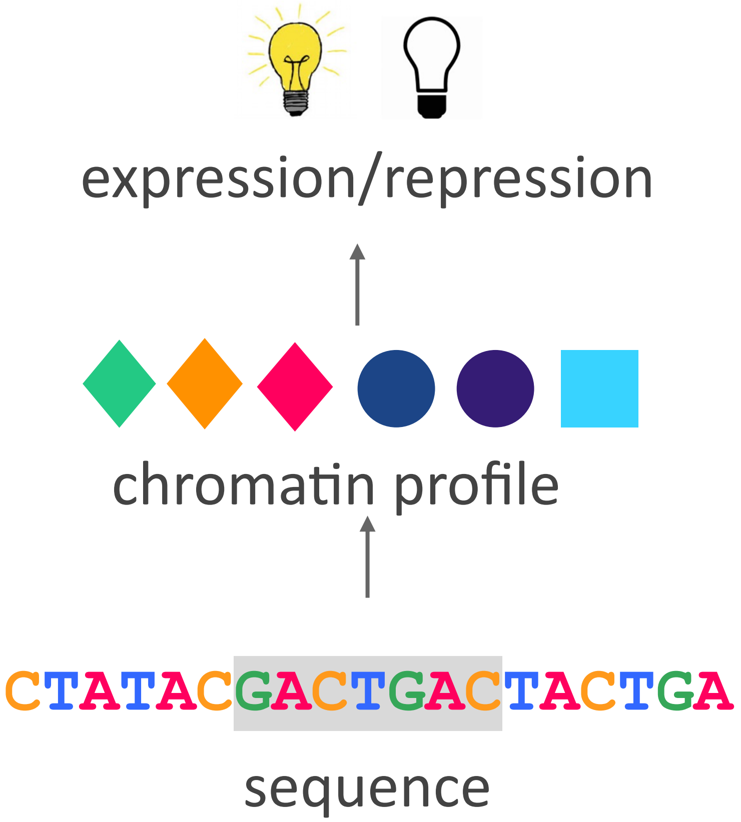
Chromatin Profile



Chromatin Profile Attributes

◆ ◆ ◆ Transcription Factors | ● ● Histone Modifications | ■ DNA Accessibility

Gene Regulation



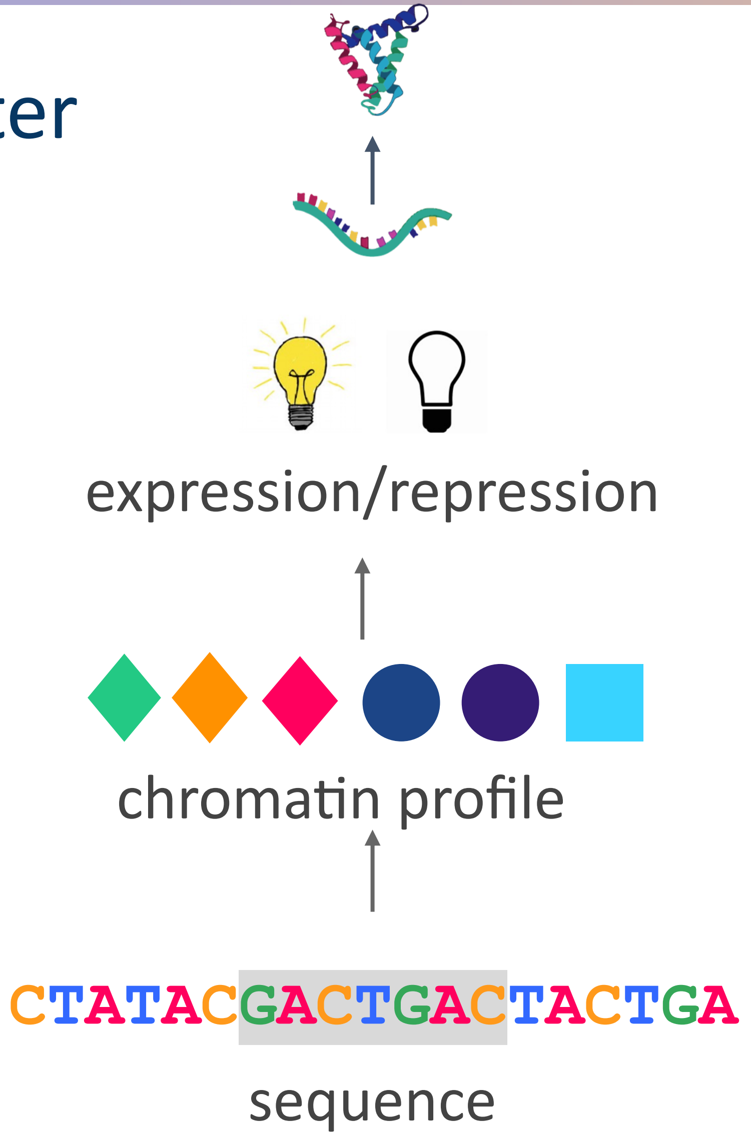
expression/repression

chromatin profile

CTATACGACTGACTACTGA

sequence

Gene Regulation and after



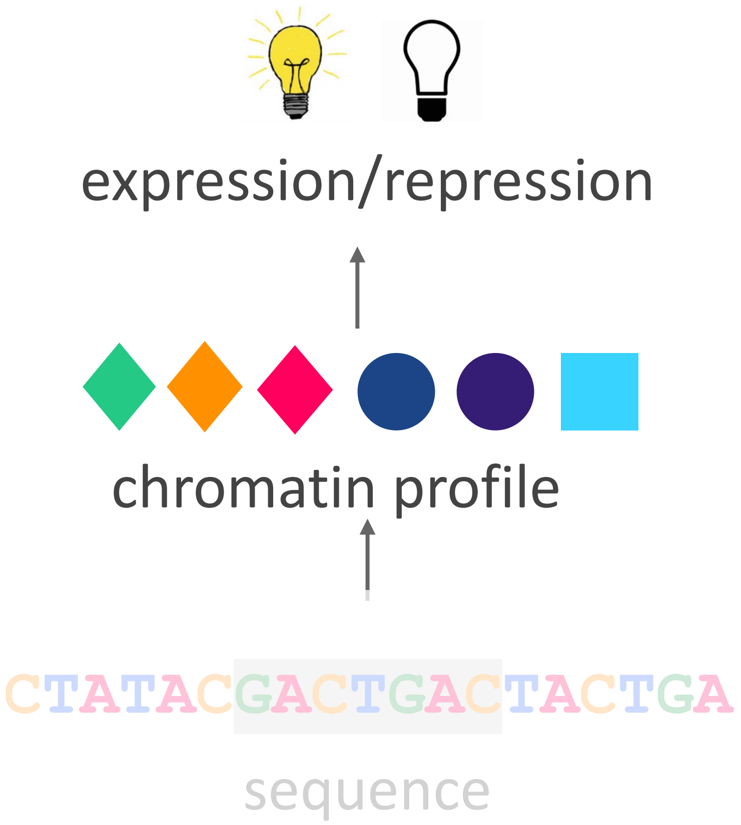
expression/repression

chromatin profile

CTATACGACTGACTACTGA

sequence

First Task:

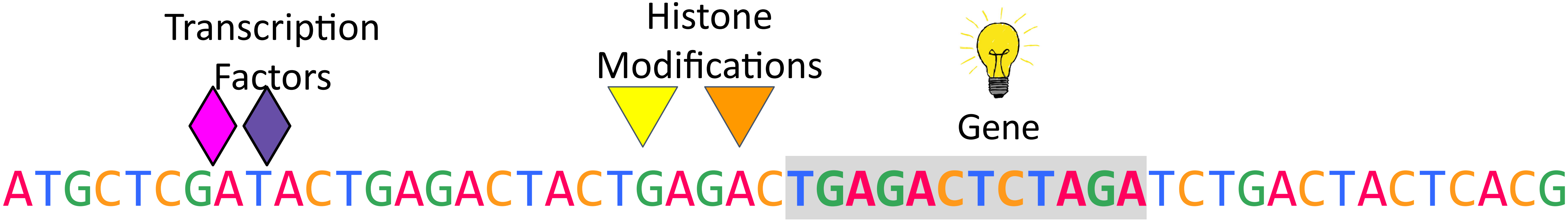


expression/repression

chromatin profile

CTATACGACTGACTACTGA

sequence



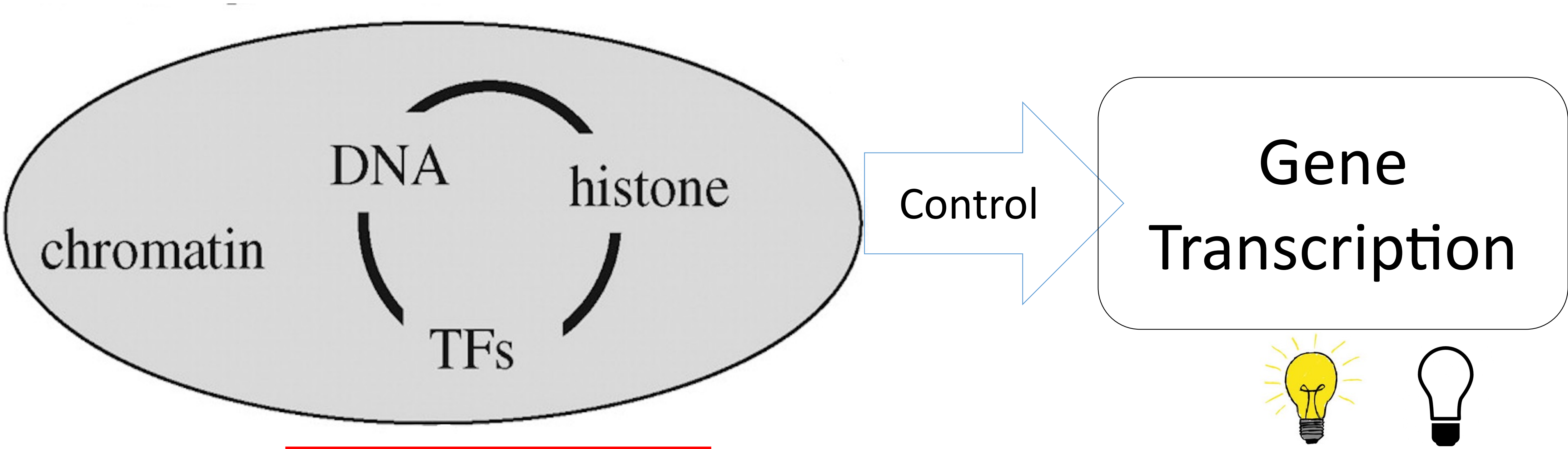
Transcription Factors

Histone Modifications

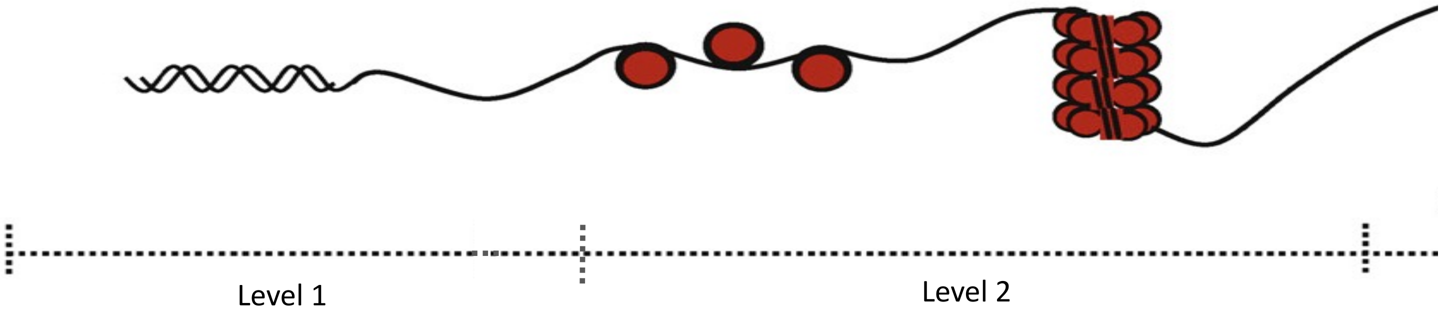
Gene

ATGCTCGATACTGAGACTACTGAGAC**TGAGACTCTAGA**TCTGACTACTCACG

Chromatin Profile as Evidence



Epigenetics
"Environment of the DNA"



Regulatory Elements

Genes
Promoters
Enhancers

Chromatin Structure

Histone Modifications
DNA methylation
Chromatin remodeling

ENCODE Project (2003-)

Describe the functional elements encoded in human DNA

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
ENCODE
GUIDEBOOK TO THE HUMAN GENOME
The ENCODE project in print and online
Science
GENES IN ACTION
22 October 2004
AAAS
nature
14 June 2007
MARS'S ANCIENT OCEAN
THE DEPTHS OF DISGUST
MENTORING
DECODING THE BLUEPRINT
The ENCODE pilot maps human genome function

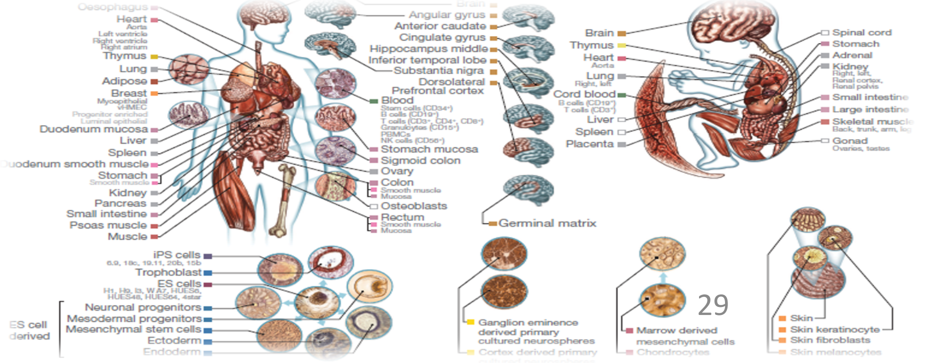
Roadmap Epigenetics Project (REMC, 2008-)

To produce a public resource of epigenomic maps for stem cells and primary ex vivo tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.

Nature

February 18, 2015

Integrative analysis of 111 reference human epigenomes (Abstract)



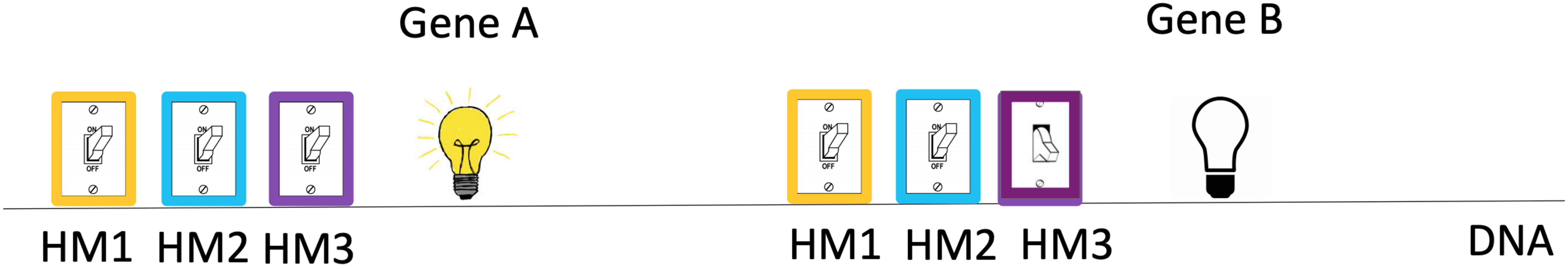
Why Study Epigenomics → Gene Expression?

- **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?

What HMs affect which genes in what cells?



Gene A

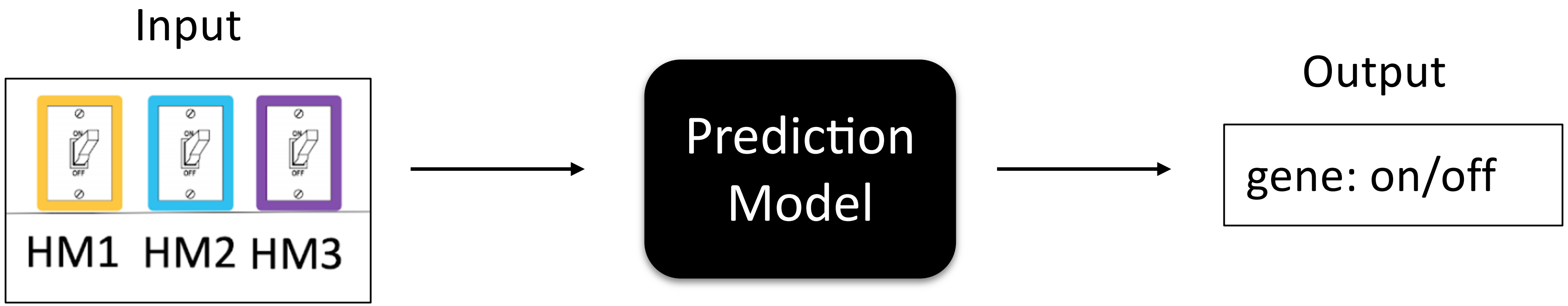
Gene B

HM1 HM2 HM3

HM1 HM2 HM3

DNA

Gene Transcription Prediction Task



Input

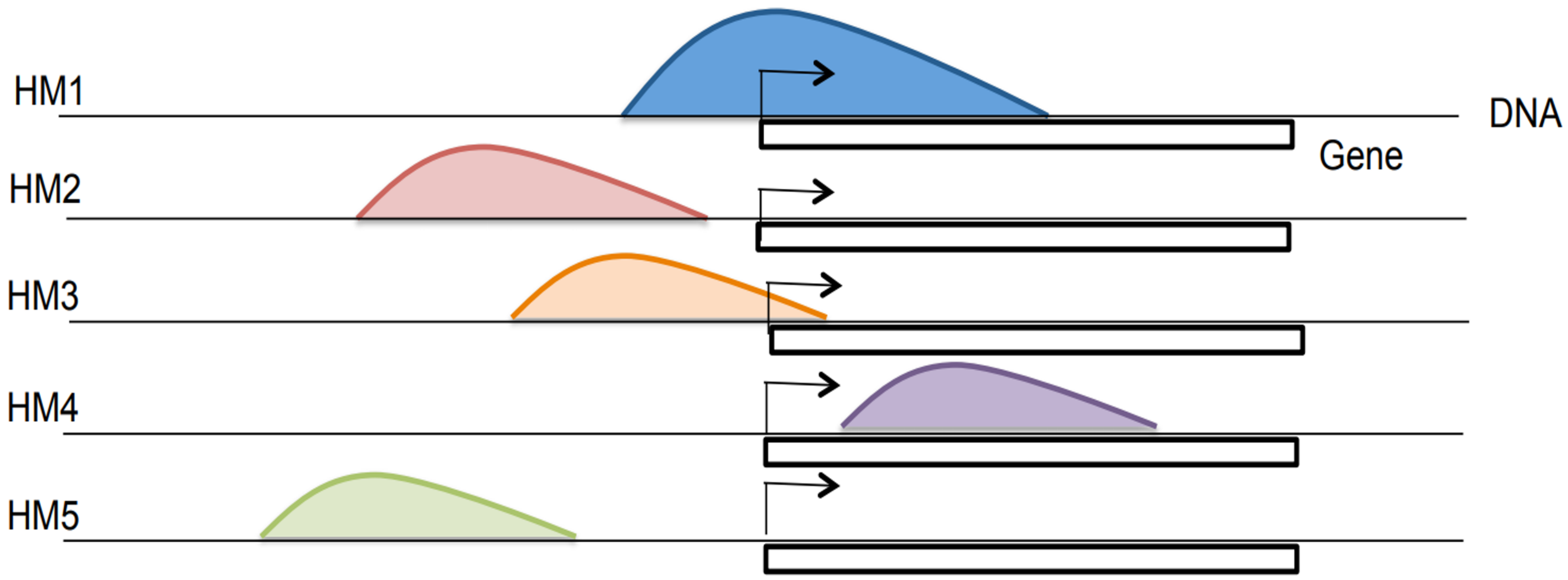
HM1 HM2 HM3

Prediction Model

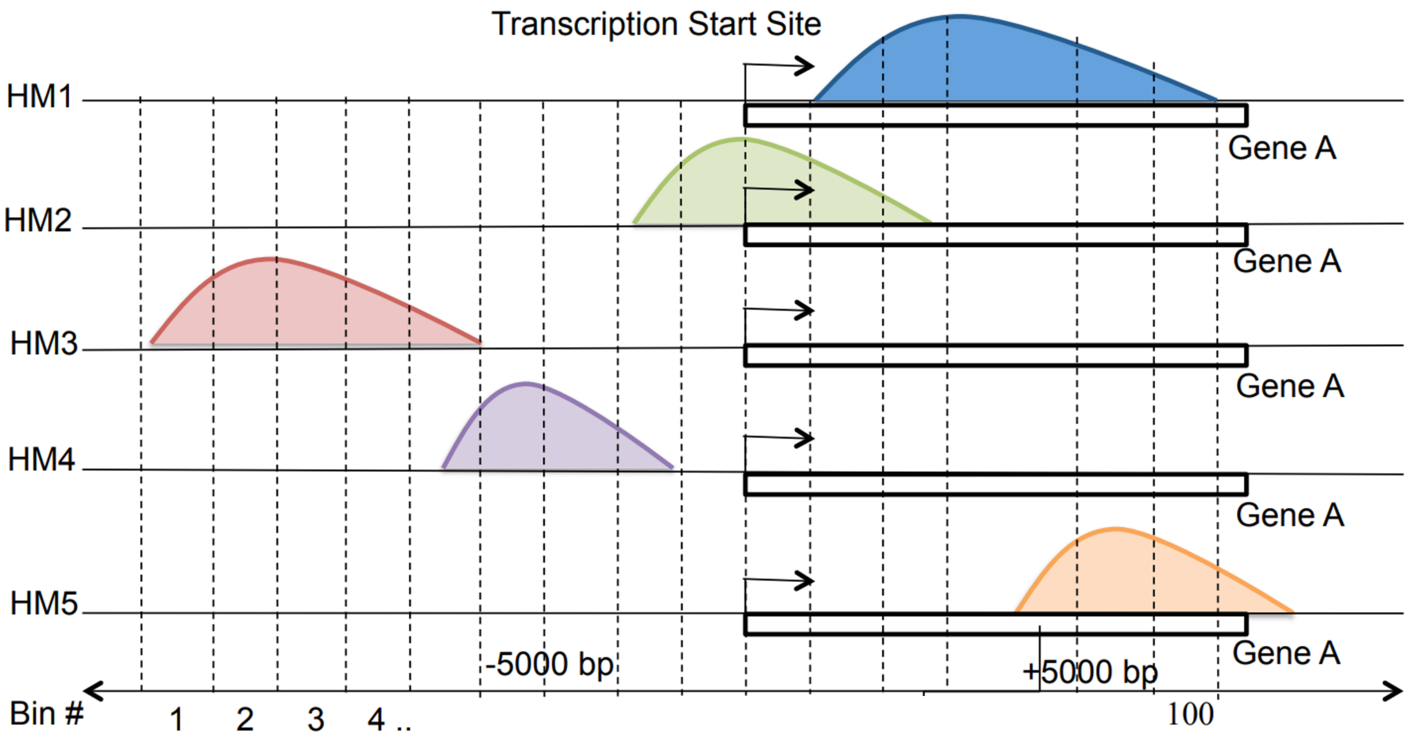
Output

gene: on/off

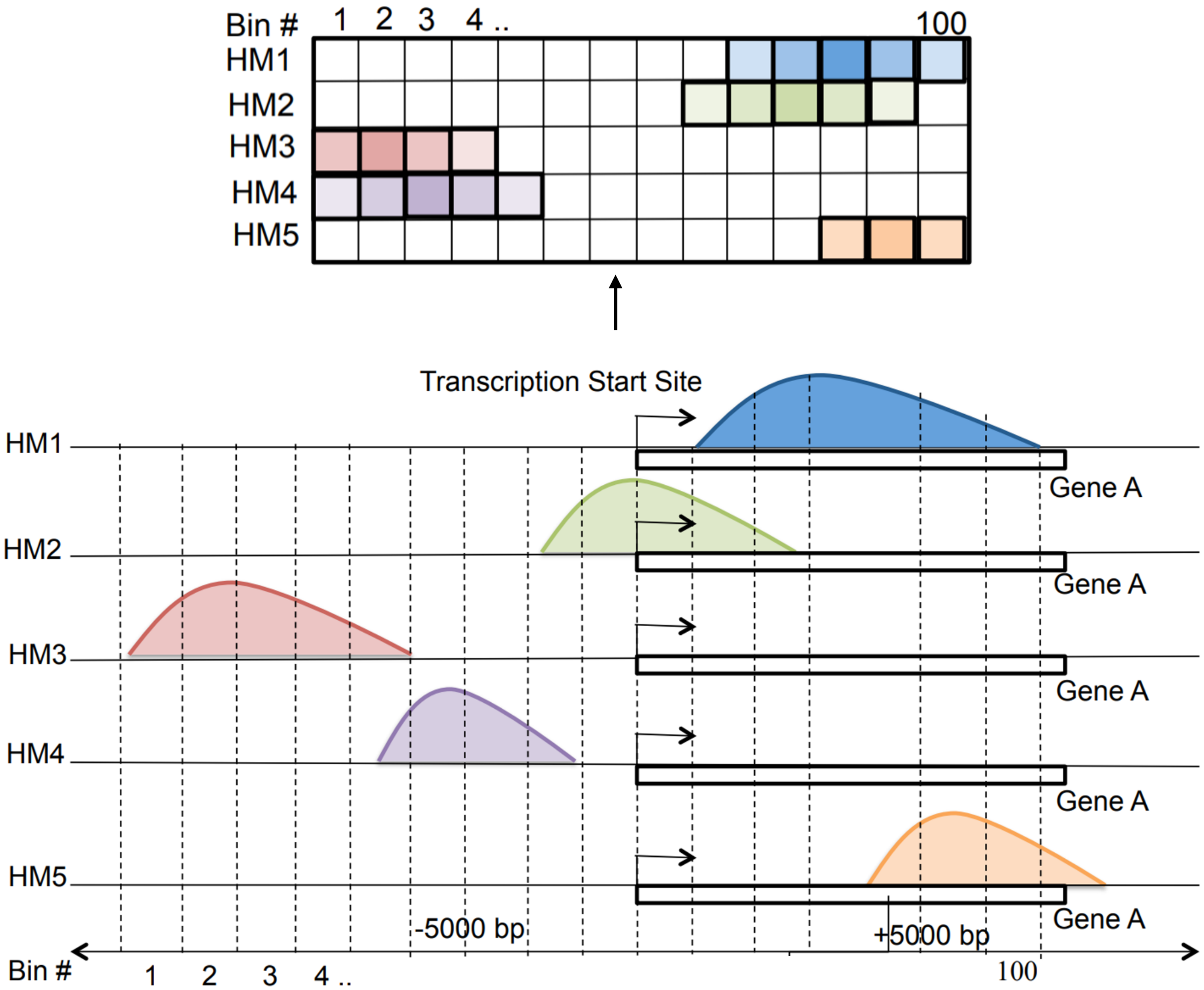
Histone Modification Input Data



Histone Modification Input Data

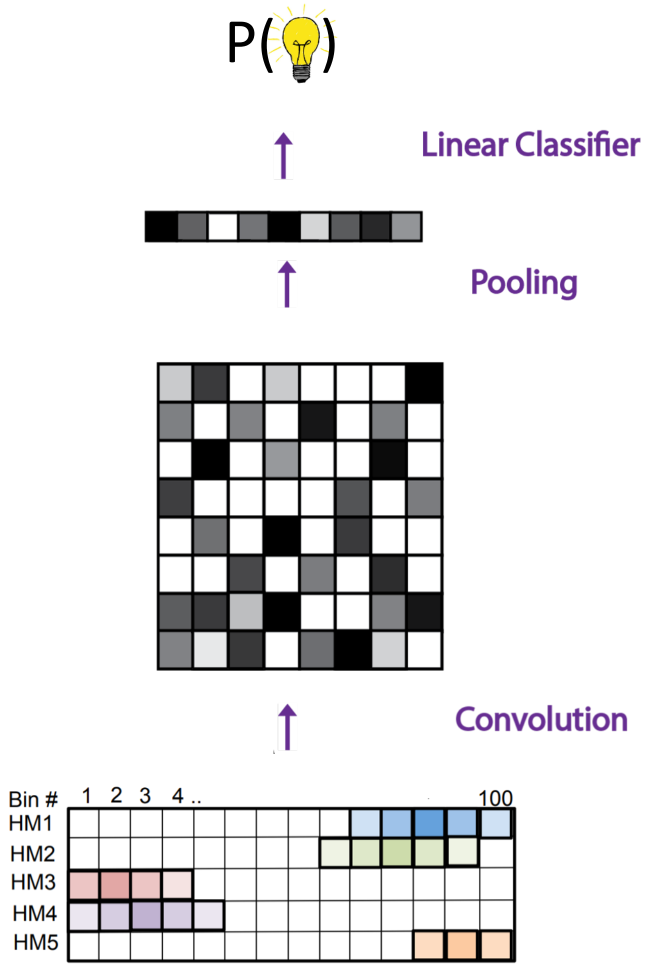


Histone Modification Input Data



DeepChrome

Singh, Lanchantin, Robins & Qi- Bioinformatics 2016



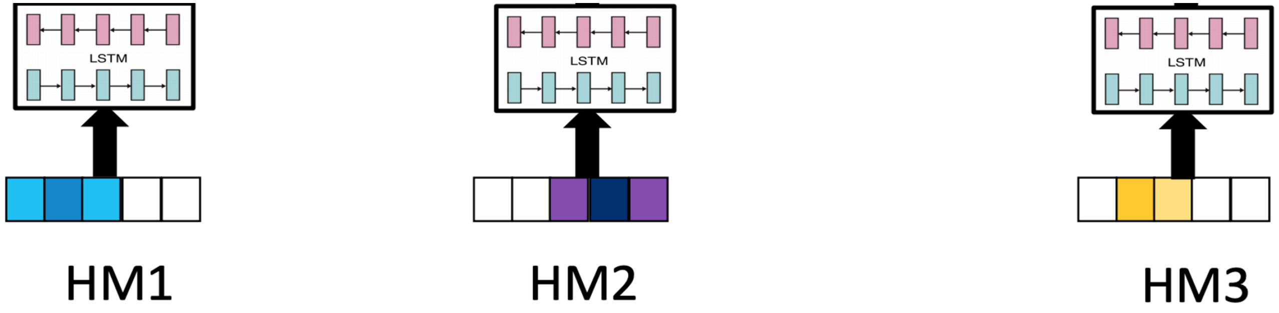
Linear Classifier

Pooling

Convolution

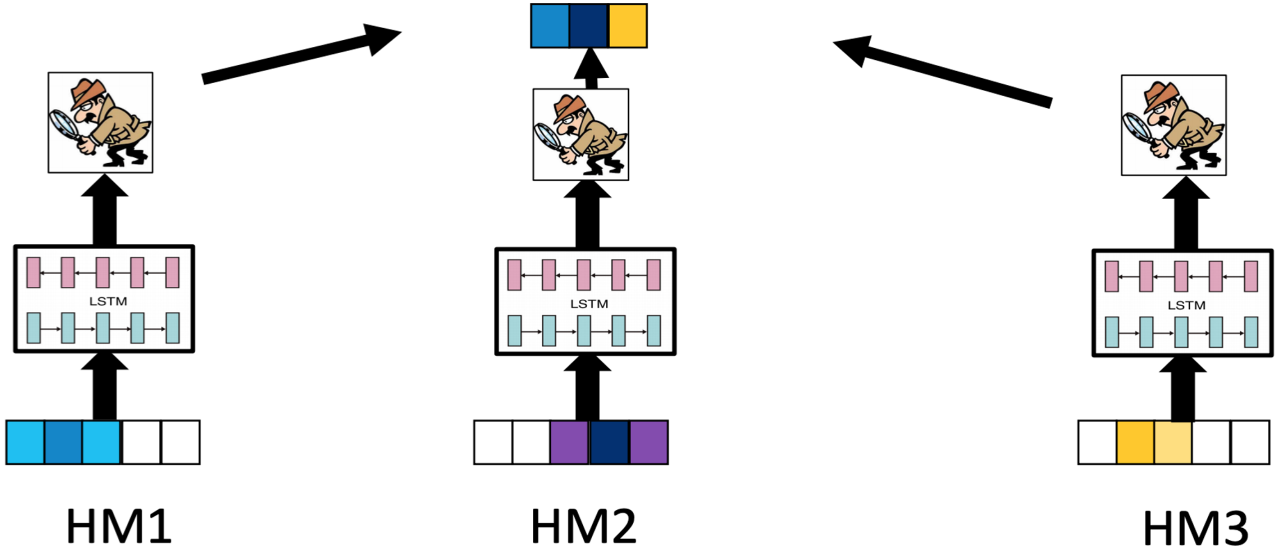
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



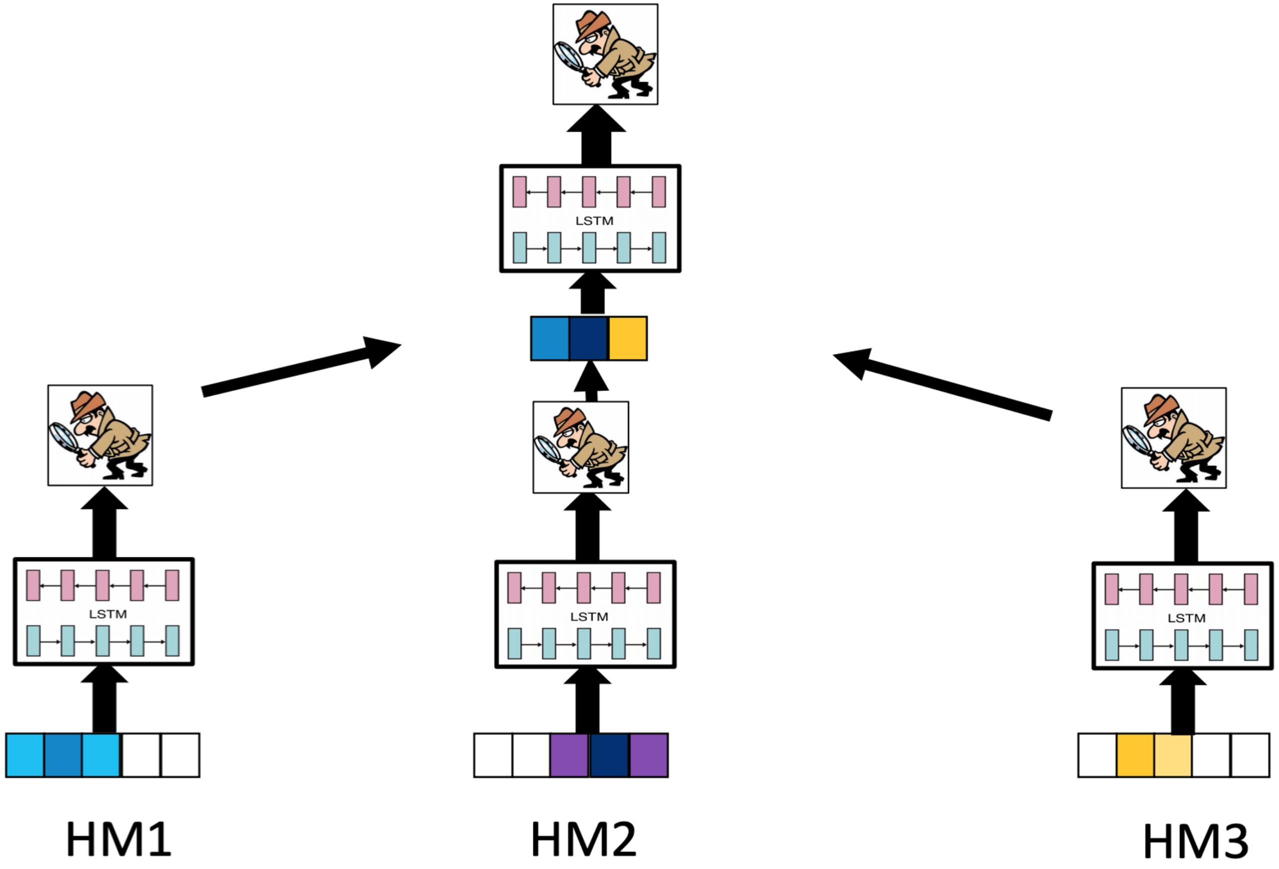
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



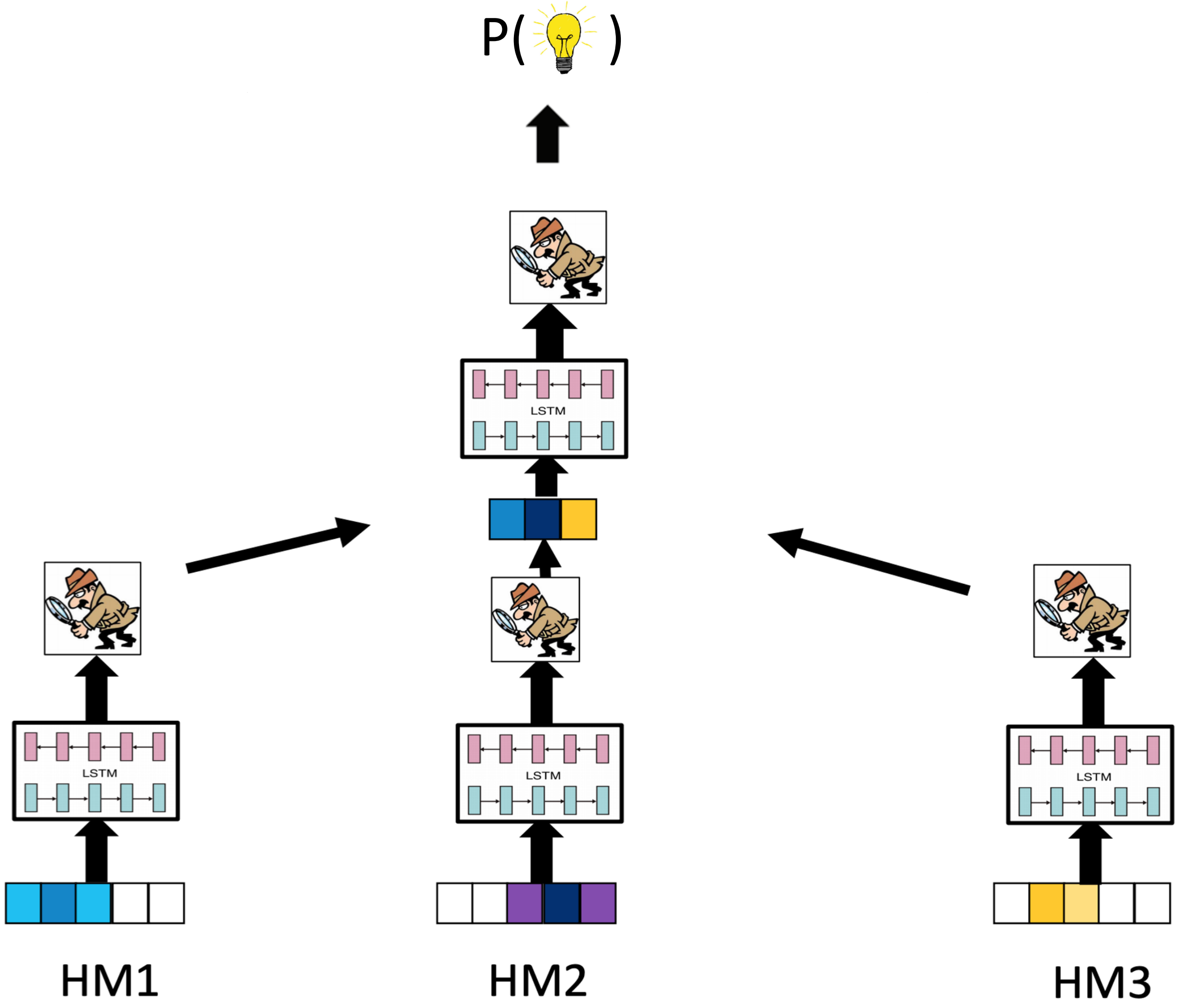
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi- NeurIPS 2017



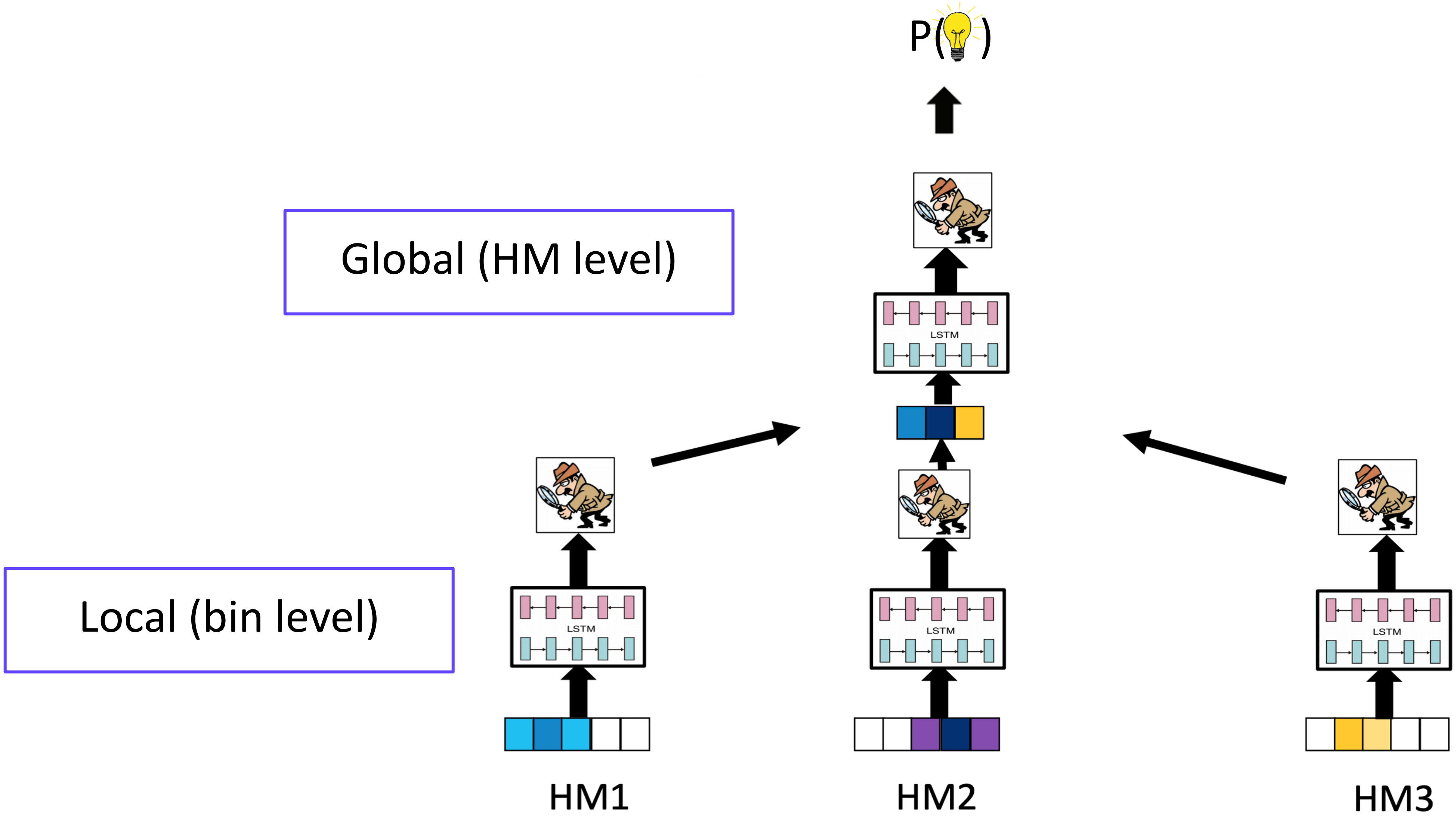
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi- NeurIPS 2017



Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017

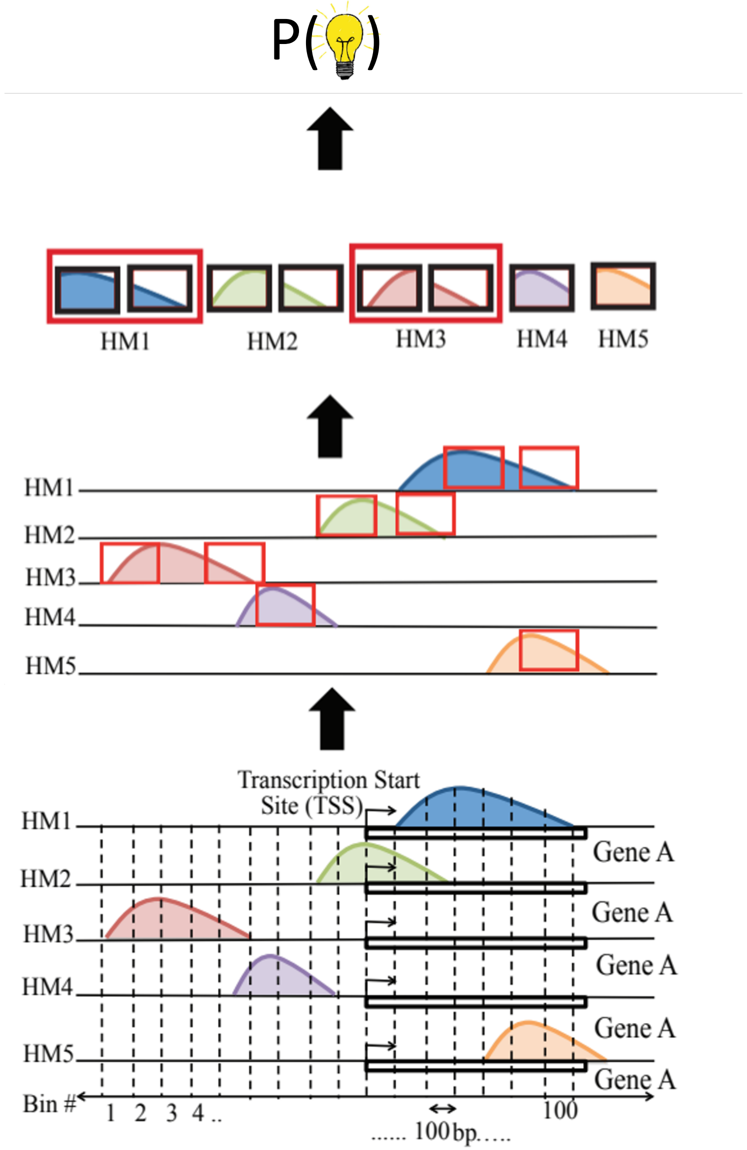


Global (HM level)

Local (bin level)

Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi- NeurIPS 2017



Interpretability by Hierarchical Attention

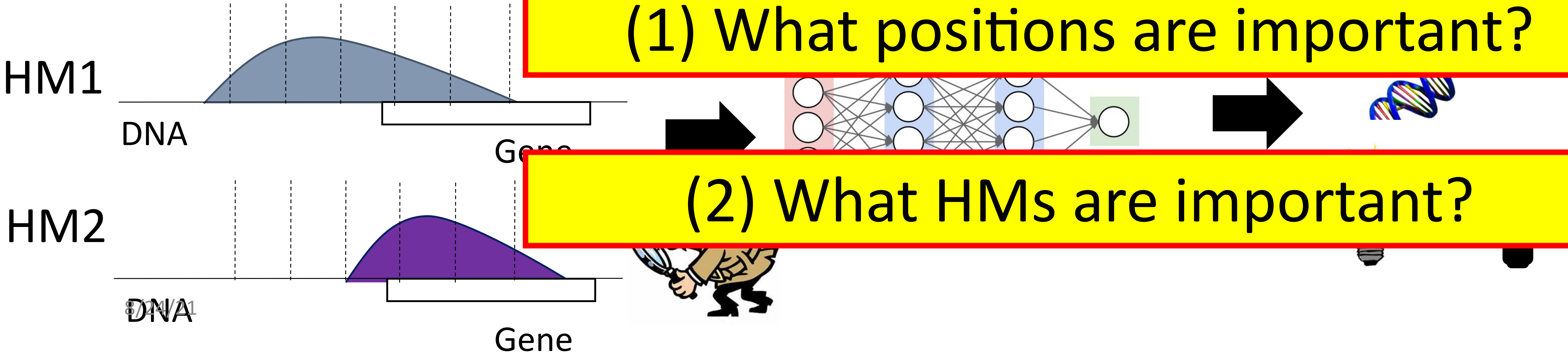
Input

Attention Mechanism

Output

Park

Gene

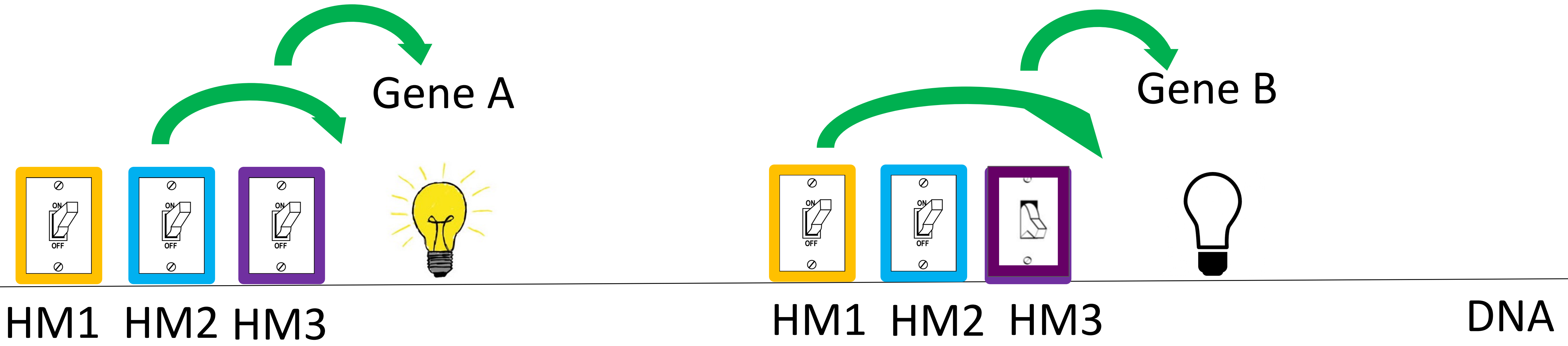


(1) What positions are important?

(2) What HMs are important?

Data Sets

~56 Cell Types



K~20 HMs

G~30,000 Genes

Experimental Setup

- Roadmap Epigenetics Project (REMC)
- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps / 5 Tier-1 HMs

Histone Mark	Functional Category
H3K27me3	Repressor
H3K36me3	Structural Promoter
H3K4me1	Distal Promoter
H3K4me3	Promoter
H3K9me3	Repressor

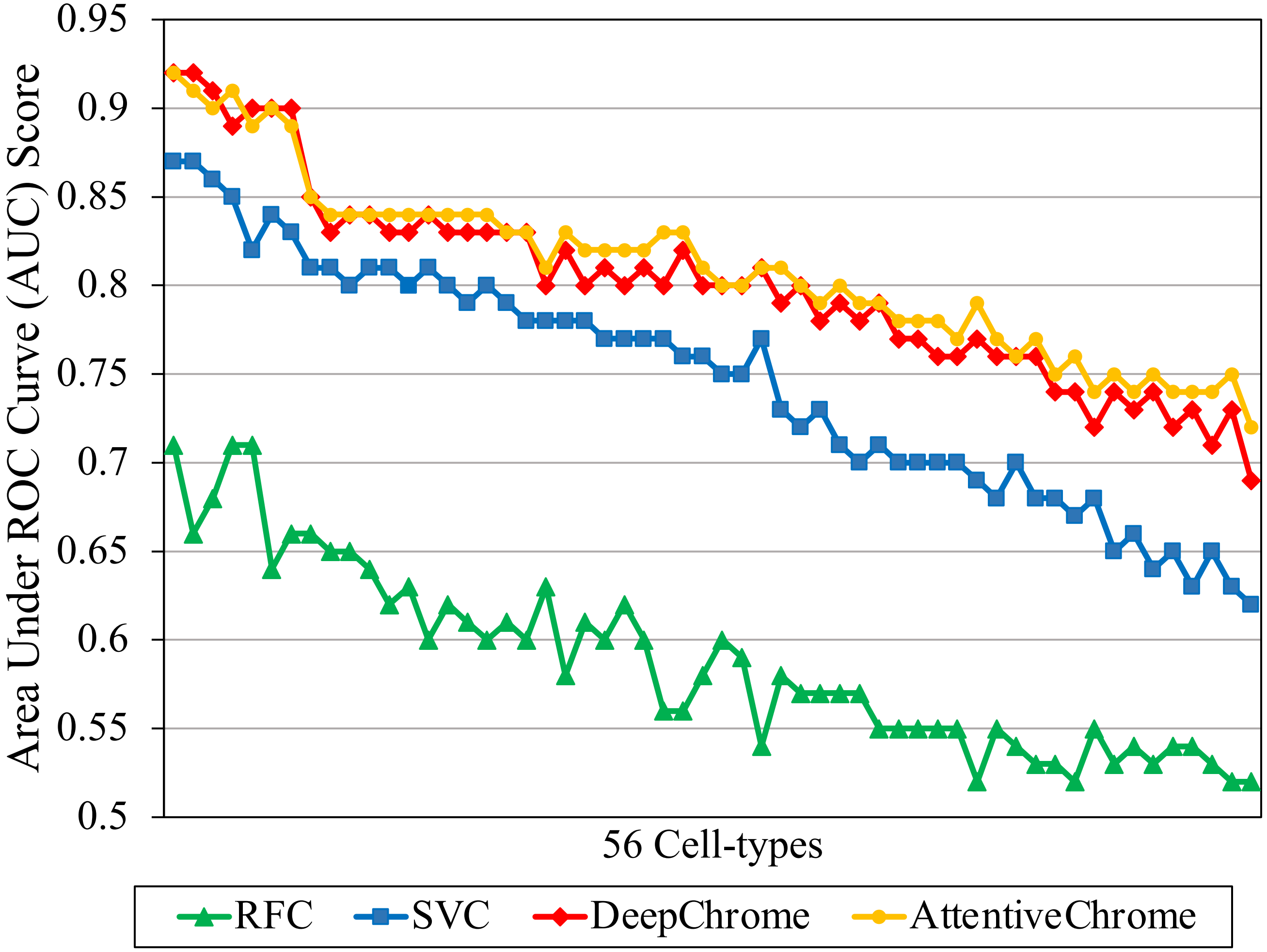
- **Output (Gene Expression):** Discretized RNA-Seq
- **Baselines:** Support Vector Classifier (SVC) and Random Forest Classifier (RFC)

Training Set
6601 Genes

Validation Set
6601 Genes

Test Set
6600 Genes

Prediction

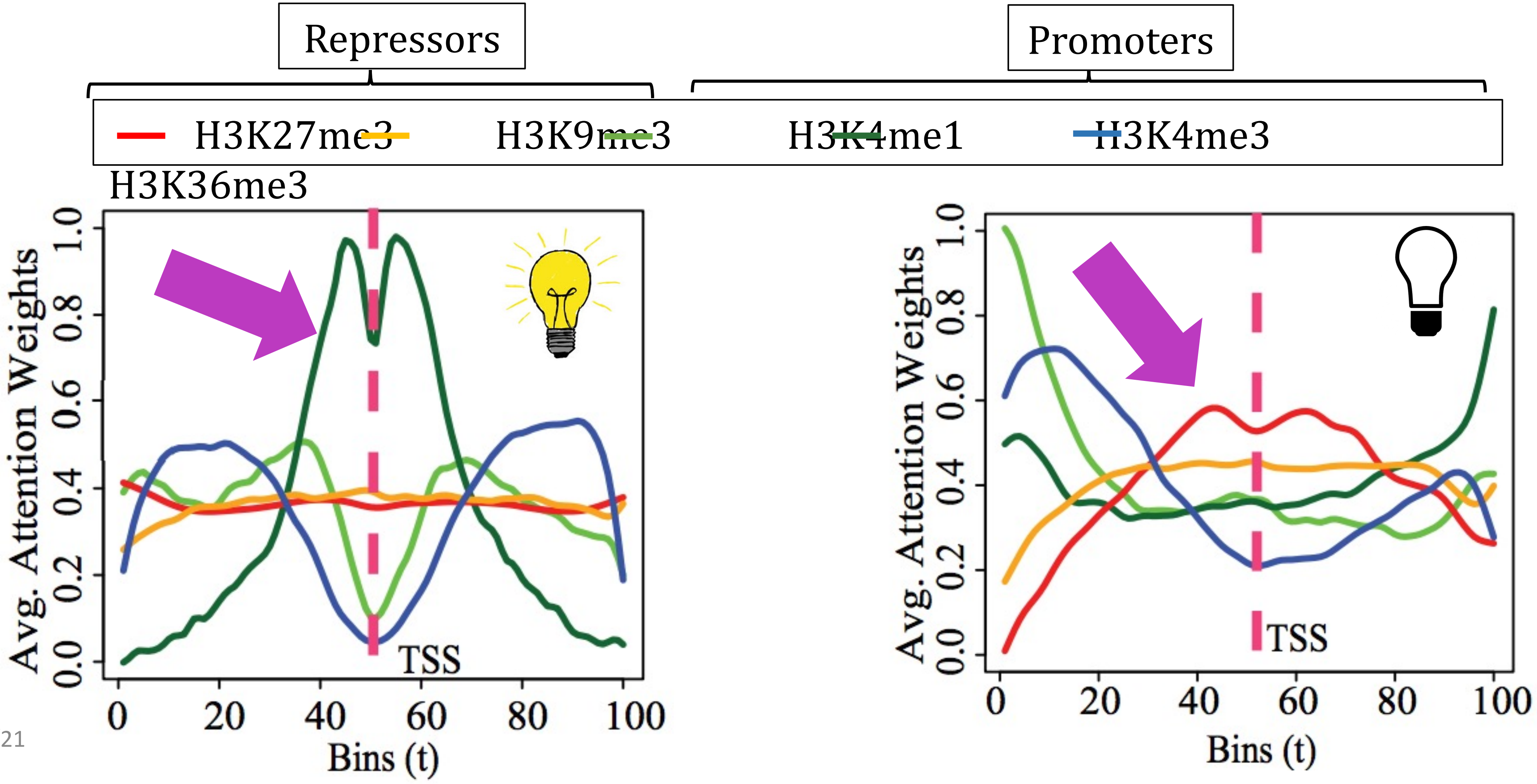


Improvement for 49/56 Cell-types

Bin-Level Visualization

(1) What positions are important?

CELL TYPE: GM12878 (Blood Cell)



Validation of Attention Weights (using one extra HM signals)

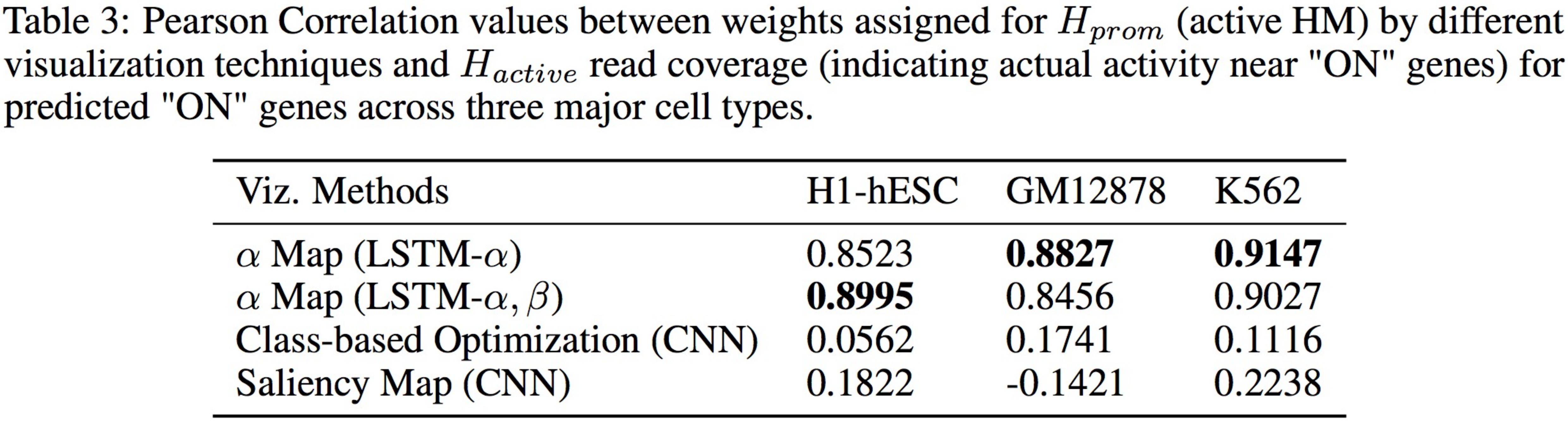


Table 3: Pearson Correlation values between weights assigned for H_{prom} (active HM) by different visualization techniques and H_{active} read coverage (indicating actual activity near "ON" genes) for predicted "ON" genes across three major cell types.

Viz. Methods	H1-hESC	GM12878	K562
α Map (LSTM-α)	0.8523	**0.8827**	**0.9147**
α Map (LSTM-α, β)	**0.8995**	0.8456	0.9027
Class-based Optimization (CNN)	0.0562	0.1741	0.1116
Saliency Map (CNN)	0.1822	-0.1421	0.2238

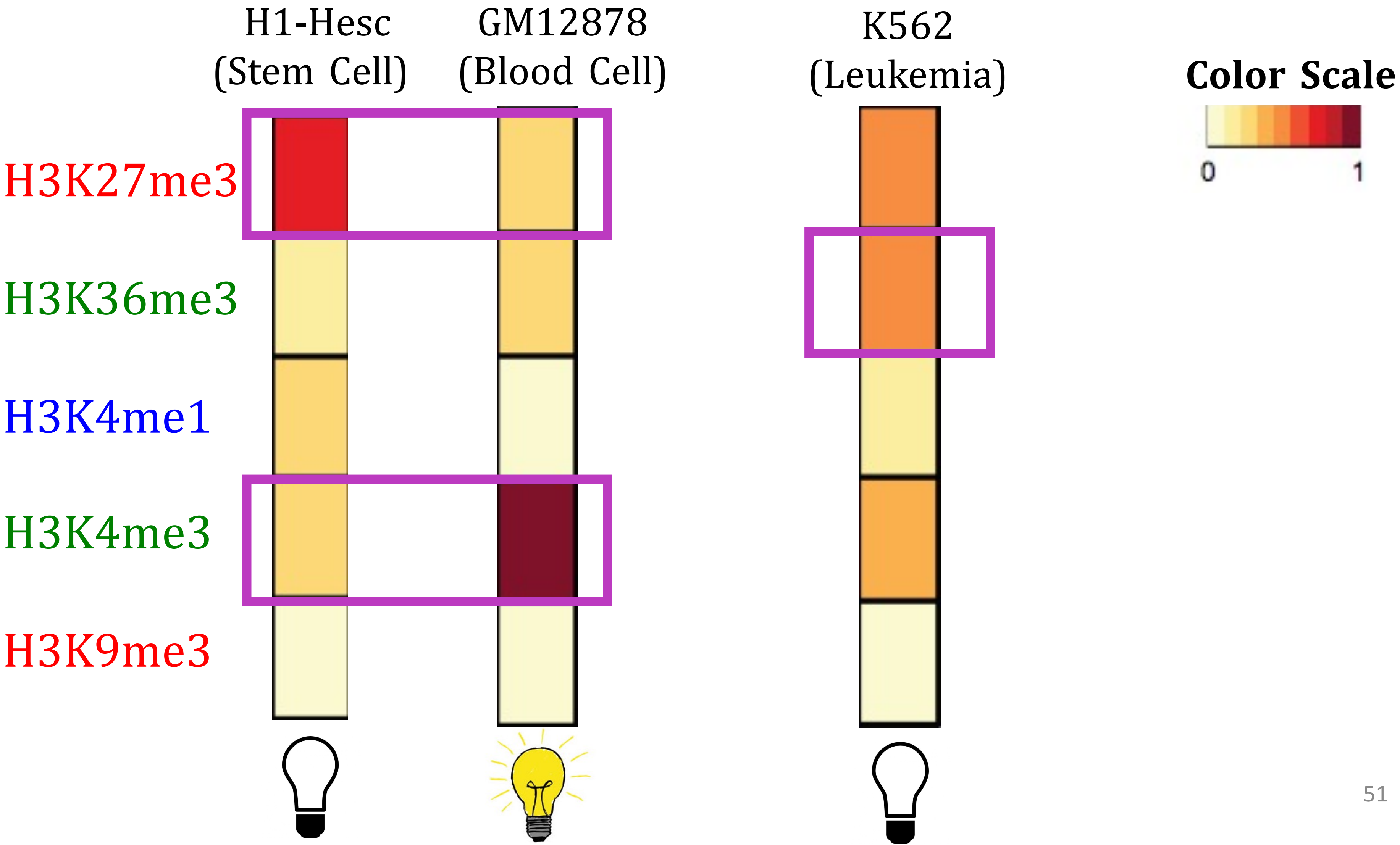
- Additional signal - H3K27ac (H-Active) from REMC
- Average local attention weights of gene=ON correspond well with H-active
- Indicating AttentiveChrome is focusing on the correct bin positions

HM-Level Visualization

(2) What HMs are important?

Cell Types:

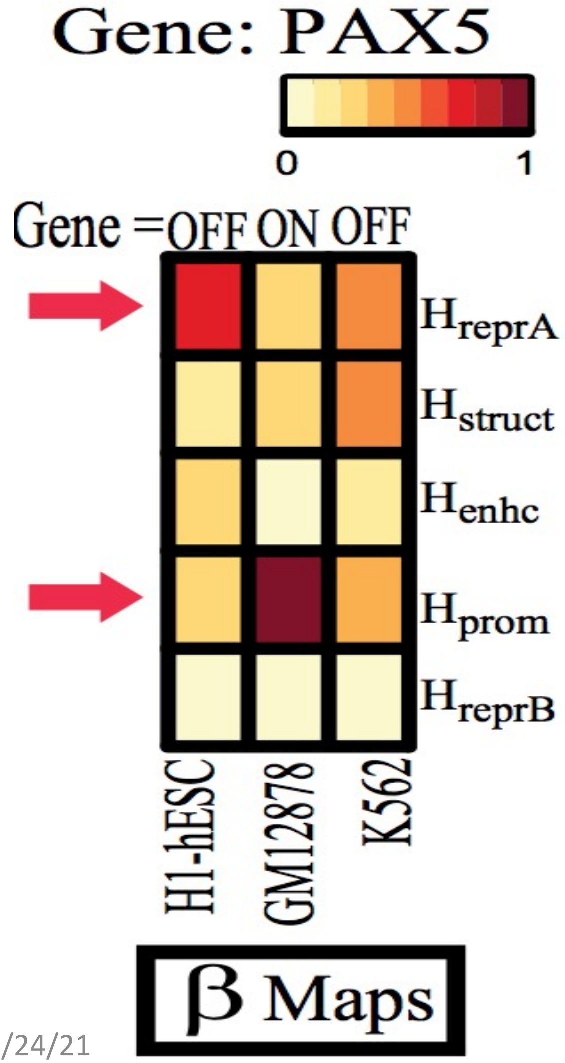
Gene: PAX5



PROMOTER
DISTAL PROMOTER
REPRESSOR

Results: HM level attention

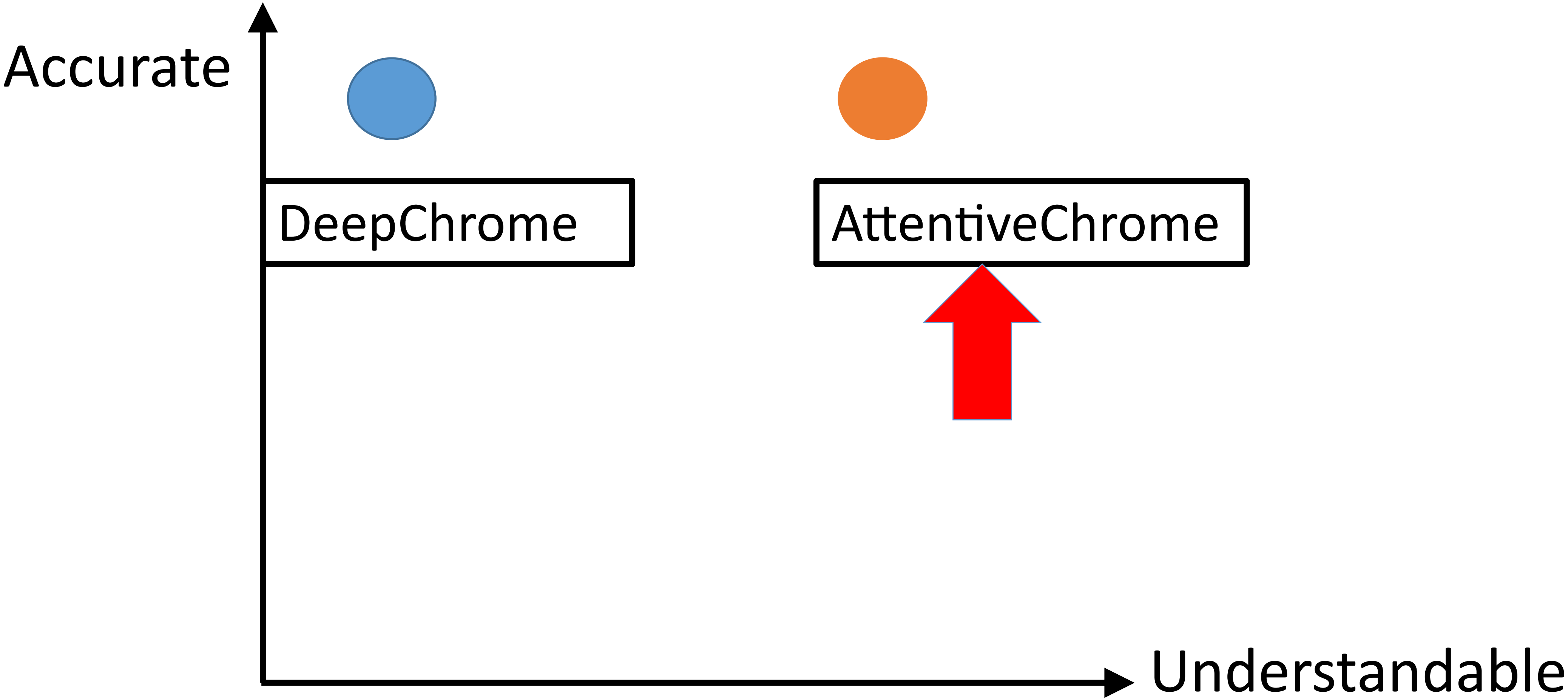
(2) What HMs are important?



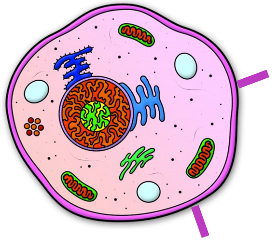
- An important differentially regulated gene (PAX5) across three blood lineage cell types:
  - H1-hESC (stem cell),
  - GM12878 (blood cell),
  - K562 (leukemia cell).
- Trend of its global weights (beta) Verified through the literature.

https://www.deepchrome.net

Summary of tools



Output (Y) Labels

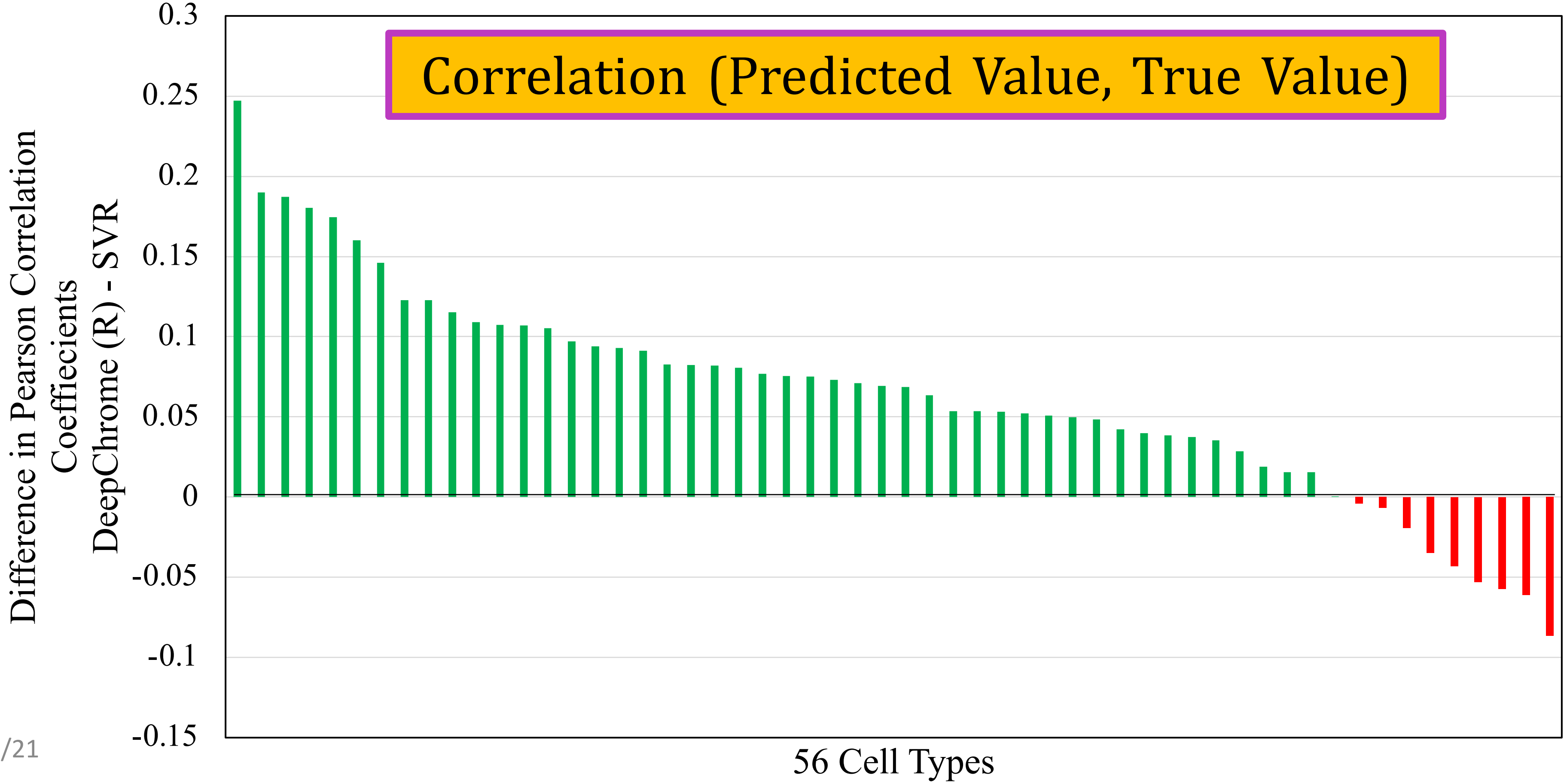


Genes	Gene Expression (RPKM)	Y Labels
RUNX1	1.296	0
SMAD2	14.902	1
MYC	3.805	0
PAX5	15.066	1
......	....	......

Threshold = 10.245 (Median)

Where we further tried?

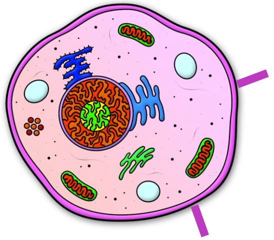
Changing Task : Classification →Regression



Correlation (Predicted Value, True Value)

Where we further tried?

Changing Task : Classification →Regression



1.770

Gene Expression

Genes	Gene Expression (RPKM)	Y log(RPKM)
RUNX1	1.296	01126
SMAD2	14.902	1.1737
MYC	3.805	0.5803
PAX5	15.066	1.779
……	…….	……

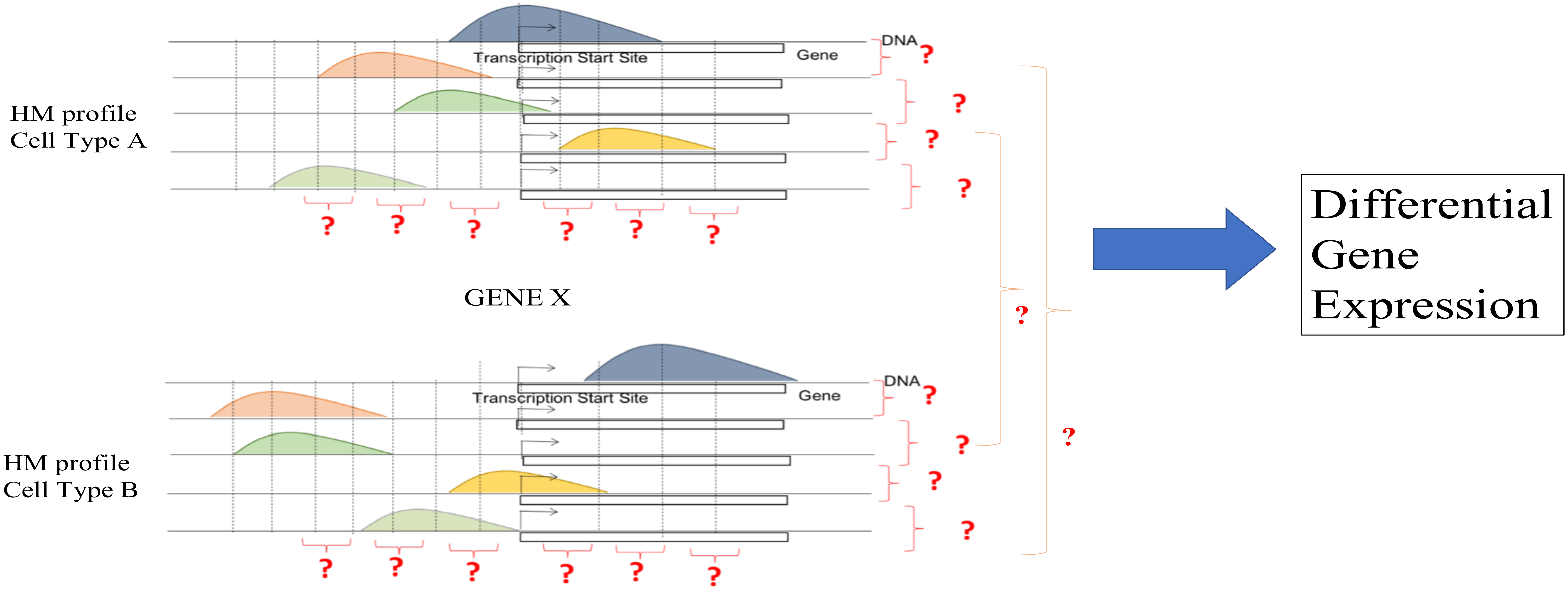
Mean Square Error Loss

$(Y - f(X))^2$

Where we further tried?

A. Sekon, R. Singh, Y. QiDeepDiff: Deep-learning for predicting Differential gene expression from histone modifications, Bioinformatics 2018

Changing Task : Cell-Specific → Cross Cell



HM profile
Cell Type A

GENE X

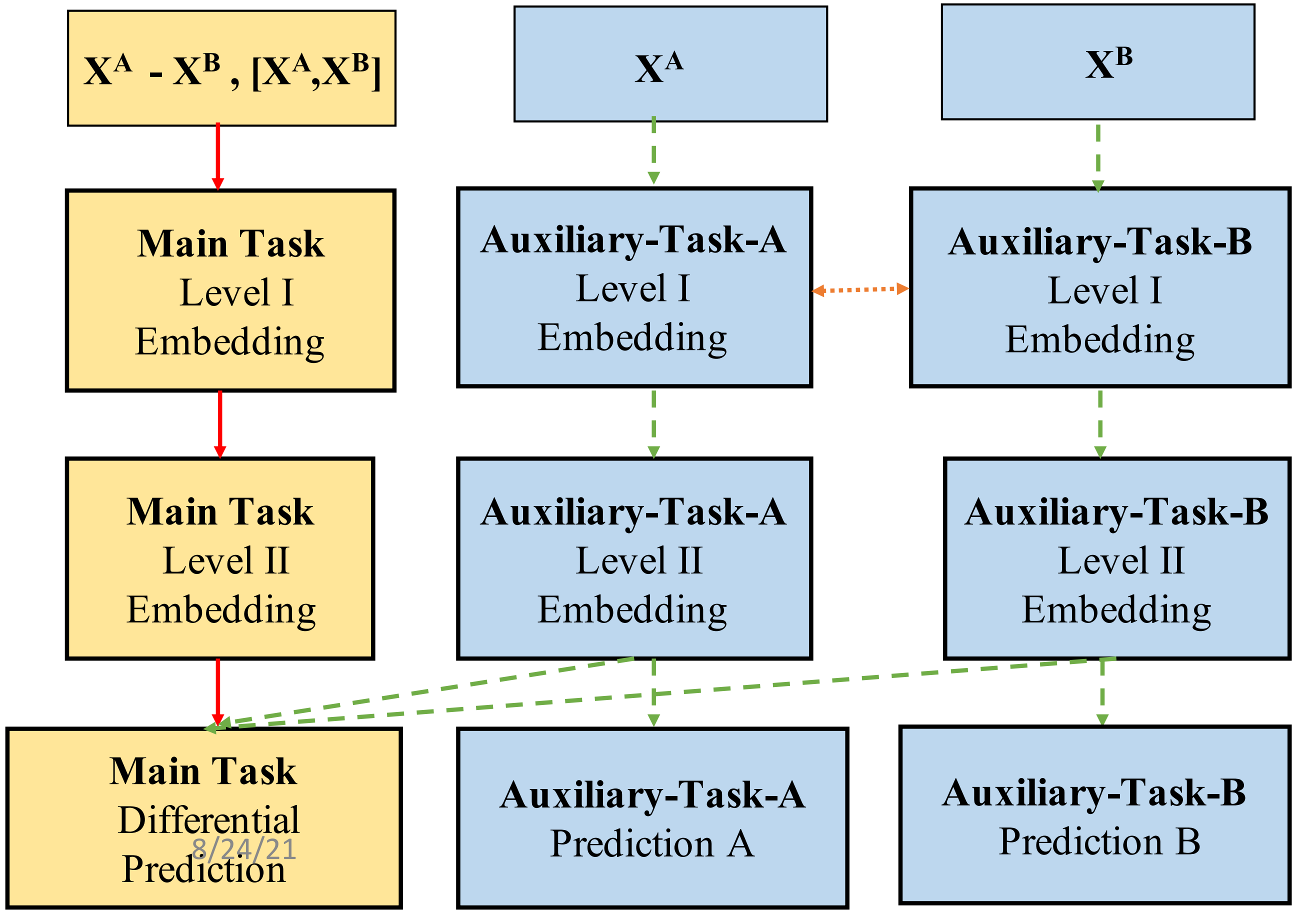
HM profile
Cell Type B

Differential
Gene
Expression

Where we further tried?

DeepDiff: Deep-learning for predicting Differential gene expression from histone modifications

Changing Task : Cell-Specific → Cross Cell



X^A - X^B , [X^A,X^B]
X^A
X^B
Main Task Level I Embedding
Auxiliary-Task-A Level I Embedding
Auxiliary-Task-B Level I Embedding
Main Task Level II Embedding
Auxiliary-Task-A Level II Embedding
Auxiliary-Task-B Level II Embedding
Main Task Differential Prediction
Auxiliary-Task-A Prediction A
Auxiliary-Task-B Prediction B

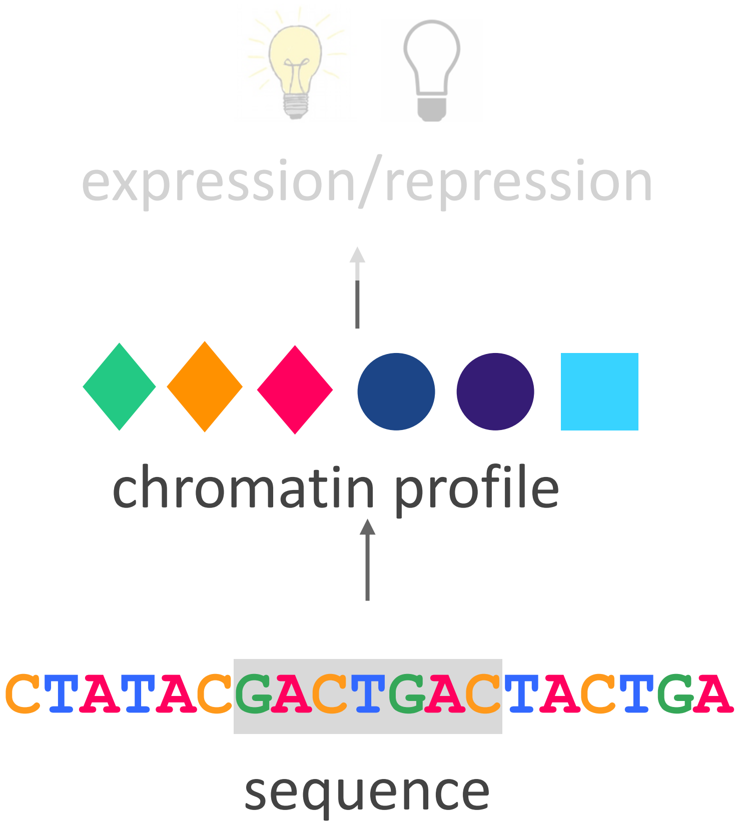
← ① **Main Task**: Differential gene expression prediction

⇠ ② **Cell-Specific Auxiliary**: Auxiliary-Task-A and Auxiliary-Task-B cell type specific prediction

↔ ③ **Siamese Auxiliary**: Siamese contrastive loss

DeepDiff Variations	Objective Loss
① Raw:d, Raw:c, Raw	ℓ_{Diff}
② Aux	$\ell_{\text{Diff}}+\ell_{\text{CellAux}}$
① + ② Raw+Aux	$\ell_{\text{Diff}}+\ell_{\text{CellAux}}$
② + ③ Aux+Siamese	$\ell_{\text{Diff}}+\ell_{\text{CellAux}}+\ell_{\text{Siamese}}$
① + ② + ③ Raw+Aux+Siamese	$\ell_{\text{Diff}}+\ell_{\text{CellAux}}+\ell_{\text{Siamese}}$

Second Task:



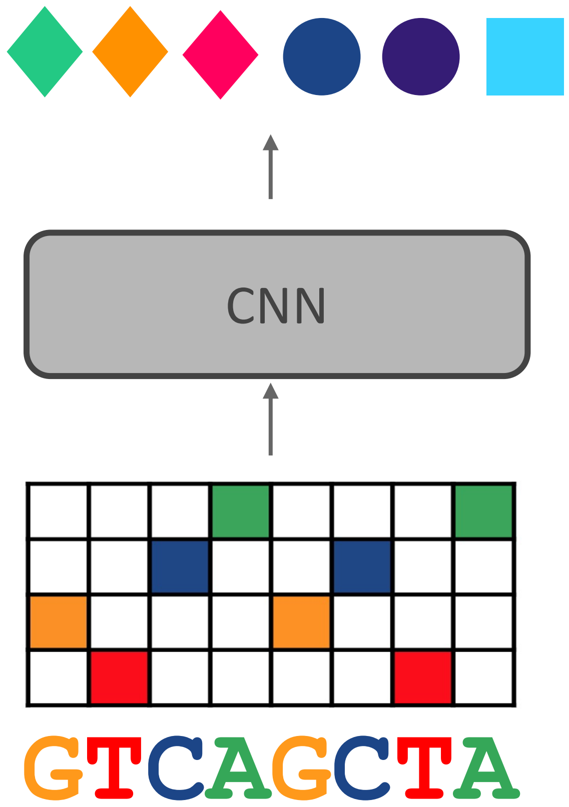
expression/repression

chromatin profile

CTATACGACTGACTACTGA

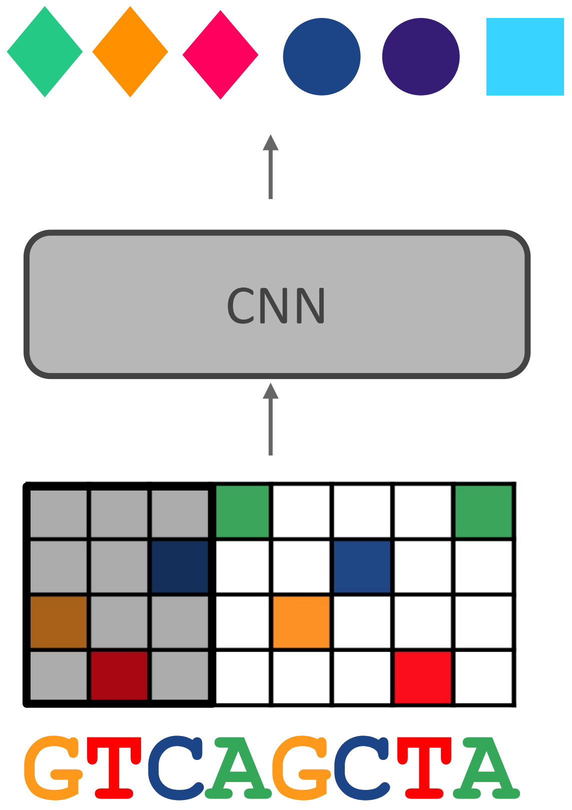
sequence

Local Sequence Chromatin Profile Prediction



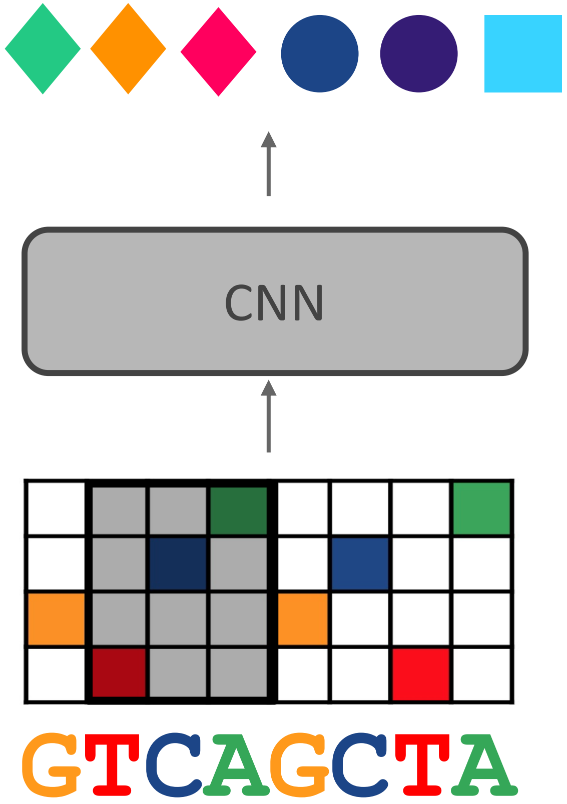
Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



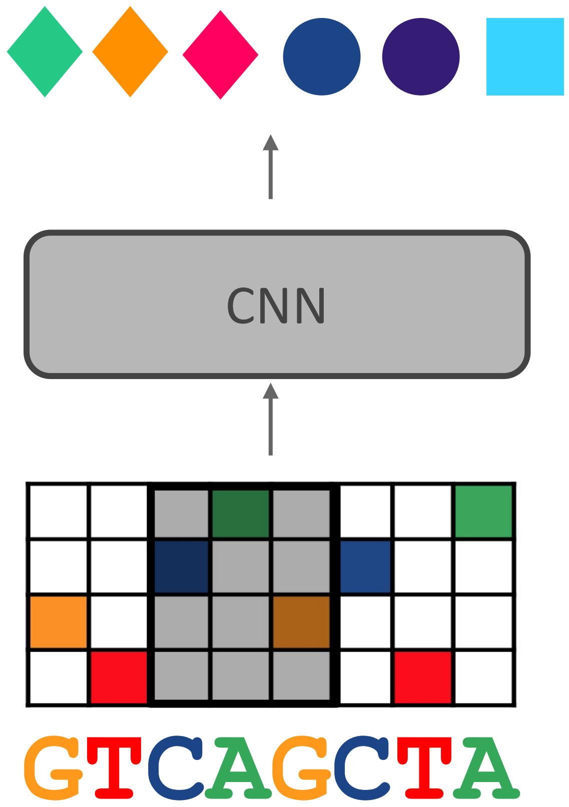
Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



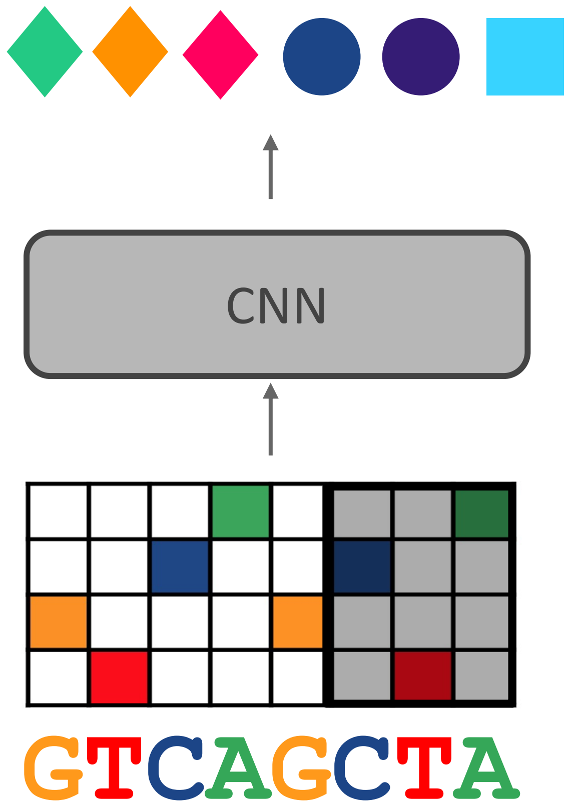
Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction

ACTGCTACCTATGACGTGATGCATCGTAGCT

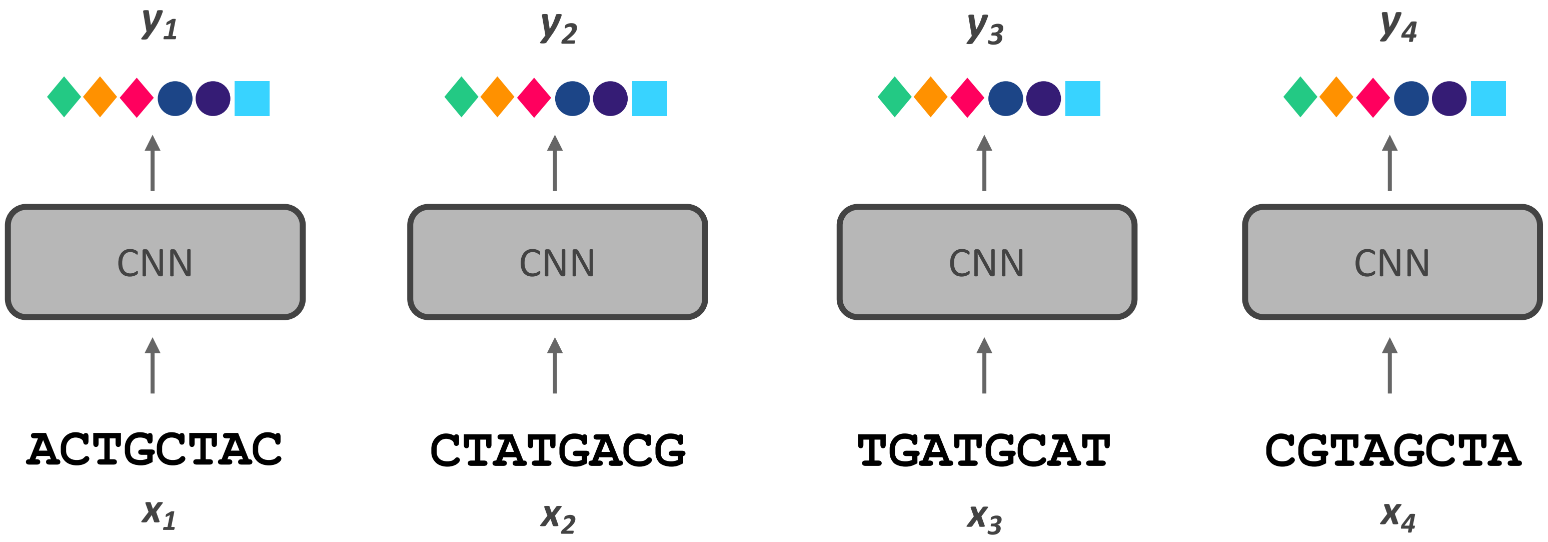
A

Local Sequence Chromatin Profile Prediction

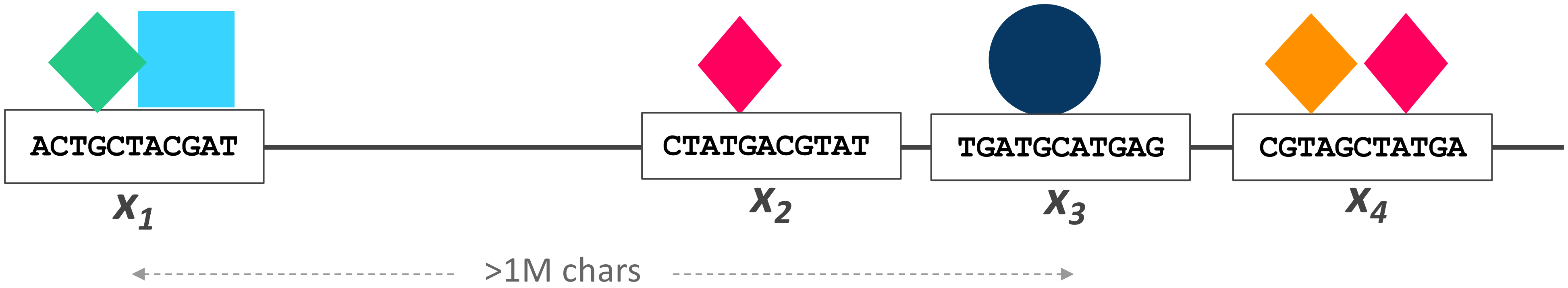
ACTGCTAC **CTATGACG** **TGATGCAT** **CGTAGCTA**

x_1 x_2 x_3 x_4

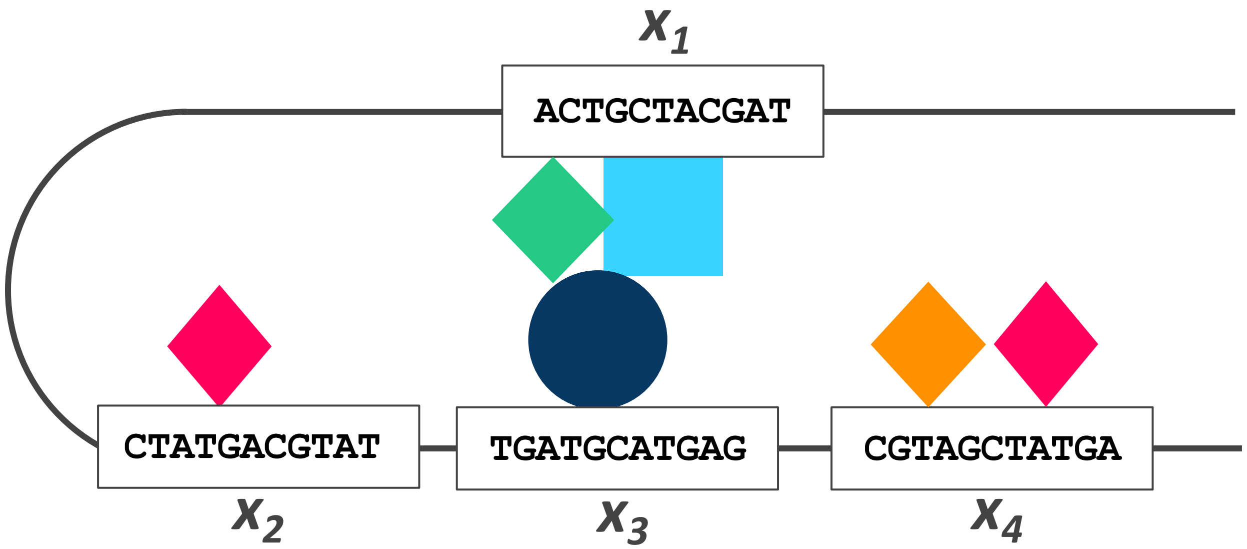
Local Sequence Chromatin Profile Prediction



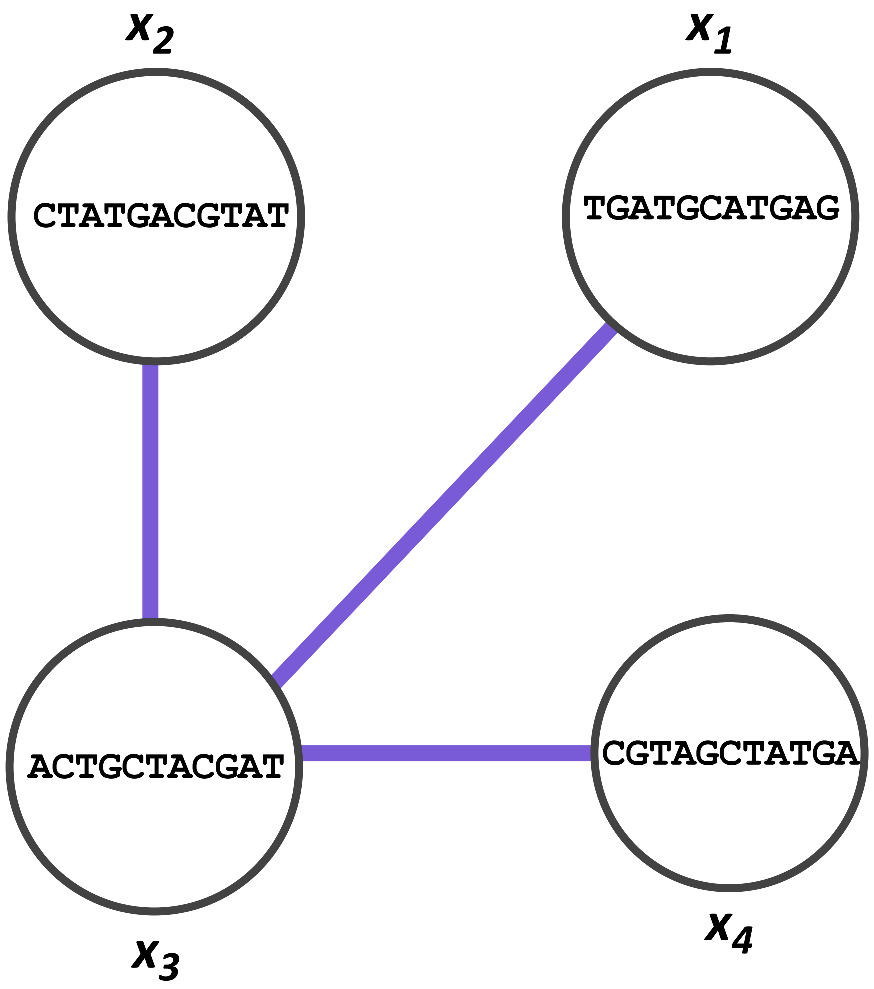
Influence of Long-Range Interactions on Chromatin Profile



Influence of Long-Range Interactions on Chromatin Profile



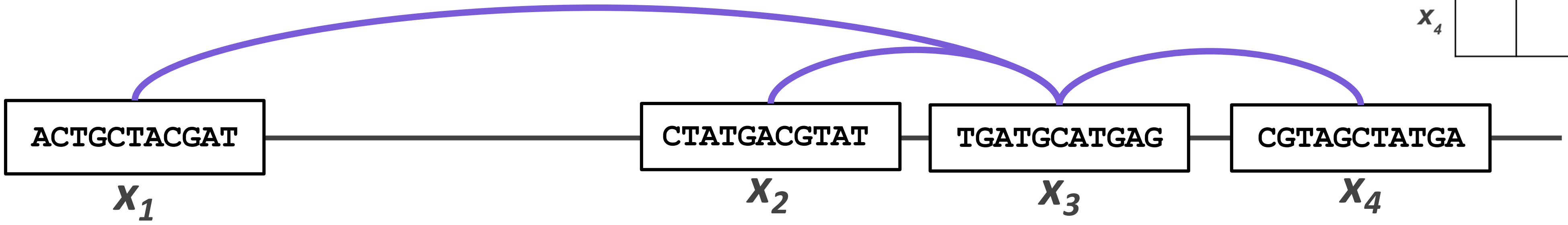
Genome: Locally a Sequence, Globally a Graph



High-throughput Chromosome Conformation Capture (Hi-C)

Hi-C Adjacency Matrix **A**

	x_1	x_2	x_3	x_4
x_1			🟪	
x_2			🟪	
x_3	🟪	🟪		🟪
x_4			🟪	

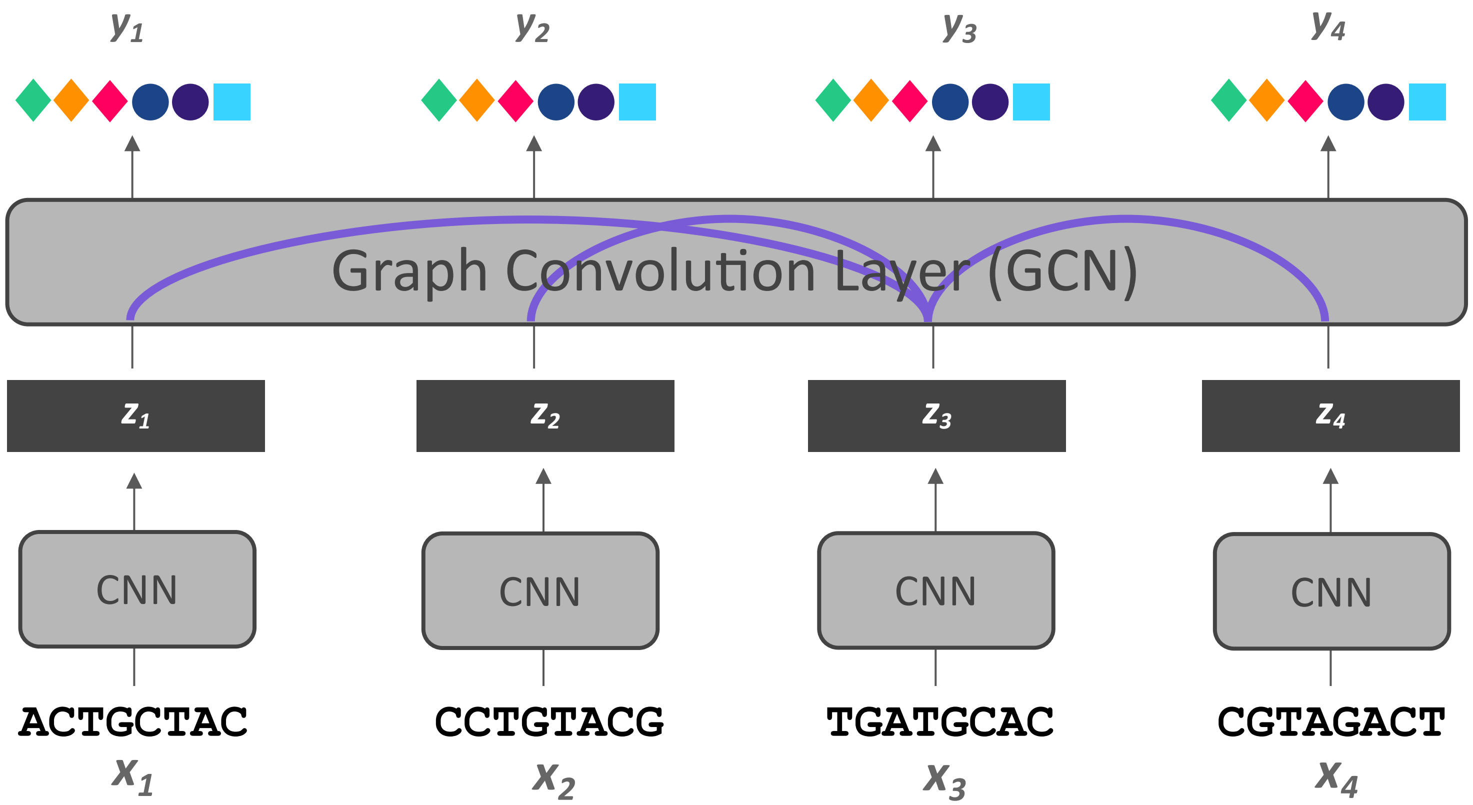


— contact between two windows

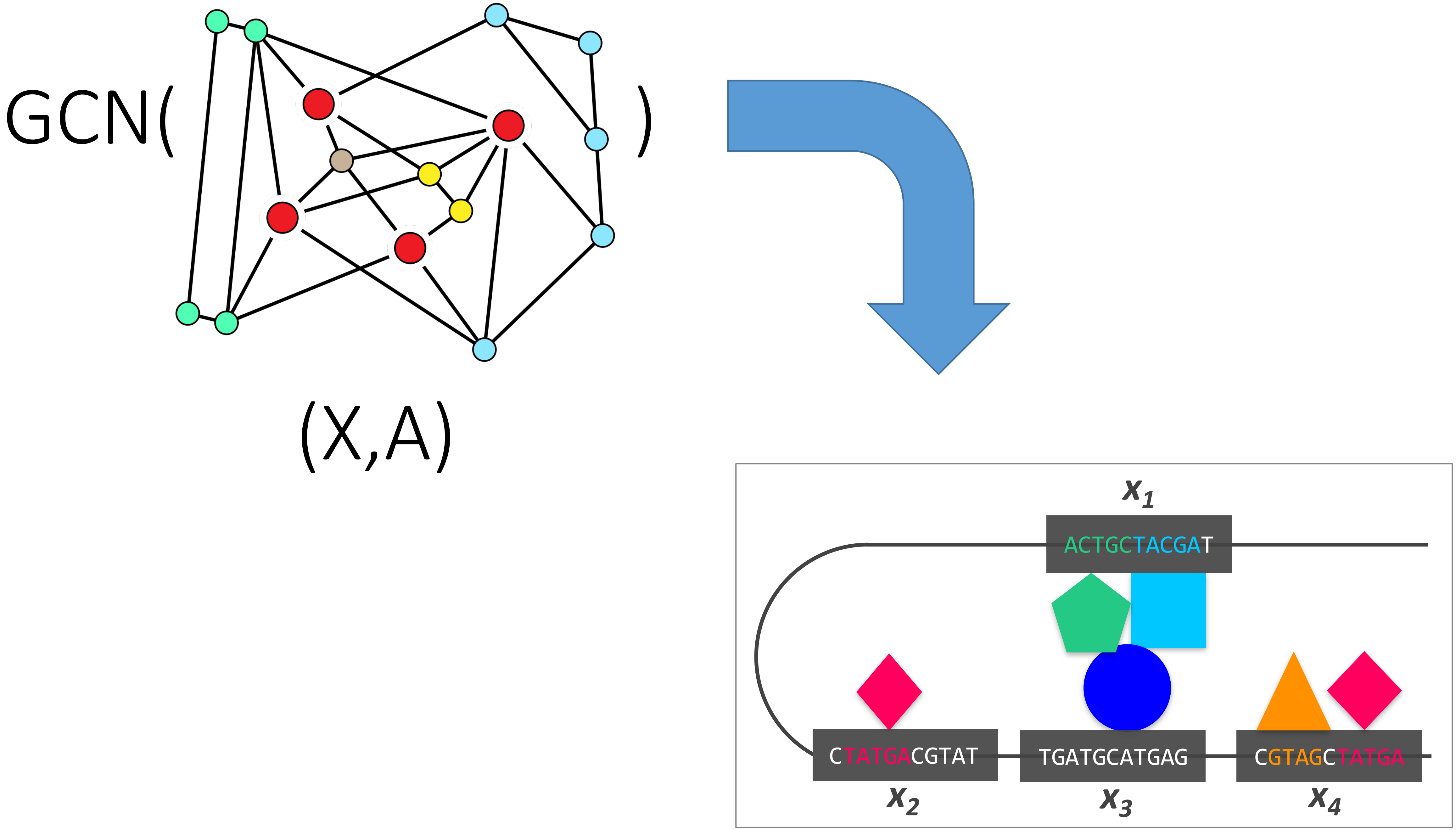
“structural blueprint” indicating interactions that may matter for regulation

Lieberman et al. - Science 2009

ChromeGCN: Combining Sequence and Graph Learning for Chromatin Profile Prediction



ChromeGCN - **Lanchantin** and Qi. Bioinformatics 2020



GCN( )

(X,A)

x_1

ACTGCTACGAT

CTATGACGTAT

TGATGCATGAG

CGTAGCTATGA

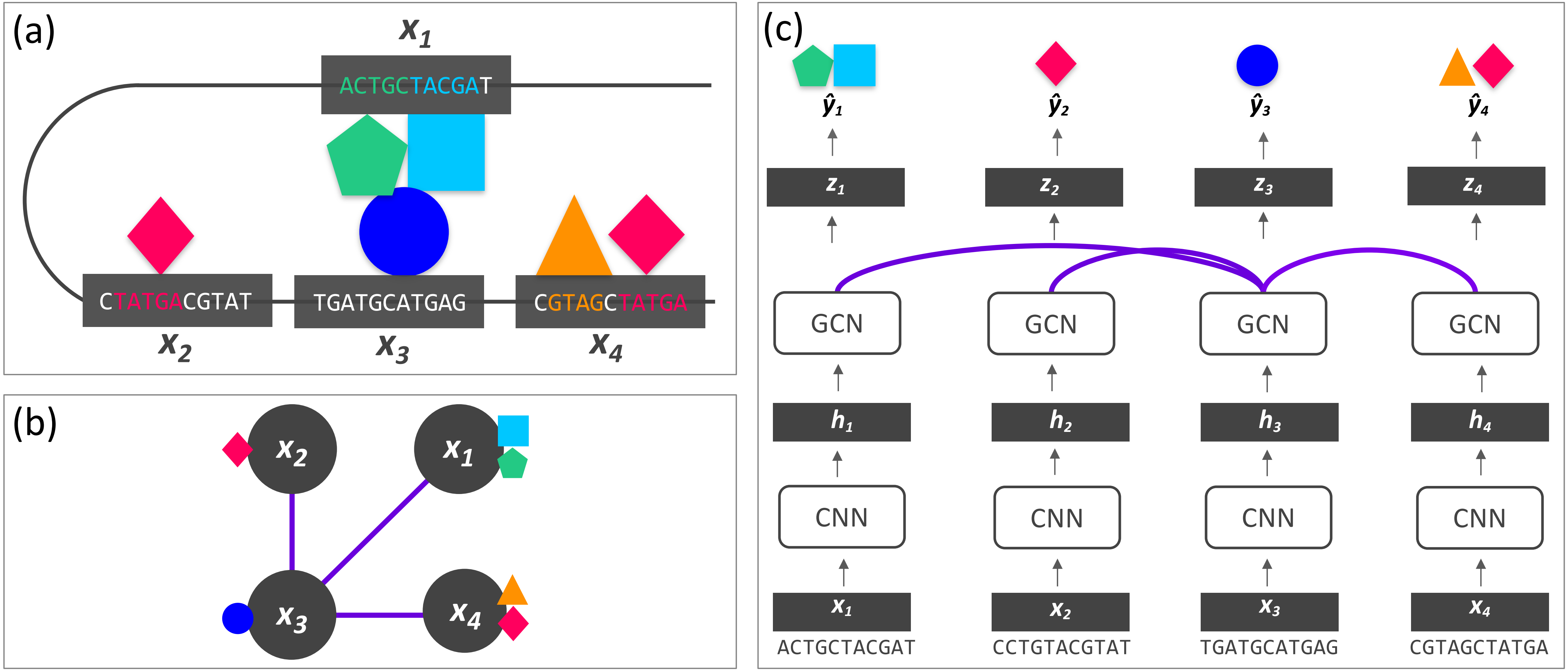
x_2

x_3

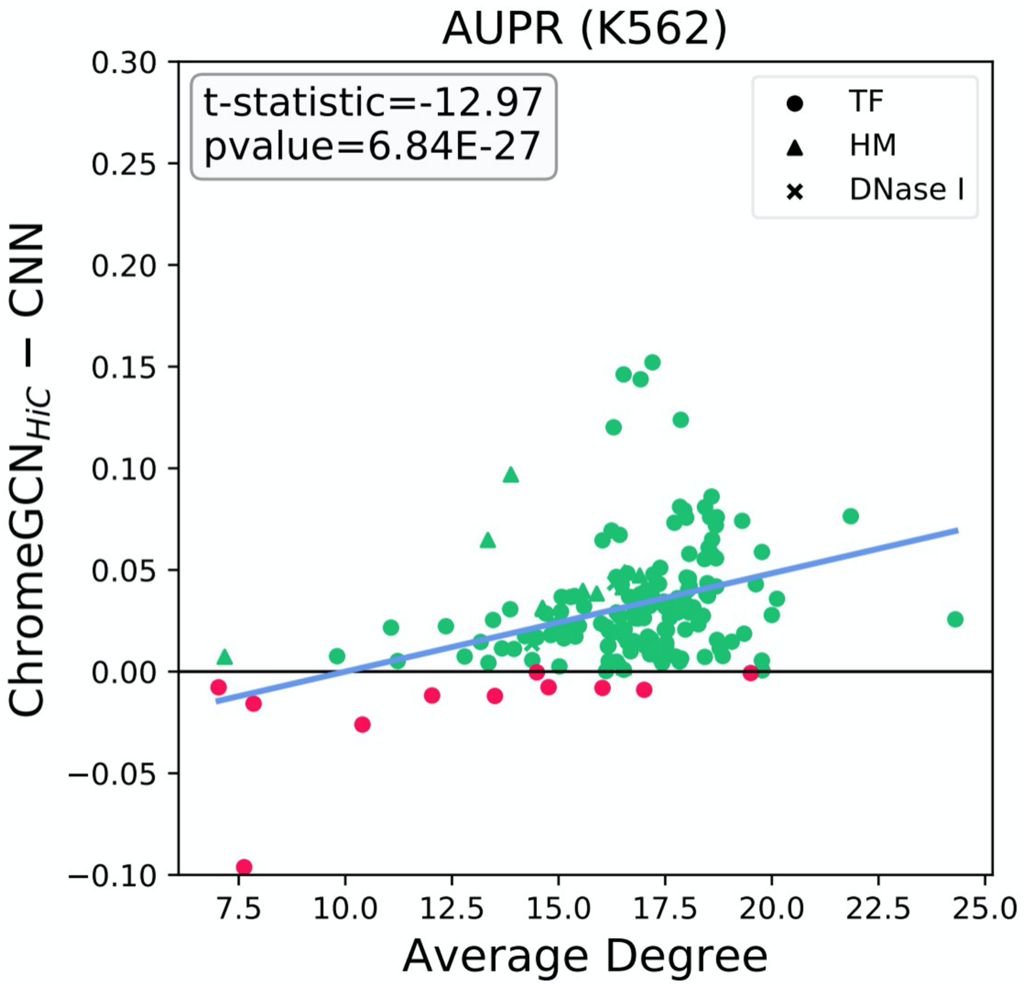
x_4

ChromeGCN: Combining Sequence and Graph Learning for Chromatin Profile Prediction

Graph Convolutional Networks for Epigenetic Activity Prediction Using Both Sequence and 3D Genome



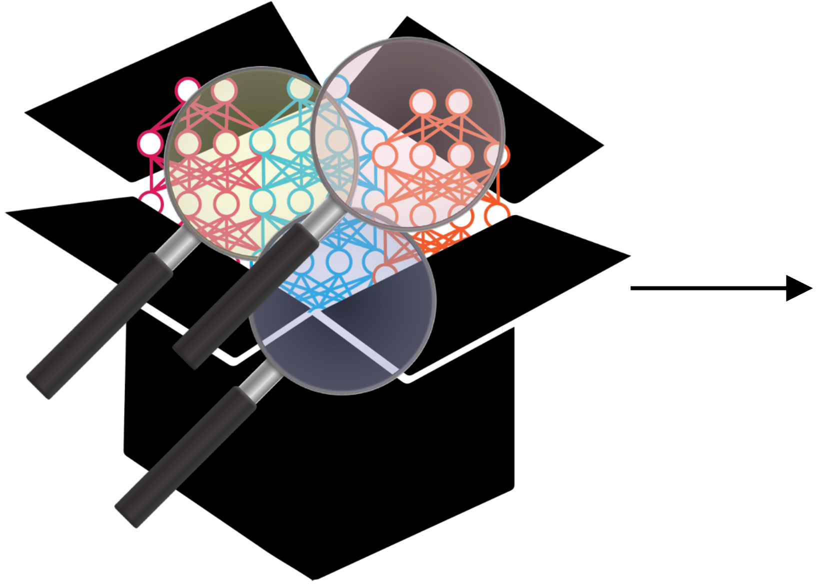
ChromeGCN: Combining Sequence and Graph Learning for Chromatin Profile Prediction

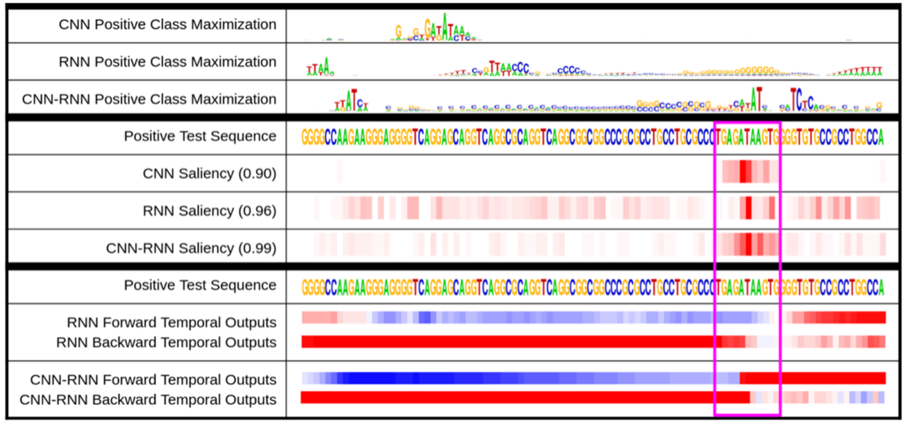


ChromeGCN - **Lanchantin** and Qi. ECCB 2020, Bioinformatics 2020

Understanding by Post Analysis

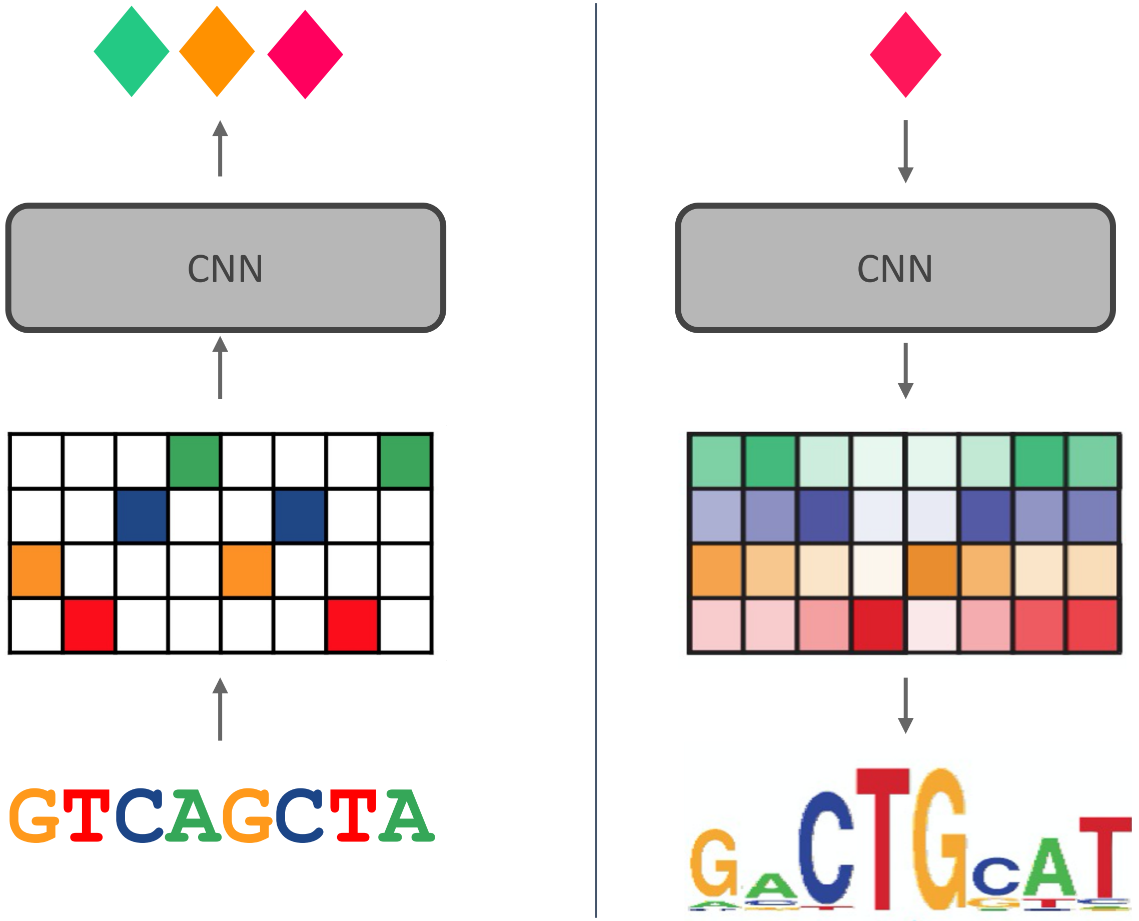
Lanchantin, Singh, Wang & Qi - Pacific Symposium on Biocomputing, 2017




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

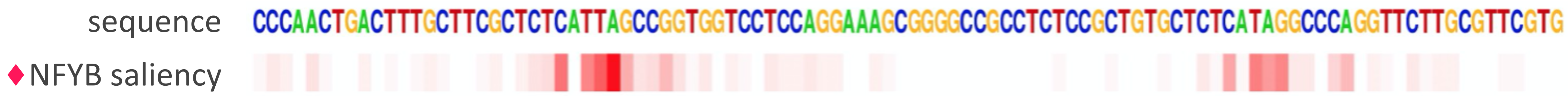
1. Saliency Maps - recommending on CNN kind
2. Class Optimization - recommending on CNN kind
3. Temporal Output Values - recommending on RNN kind

Interpreting Sequence Syntax with Class Optimization



Deep Motif Dashboard - **Lanchantin**, Singh, Wang, Qi. ICLR Workshop Track 2016, PSB 2017

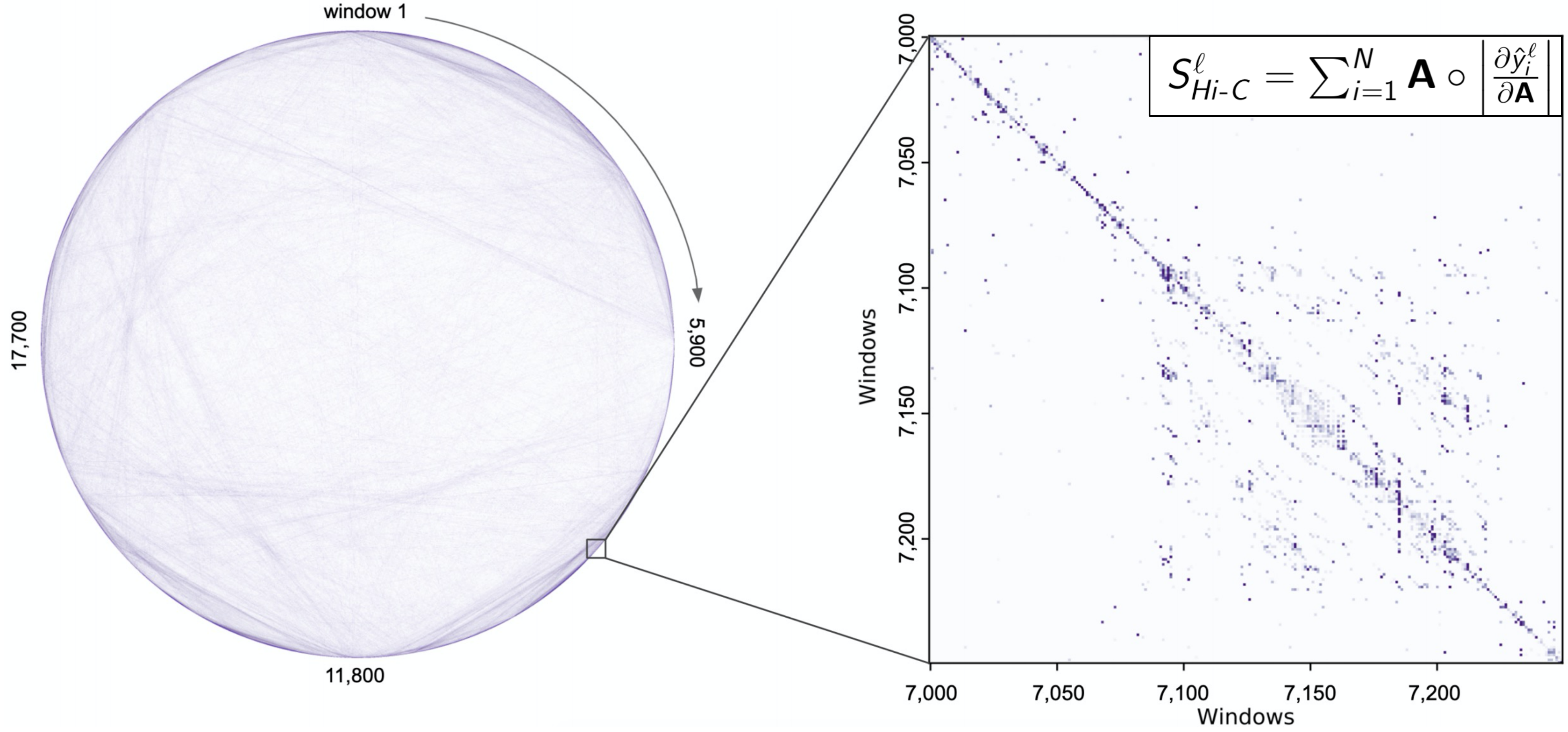
Interpreting Sequence Syntax with Saliency Maps



🟥 = important nucleotide for prediction

Deep Motif Dashboard - **Lanchantin**, Singh, Wang, Qi. ICLR Workshop Track 2016, PSB 2017

Interpreting Long Range Interactions with Hi-C Saliency Maps

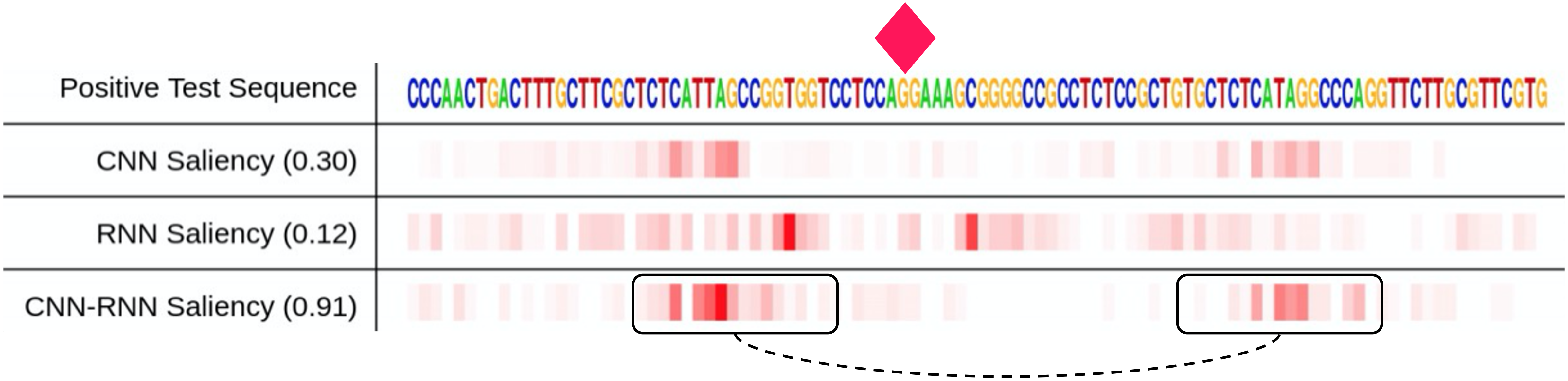


$$S^{\ell}_{Hi\text{-}C} = \sum_{i=1}^{N} \mathbf{A} \circ \left| \frac{\partial \hat{y}_i^{\ell}}{\partial \mathbf{A}} \right|$$

YY1 Hi-C saliency map for all 500k edges in **A**

ChromeGCN - **Lanchantin** and Qi. Bioinformatics 2020

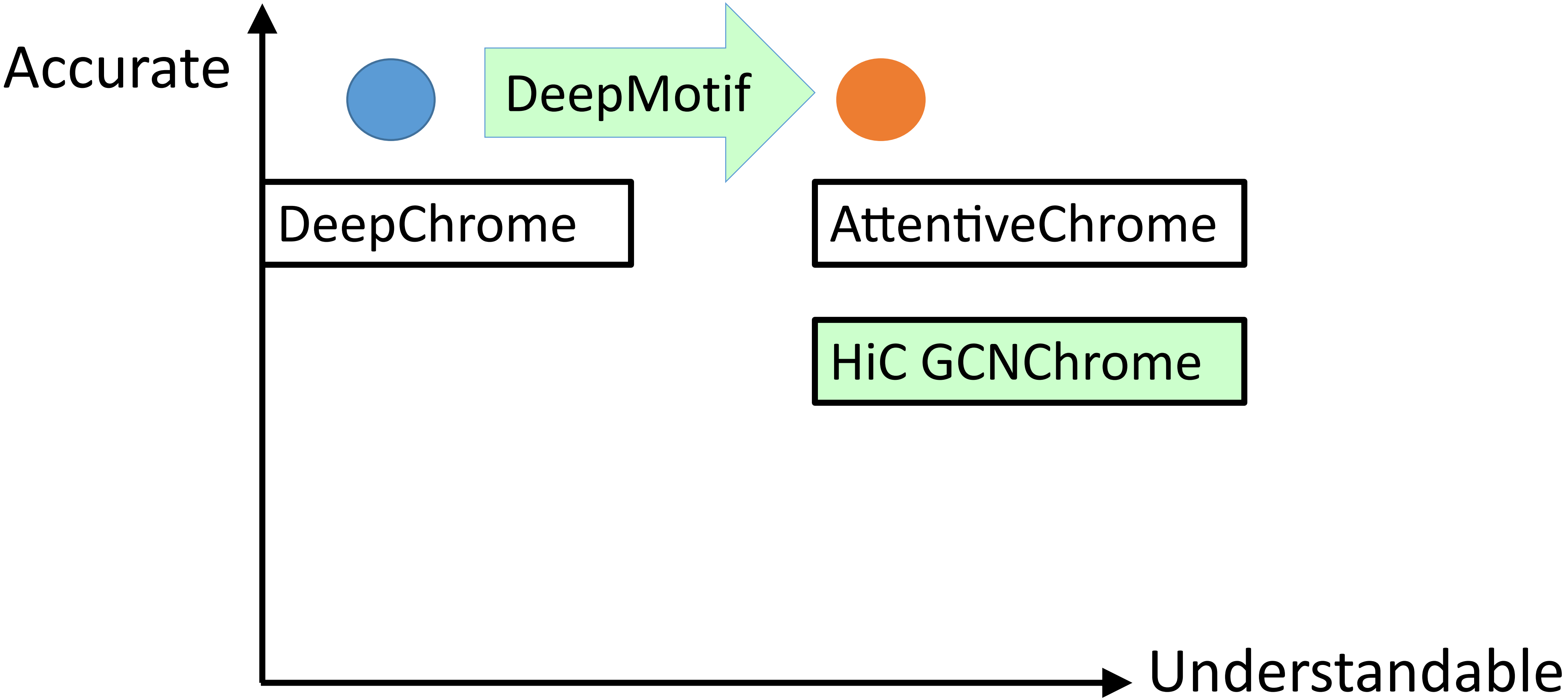
Local sequence interactions



Deep Motif Dashboard - **Lanchantin**, Singh, Wang, Qi. ICLR Workshop Track 2016, PSB 2017

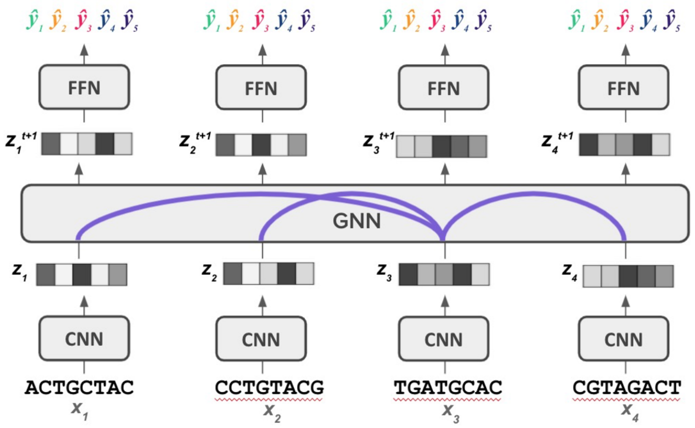
https://www.deepchrome.net

Summary of tools

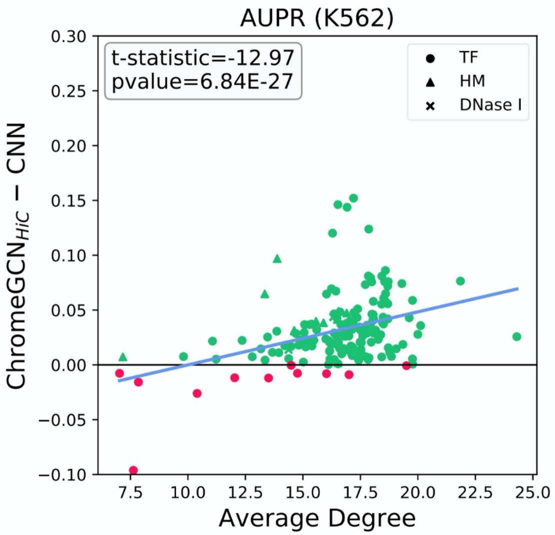


Contributions

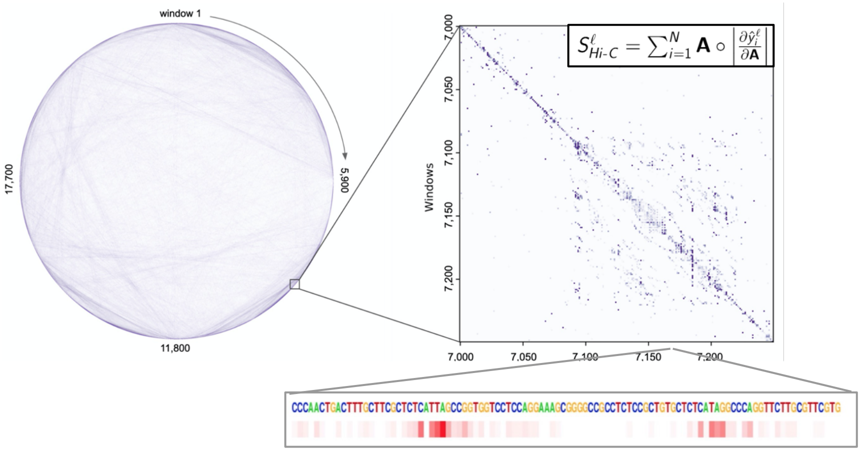
1. Cohesive framework: we **fuse local sequence features and long range interactions** for chromatin profile prediction



2. Accurate: incorporating **long range interactions outperforms** the baselines

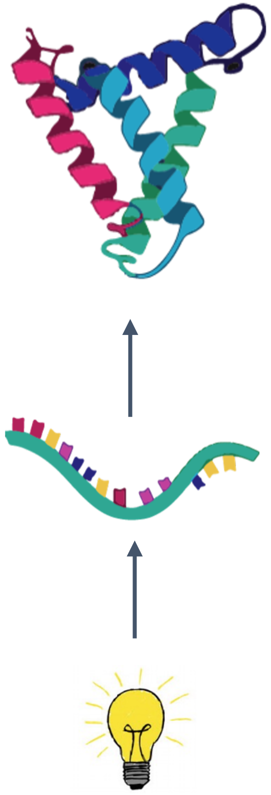


3. Interpretable: we introduce **Hi-C saliency maps** to find important interactions, and **deep motif dashboard** to interpret local features



window 1
5,900
11,800
17,700
$S^{\ell}_{Hi\text{-}C} = \sum_{i=1}^{N} \mathbf{A} \circ \left|\frac{\partial \hat{y}_i^{\ell}}{\partial \mathbf{A}}\right|$
Windows
7,000
7,050
7,100
7,150
7,200
CCCAACTGACTTTGCTTCGCTCTCATTAGCCGGTGGTCCTCCAGGAAAGCGGGGCCGCCTCTCCGCTGTGCTCTCATAGGCCCAGGTTCTTGCGTTCGTG

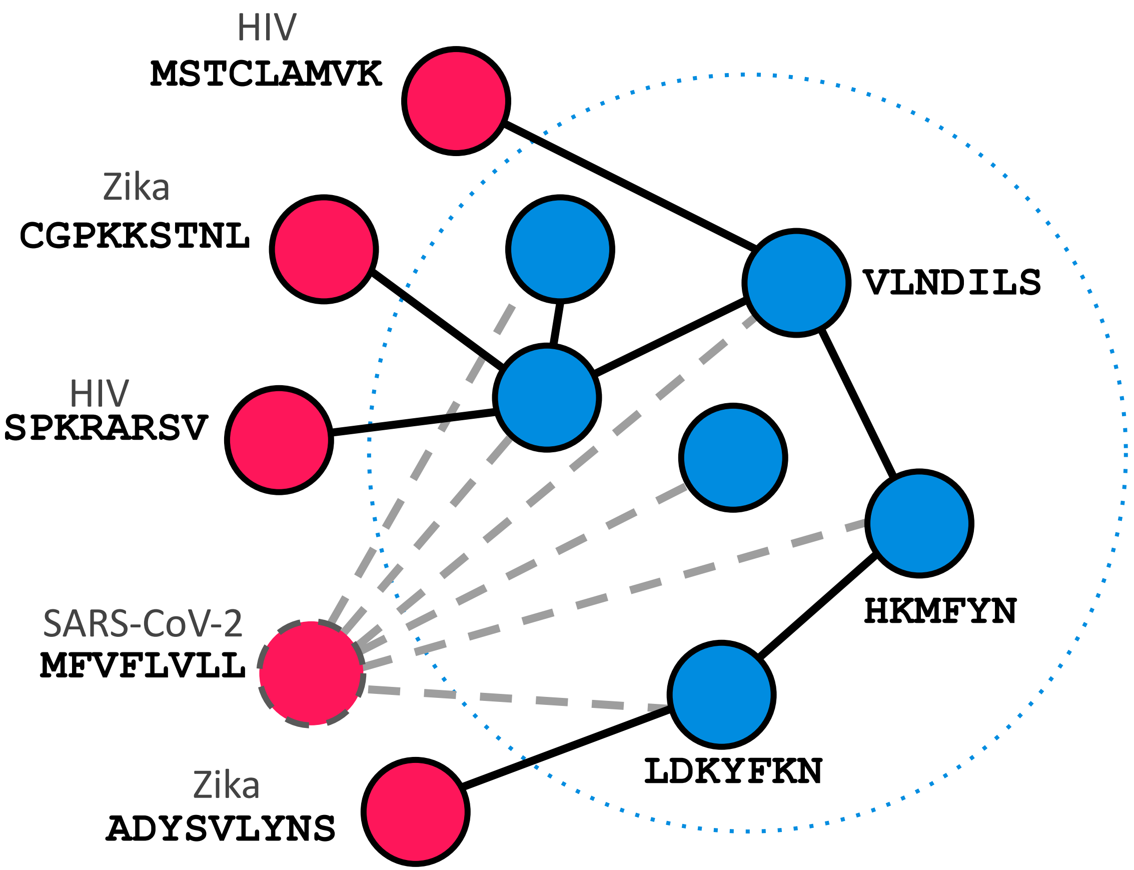
Third Task:

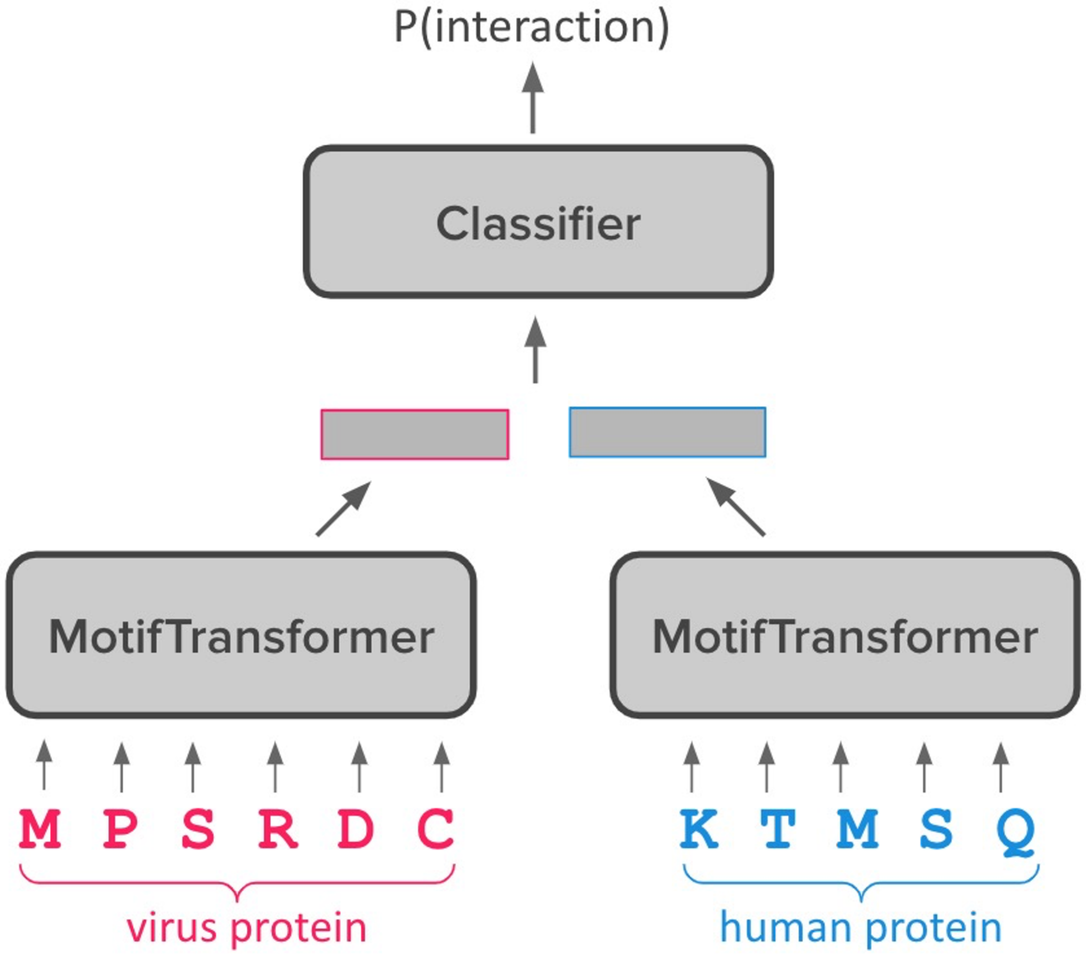


gene expressed

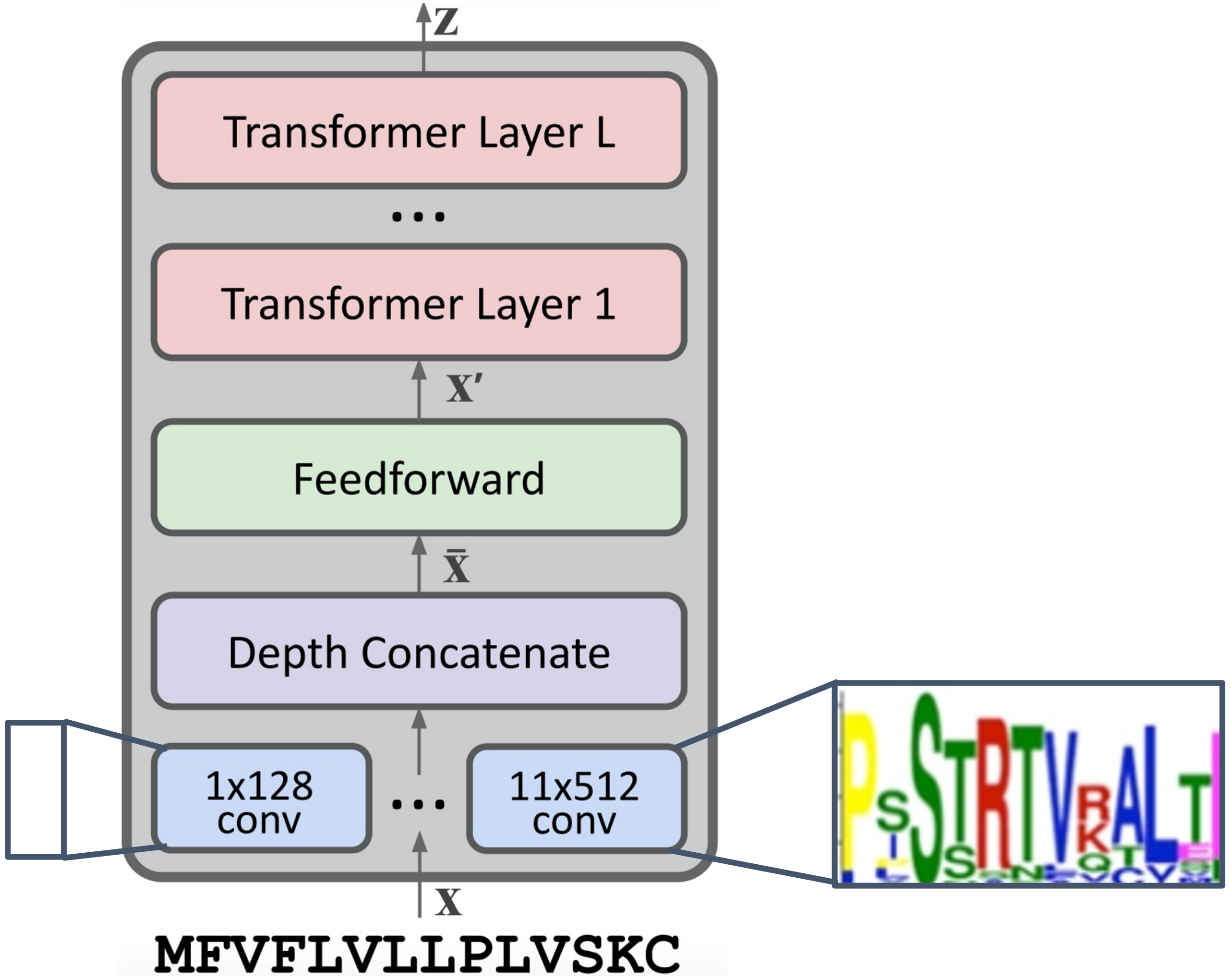
ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

Interaction Prediction





Motif Transformer

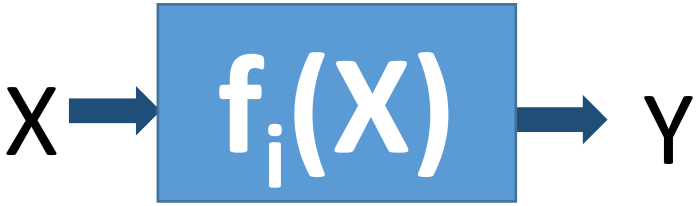


What we have tried: *Using Deep Learning to Read the Genome, Epigenome and Proteome*

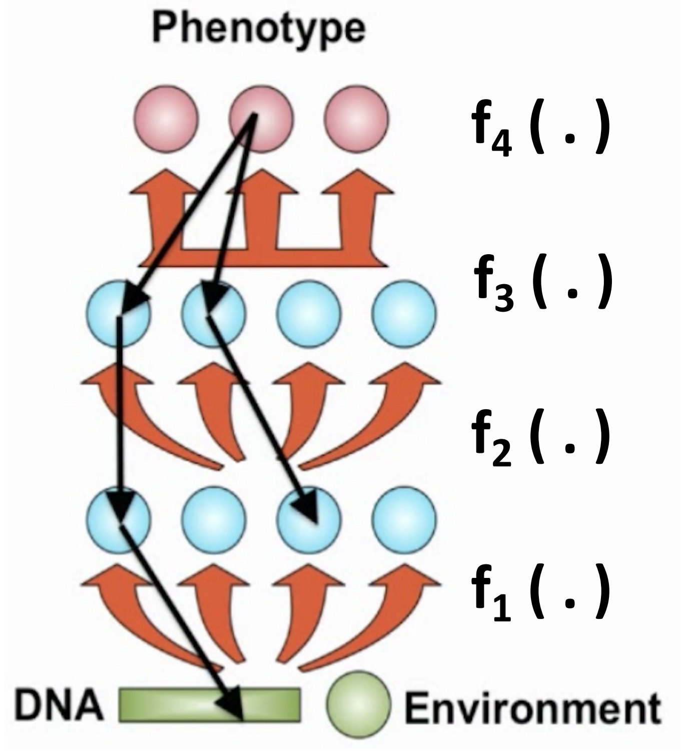
1. Deep Learning module to reflect biological modules

2. Compose modules to reflect biology

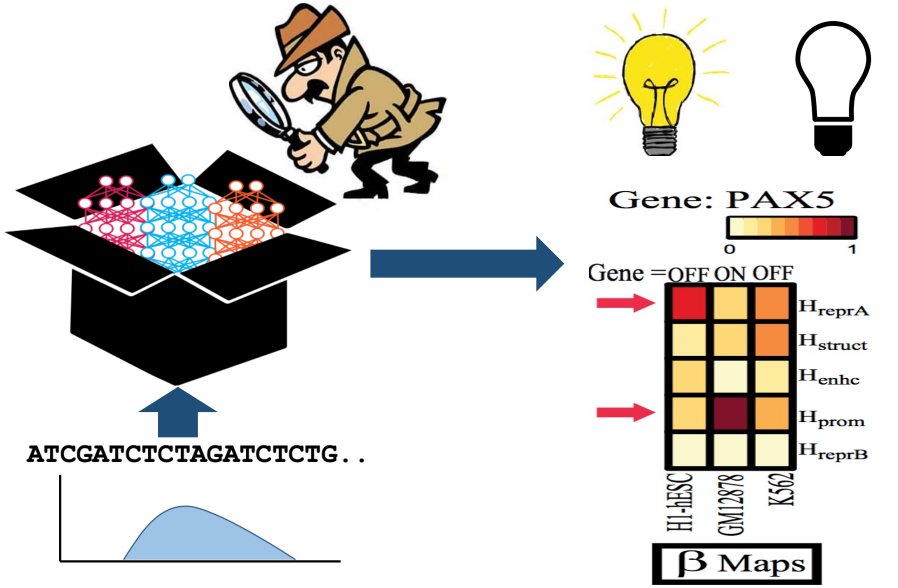
3. Open DNN black-box and provide domain explanations

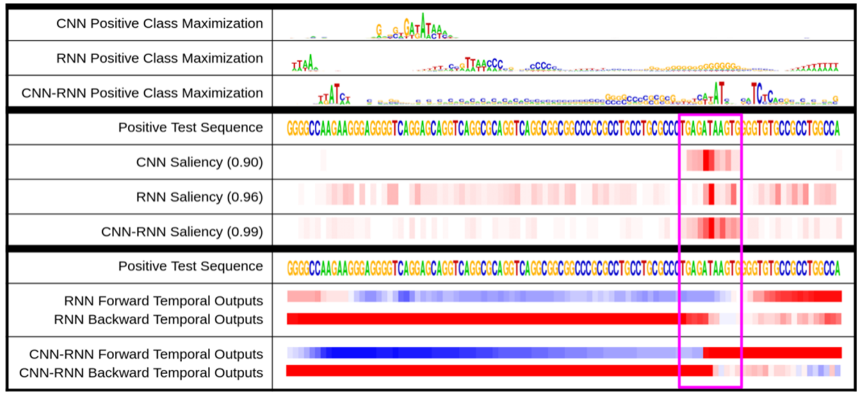


X	Y
DNA	RNA / Func
Epigenetic	RNA
DNA	Interaction to Protein (TF)
Protein	Funcs
Protein	Interaction to DNA/RNA
…	…



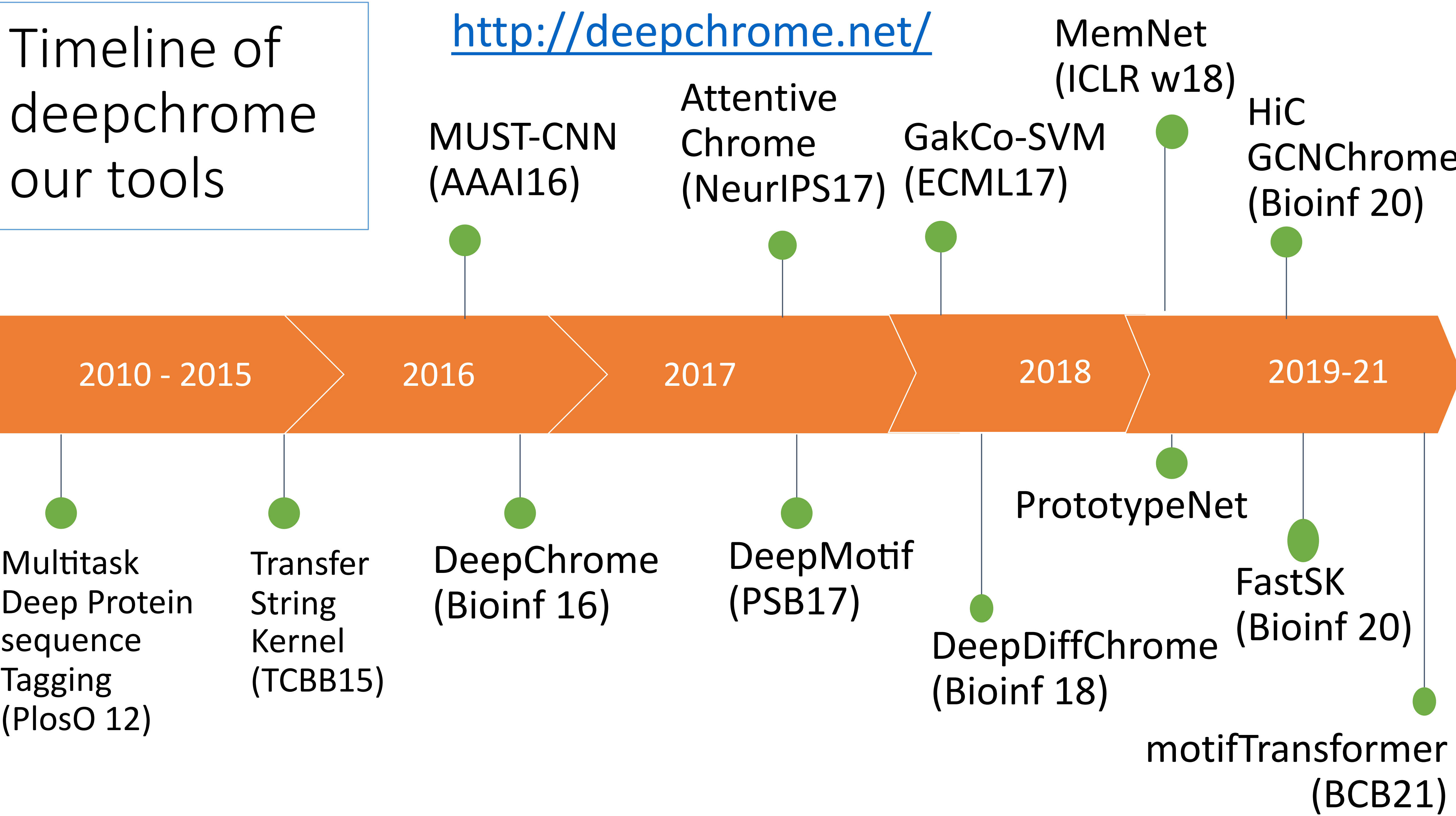
http://deepchrome.net





Timeline of deepchrome our tools

http://deepchrome.net/



2010 - 2015

Multitask Deep Protein sequence Tagging (PlosO 12)

Transfer String Kernel (TCBB15)

2016

MUST-CNN (AAAI16)

DeepChrome (Bioinf 16)

2017

Attentive Chrome (NeurIPS17)

DeepMotif (PSB17)

GakCo-SVM (ECML17)

2018

MemNet (ICLR w18)

DeepDiffChrome (Bioinf 18)

PrototypeNet

2019-21

HiC GCNChrome (Bioinf 20)

FastSK (Bioinf 20)

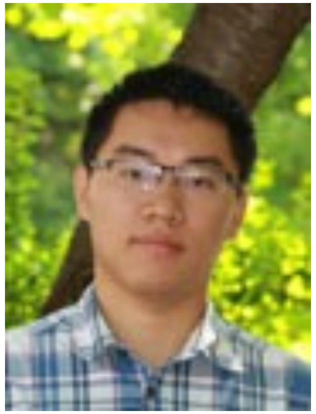
motifTransformer (BCB21)

Acknowledgements

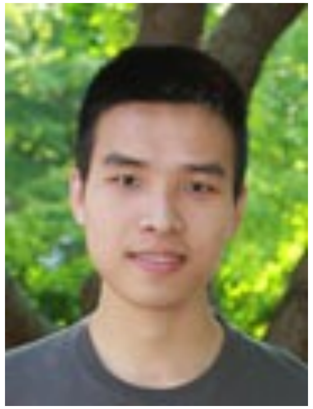
Ritambhara Singh
Now Assistant
Professor @Brown

Jack Lanchantin
Now @ Facebook
Research

Arshdeep Sekhon



Beilun Wang
Now Associate
Professor @
Southeastern Univ.

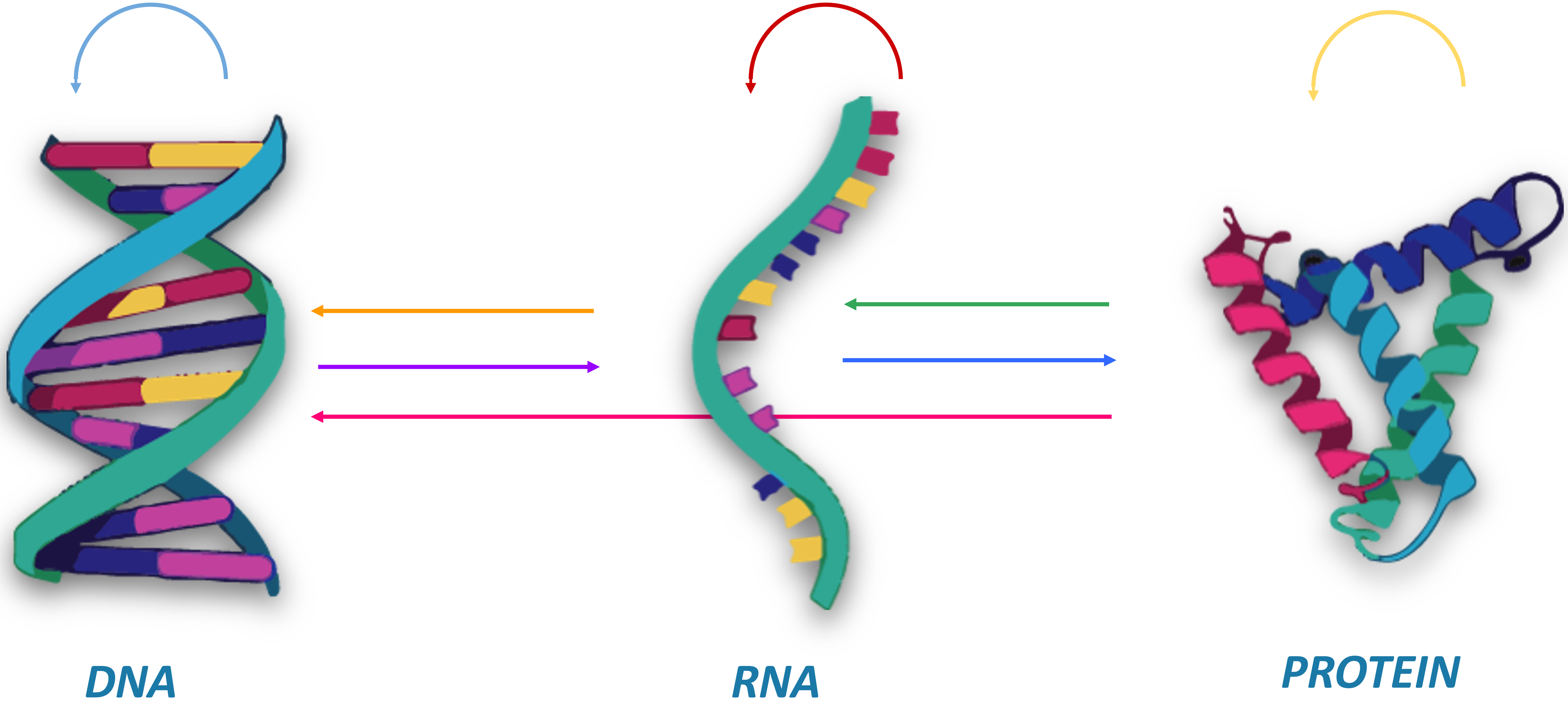


Weilin Xu, Now
Research Staff @
Intel Labs

UVA Department of Biochemistry and Molecular Genetics: Dr. Mazhar Adli

Thank you

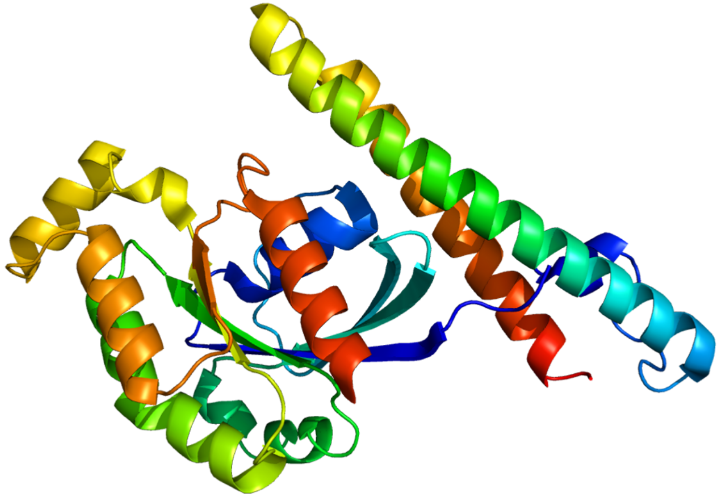
What we have tried: ***Using Deep Learning to Read the Genome, Epigenome and Proteome***



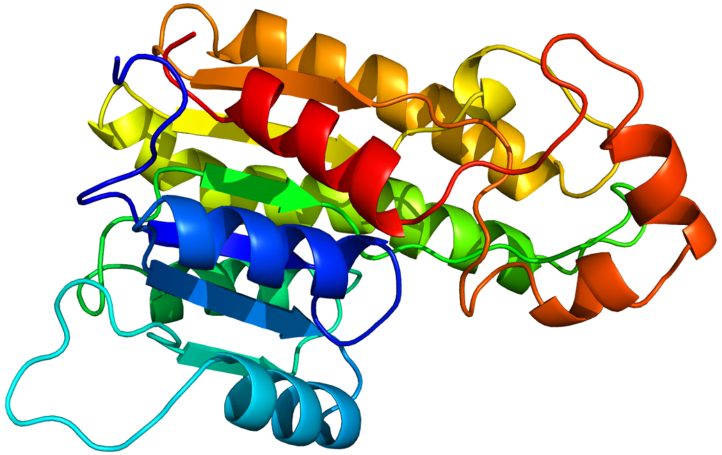
Third Task:

gene expressed
ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

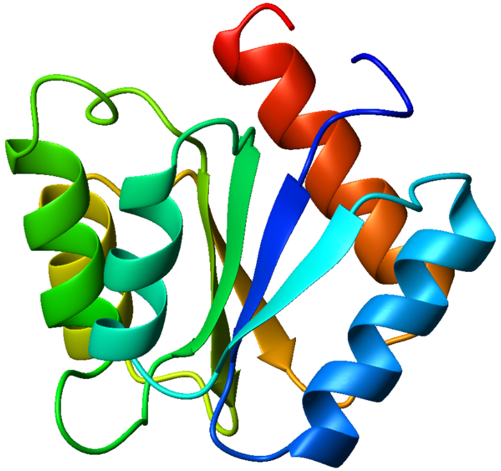
Proteins: the building blocks of life



oxygen transportation



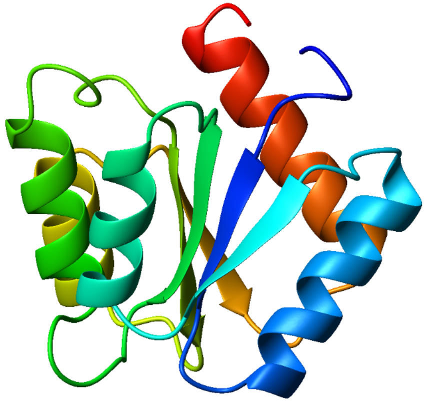
antibodies



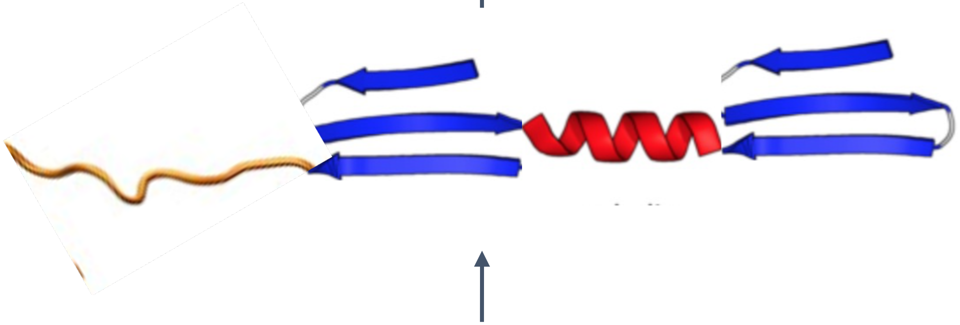
digestive enzymes

Protein Structures

Tertiary Structure



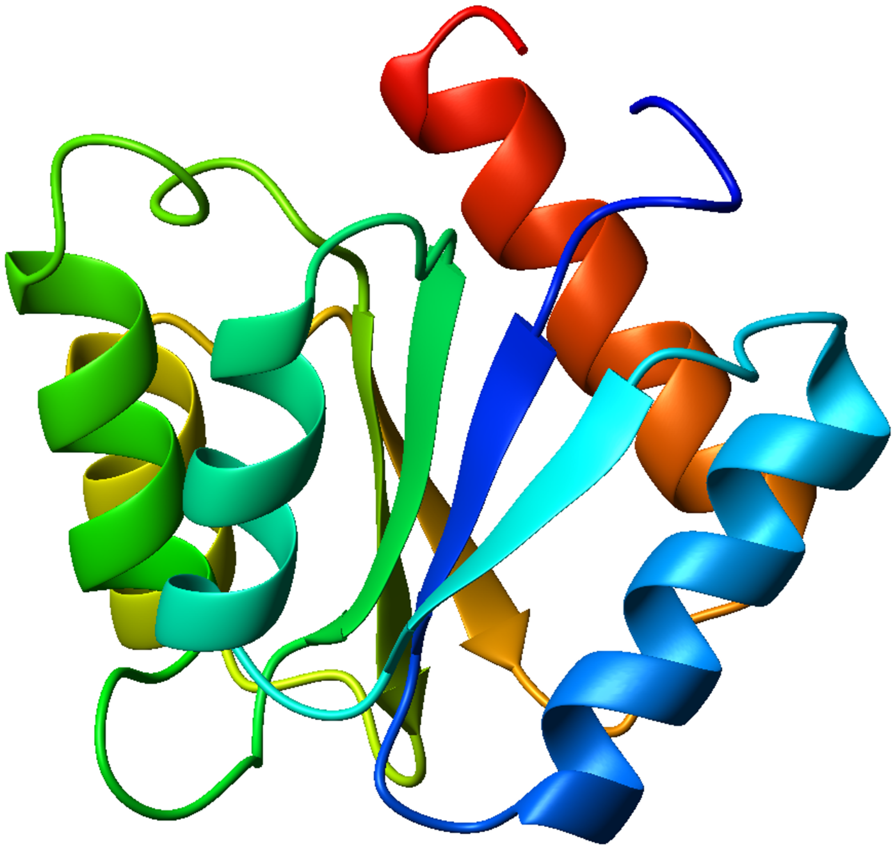
Secondary Structure



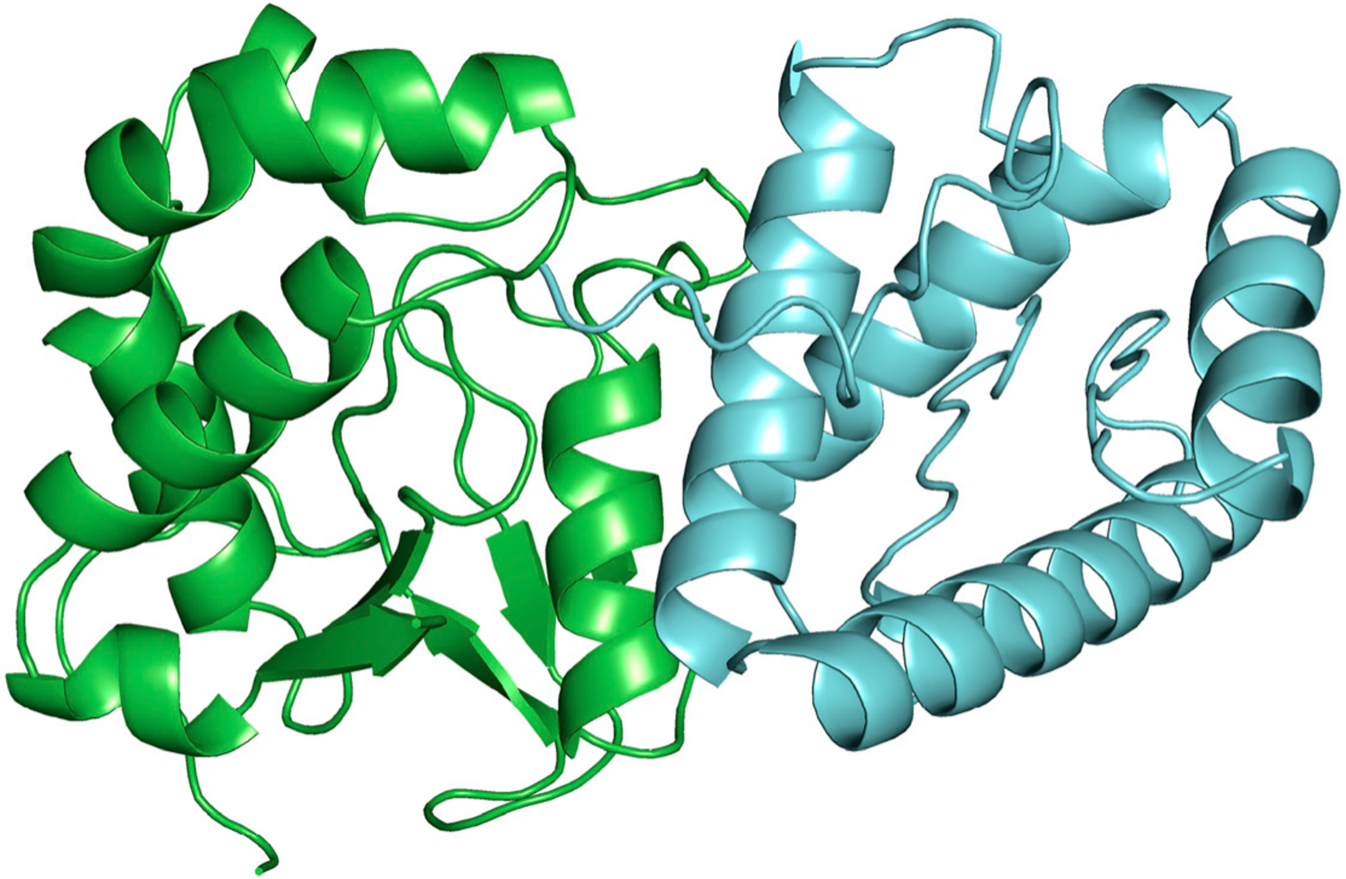
Primary Sequence

MHFTEDKATILWGKVNVEGETLGRVYPWQ

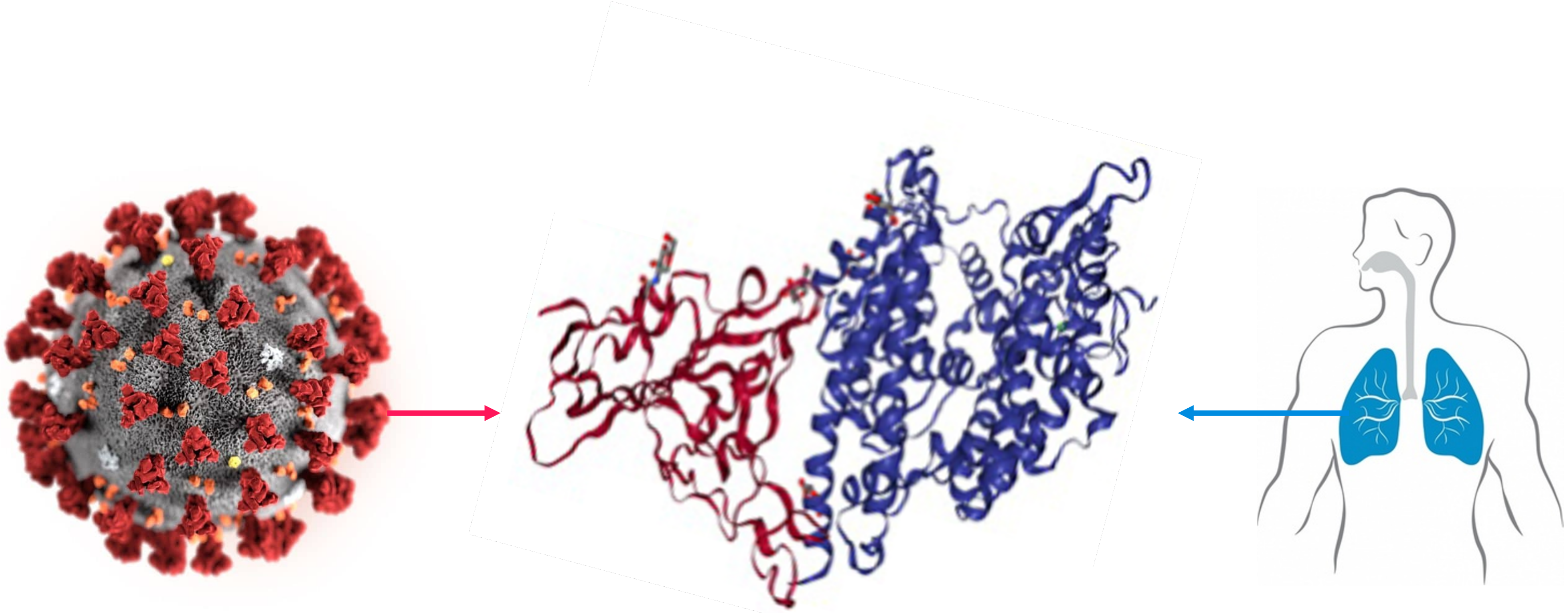
Structure Determines Function



One primary function: Protein-Protein Interactions (PPIs)



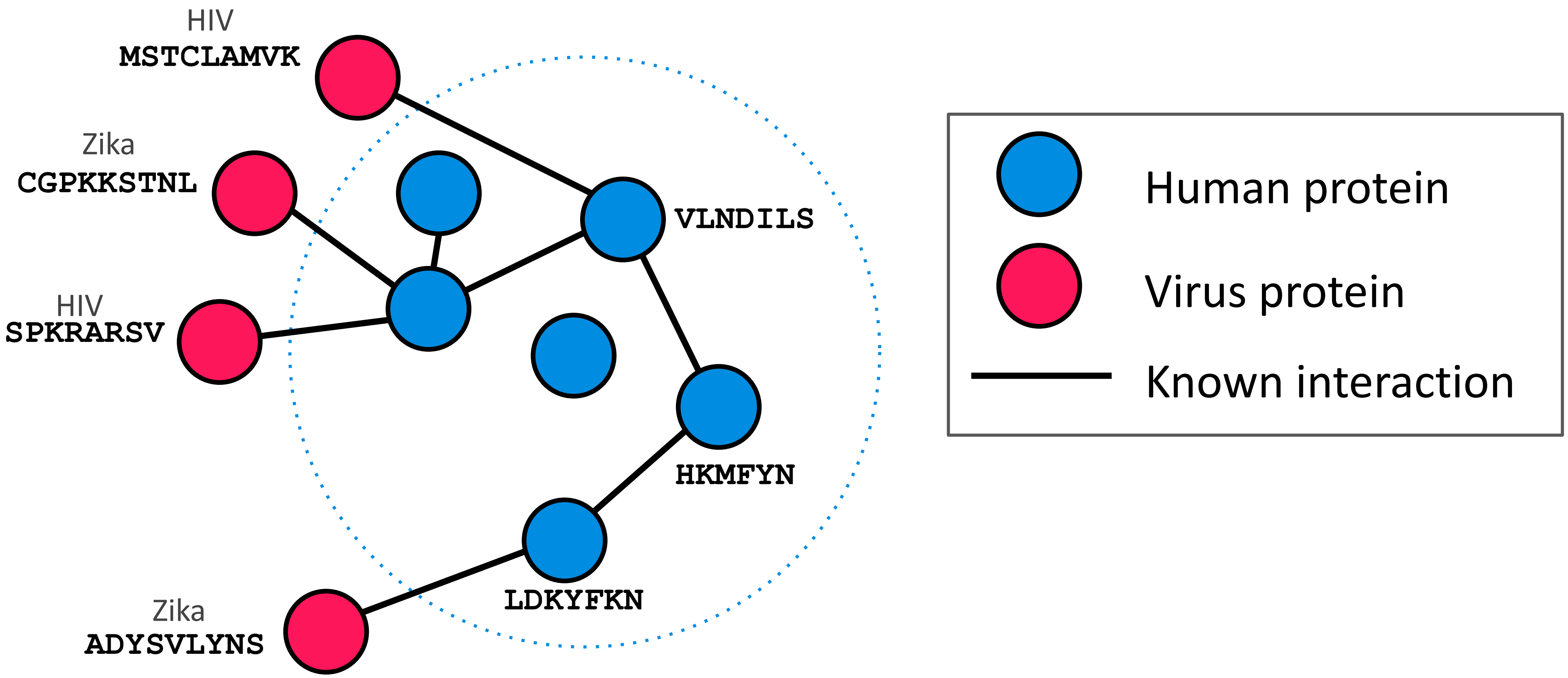
Our Task: To Discover Human-Virus Protein-Protein Interactions



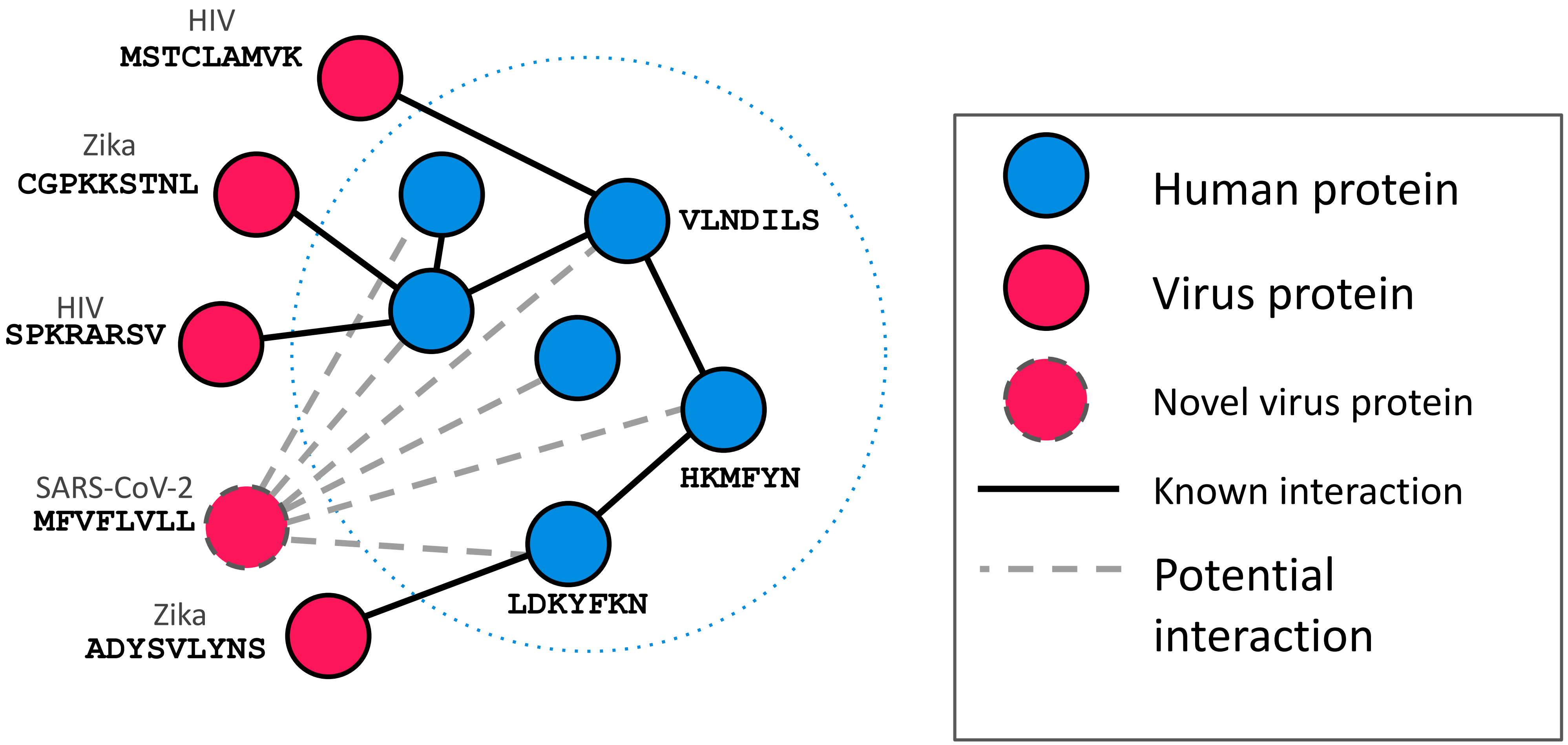
SARS-CoV-2
Spike Protein

Human
ACE2 Protein

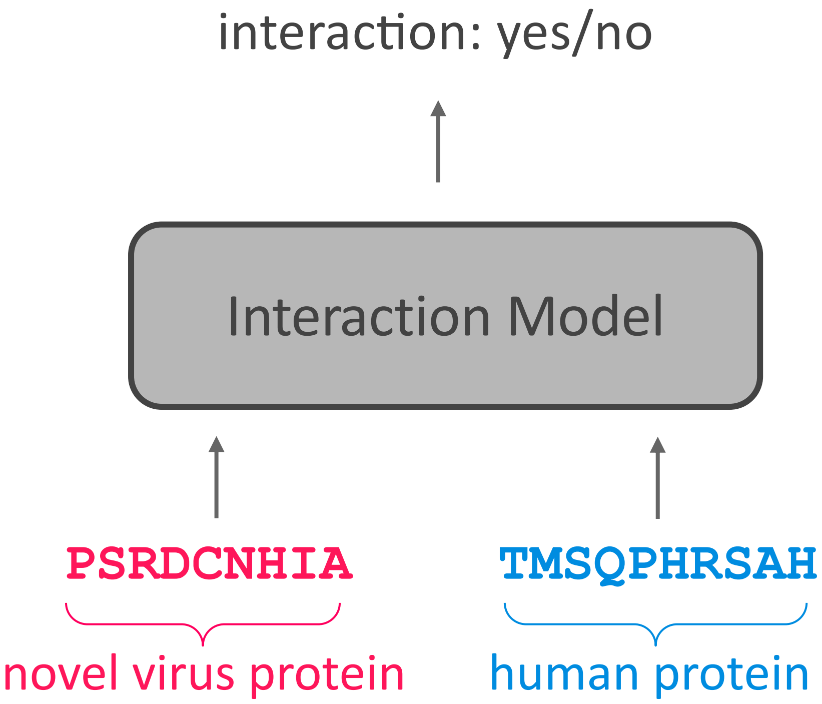
Human-Virus Protein-Protein Interactions



Human-Virus Protein-Protein Interactions



Novel Virus-Human Protein Interaction Prediction from Sequence



interaction: yes/no

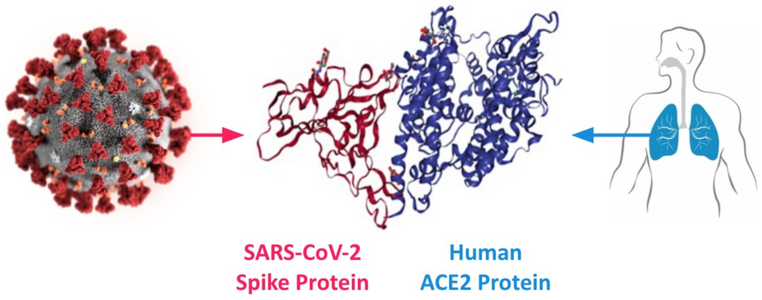
Interaction Model

PSRDCNHIA

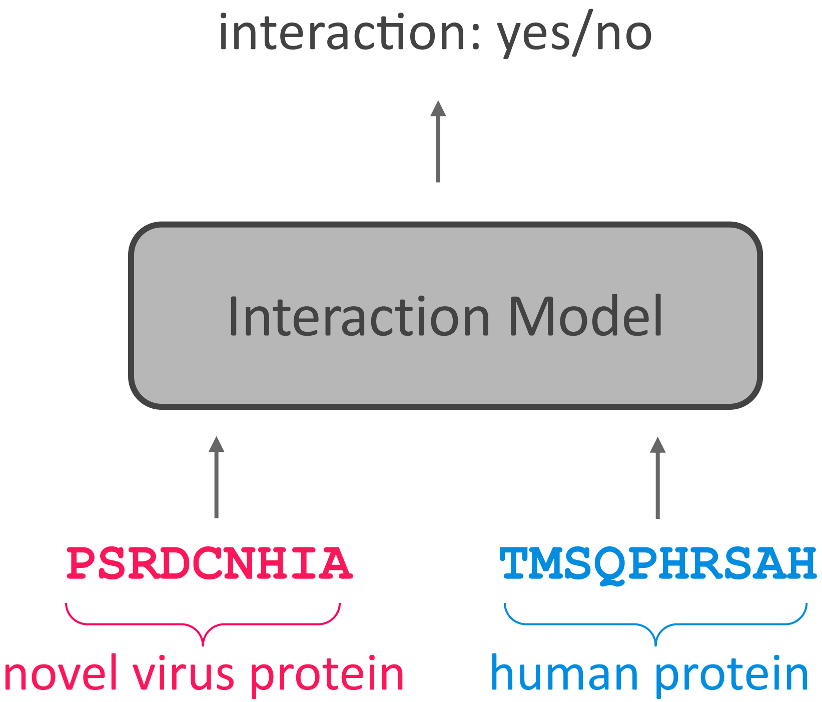
novel virus protein

TMSQPHRSAH

human protein

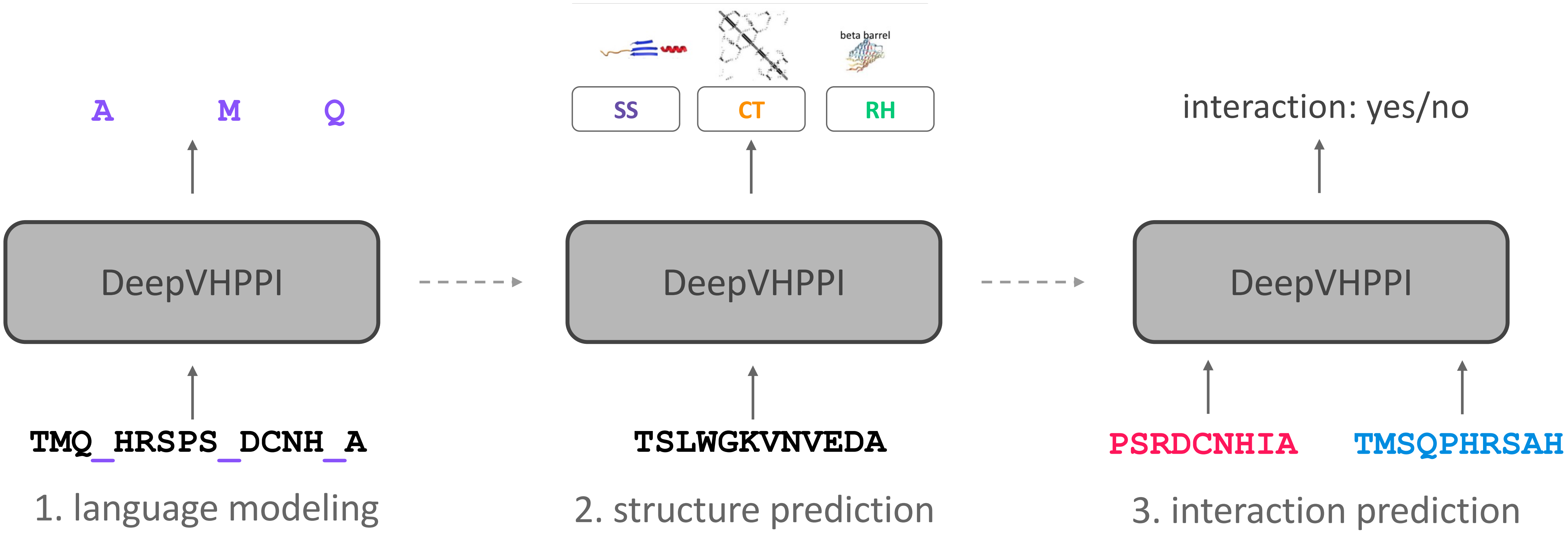


Novel Virus-Human Protein Interaction Prediction from Sequence



1. Limited interaction data available
2. Interactions are largely determined by structure

Transfer Learning for Sequence-Based Interaction Prediction



A M Q

DeepVHPPI

TMQ_HRSPS_DCNH_A

1. language modeling

SS CT RH

DeepVHPPI

TSLWGKVNVEDA

2. structure prediction

interaction: yes/no

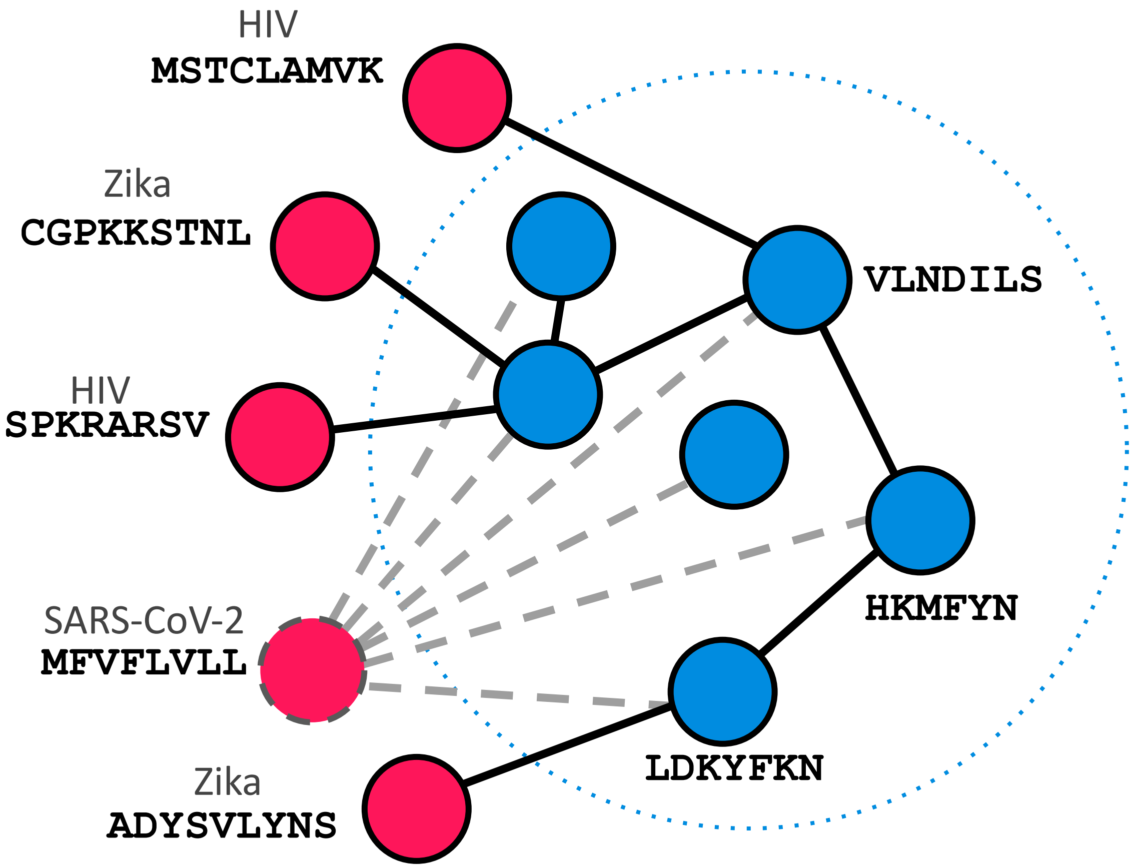
DeepVHPPI

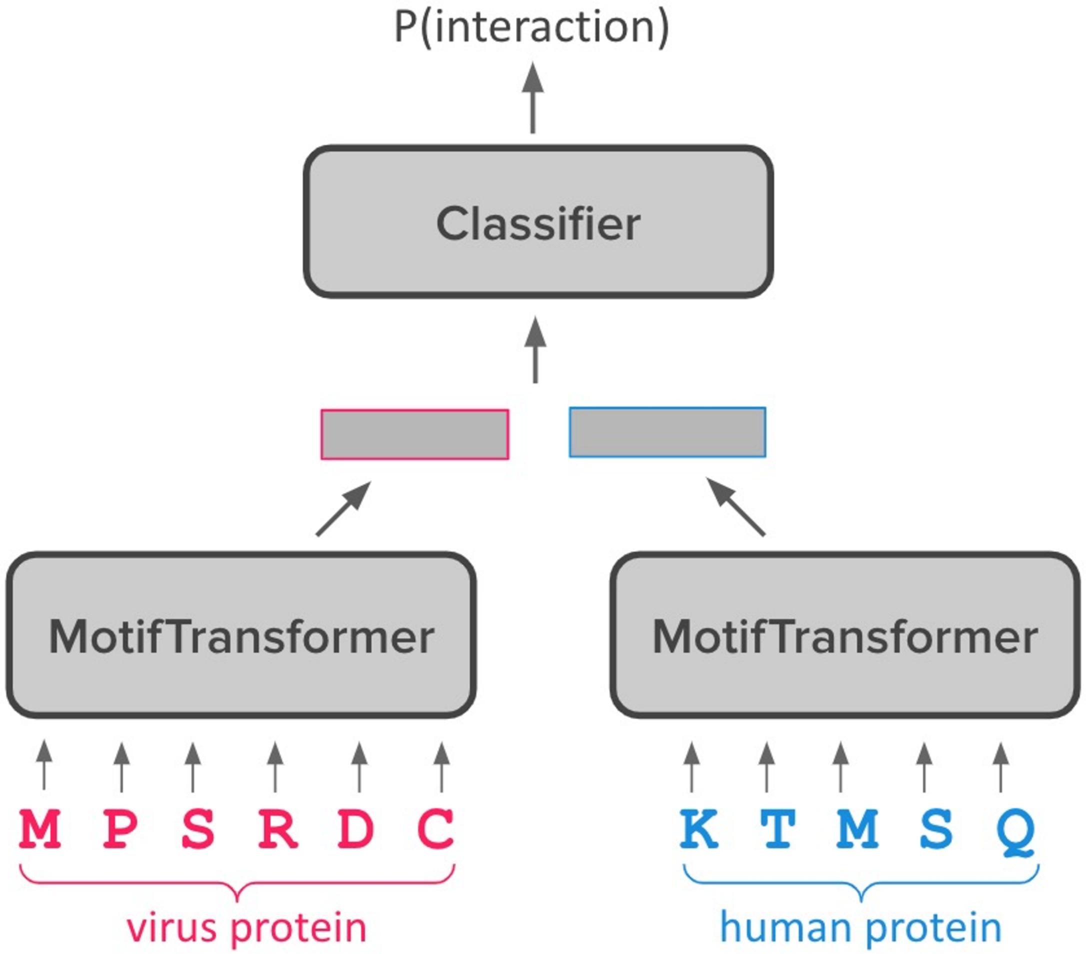
PSRDCNHIA TMSQPHRSAH

3. interaction prediction

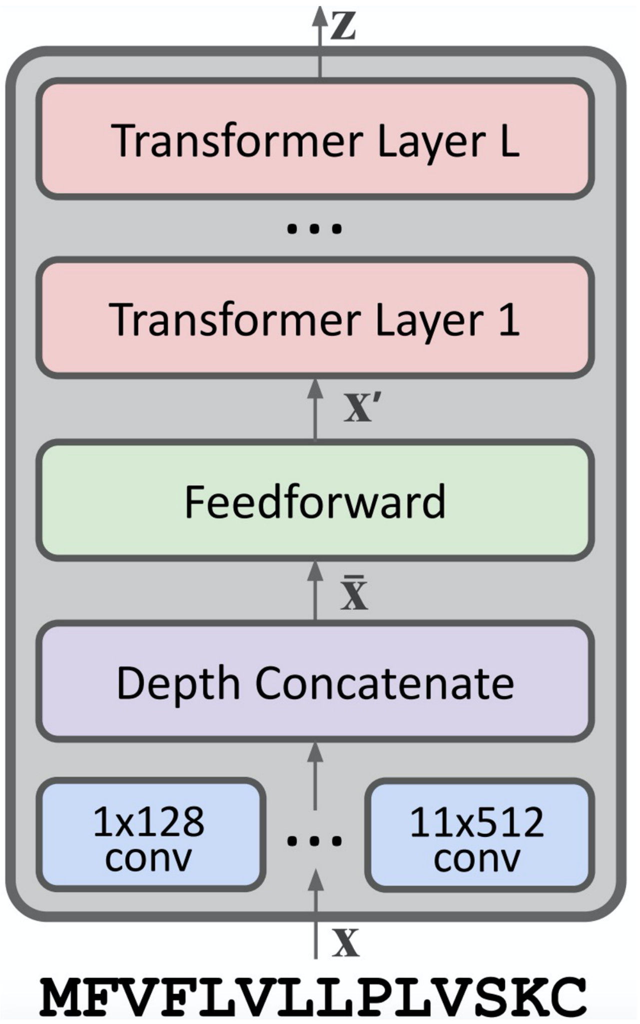
Lanchantin, Weingarten, Sekhon, Miller, Qi - ACM BCB 2021

Interaction Prediction

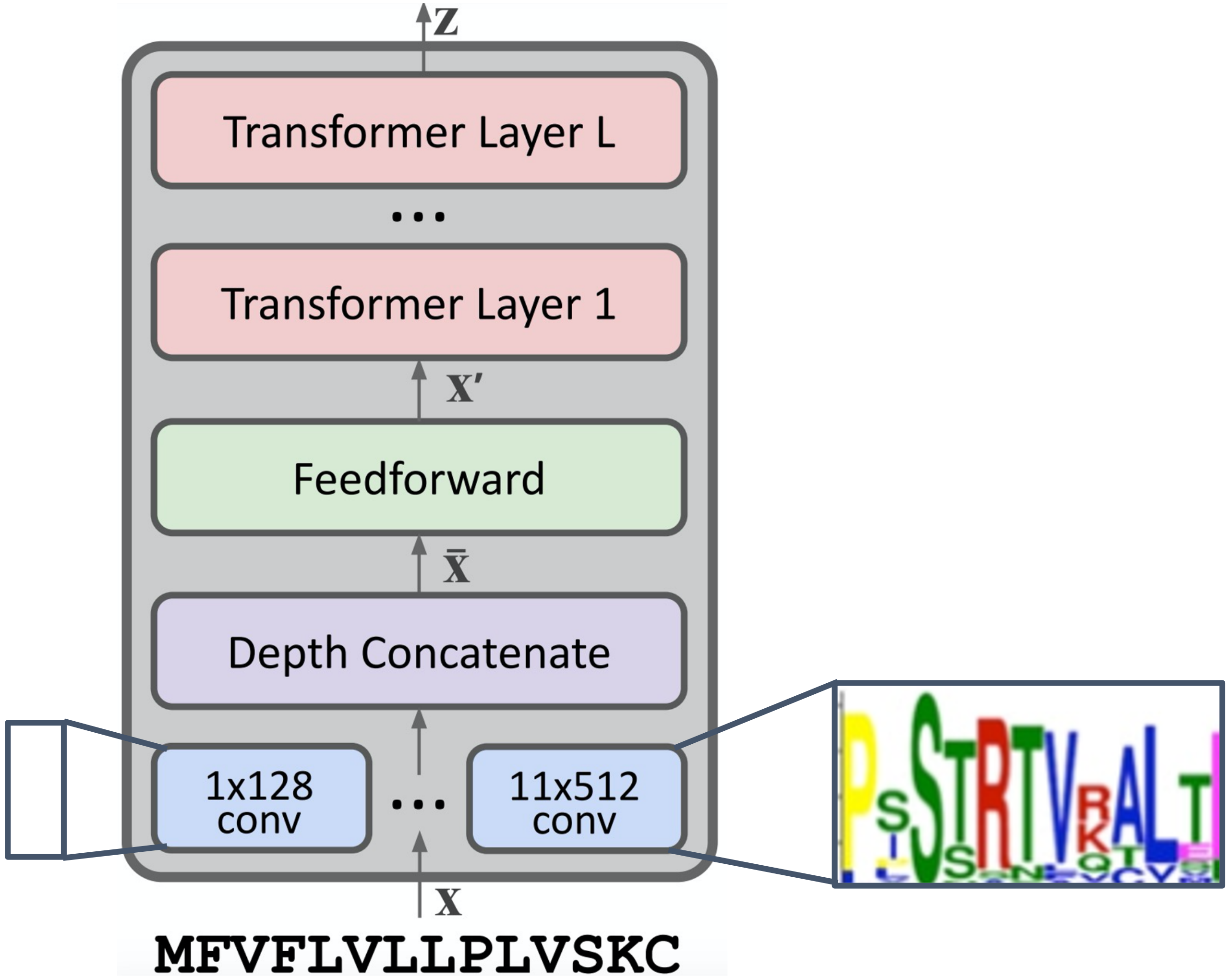




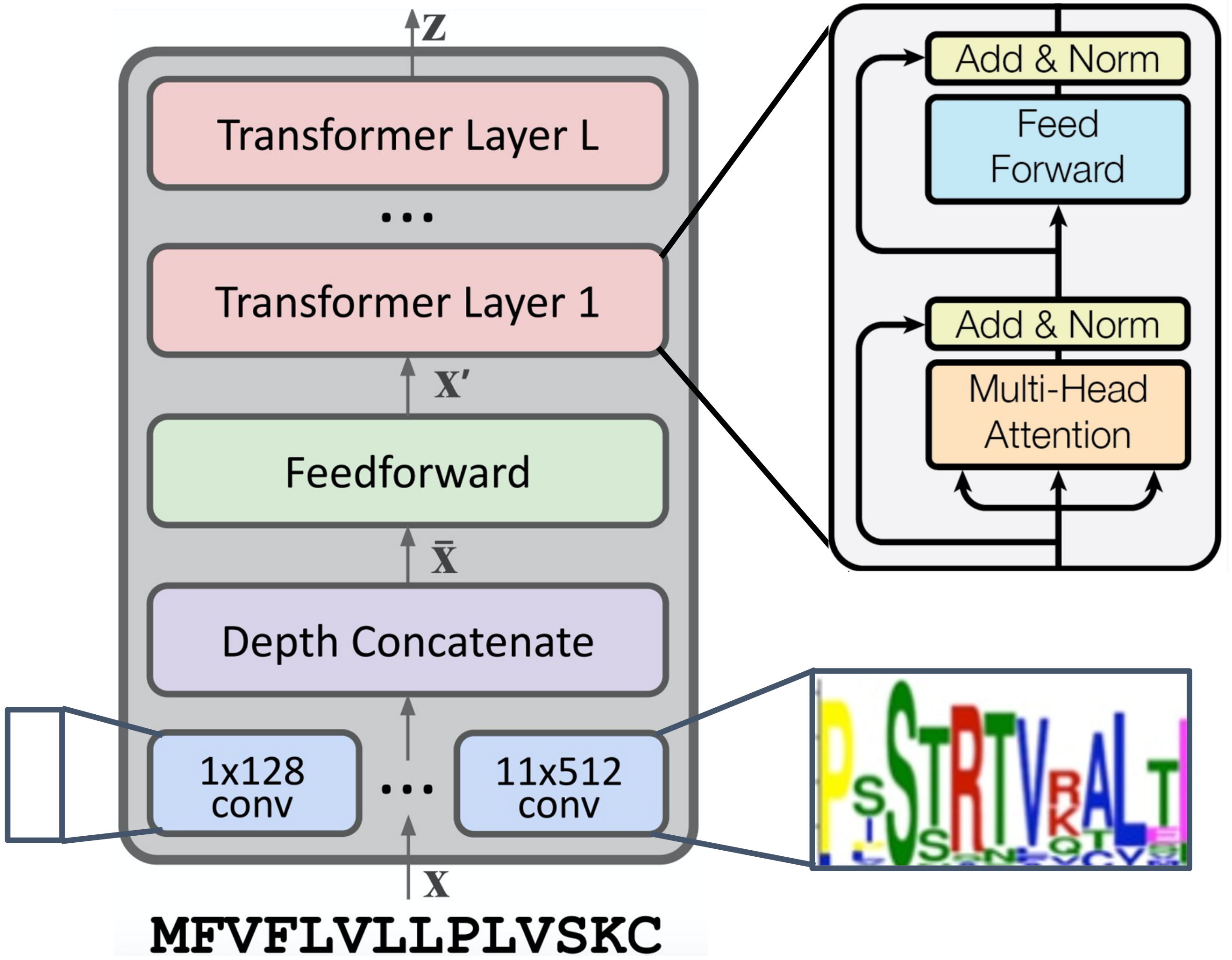
Motif Transformer



Motif Transformer



Motif Transformer



Experimental Setup

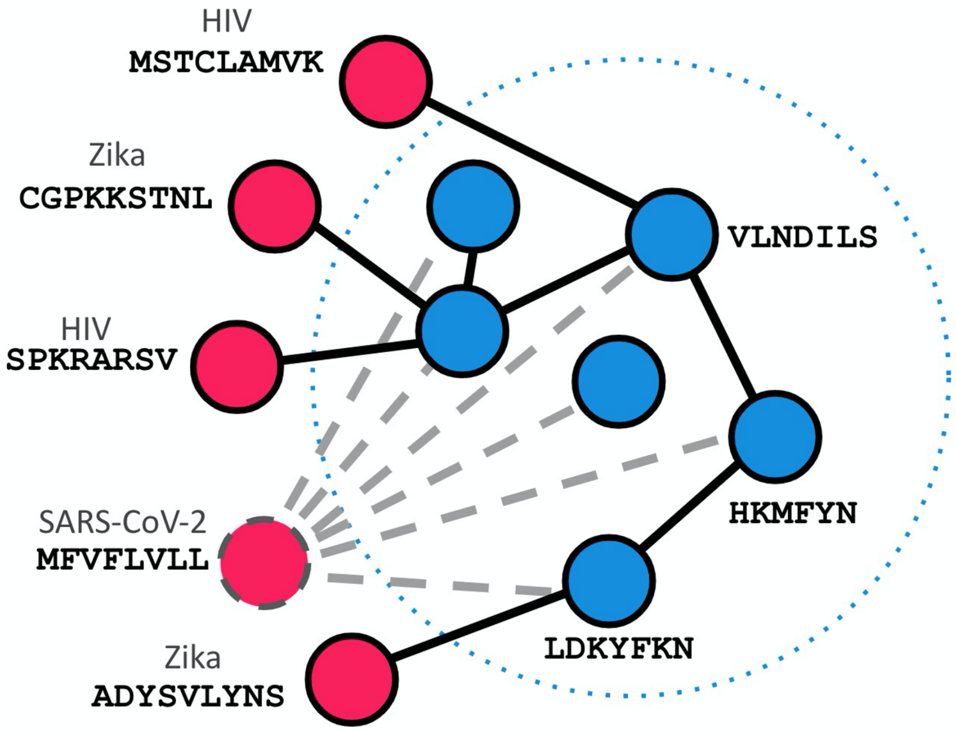
- **Training Data: HPIDB 3.0 Dataset**
  - 22,000 positive interactions, 226,000 negative interactions
  - 1,100k virus proteins, 20,000 host (human) proteins
- **Testing Data:**
  - HIV, Ebola interactions from Zhou et al.
  - Our own curated SARS-CoV-2 interactions collected from BioGrid

Protein-Protein Interaction Prediction Results

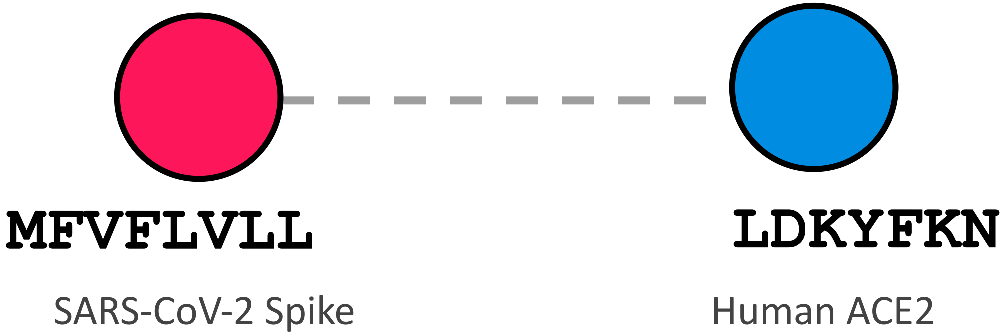
	H1N1		Ebola		SARS-CoV-2	
Method	AUROC	F1	AUROC	F1	AUROC	F1
SVM (Zhou et al.)	0.886	0.762	0.867	0.760	-	-
Embedding + RF (Yang et al)	-	-	-	-	0.748	**0.115**
MotifTransformer	0.894	0.819	0.927	0.836	0.726	0.089
MotifTransformer + LM	0.910	0.837	0.943	0.867	0.735	0.095
MotifTransformer + LM + SP	**0.926**	**0.848**	**0.979**	**0.895**	**0.767**	0.105

Use Cases of Sequence Based Interaction Predictors

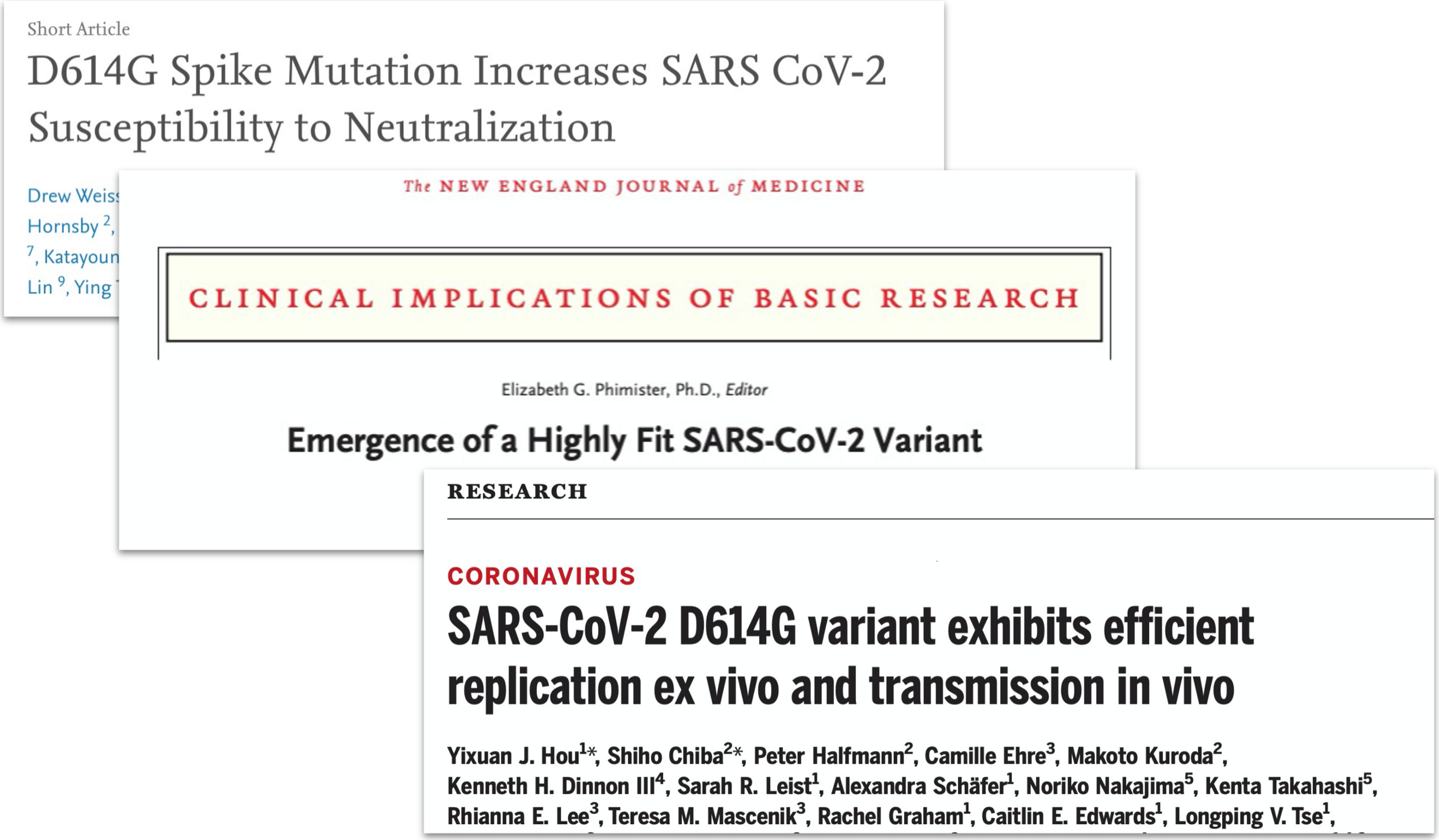
1. predict novel virus interactions



2. analyze how mutations affect interactions



Perturbation Analysis: Investigating Mutations



Short Article

D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization

Drew Weiss
Hornsby [2],
[7], Katayoun
Lin [9], Ying

The NEW ENGLAND JOURNAL of MEDICINE

CLINICAL IMPLICATIONS OF BASIC RESEARCH

Elizabeth G. Phimister, Ph.D., *Editor*

Emergence of a Highly Fit SARS-CoV-2 Variant

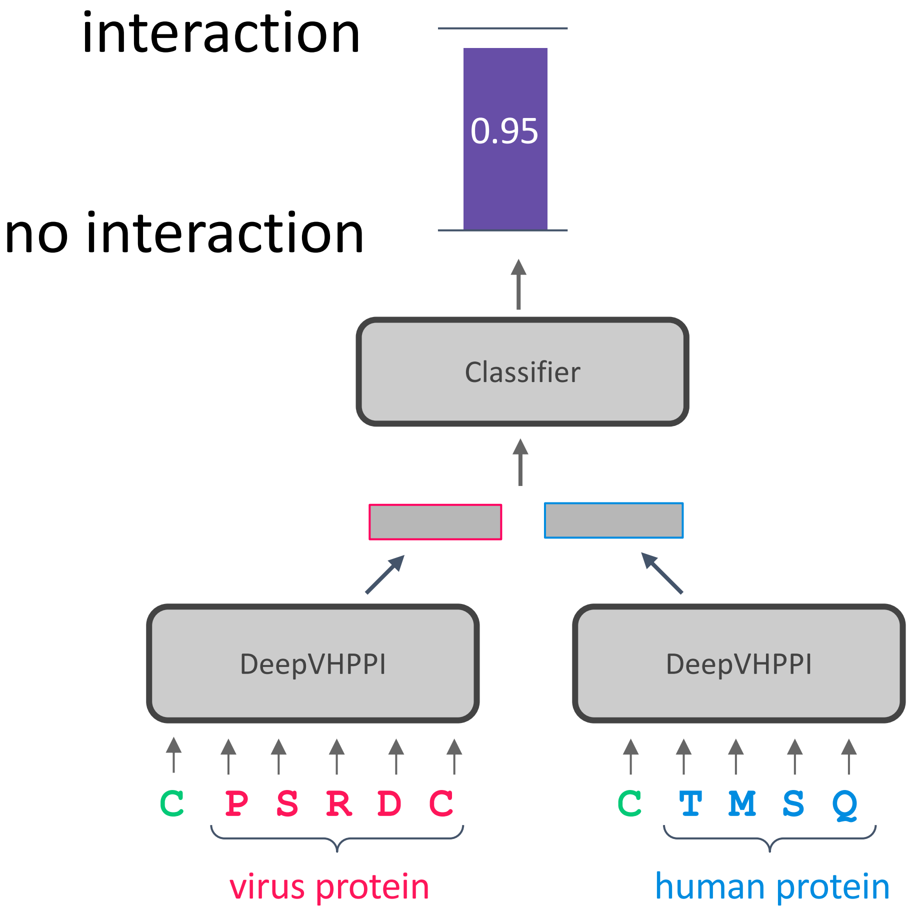
RESEARCH

CORONAVIRUS

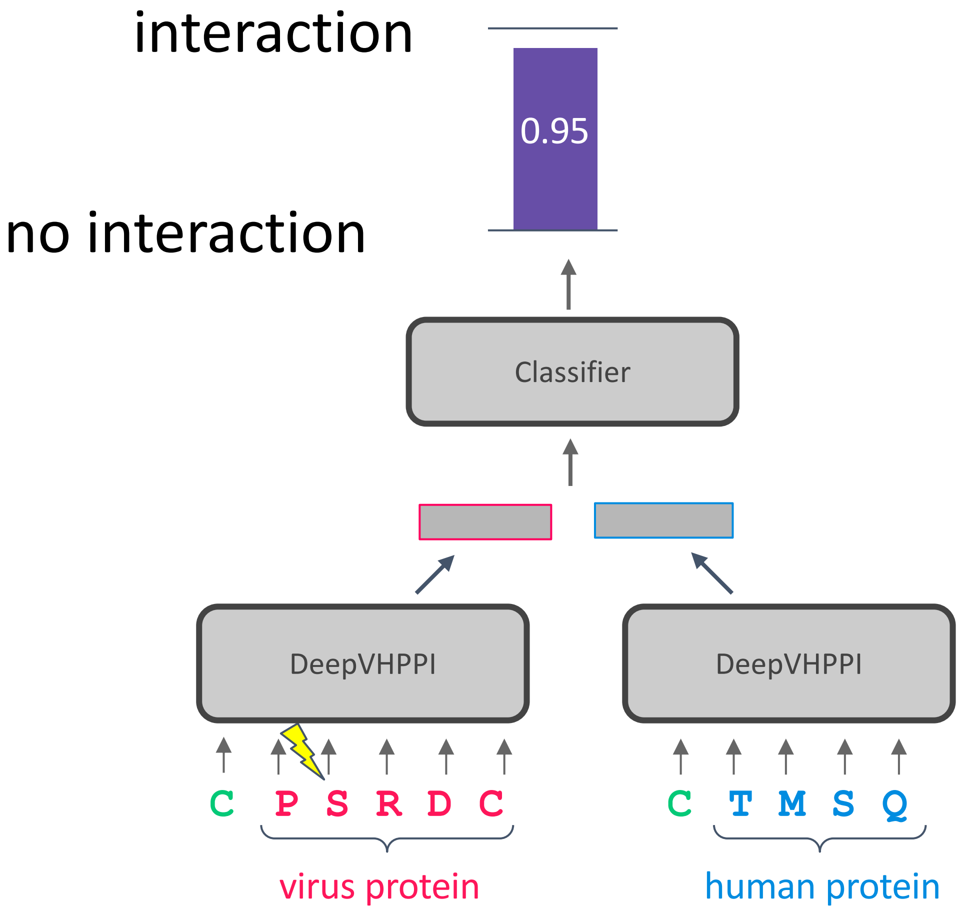
SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo

Yixuan J. Hou[1]*, Shiho Chiba[2]*, Peter Halfmann[2], Camille Ehre[3], Makoto Kuroda[2], Kenneth H. Dinnon III[4], Sarah R. Leist[1], Alexandra Schäfer[1], Noriko Nakajima[5], Kenta Takahashi[5], Rhianna E. Lee[3], Teresa M. Mascenik[3], Rachel Graham[1], Caitlin E. Edwards[1], Longping V. Tse[1],

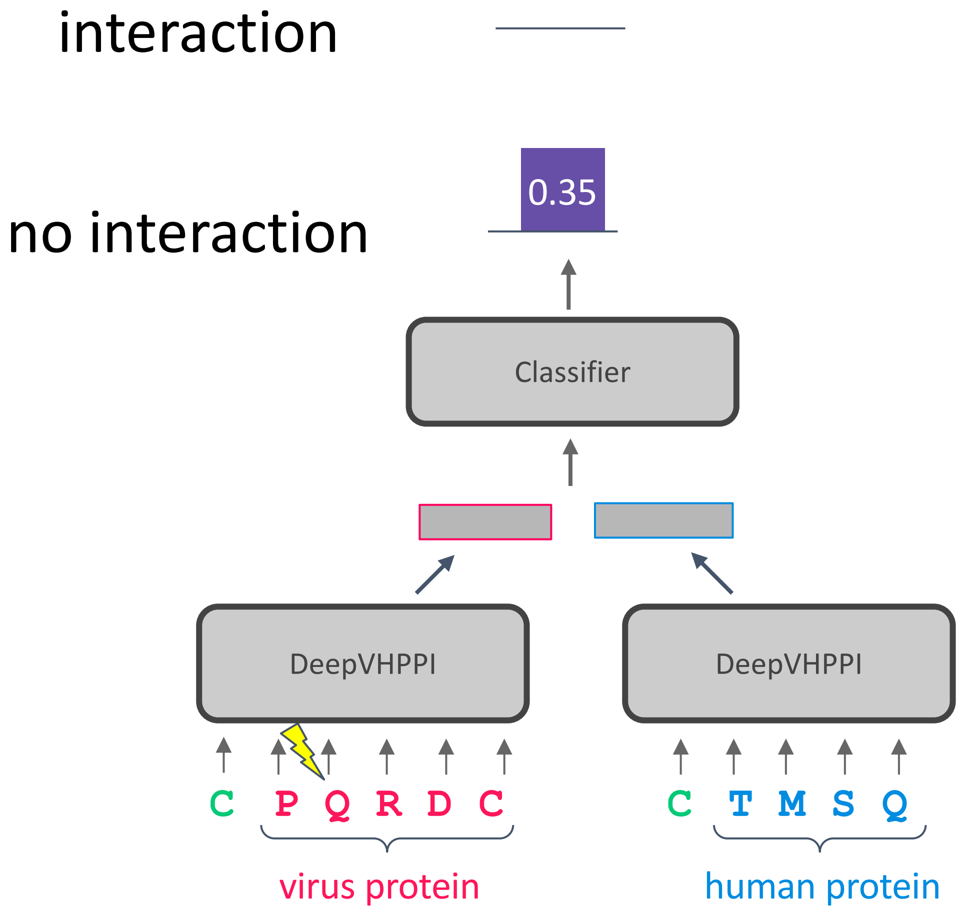
Perturbation Analysis



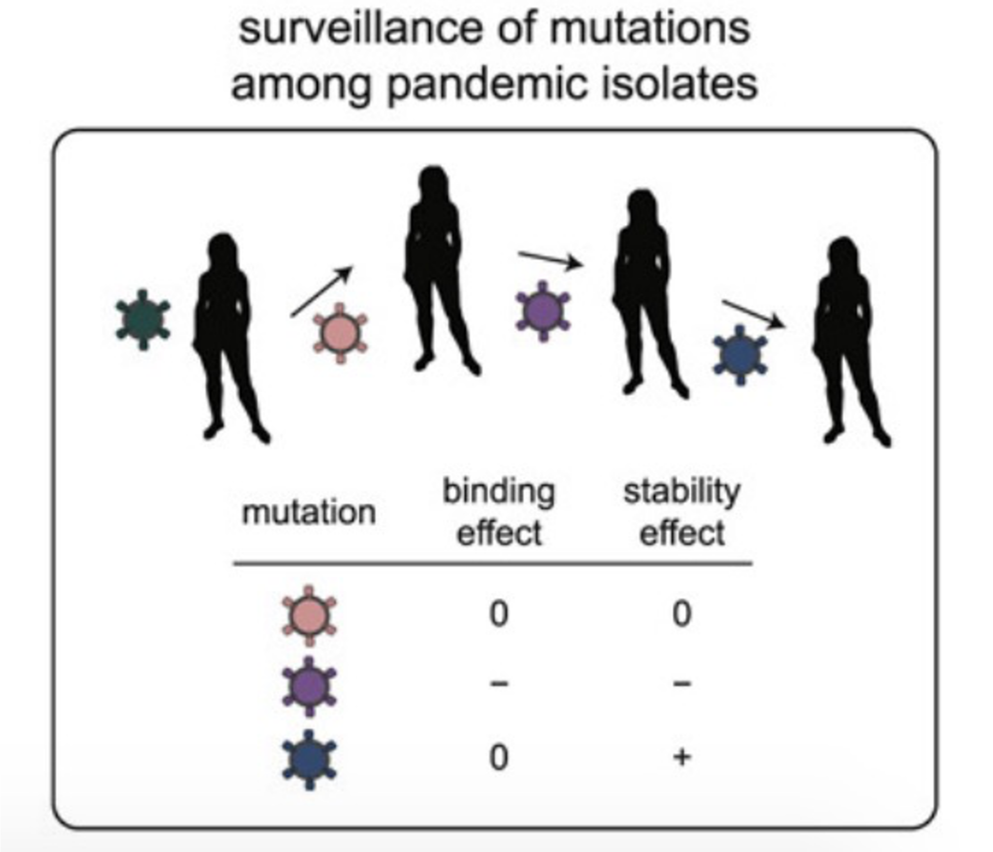
Perturbation Analysis



Perturbation Analysis

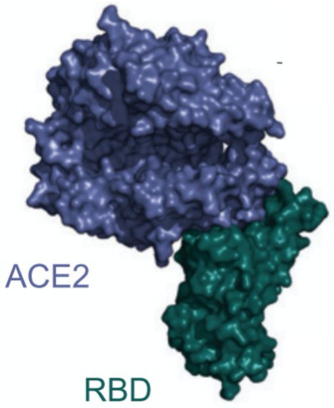


Perturbation Analysis: Investigating Mutations



surveillance of mutations among pandemic isolates

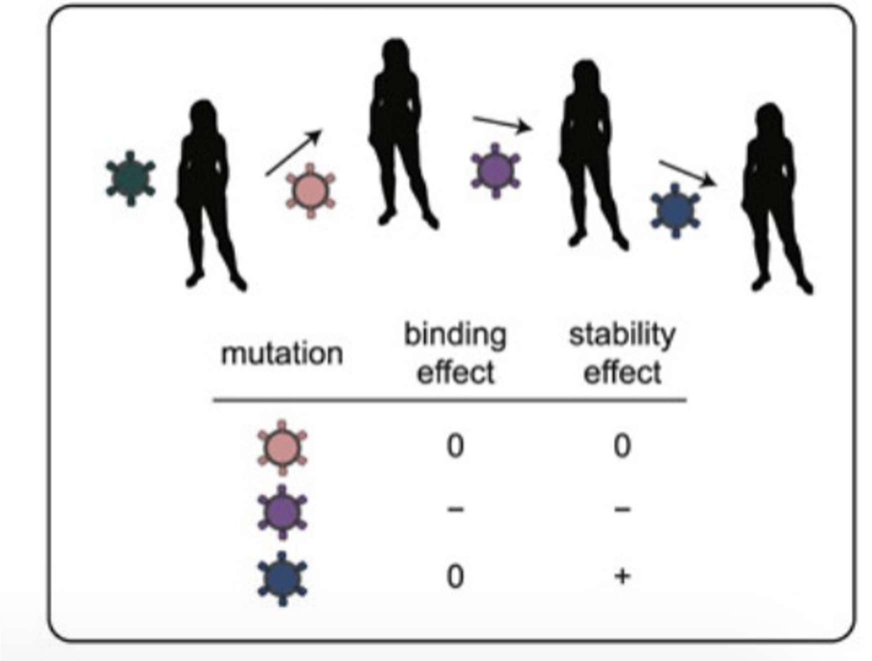
mutation	binding effect	stability effect
●	0	0
●	–	–
●	0	+



Starr et al. Cell 2020

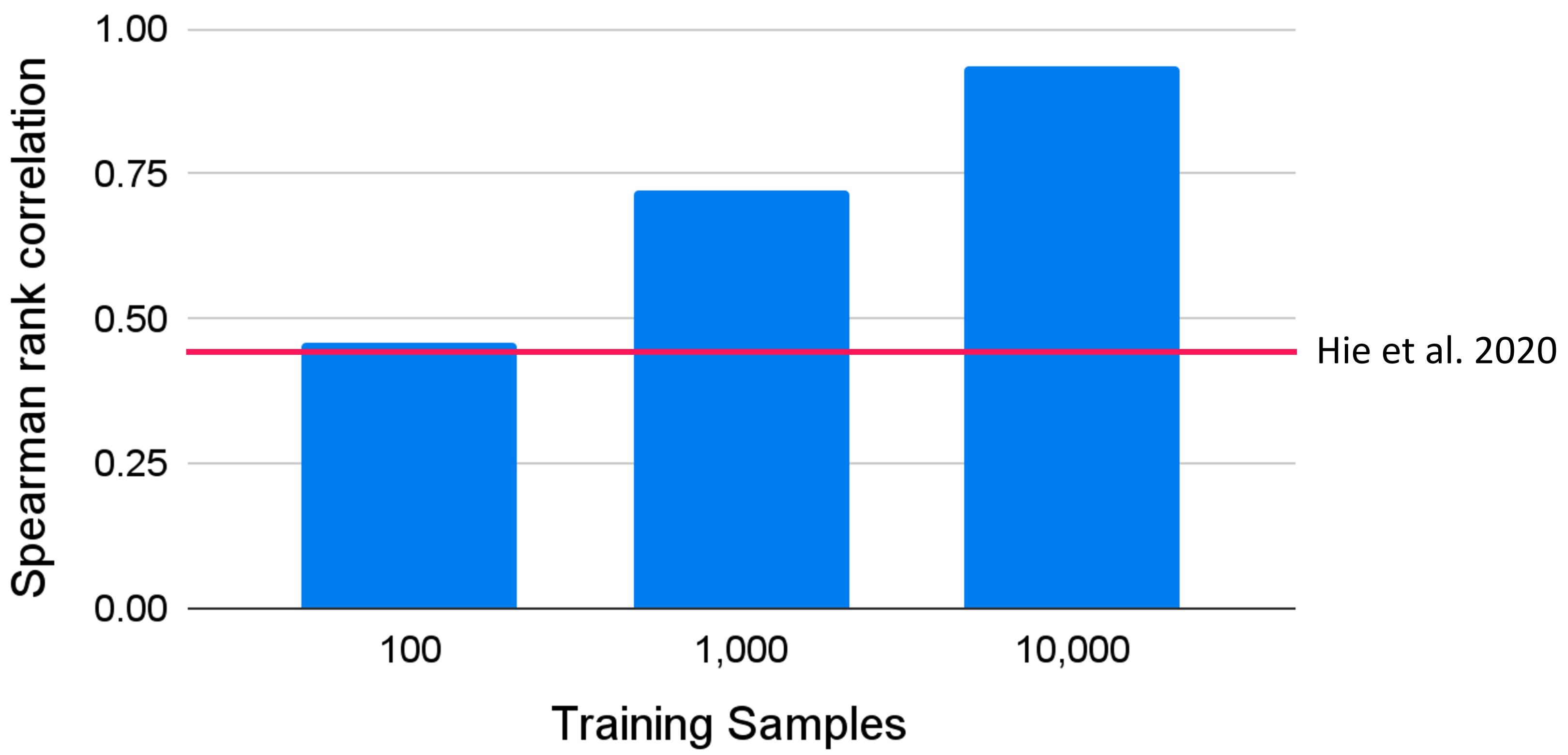
Experimental Setup

- 105,528 mutated Spike sequences and their corresponding ACE2 binding affinities from Starr et al. 2020
- **Training / Test splits**
  - 100 training, 105,428 testing
  - 1,000 training, 104,528 testing
  - 10,000 training, 95,528 testing



Perturbation Analysis: Mutated Spike and ACE2 Interactions

Spearman rank correlation between DeepVHPPI binding prediction and dissociation constant

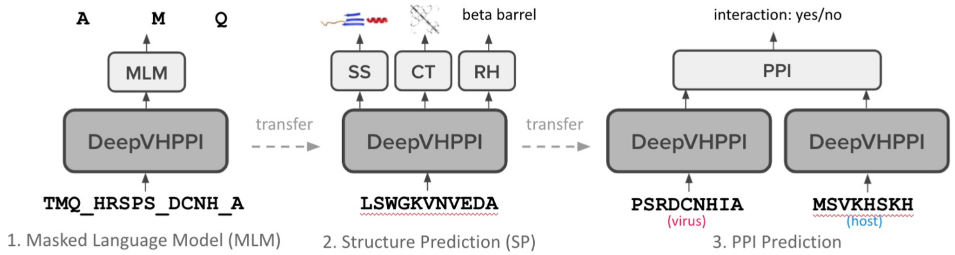


Contributions

1. Flexible transfer learning framework for protein-protein interaction prediction

2. Accurate novel virus interaction predictions

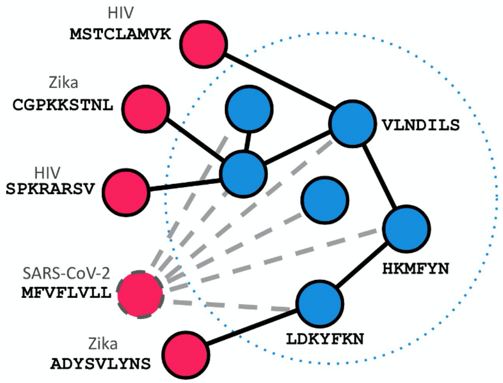
3. Interpretable and interactive mutation perturbation analysis



1. Masked Language Model (MLM)

2. Structure Prediction (SP)

3. PPI Prediction



MFVFLVLL
SARS-CoV-2 Spike
LDKYFKN
Human ACE2

Journey Ahead

- Deeply interested in analyzing this group of amazingly complicated and large-scale datasets
- Realized that finding mutual interests is hard
  - Computational impacts
  - Biomedical impacts

- Need help in biology
- Need help in medicine
- Need help in figuring out NIH grant applications