http://deepchrome.net

Make Deep Learning Interpretable for Sequential Data Analysis in Biomedicine

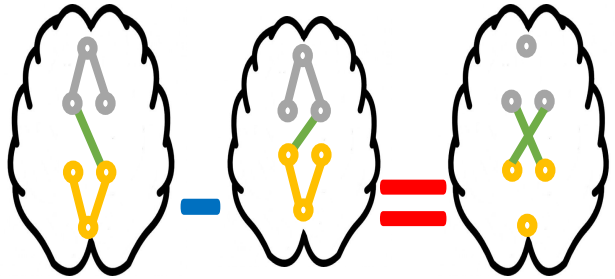
Dr. Yanjun Qi

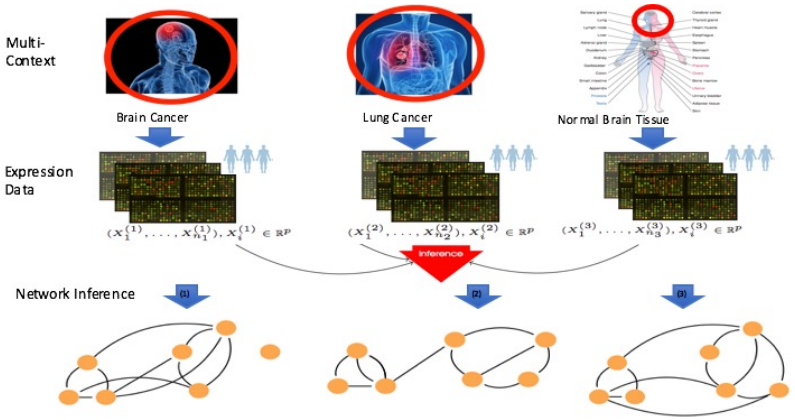
Department of Computer Science

University of Virginia

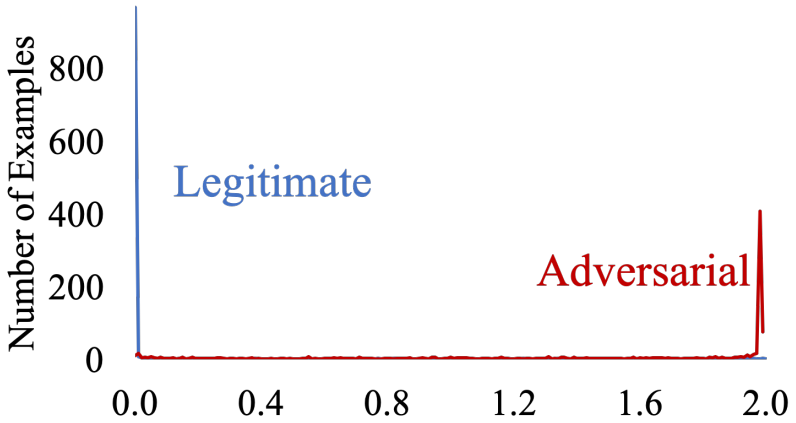
Research Highlights (three fronts)– Yanjun Qi Team

1. Fast and Scalable Learning Algorithms to Extract Networks from Samples

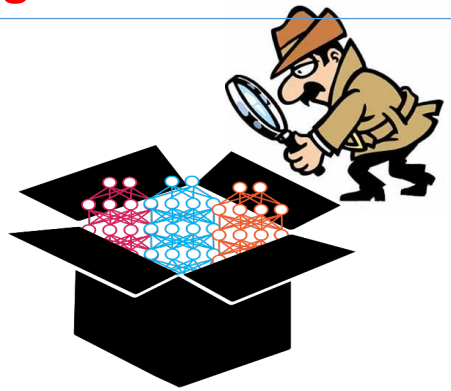




2. Making Deep Learning Robust and trustworthy

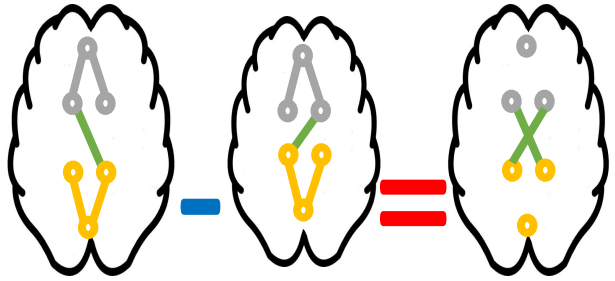


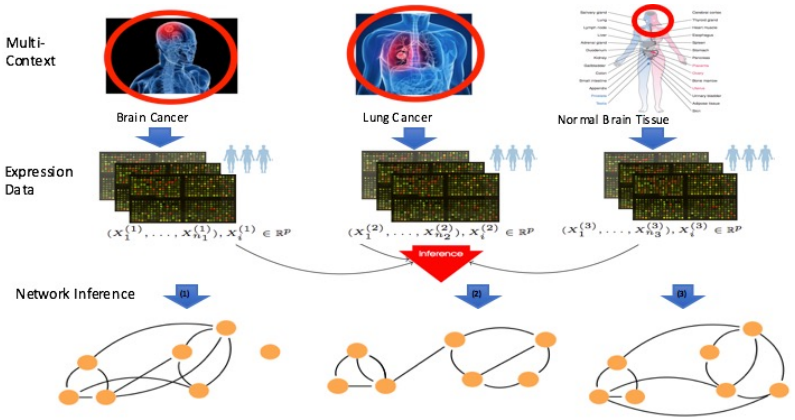
3. Making Explainable Deep Learning for Biomedical Data



Research Highlight (three fronts)– Yanjun Qi

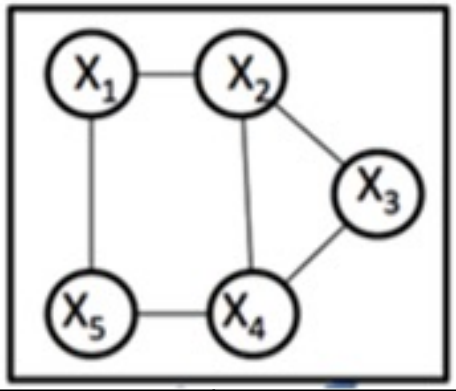
1. Fast and Scalable Learning Algorithms to Extract Networks from Samples





First Front: from Data to Connectome

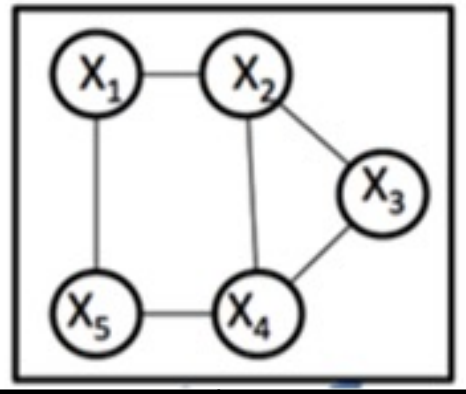
1. Graphical Models to extract interactions among important variables



Xi	Xj
Protein	Protein
Gene	Gene
Protein	DNA/RNA
Neuron Region	Neuron Region
…	….

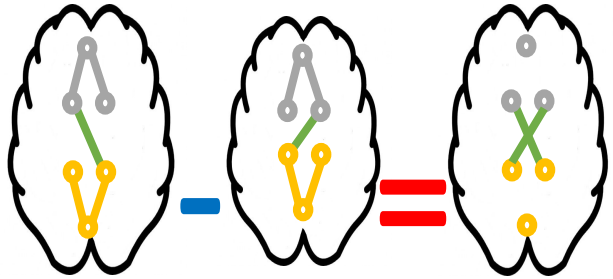
First Front: from Data to Connectome

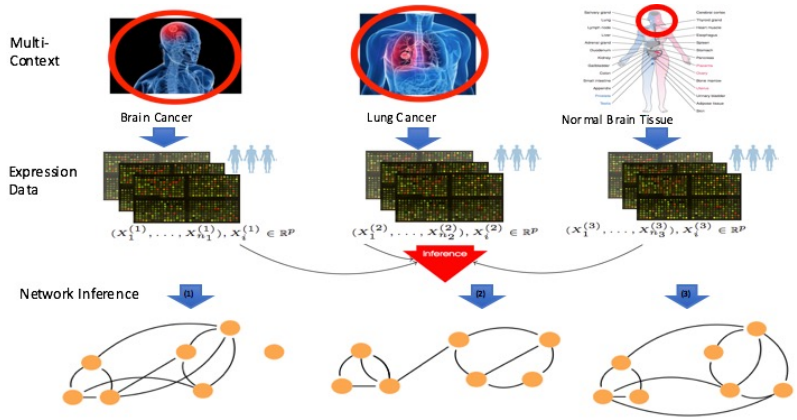
1. Graphical Models to reflect interactions among important variables



Xi	Xj
Protein	Protein
Gene	Gene
Protein	DNA/RNA
Neuron Region	Neuron Region
…	….

2. Consider Sample Heterogeneity from many contexts

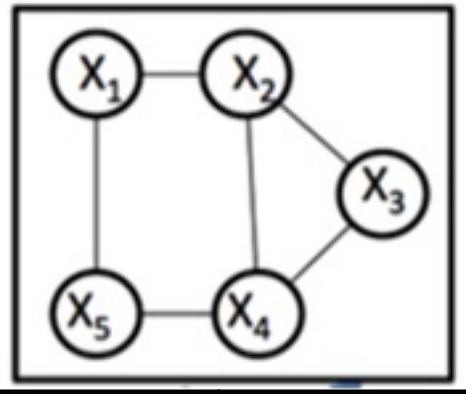




jointnets.org

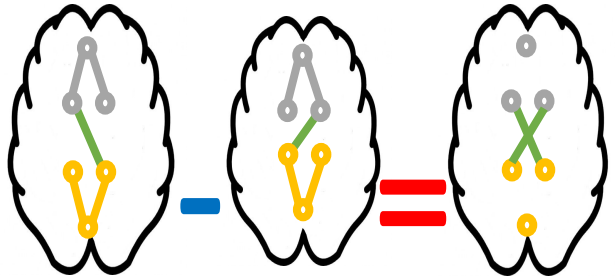
First Front: from Data to Connectome

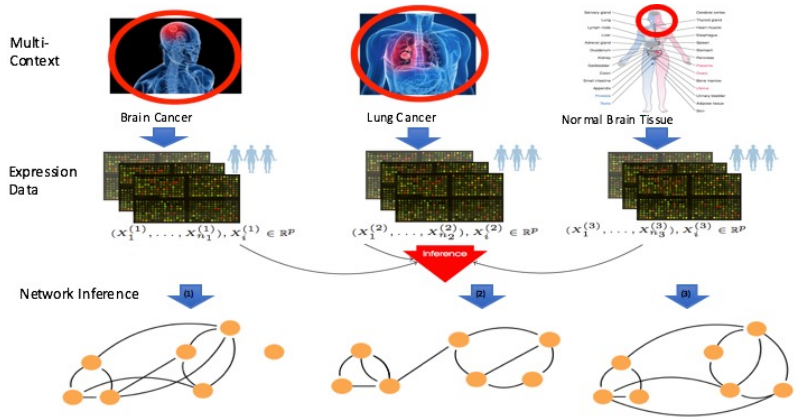
1. Graphical Models to reflect interactions among important variables



Xi	Xj
Protein	Protein
Gene	Gene
Protein	DNA/RNA
Neuron Region	Neuron Region
…	….

2. Consider Sample Heterogeneity to network for many contexts



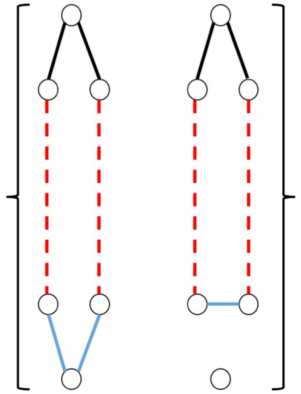


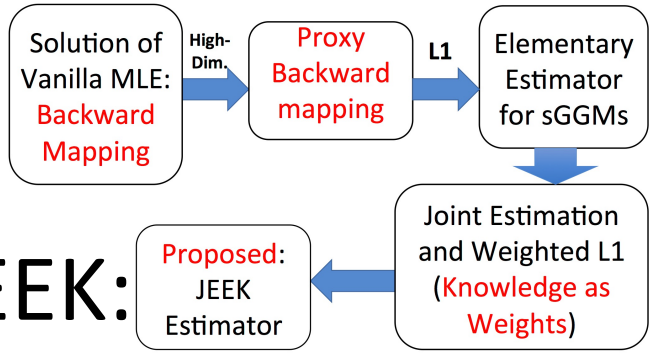
jointnets.org

- Joint graph discovery from heterogeneous samples
  - Fast and scalable graph estimators
  - Parallelizable method (GPU, multi-threading)
  - Sharp convergence rate (sharp error bounds)

Timeline of JointNets

http://jointnets.org/



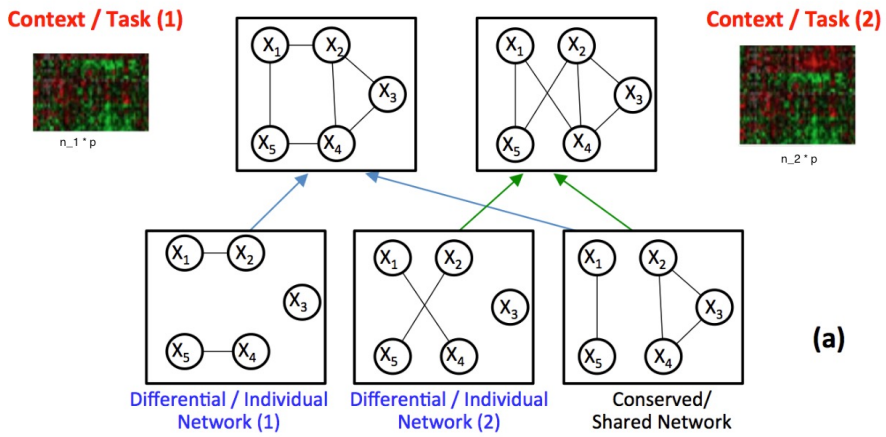


WSIMULE:

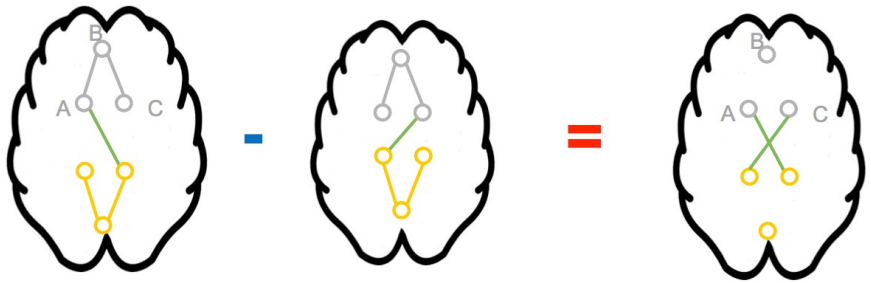
JEEK:

2015 | 2016 | 2017 | 2018 | 2019-21

SIMULE:



DIFFEE:

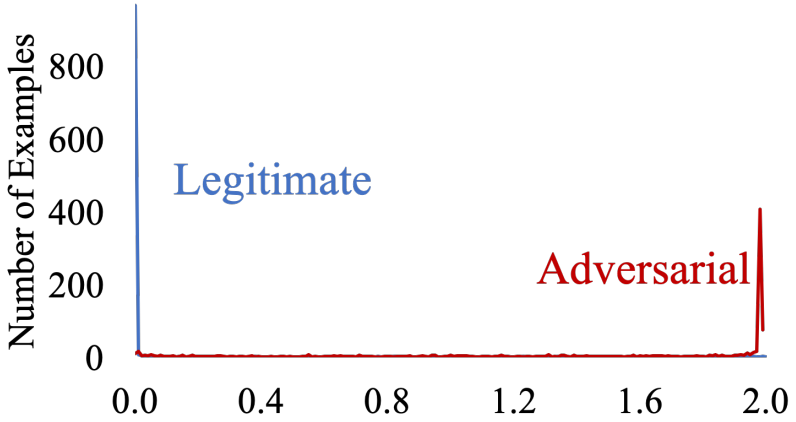


kDIFFNet:

Aligning Differential Network Discovery with Scientific Knowledge

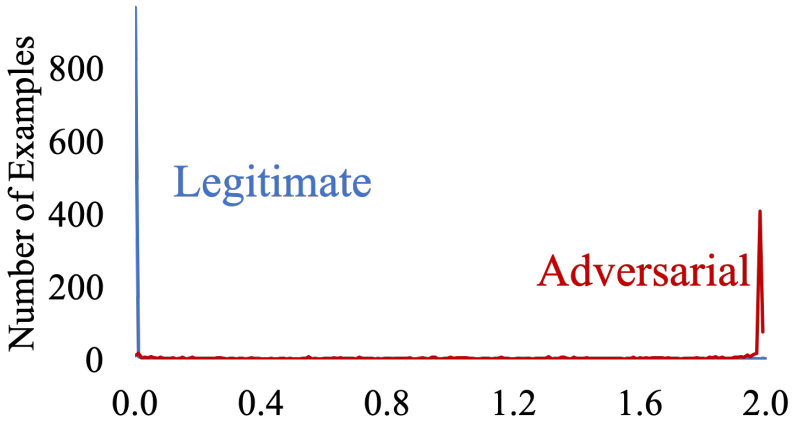
Research Highlight (three fronts)– Yanjun Qi

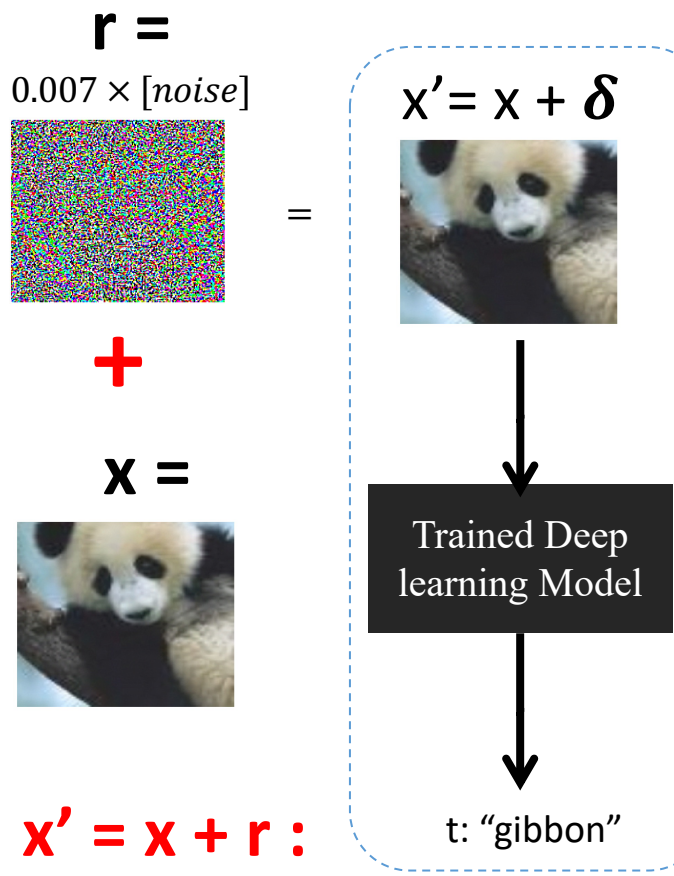
2. Making Deep Learning Robust and trustworthy



Research Highlight (three fronts)– Yanjun Qi

2. Making Deep Learning Robust and trustworthy





x' = x + r : adversarial sample

Second Front: Making Deep Learning Robust

http://trustworthymachinelearning.org

1. To Fool / Evade Learned Models

2. To Detect fooling/ Evasion

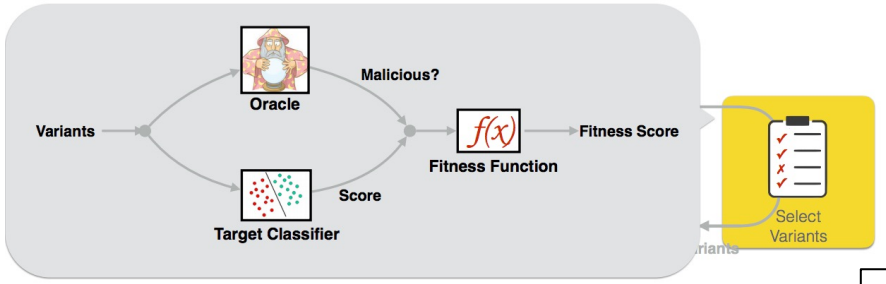
3. To Defend Against Evasion

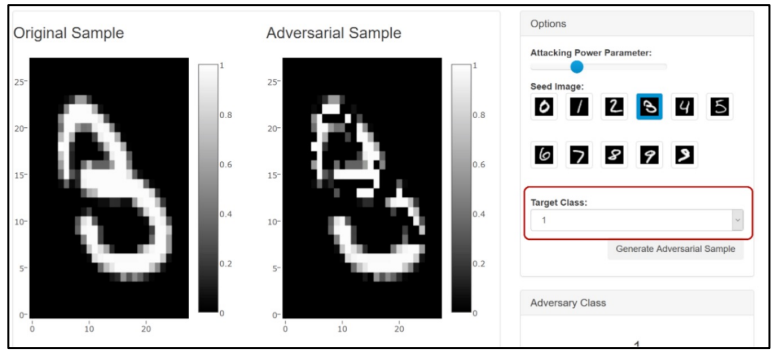
4. To Visualize and Benchmarking

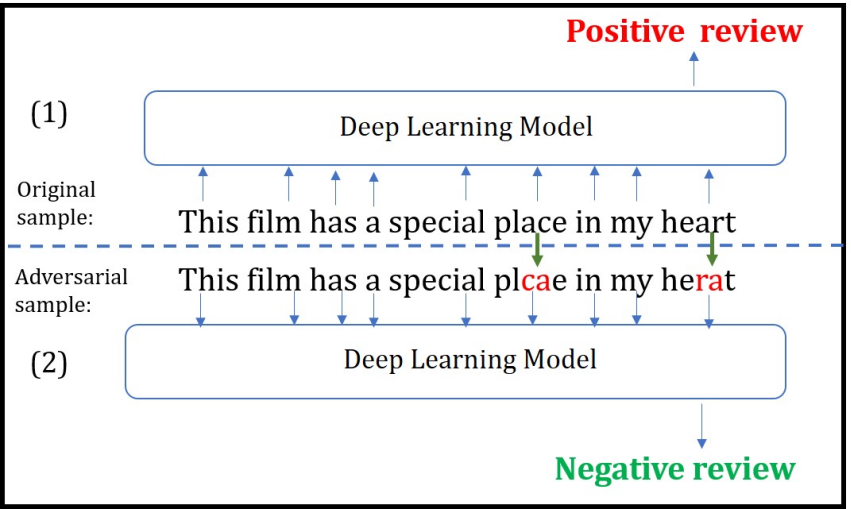
5. To Understand Theoretically

Automated Evasion Approach

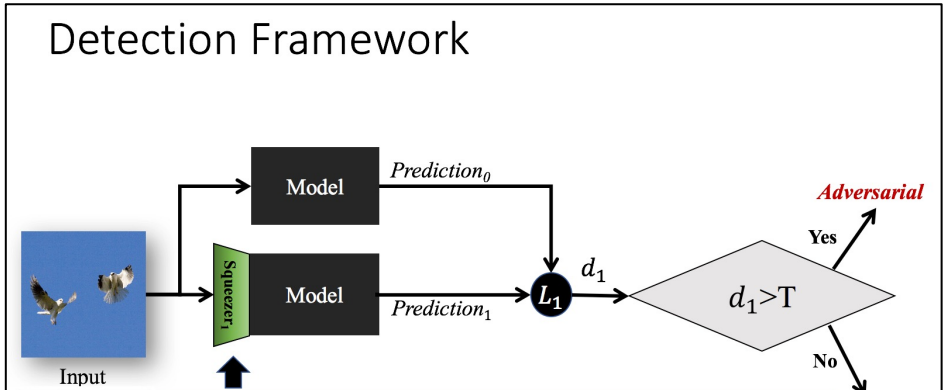
Based on Genetic Programming





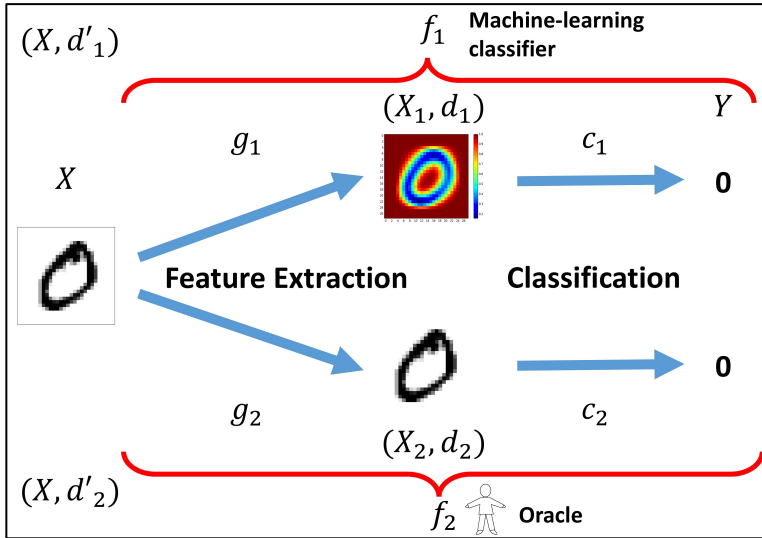


Detection Framework



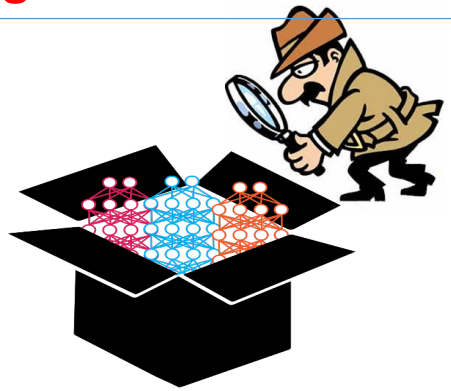
Feature Squeezer coalesces similar samples into a single one.

- Barely change legitimate input.
- Destruct adversarial perturbations.

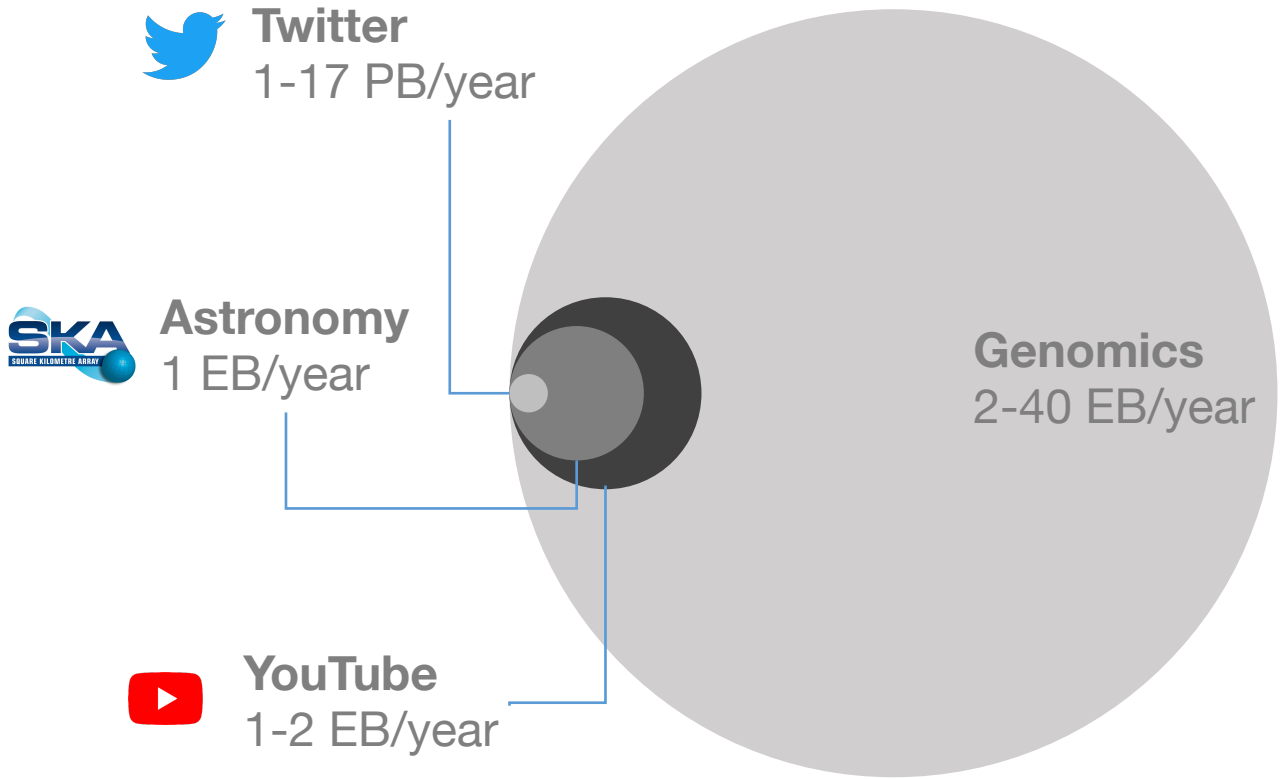


Research Highlight (three fronts)– Yanjun Qi

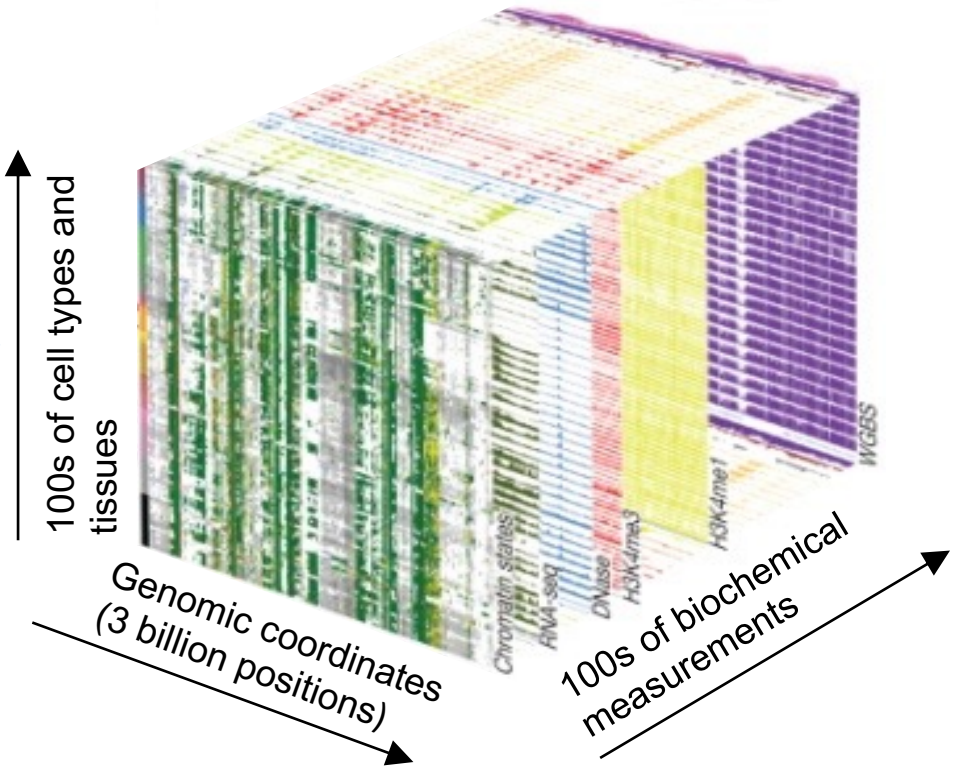
3. Making Explainable Deep Learning for Biomedicine



Big Data: Large-scale genomics measurements



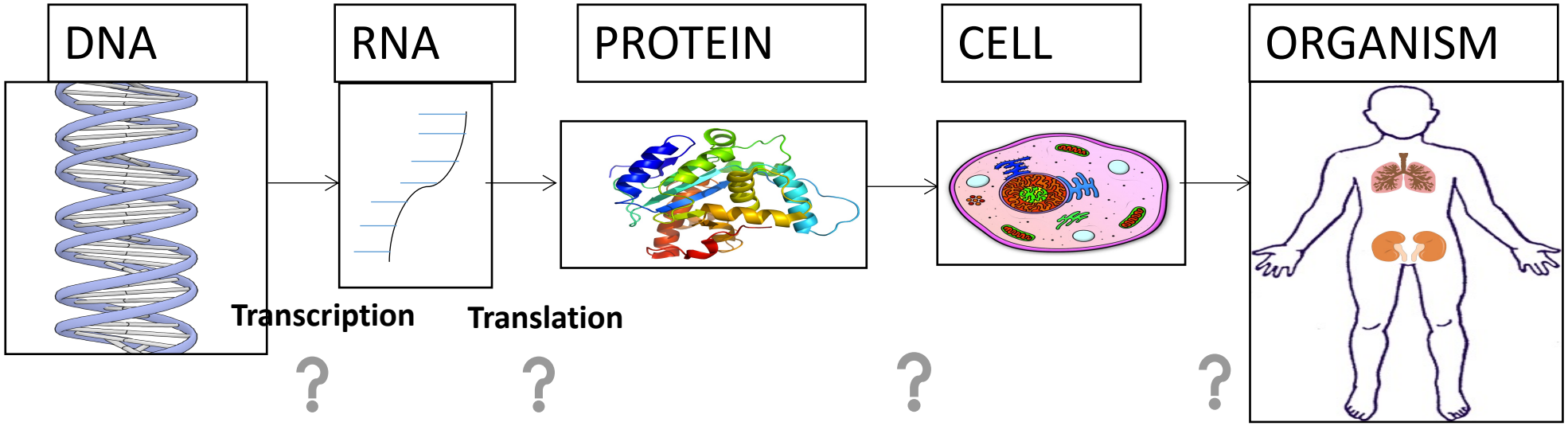
e.g.



Adapted from Stephens ZD et al PLOS Biol 2015

http://deepchrome.net

Biological Modules

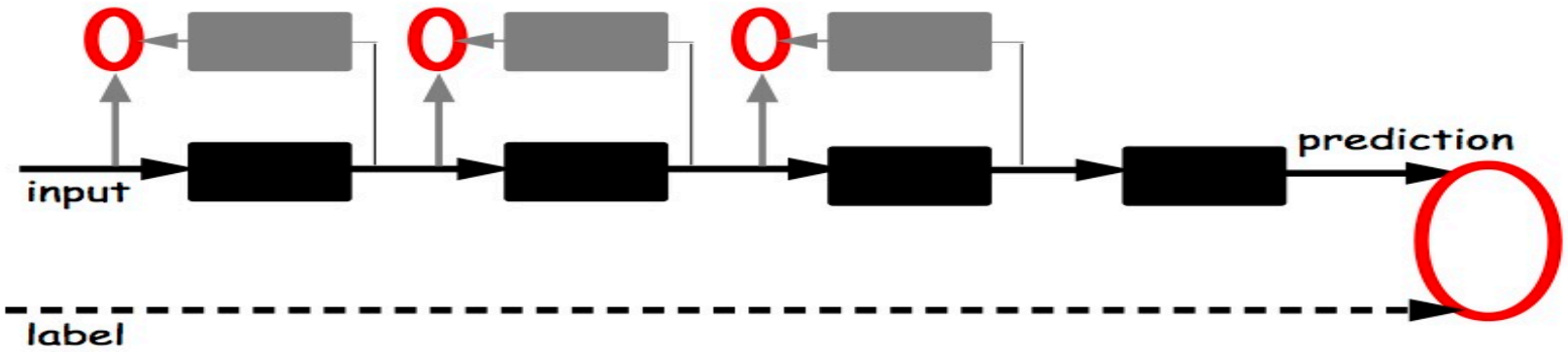


CATGACTG
CATGCCTG

Genetic Variant

Disease

Deep Learning Modules (composable)

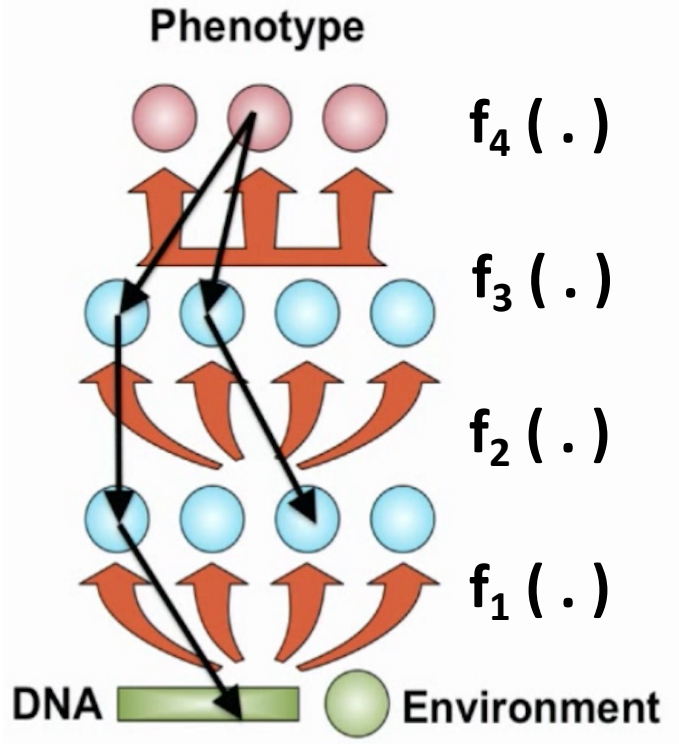


Third Front: Deep Learning on Biomedicine

1. Deep Learning module to reflect biological modules

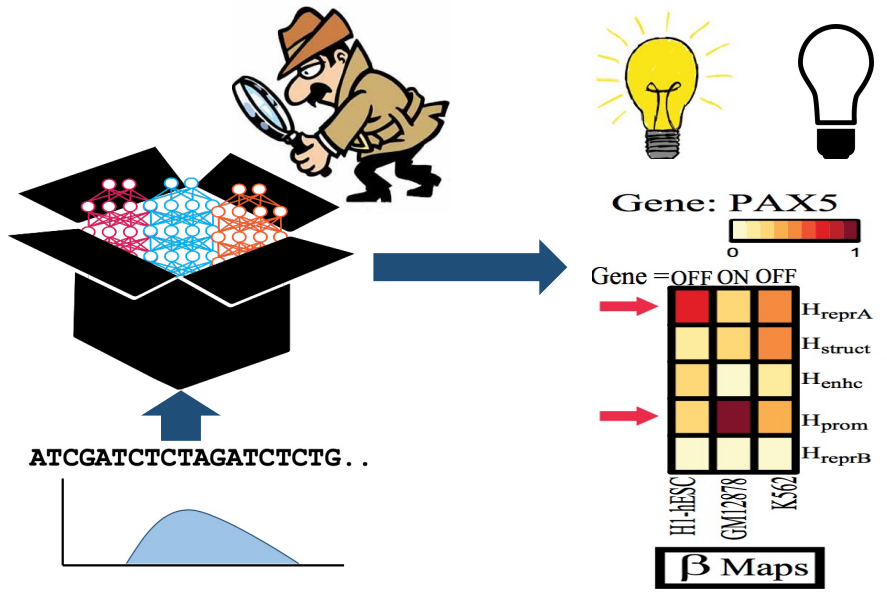
X	Y
DNA	RNA / Func
Epigenetic	RNA
DNA	Interaction to Protein (TF)
Protein	Funcs
Protein	Interaction to DNA/RNA
…	…

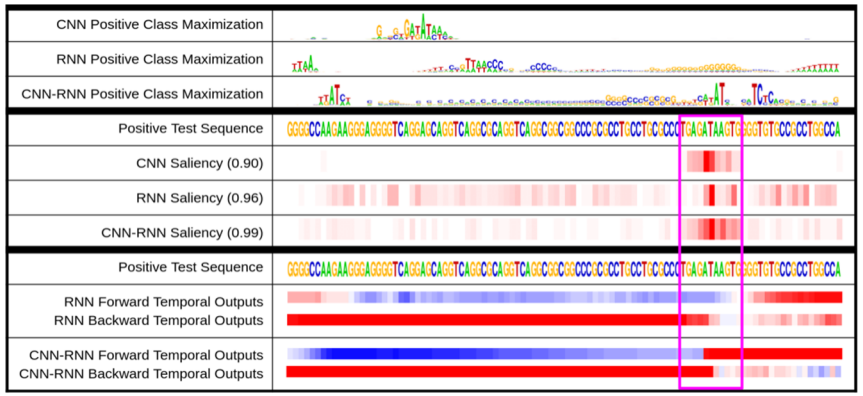
2. Compose modules to reflect biology



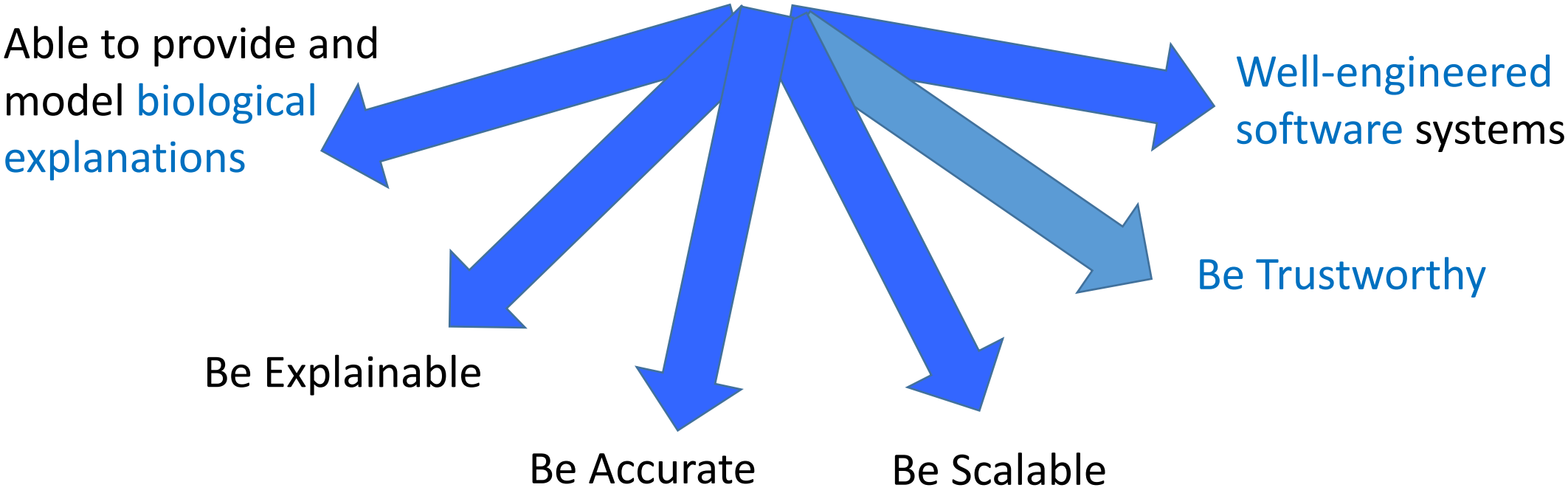
http://deepchrome.net

3. Open DNN black-box and provide domain explanations





Machine learning for Biomedicine
Our Research Philosophy:

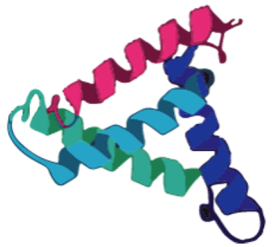


Objective of this talk

- To check if our preliminary solutions make sense
- To get feedbacks / recommendations of next dataset to touch
- To get helps finding mutual interests / Foundation, NIH grant app?

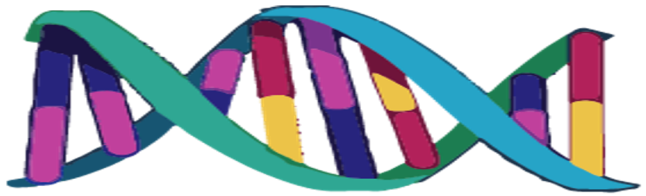
Please, please ask questions!

PROTEIN



RNA

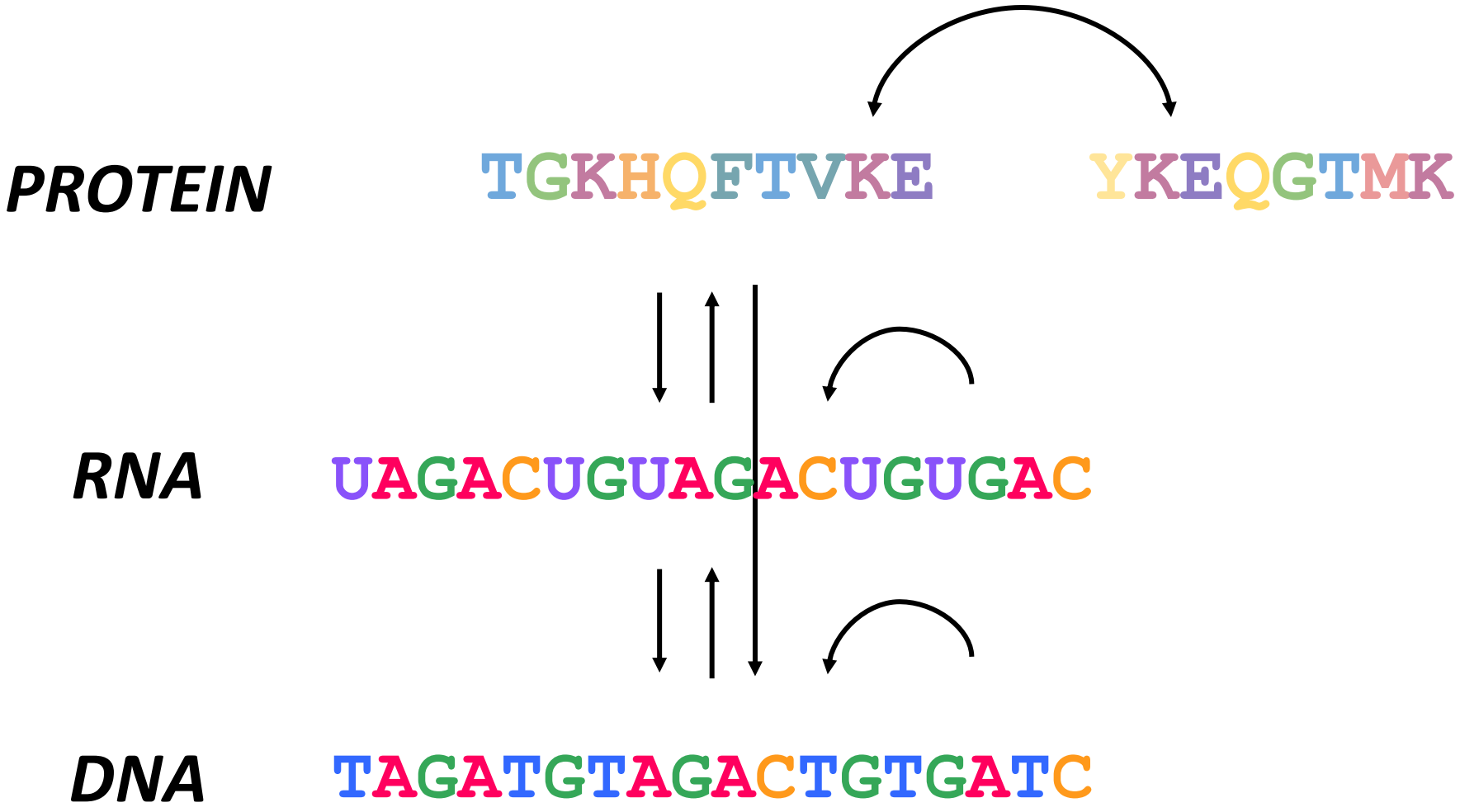
DNA



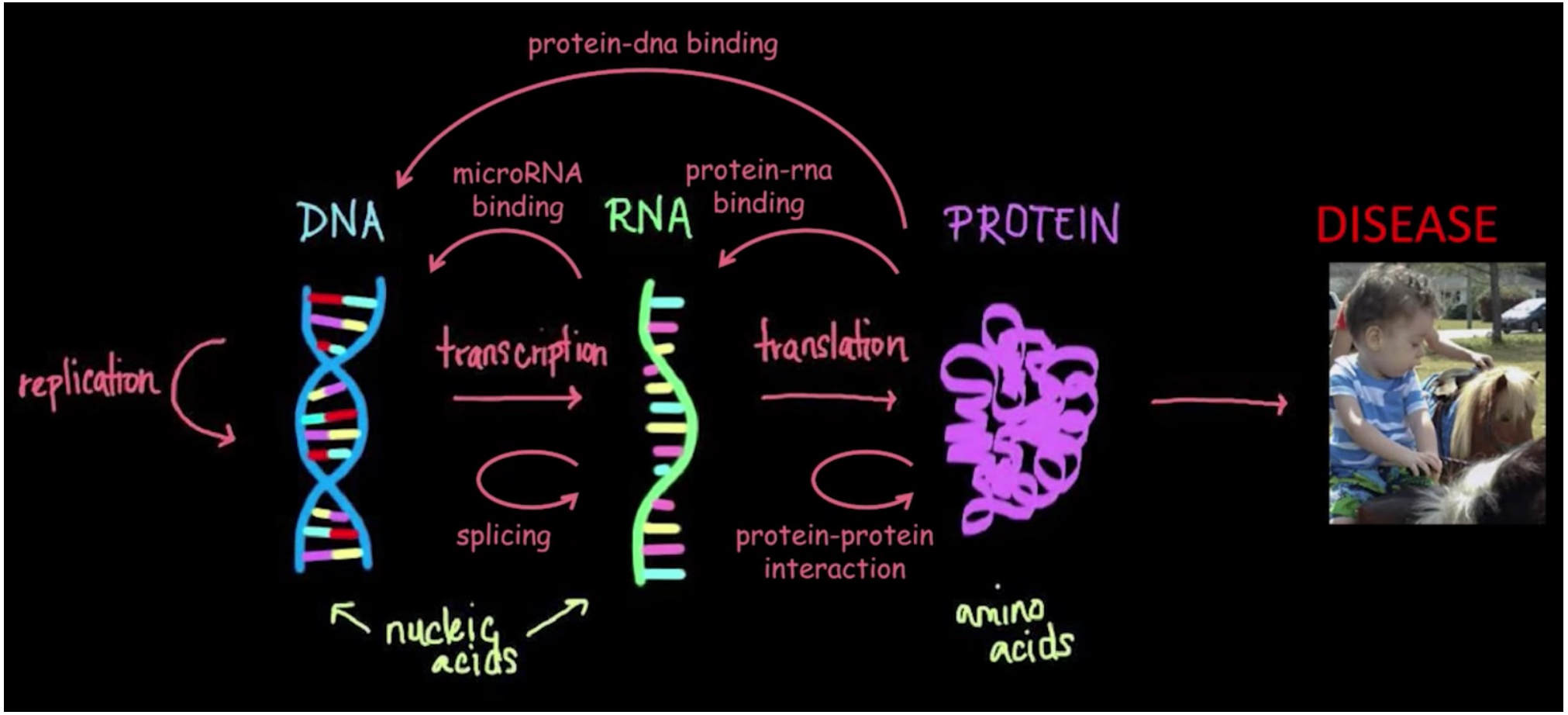
PROTEIN TGKHQFTVKE

RNA UAGACUGUAGACUGUGAC

DNA TAGATGTAGACTGTGATC


PROTEIN
TGKHQFTVKE
YKEQGTMK
RNA
UAGACUGUAGACUGUGAC
DNA
TAGATGTAGACTGTGATC

Biology has so many for a computer scientist to learn



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

Image Credit: Brendan Frey

Current Methods are not enough:

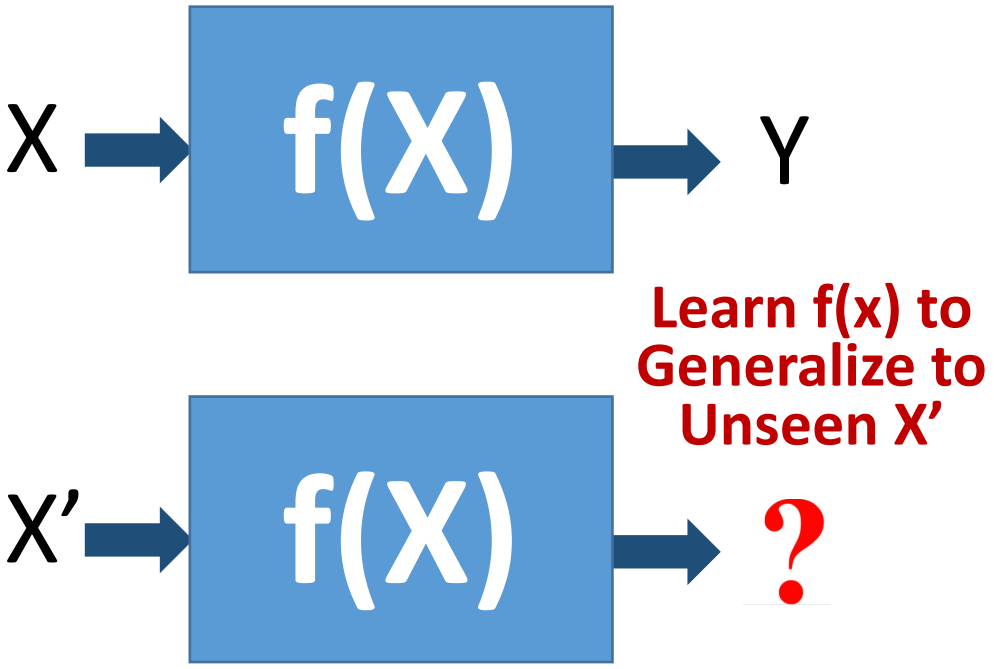
- Correlation based,
- Single-variable based hypothesis testing
- Linear models:
- Differential equations based:
- Shallow machine learning based:
- More?

Current Methods are not enough:

- 0. Correlation based strategy are not inductive

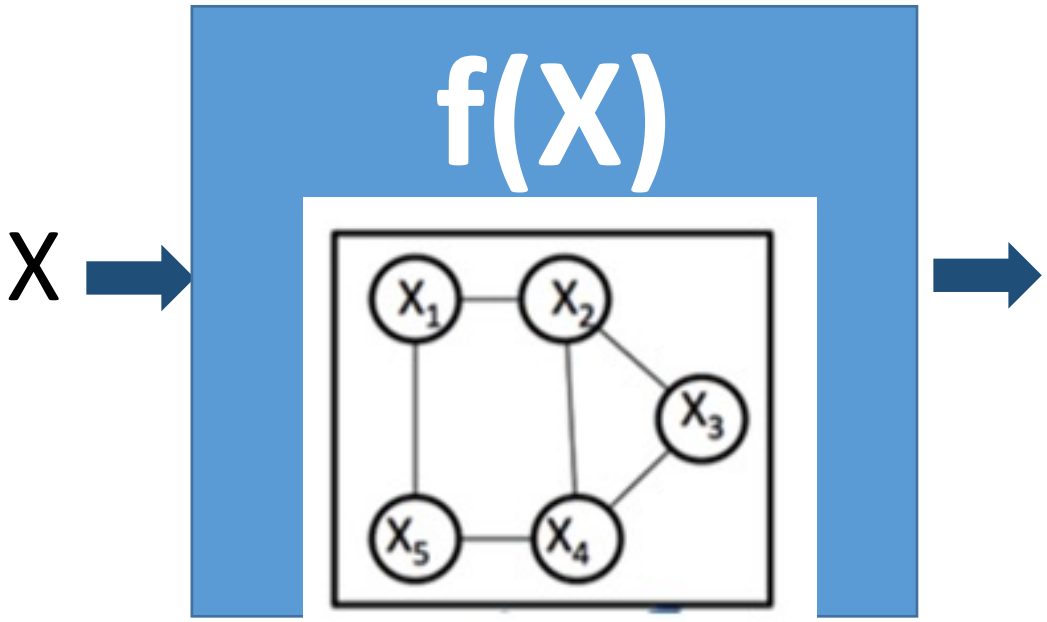
Impactful data-driven computational model should be able to generalize, aka.
➔ To predict effect of therapy / drug / variations

- Need inductive reasoning
  - Generalizations from observed data to unseen data



Current Methods are not enough:

- 1. Single-variable based hypothesis testing can not model interactions



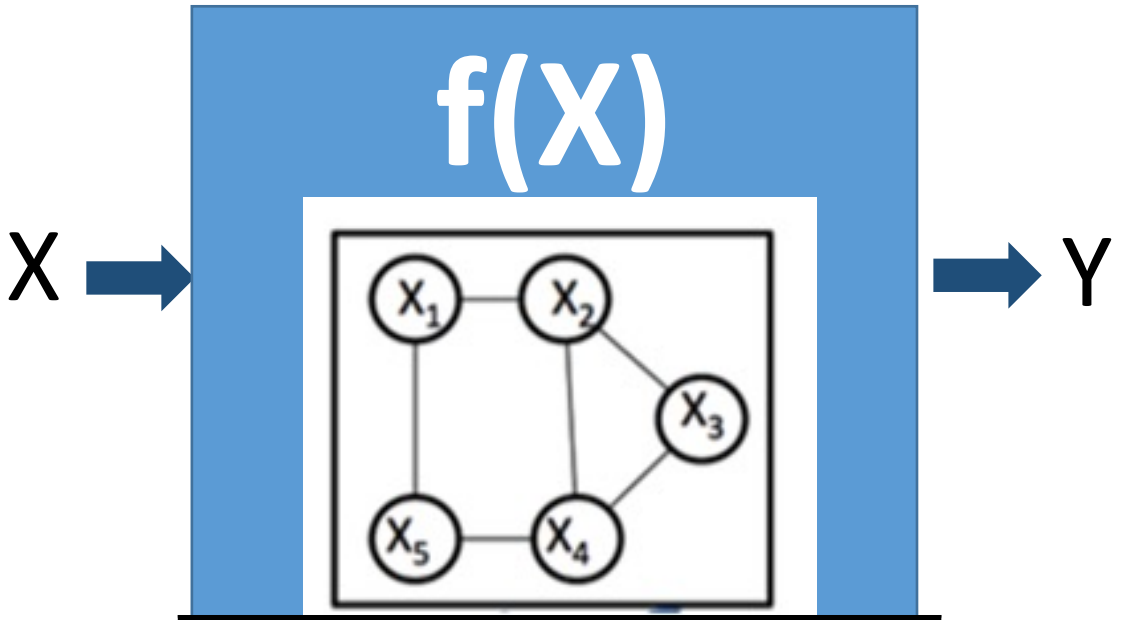
Biology is complicated, interactions among players need to be considered

Current Methods are not enough:

- 2. Linear models can not model interactions

$$f(x) = w \cdot x$$

Biology is complicated, interactions among players need to be considered

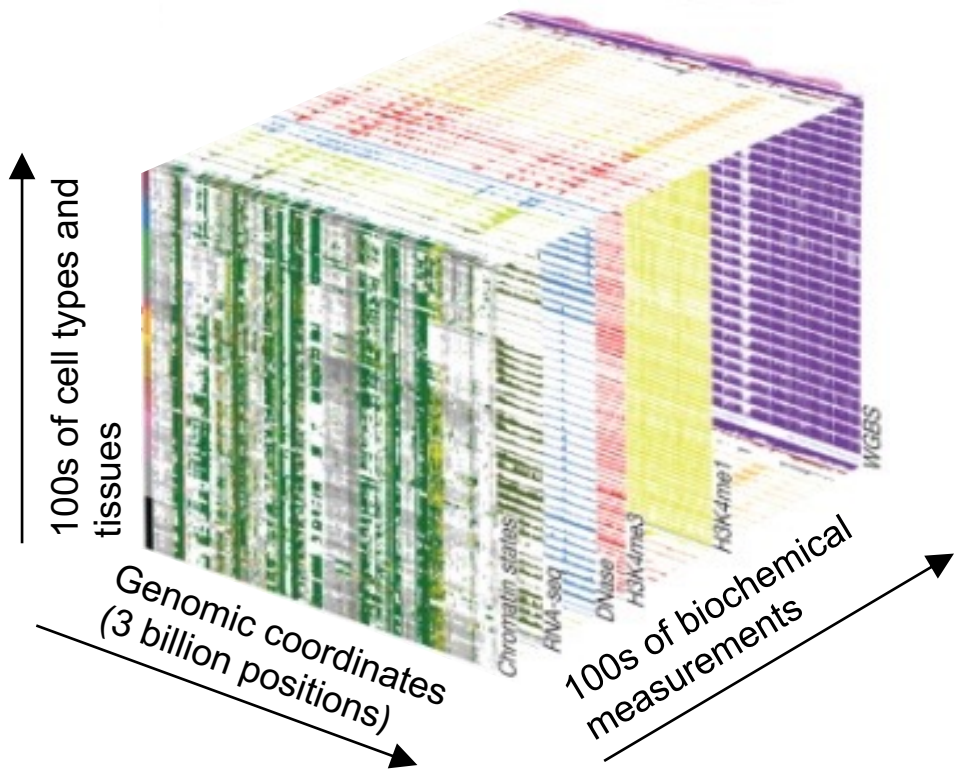


Xi	Xj
Protein	Protein
Gene	Gene
Protein	DNA/RNA
Neuron Region	Neuron Region
…	….

Current Methods are not enough:

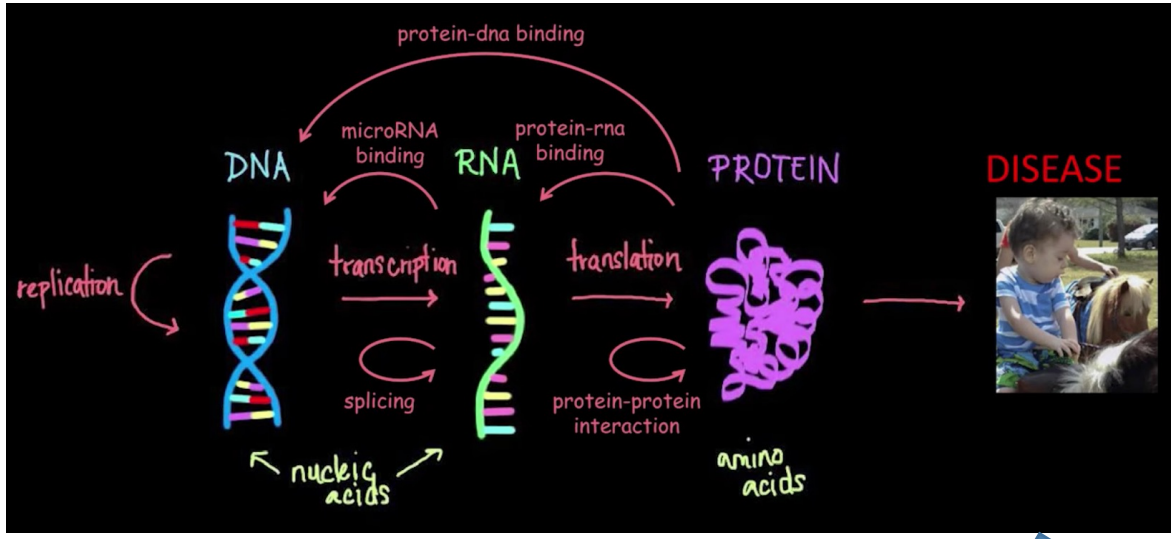
- 3. differential equation based models are not scalable and can not model hidden factors

Hidden/latent/cofounding factors are intrinsic in biological datasets

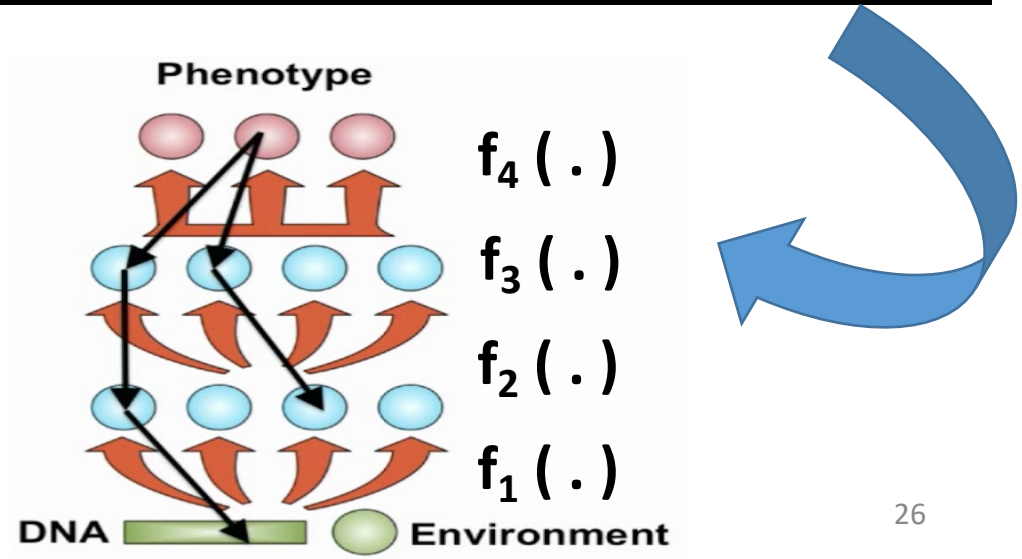


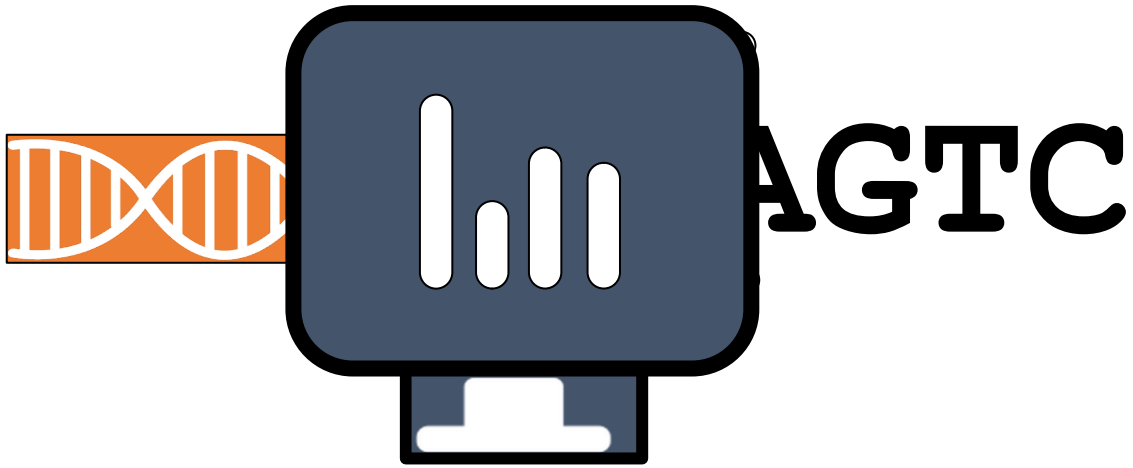
Current Methods are not enough:

- 4. Shallow machine learning models hard to compose with each other and need intensive feature engineering , e.g. random forest, SVM, …

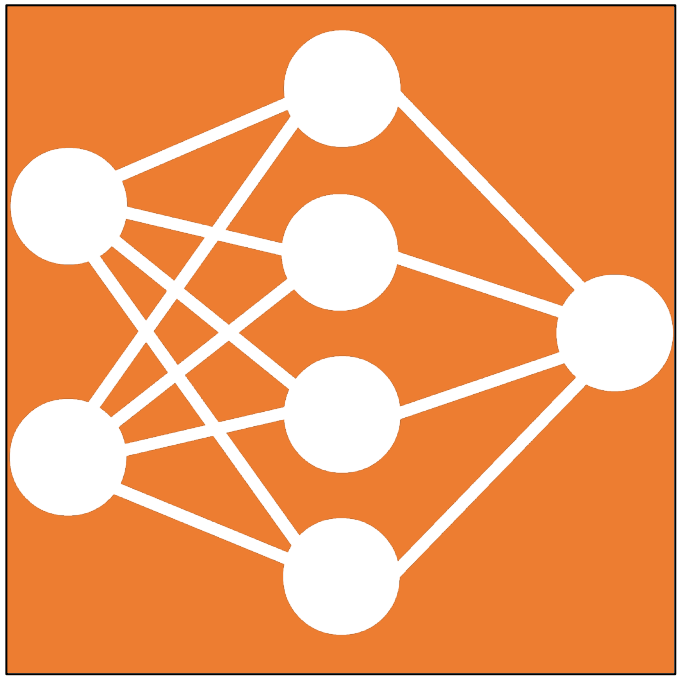


Possible to learn good features / representations automatically?



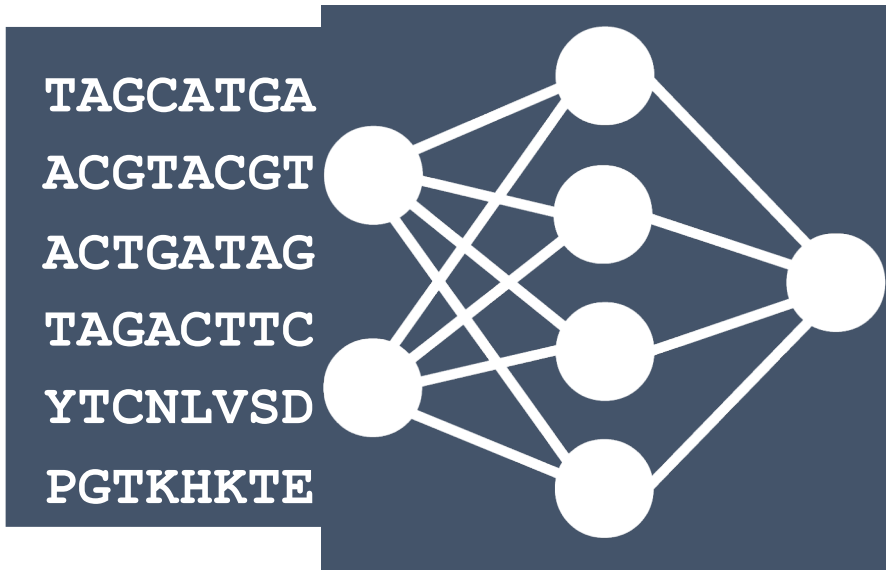


Next-Gen Sequencing



Deep Learning

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

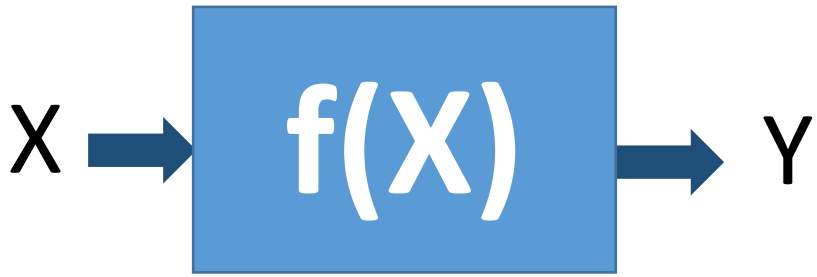


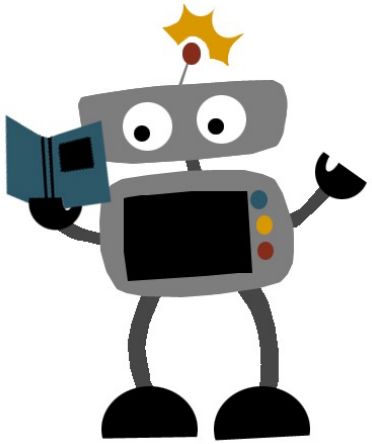
1. Predict

2. Interpret

Basics of Machine Learning

Training Stage





Testing Stage

?

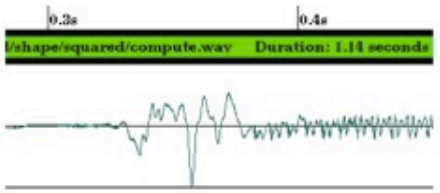
Learn f(x) to Generalize to Unseen X'

Supervised Learning

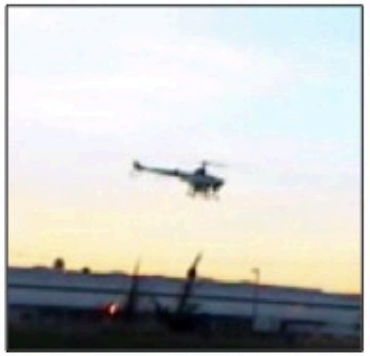
Generalisation: learn model f(x) from past data in order to "explain", "predict", "model" or "control" new data examples

Deep Learning is Changing the World

How may I help you, human?



Speech Recognition



Control learning

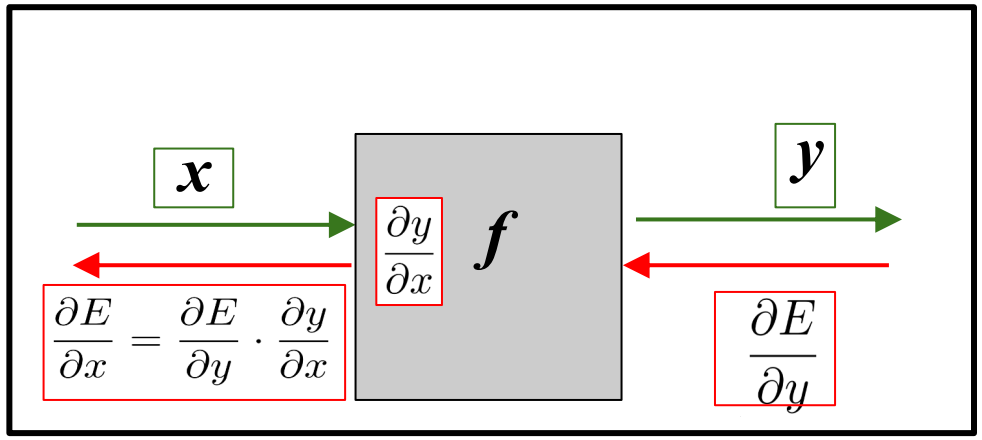
Text analysis

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 Semtech Corp 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

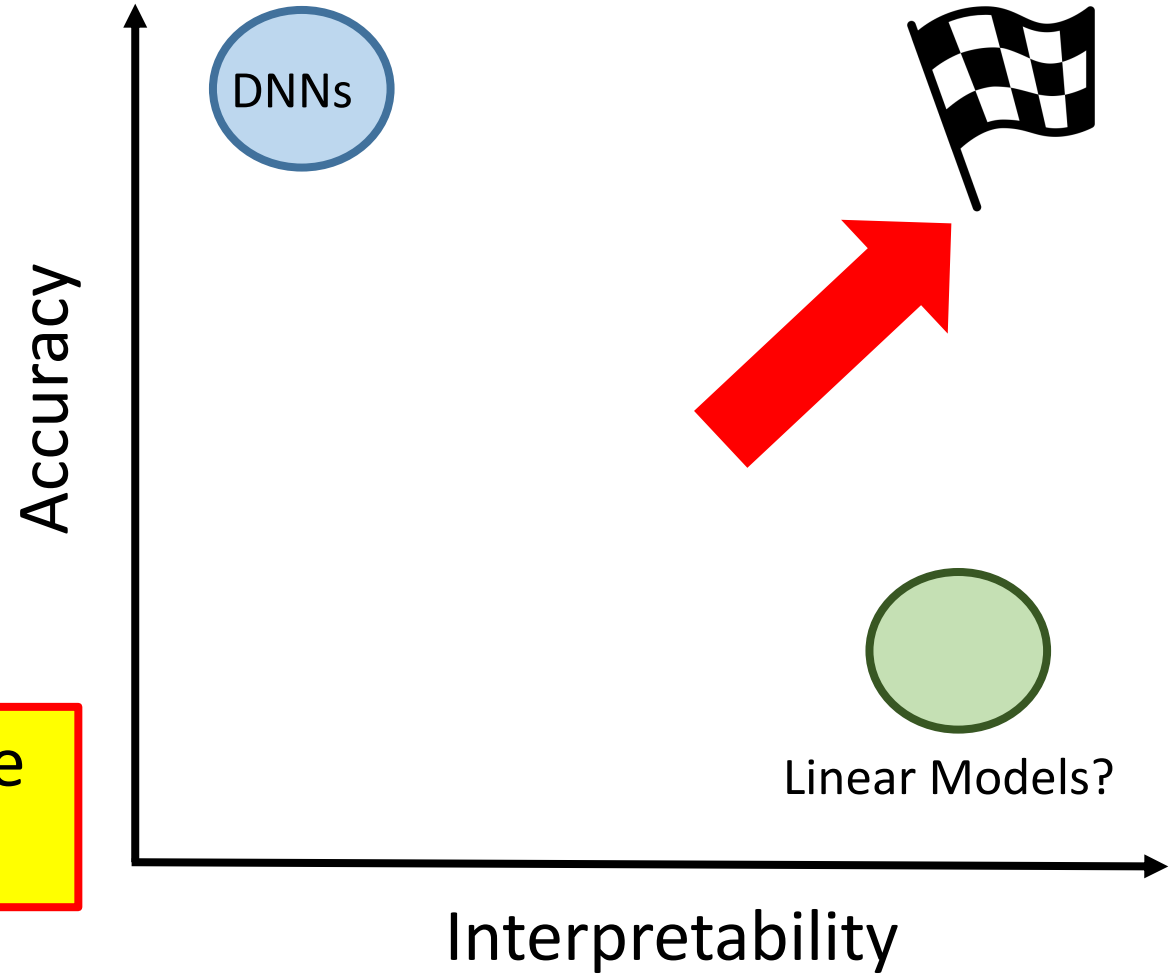
Many more !

Building Deep Neural Nets



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

Our Goal: Interpretable DNNs



Challenge : DNNs are hard to Interpret

Gene

ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

Gene Expressed

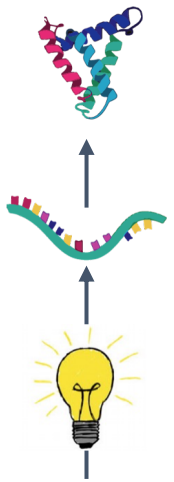
ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

Gene Expressed

ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

what causes a gene to be expressed?

To understand gene regulation



gene expressed

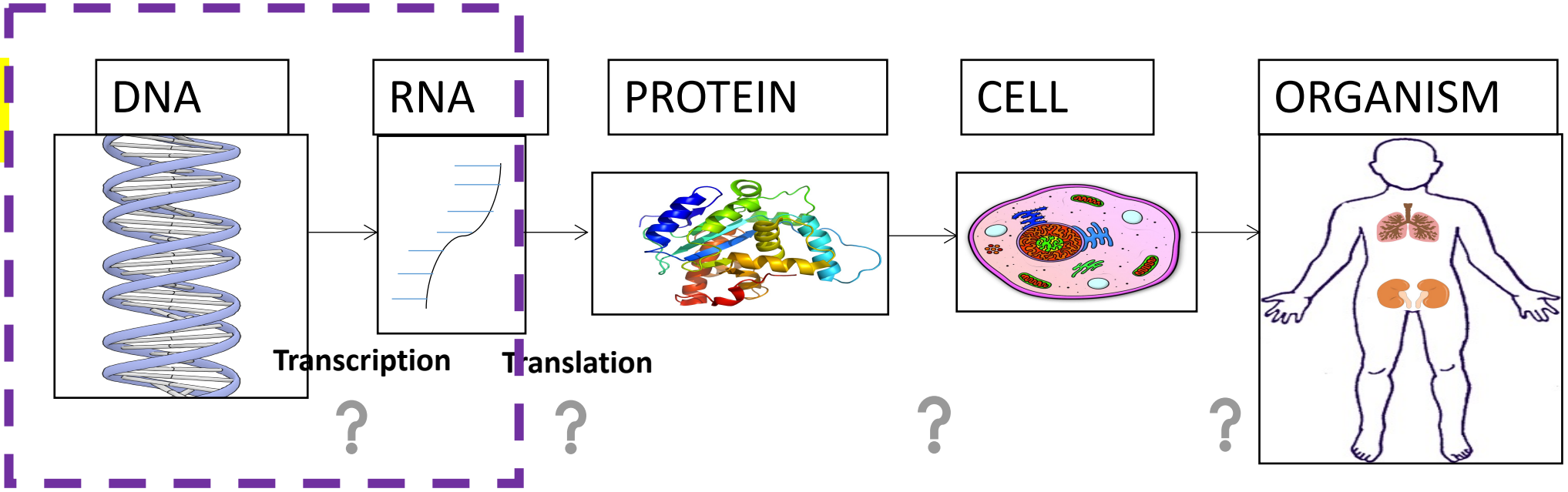
ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

To understand gene regulation

gene repressed

ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

Biological Modules

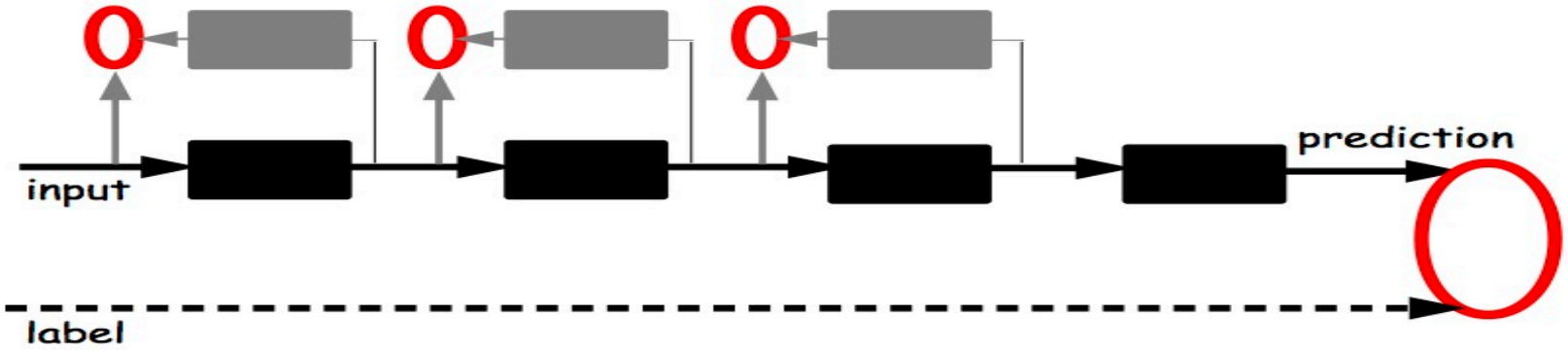


CATGACTG
CATGCCTG

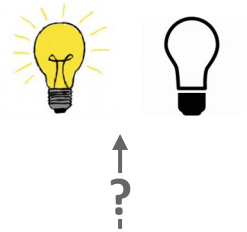
Genetic Variant

Disease

Deep Learning Modules (composable)



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

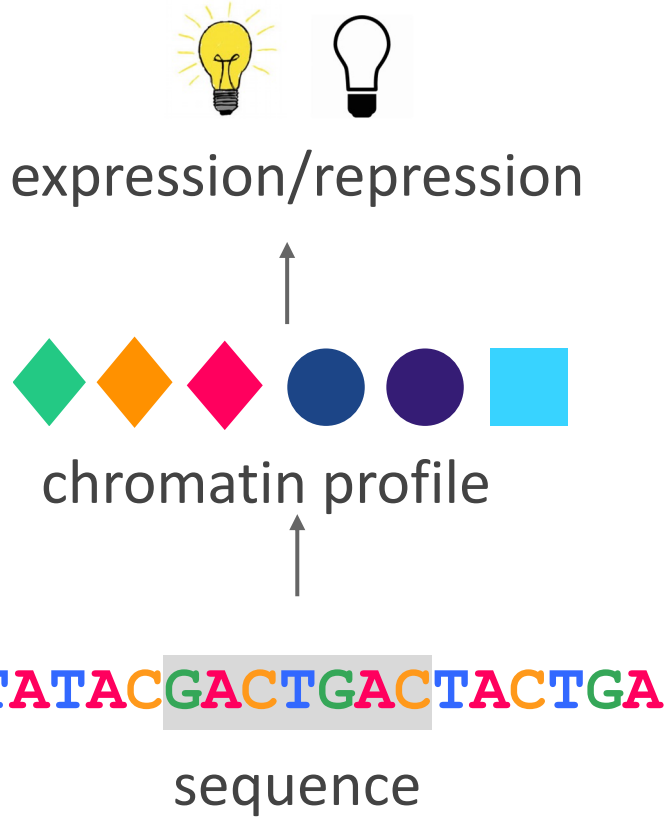
Transcription Factors | Histone Modifications | DNA Accessibility

Chromatin Profile

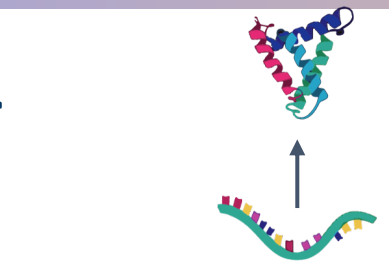
Chromatin Profile Attributes

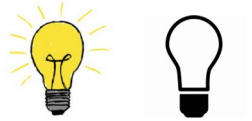
Transcription Factors | Histone Modifications | DNA Accessibility

Gene Regulation

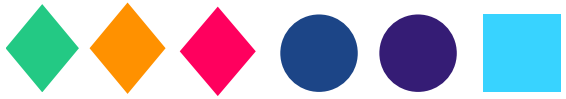


Gene Regulation and after





expression/repression

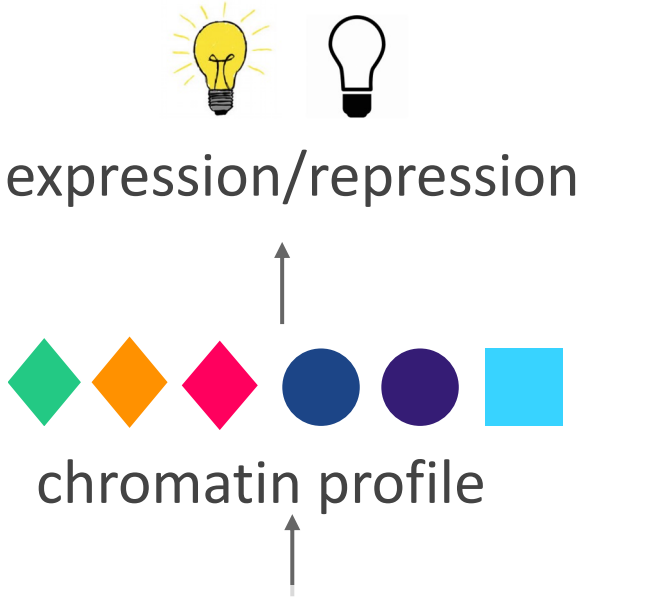


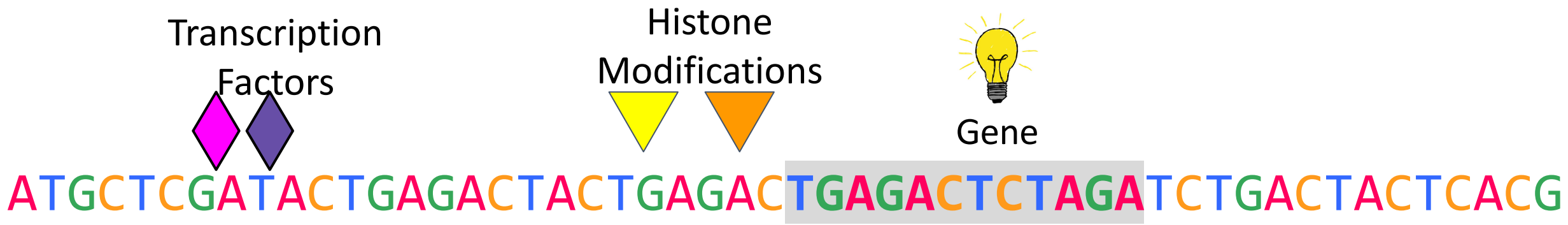
chromatin profile

CTATACGACTGACTACTGA

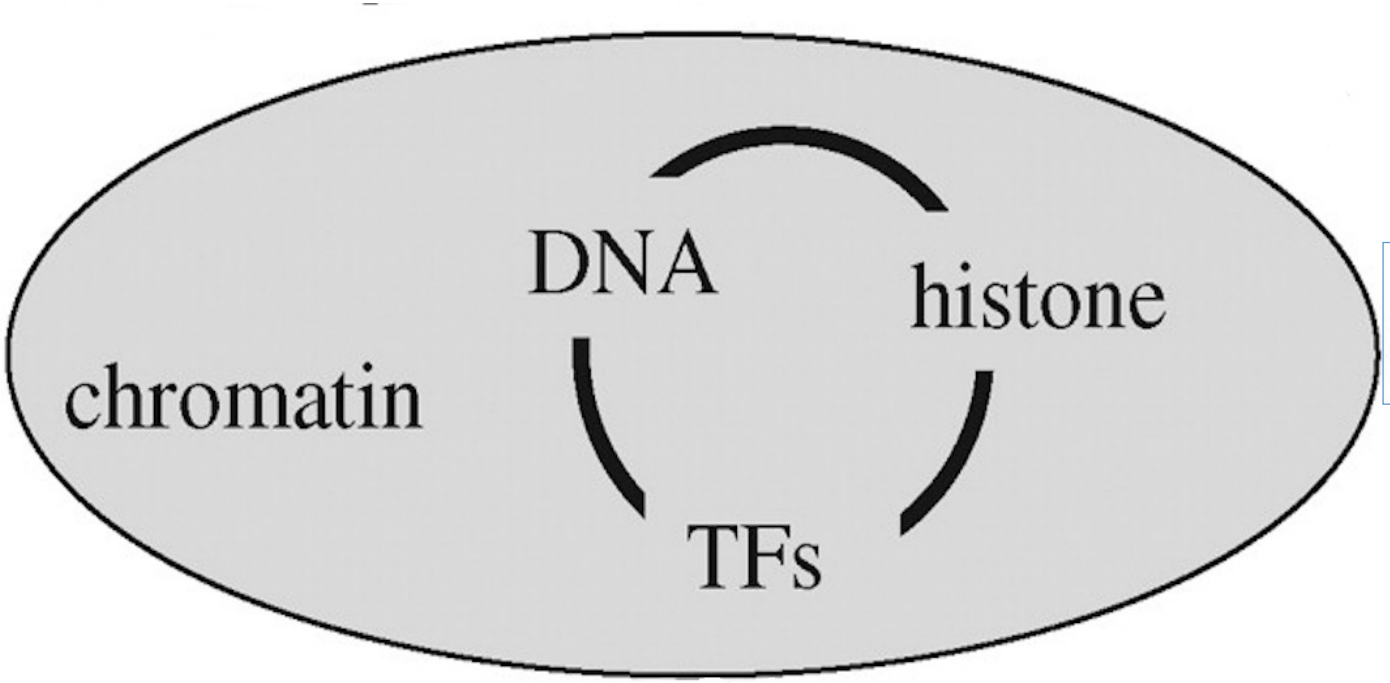
sequence

First Task:


expression/repression
chromatin profile
CTATACGACTGACTACTGA
sequence

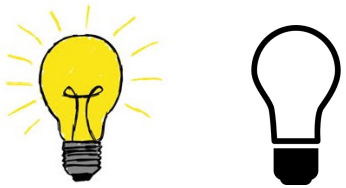

Transcription Factors
Histone Modifications
Gene
ATGCTCGATACTGAGACTACTGAGACTGAGACTCTAGATCTGACTACTCACG

Chromatin Profile as Evidence

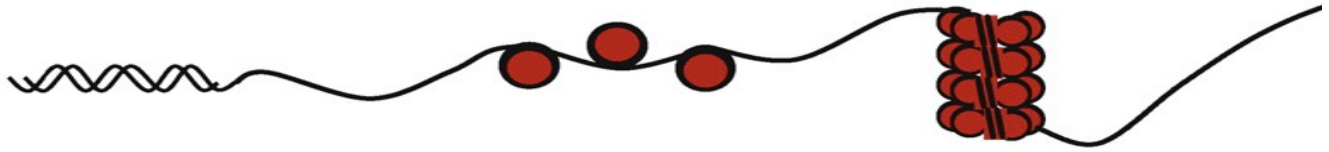


Control

Gene Transcription



Epigenetics "Environment of the DNA"



Level 1

Level 2

Regulatory Elements

Genes
Promoters
Enhancers

Chromatin Structure

Histone Modifications
DNA methylation
Chromatin remodeling

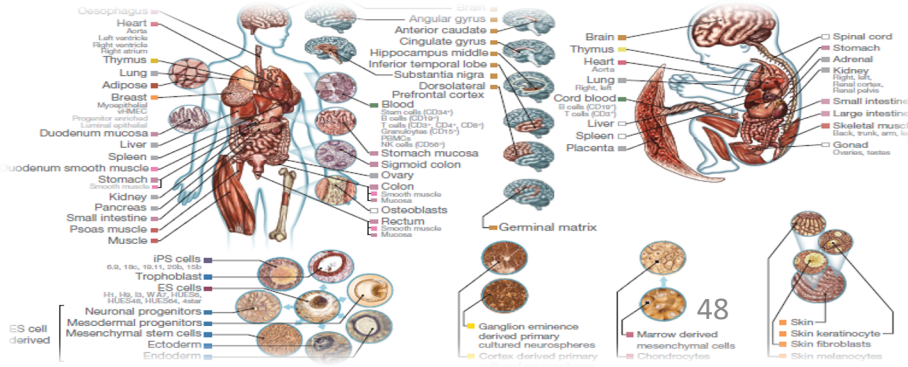
ENCODE Project (2003-)

Describe the functional elements encoded in human DNA

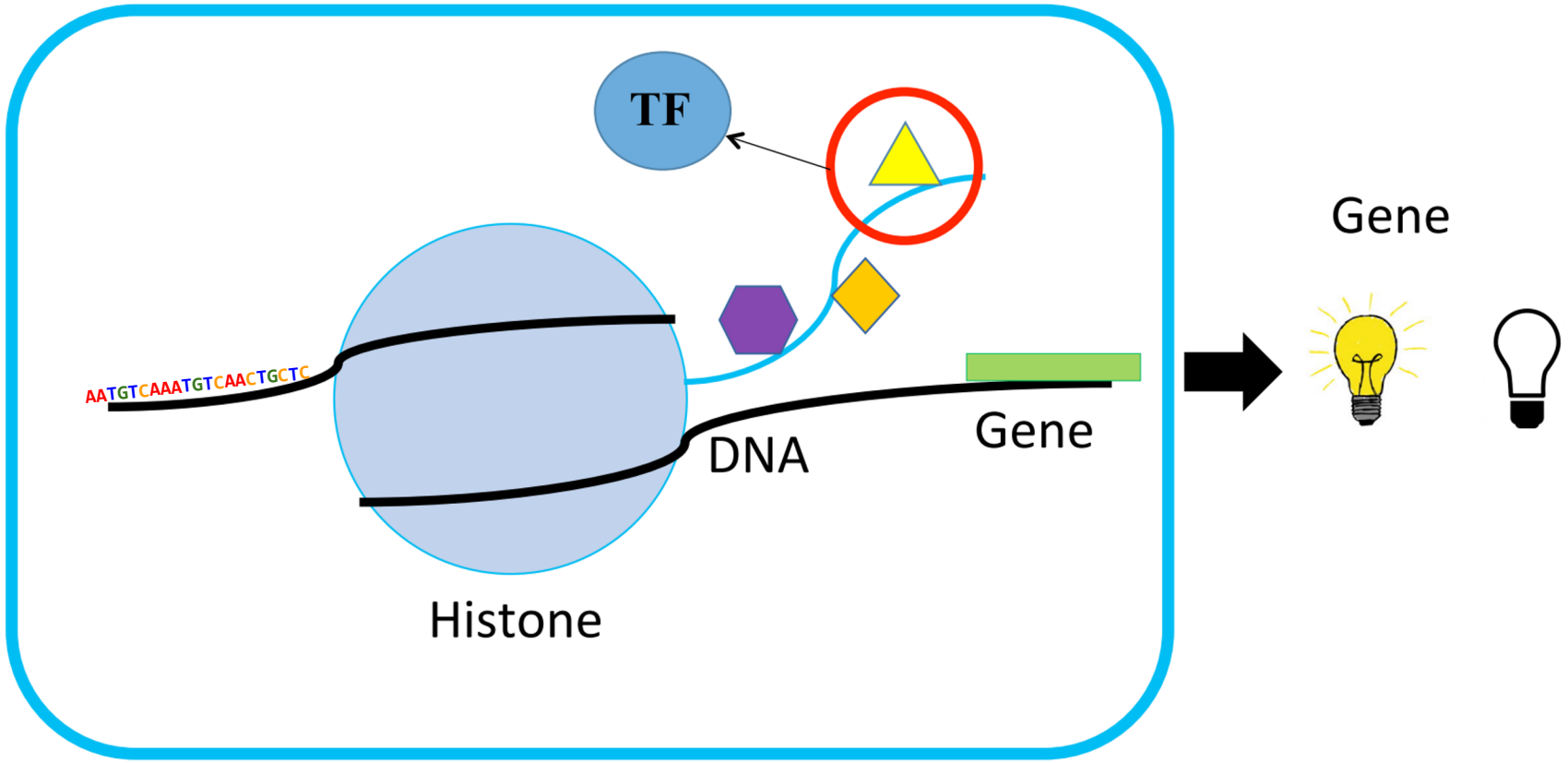
Roadmap Epigenetics Project (REMC, 2008-)

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.

Integrative analysis of 111 reference human epigenomes (Abstract)

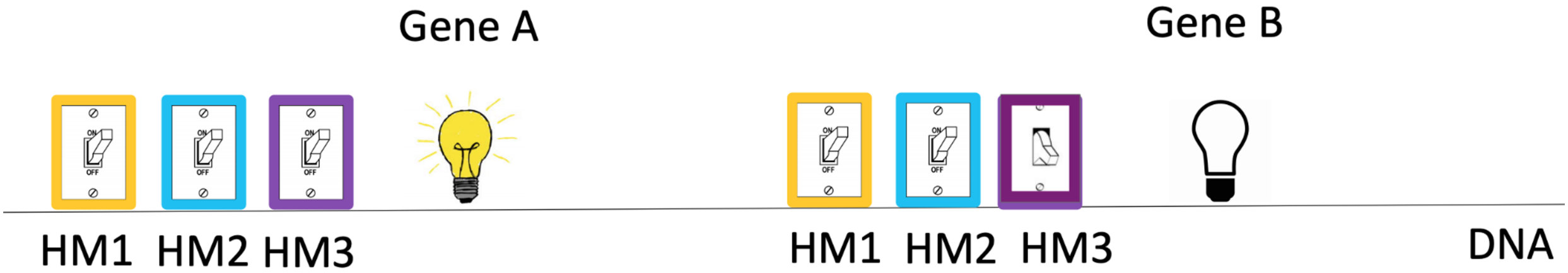


Histone Modifications (HMs)



Can we predict gene expression from histone modification signals?

What HMs affect which genes in what cells?



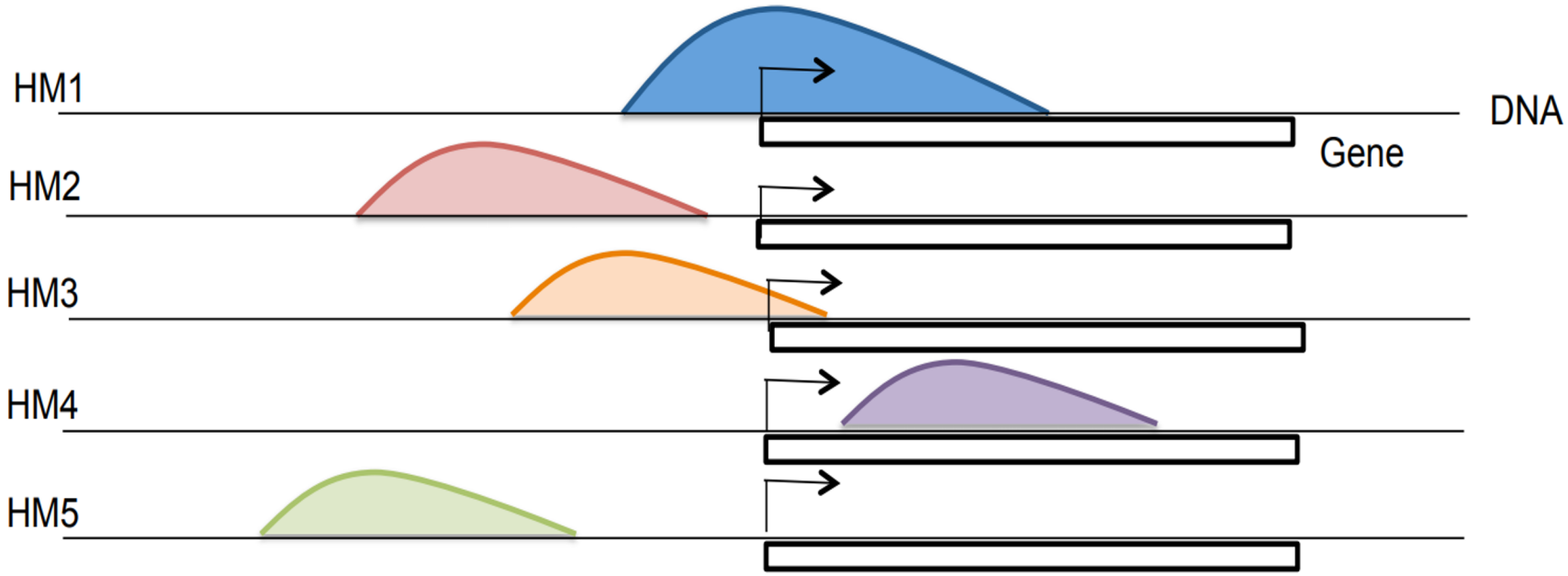
Gene Transcription Prediction Task

Why Study HM → Gene Expression?

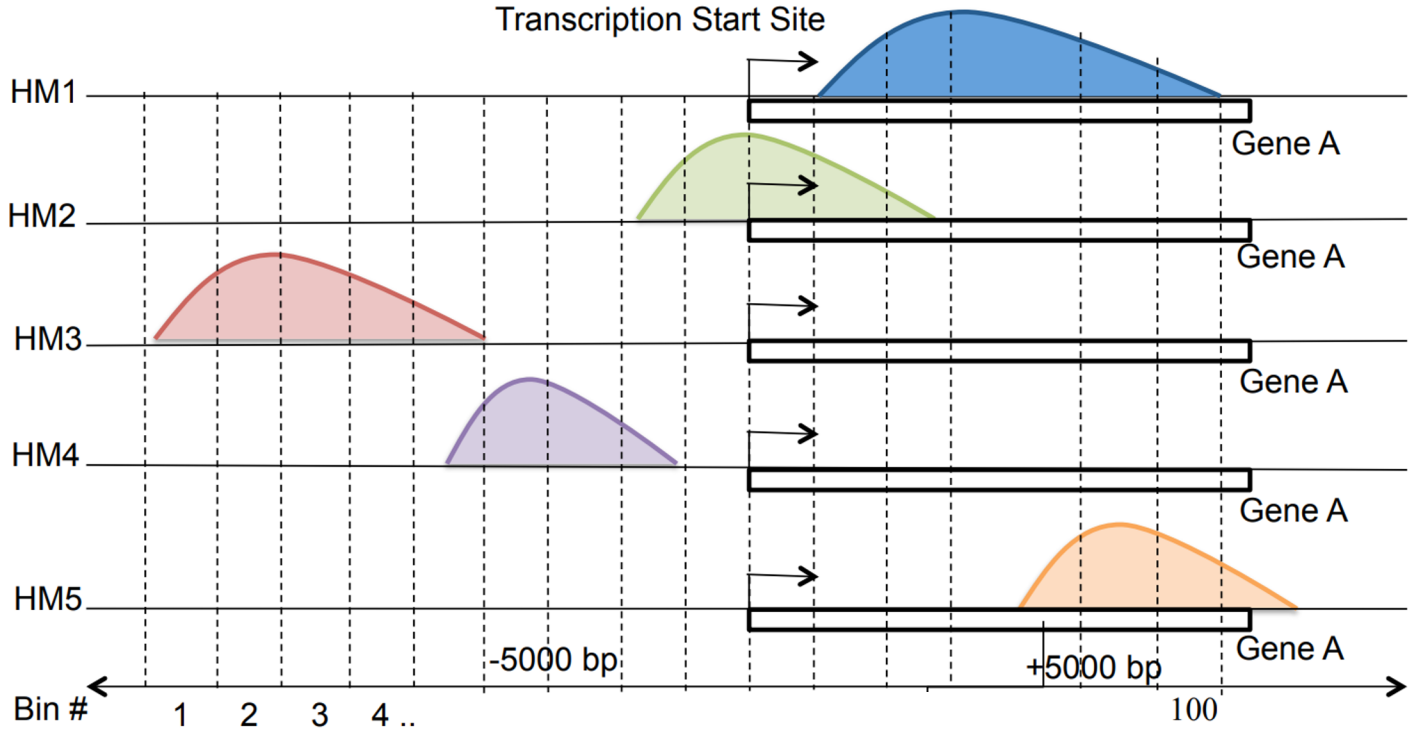
- **Epigenomics:** study of chemical changes in DNA and histones (without altering DNA sequence)
- **Epigenome is dynamic:** can be altered by environmental conditions.

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

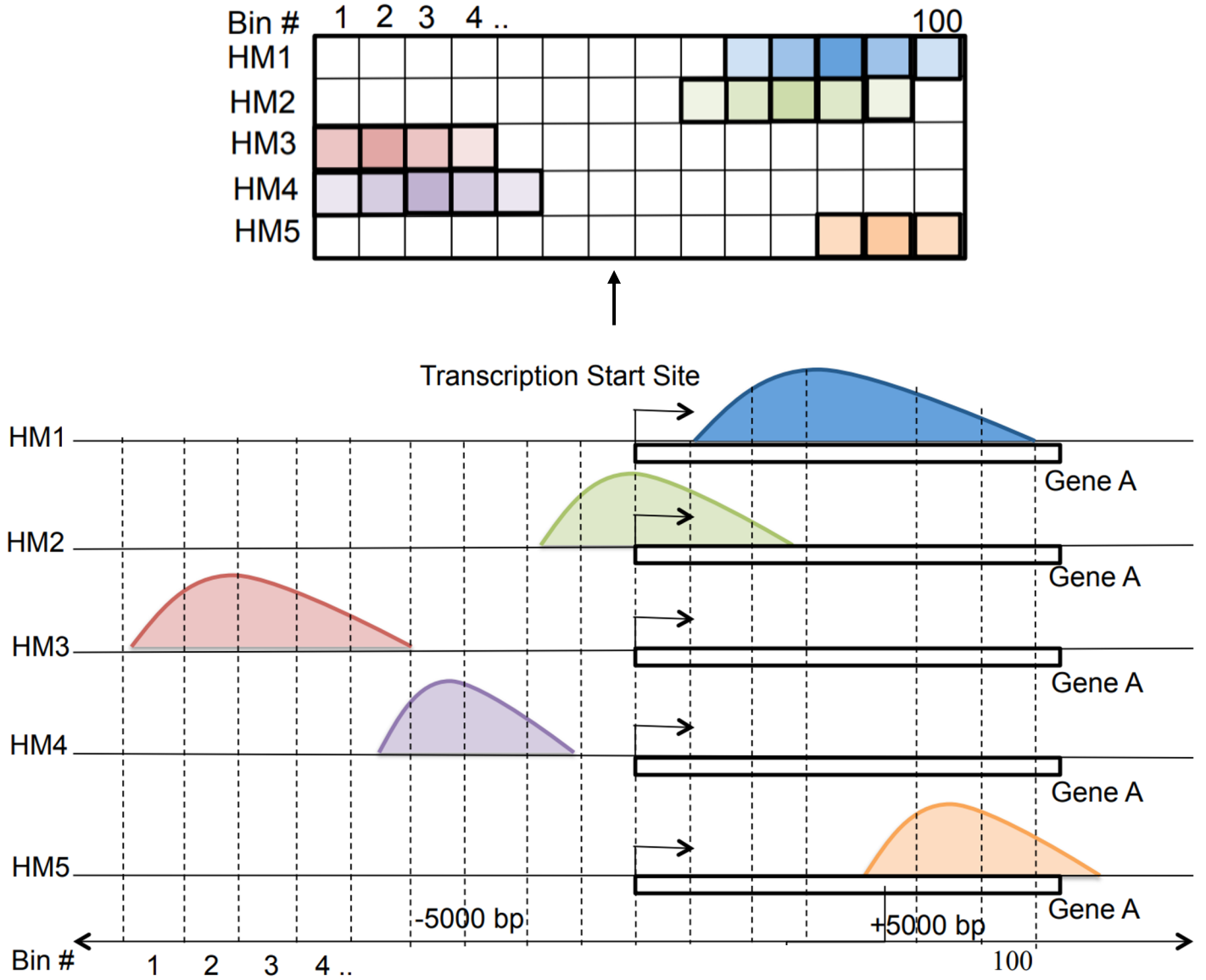
Histone Modification Input Data



Histone Modification Input Data

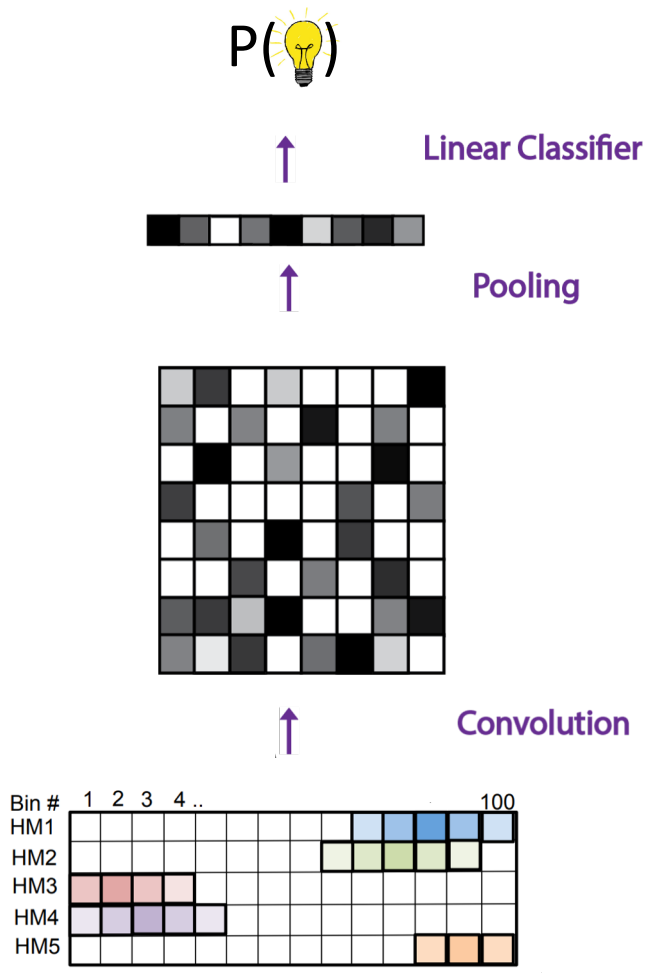


Histone Modification Input Data



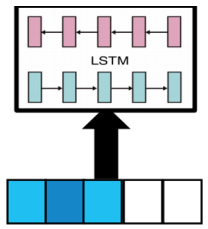
DeepChrome

Singh, Lanchantin, Robins & Qi - Bioinformatics 2016

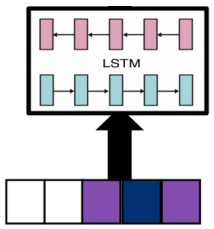


Attentive Chrome

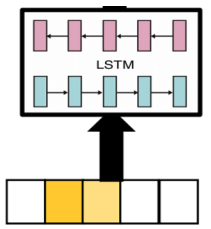
Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



HM1



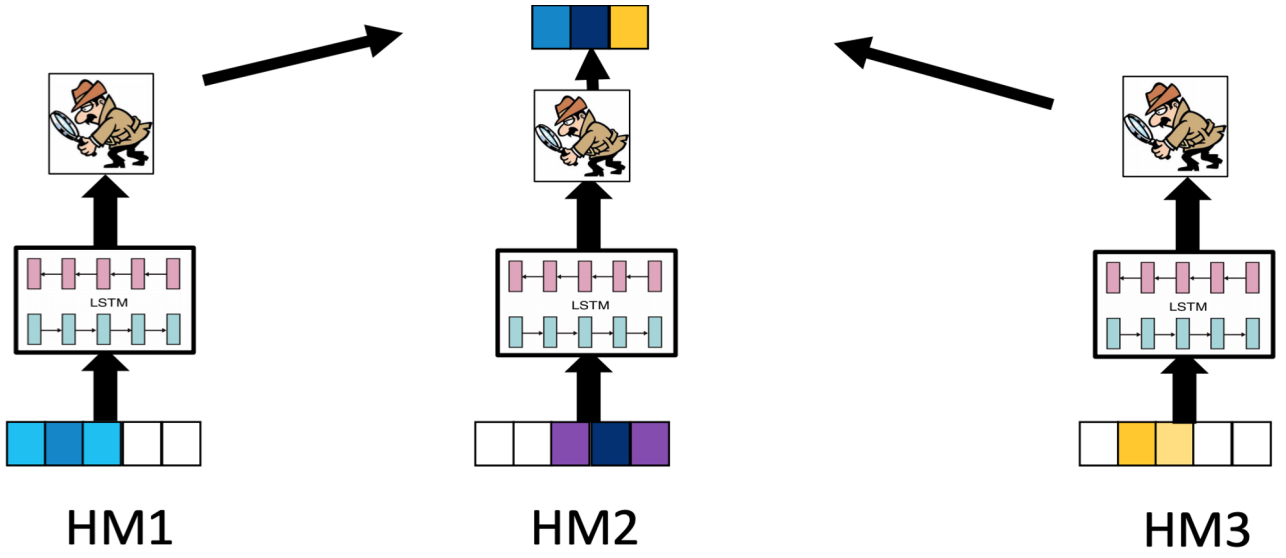
HM2



HM3

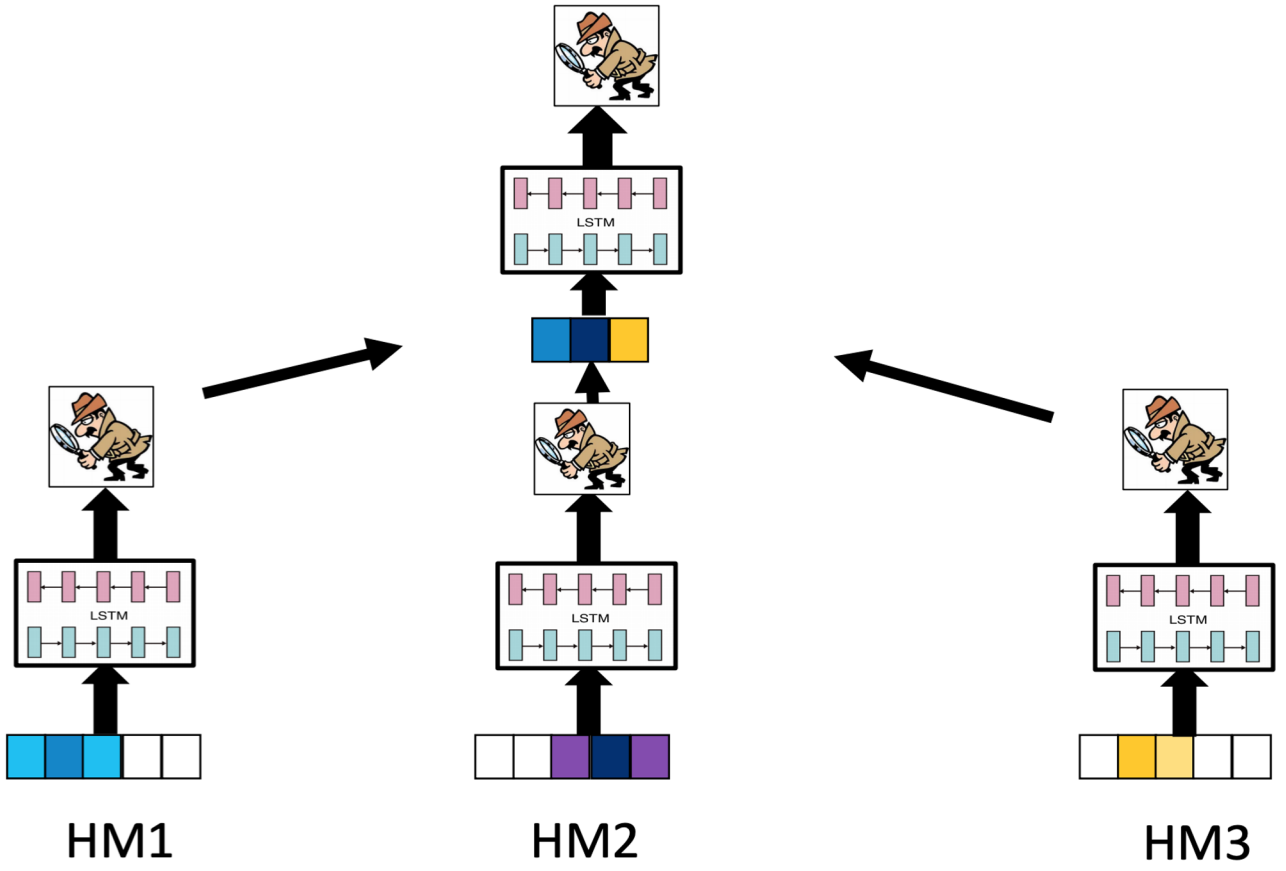
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



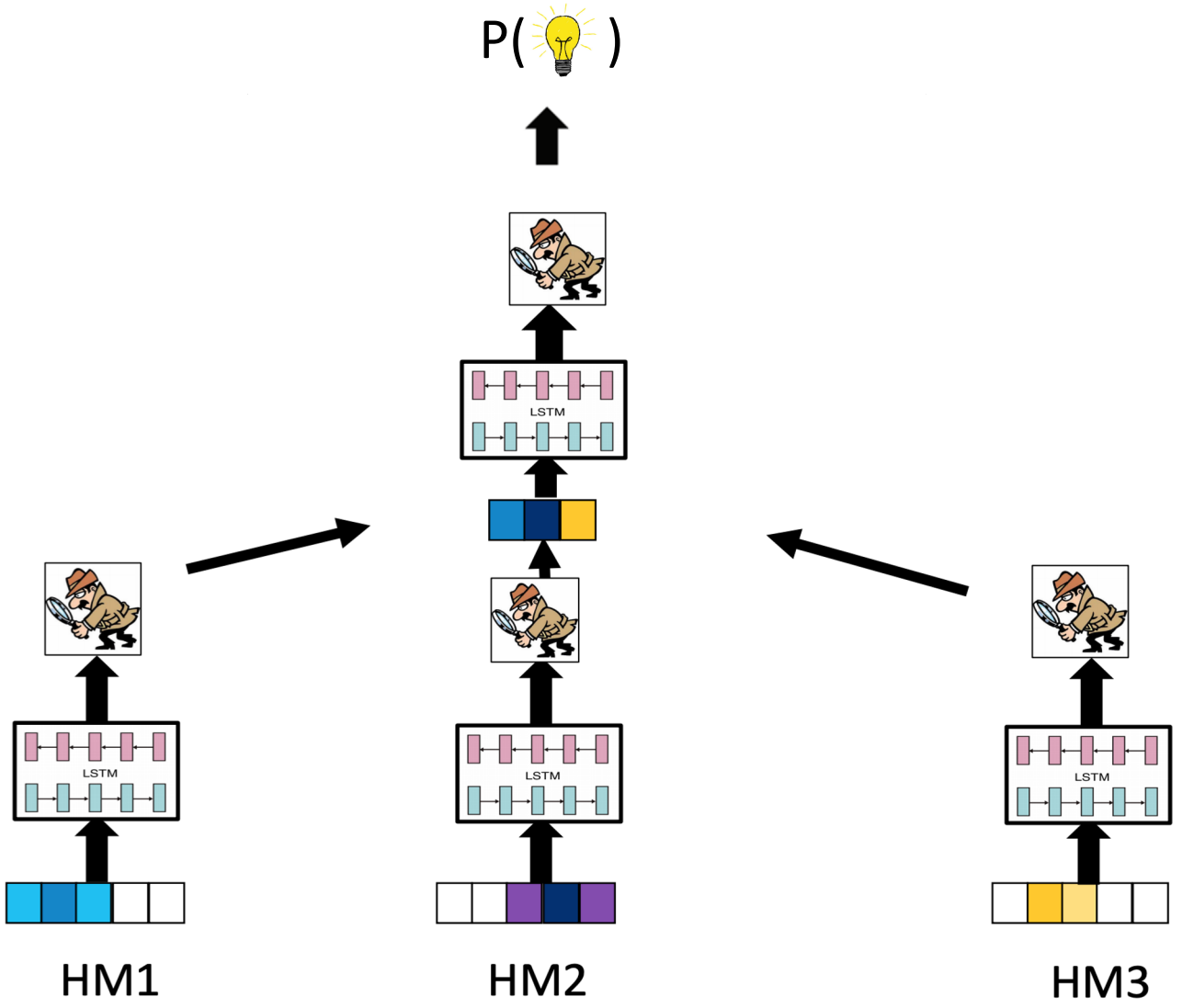
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



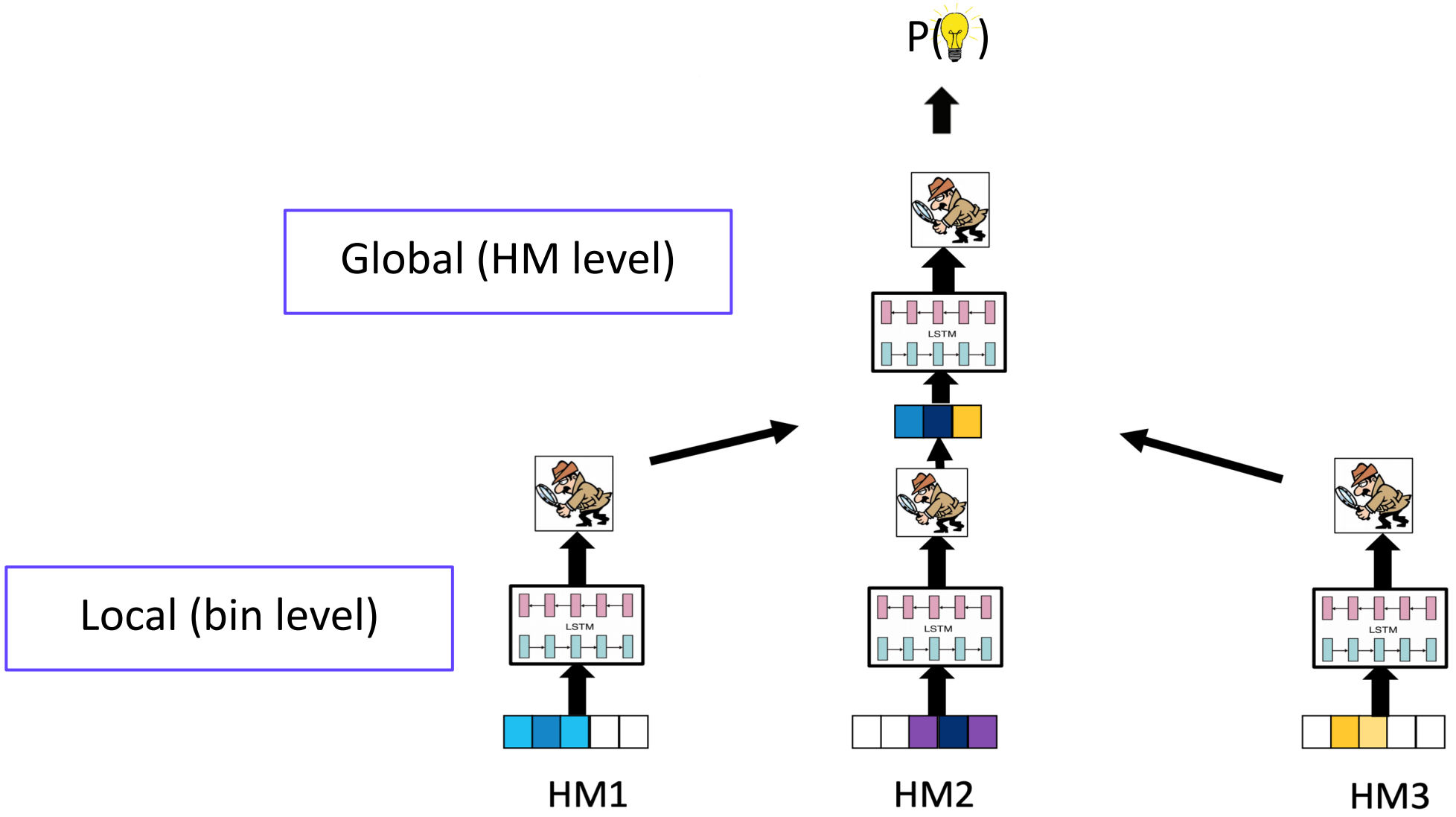
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



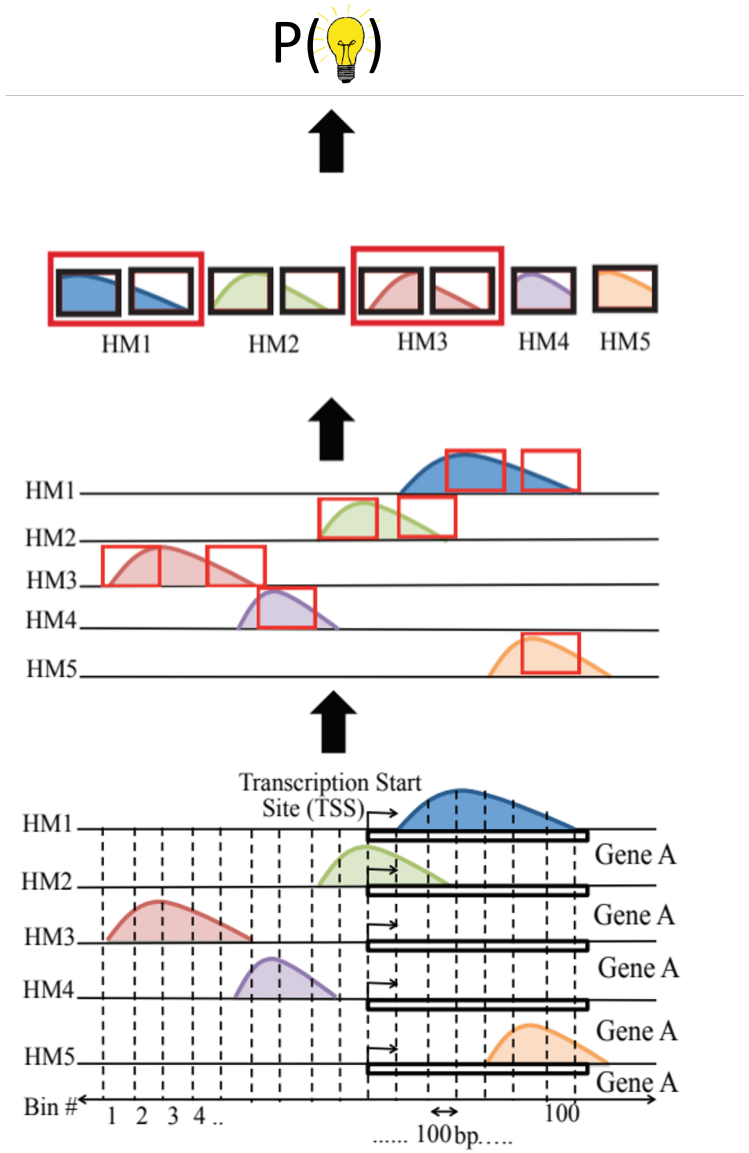
Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



Attentive Chrome

Singh, Lanchantin, Sekhon, & Qi - NeurIPS 2017



Interpretability by Hierarchical Attention

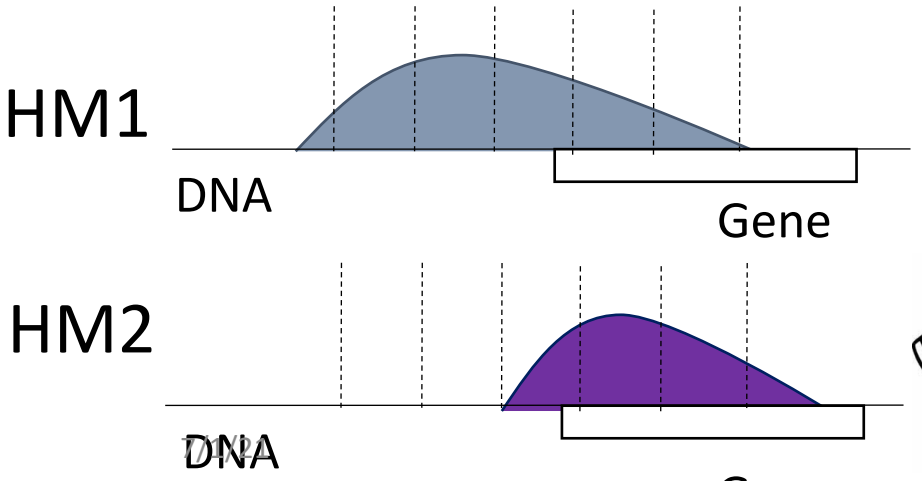
Input

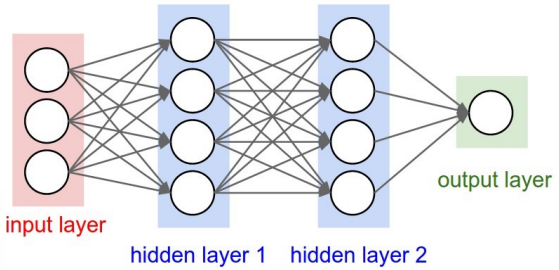
Attention Mechanism

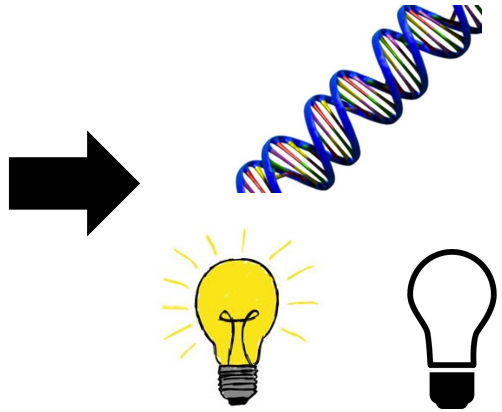
Output

Park

Gene







Interpretability by Hierarchical Attention

Input

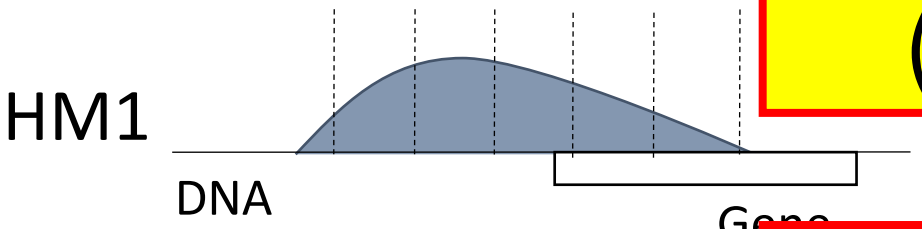
Output

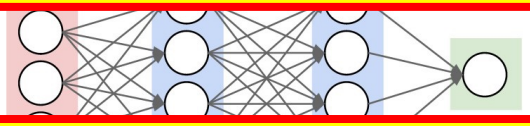
Attention Mechanism

Park

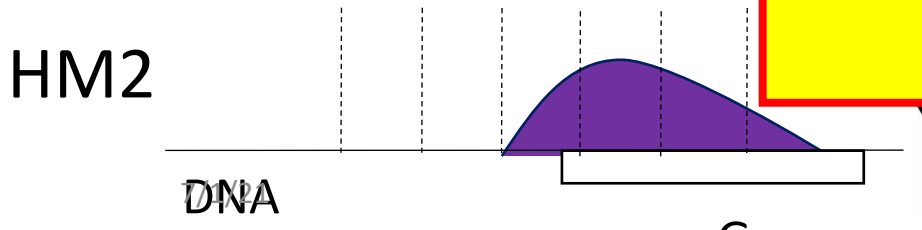
Gene

(1) What positions are important?



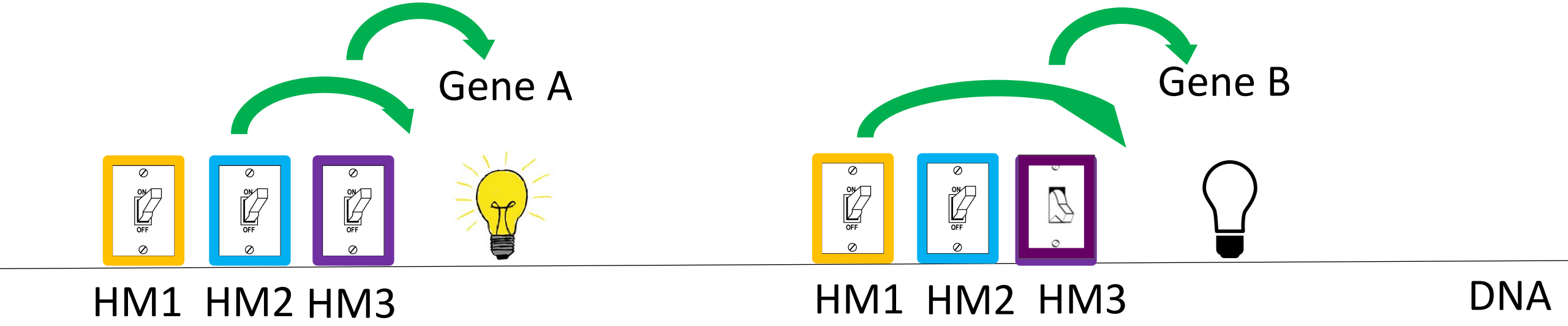


(2) What HMs are important?



Data Sets

~56 Cell Types



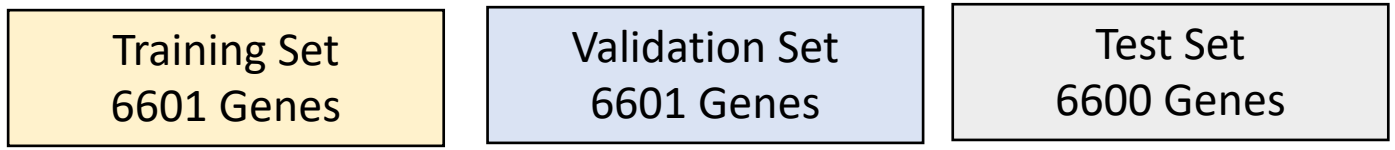
K~20 HMs
G~30,000 Genes

Experimental Setup

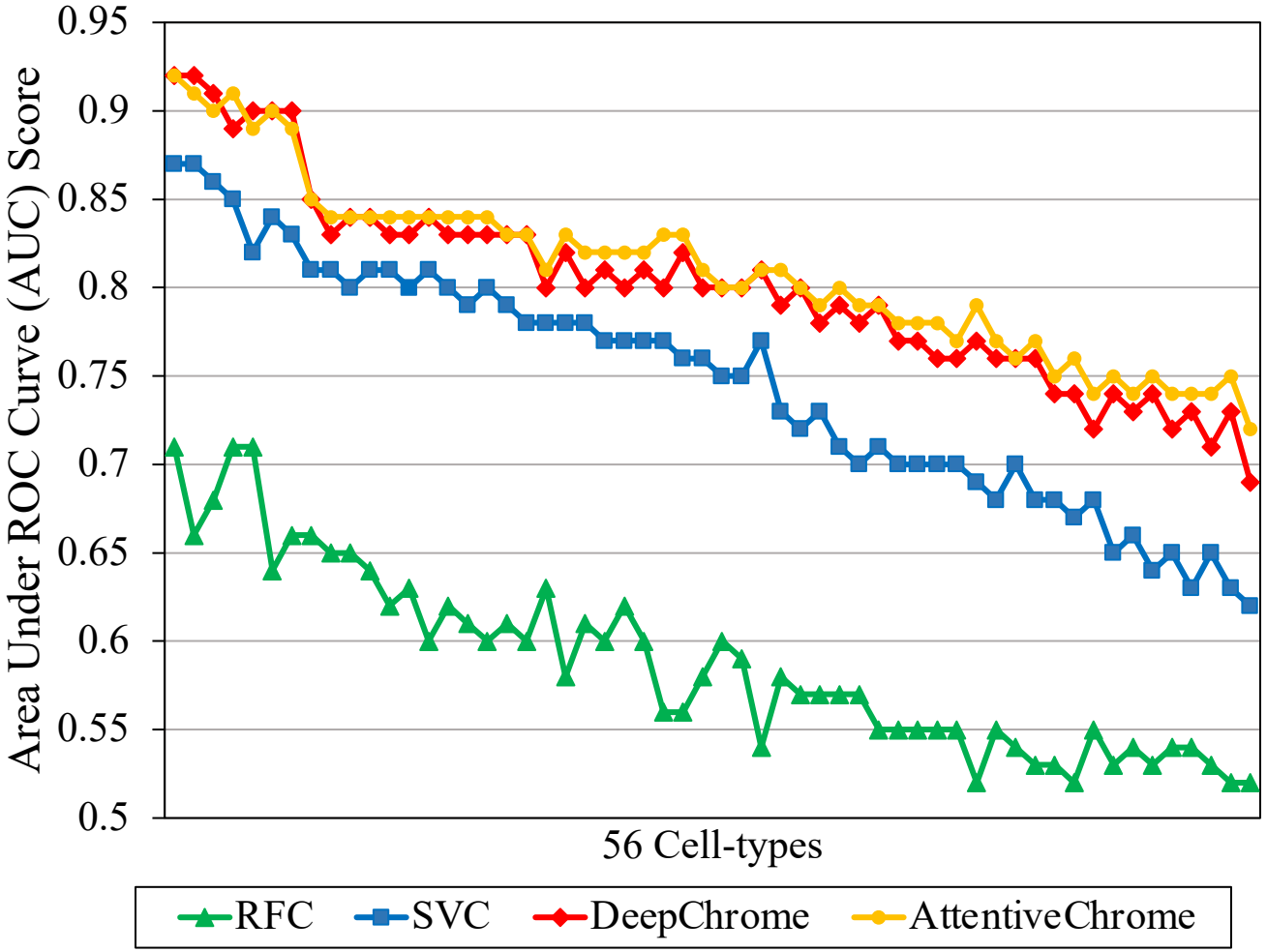
- Roadmap Epigenetics Project (REMC)
- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps / 5 Tier-1 HMs

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)



Prediction



Improvement for 49/56 Cell-types

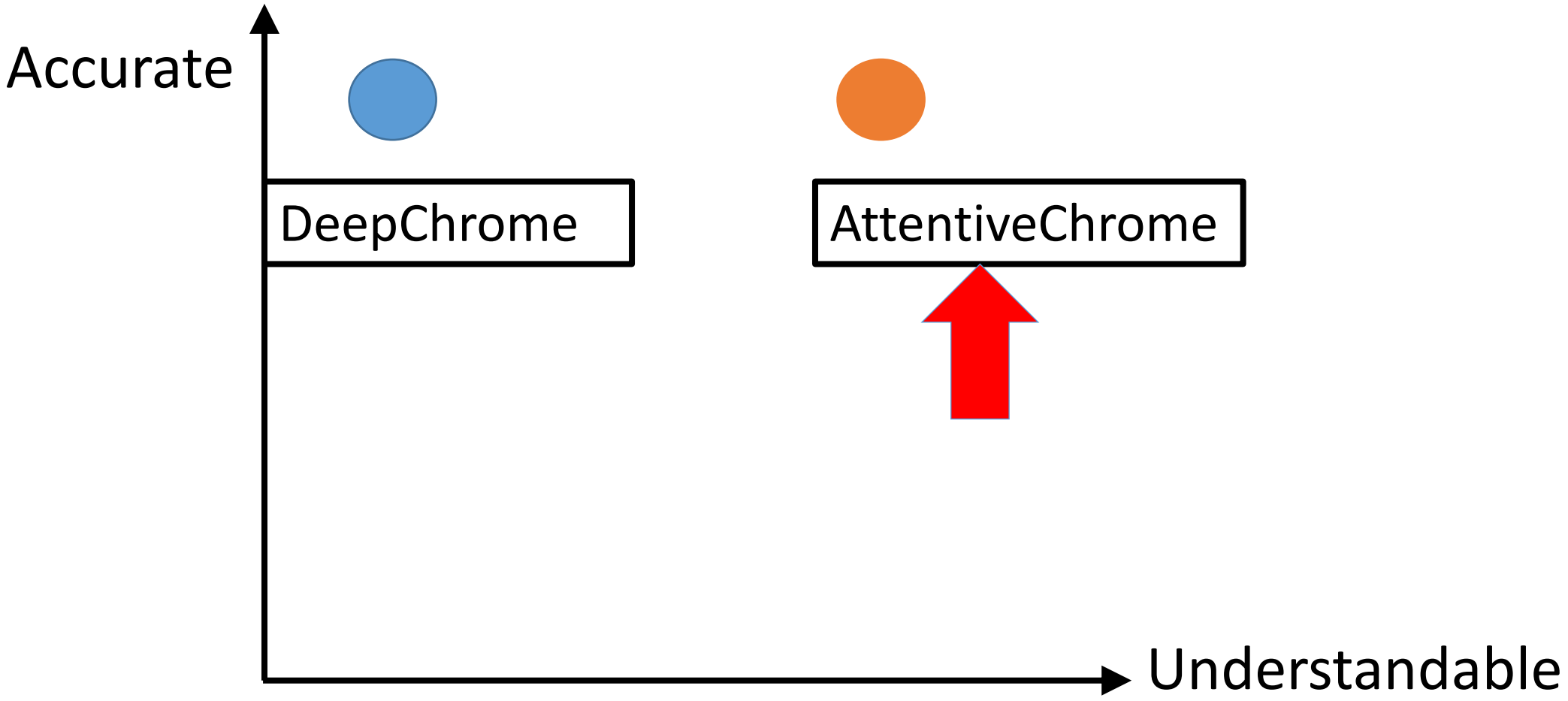
Validation of Attention Weights (using one extra HM signals)

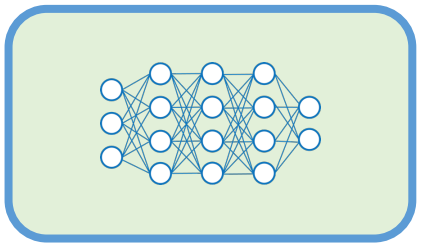
Table 3: Pearson Correlation values between weights assigned for H_{prom} (active HM) by different visualization techniques and H_{active} read coverage (indicating actual activity near "ON" genes) for predicted "ON" genes across three major cell types.

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

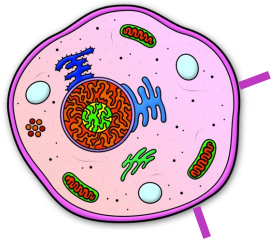
- Additional signal - H3K27ac (H-Active) from REMC
- Average local attention weights of gene=ON correspond well with H-active
- Indicating AttentiveChrome is focusing on the correct bin positions

Summary of tools

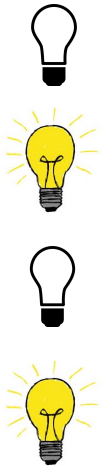




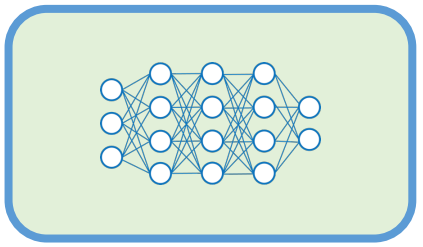
Output (Y) Labels



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

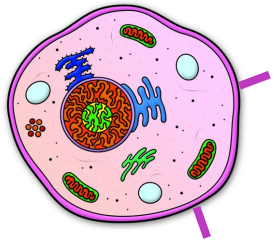


Threshold = 10.245 (Median)



Where are we heading?

Changing Task : Classification →Regression



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
……	…….	……

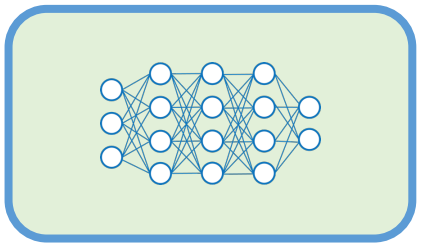
1.770

Gene Expression

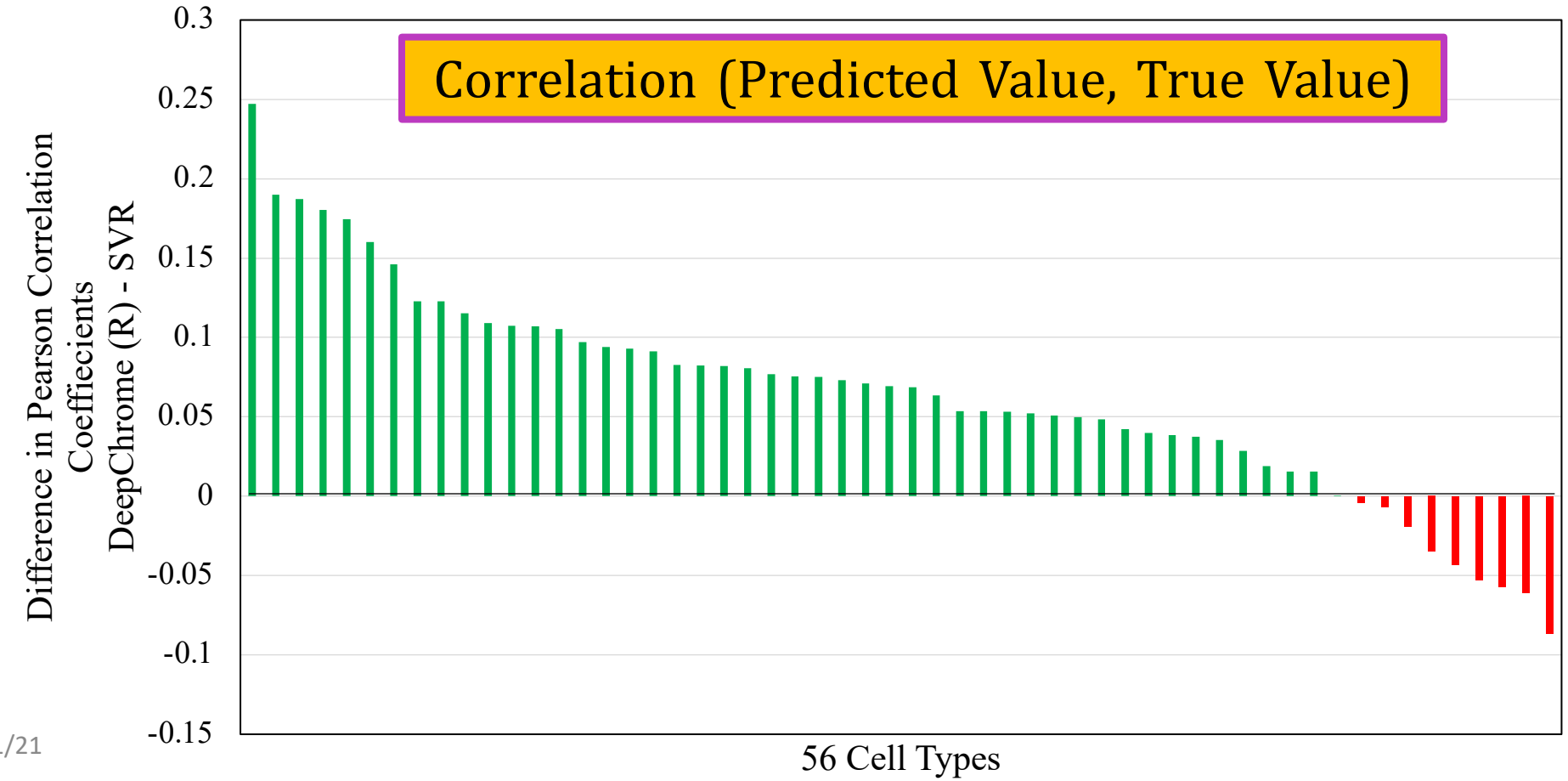
Mean Square Error Loss

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

Where are we heading?



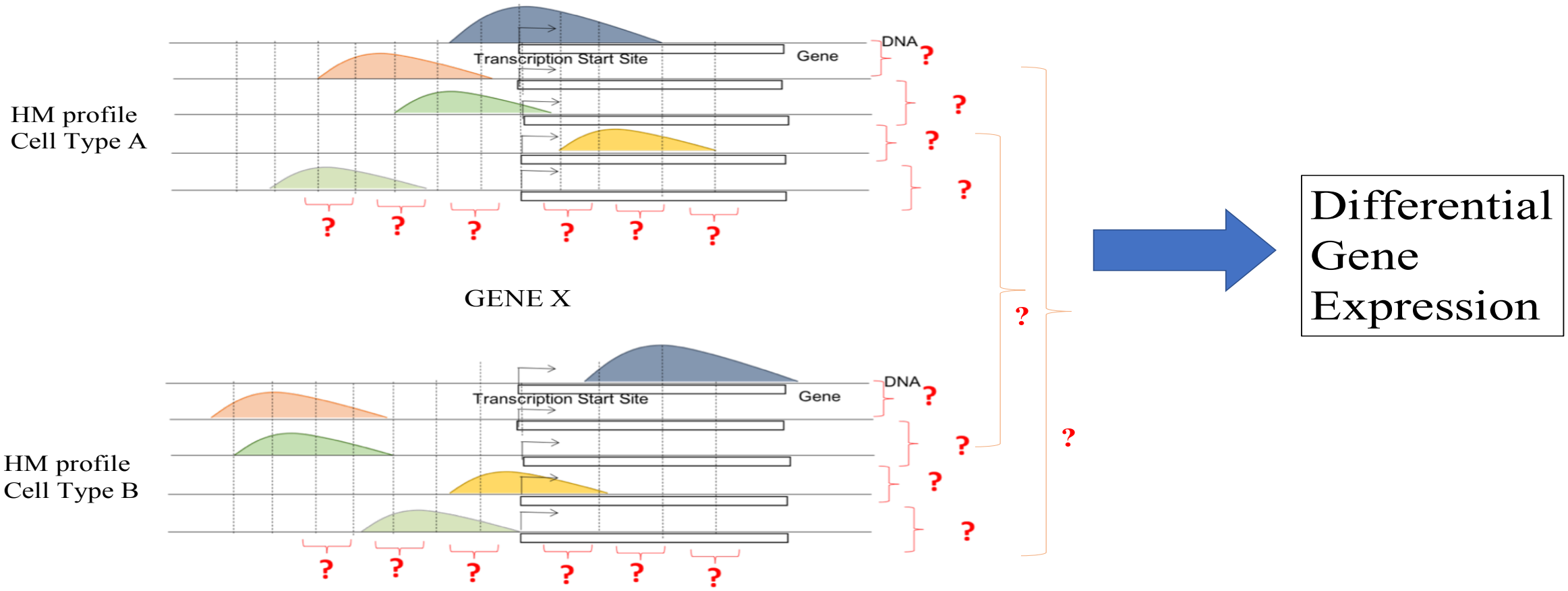
Changing Task : Classification →Regression

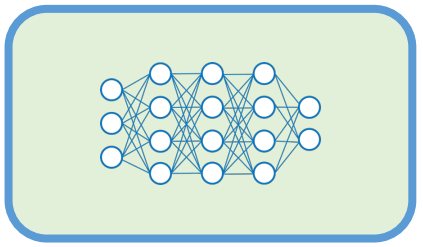


Where are we heading?

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

Changing Task : Cell-Specific → Cross Cell

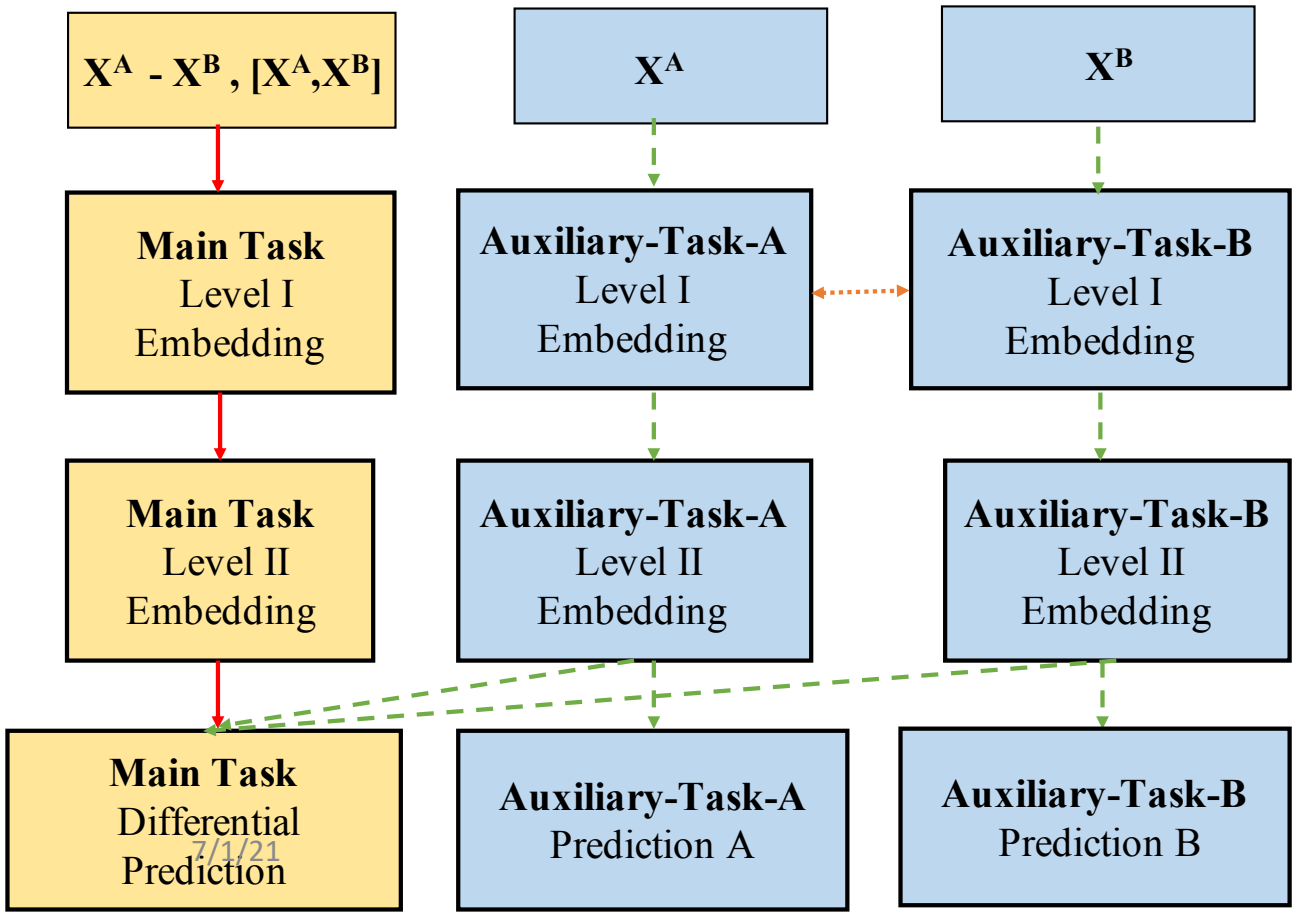




Where are we heading?

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

Changing Task : Cell-Specific → Cross Cell



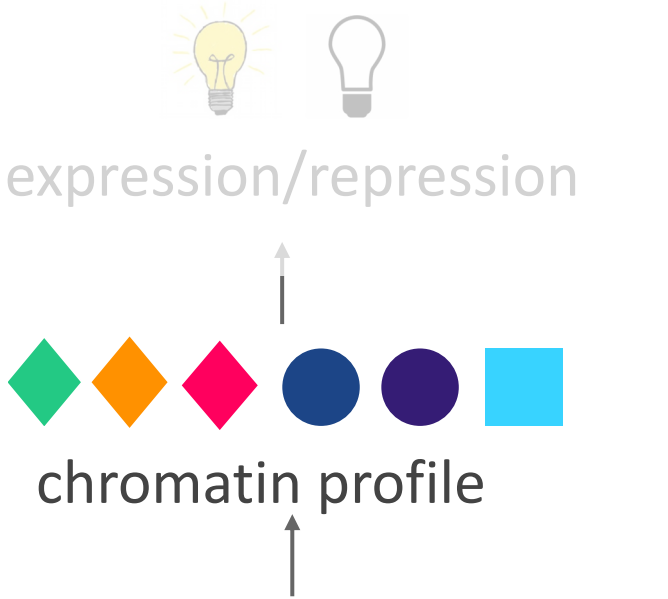
1 **Main Task**: Differential gene expression prediction

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

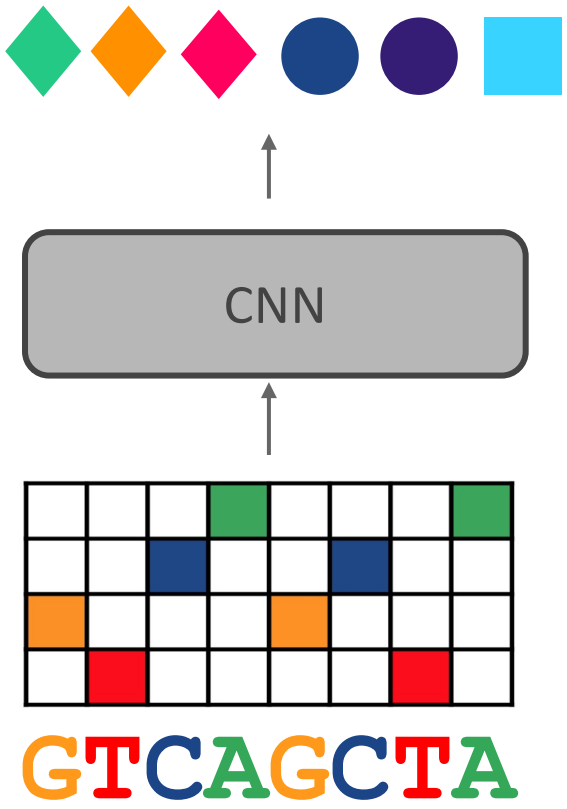
3 **Siamese Auxiliary**: Siamese contrastive loss

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

Second Task:

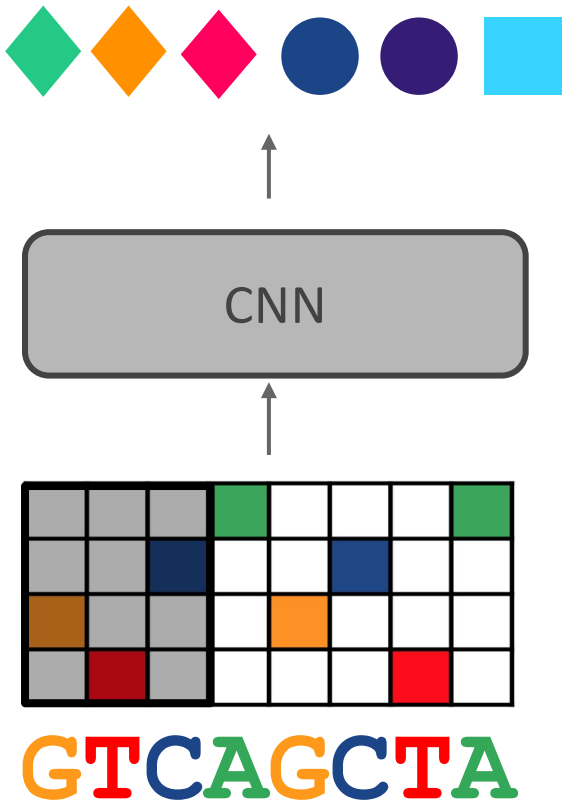

expression/repression
chromatin profile
CTATACGACTGACTACTGA
sequence

Local Sequence Chromatin Profile Prediction



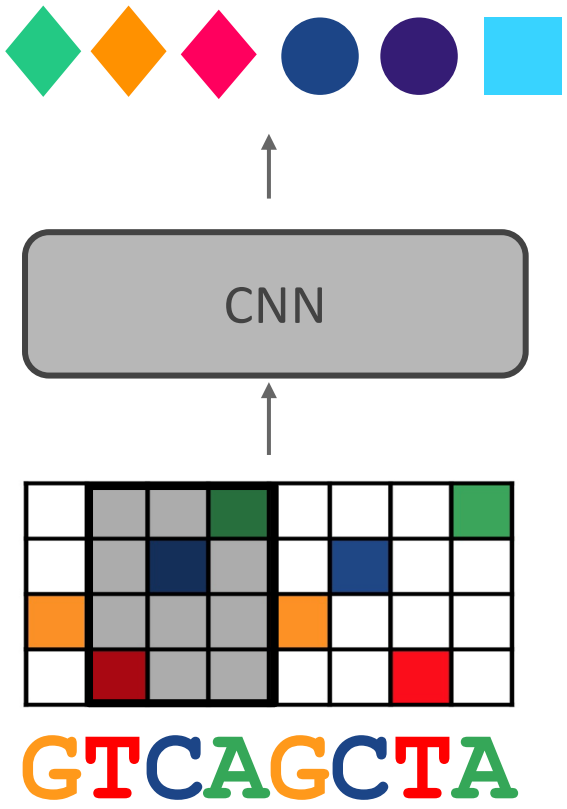
Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



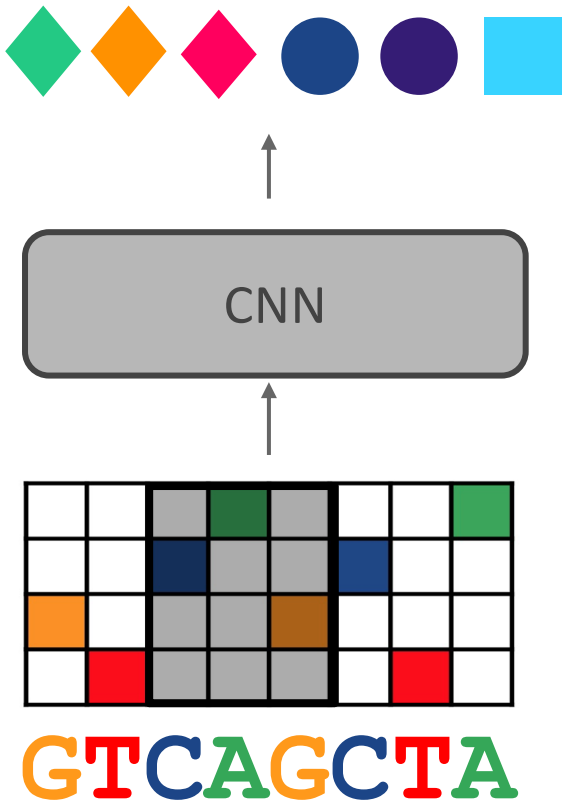
Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



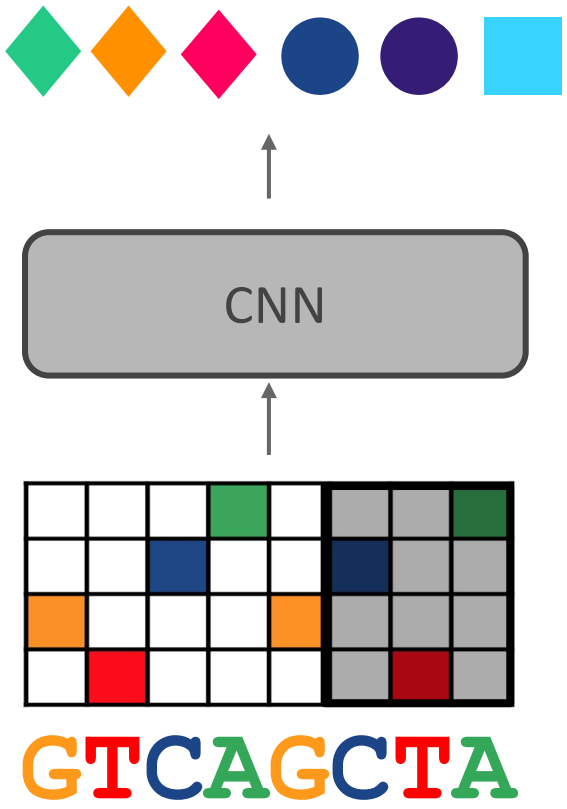
Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



Zhou and Troyanskaya Nature Methods 2015

Local Sequence Chromatin Profile Prediction



Zhou and Troyanskaya Nature Methods 2015

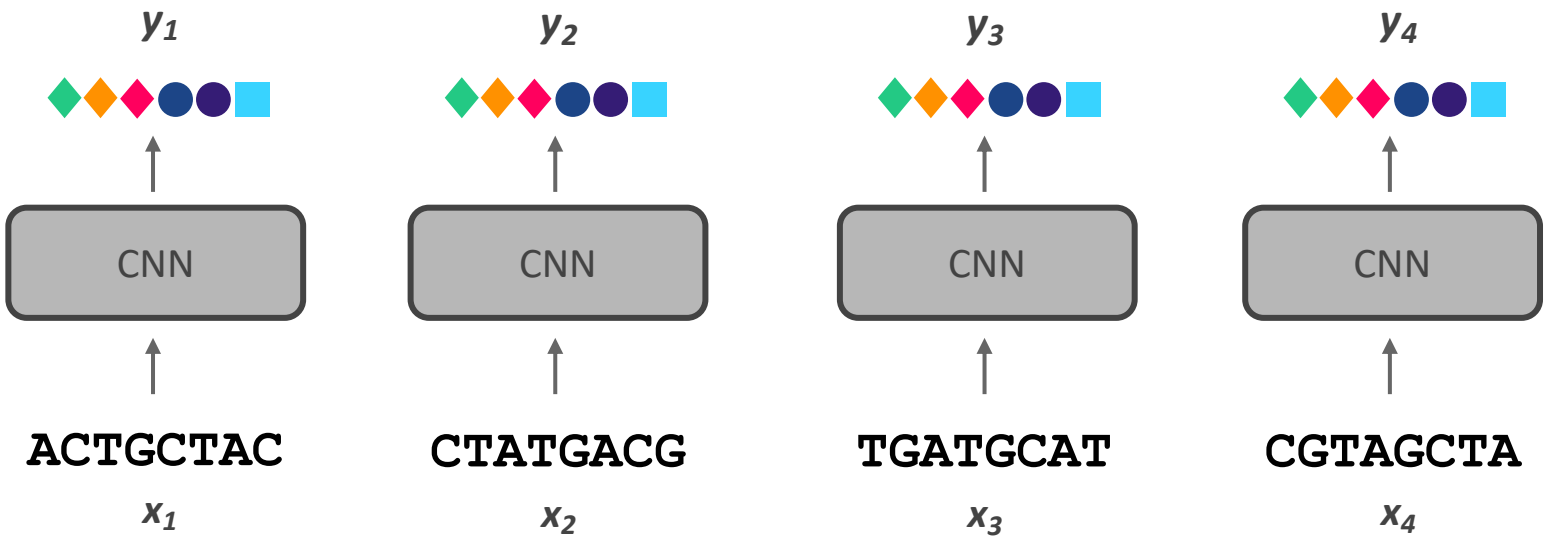
Local Sequence Chromatin Profile Prediction

ACTGCTACCTATGACGTGATGCATCGTAGCT
A

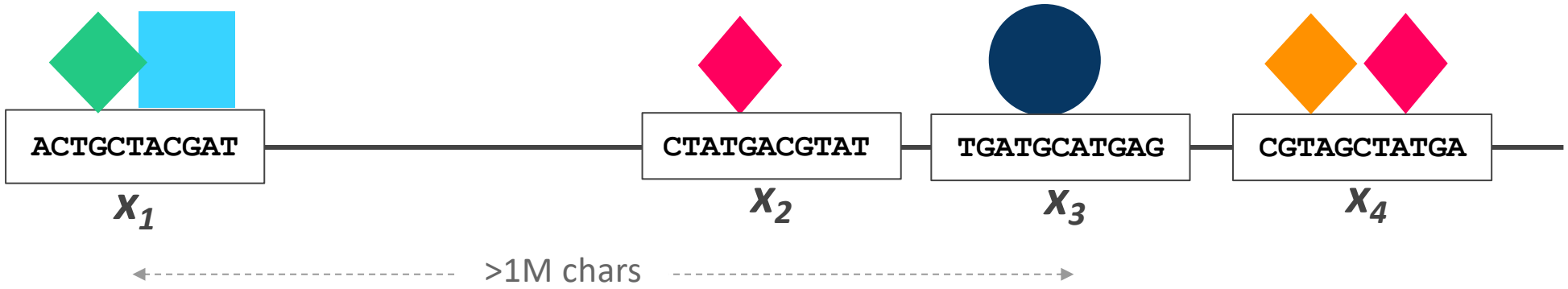
Local Sequence Chromatin Profile Prediction

ACTGCTAC	CTATGACG	TGATGCAT	CGTAGCTA
x_1	x_2	x_3	x_4

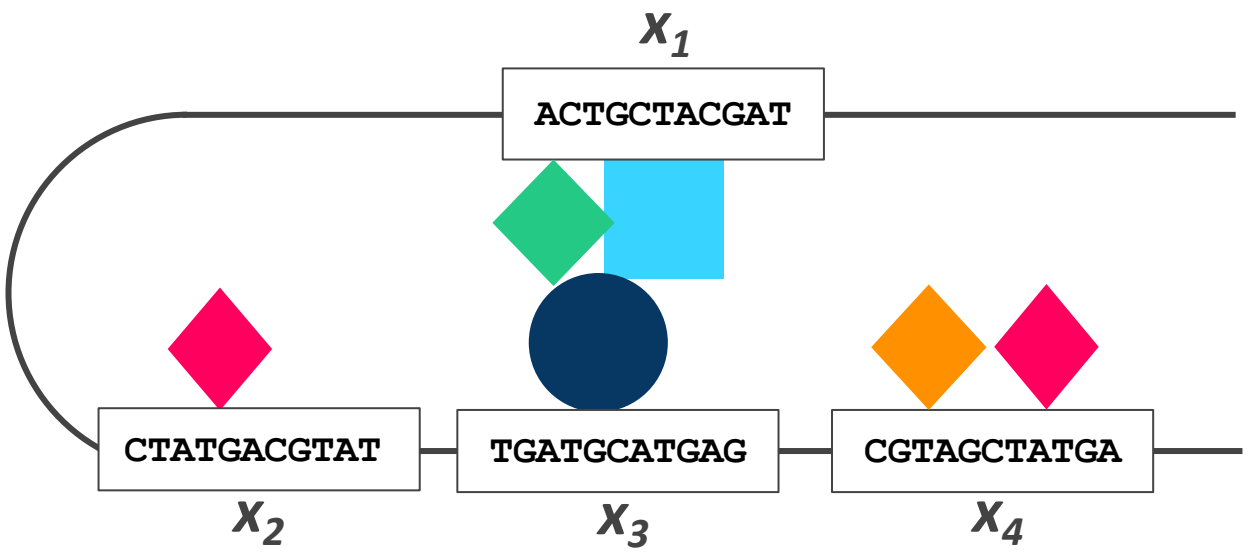
Local Sequence Chromatin Profile Prediction



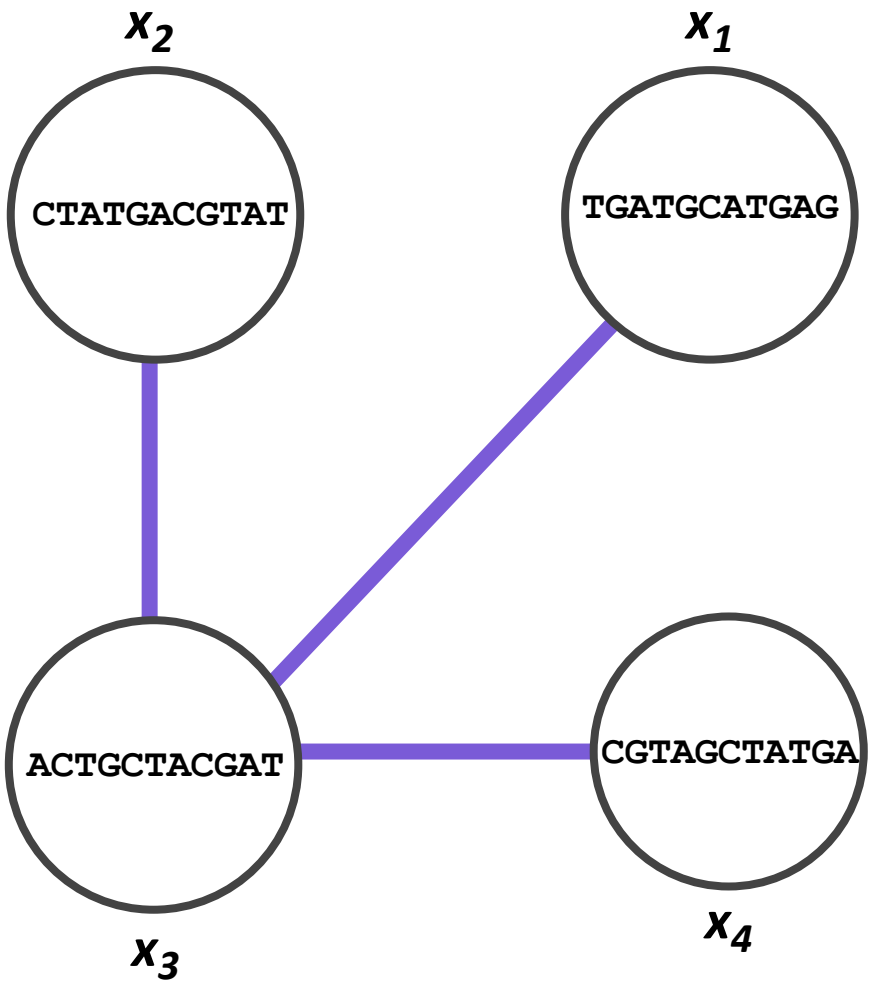
Influence of Long-Range Interactions on Chromatin Profile



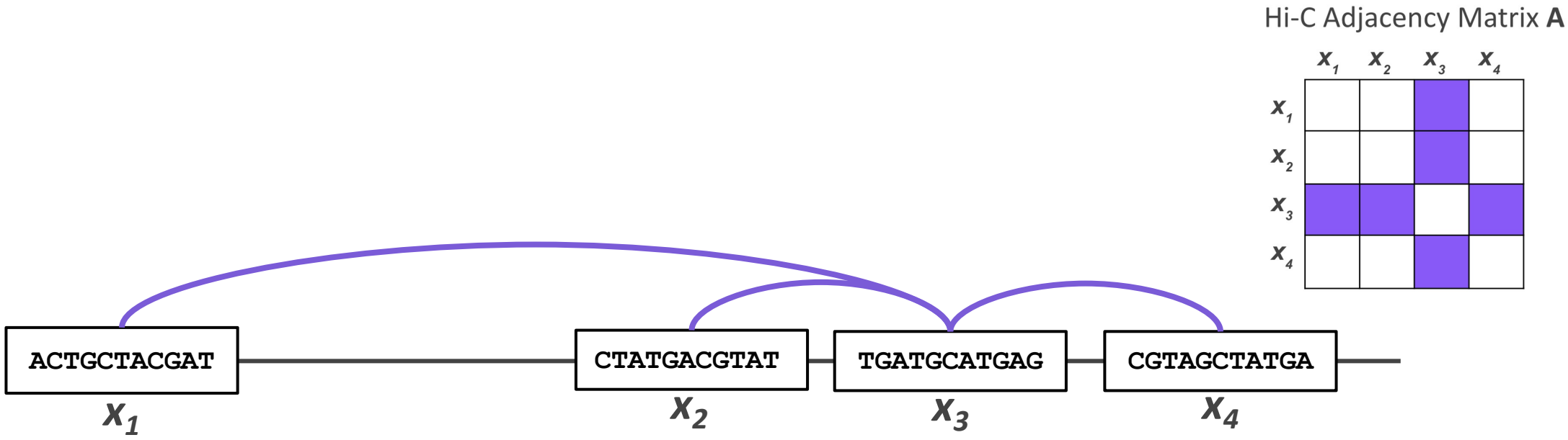
Influence of Long-Range Interactions on Chromatin Profile



Genome: Locally a Sequence, Globally a Graph



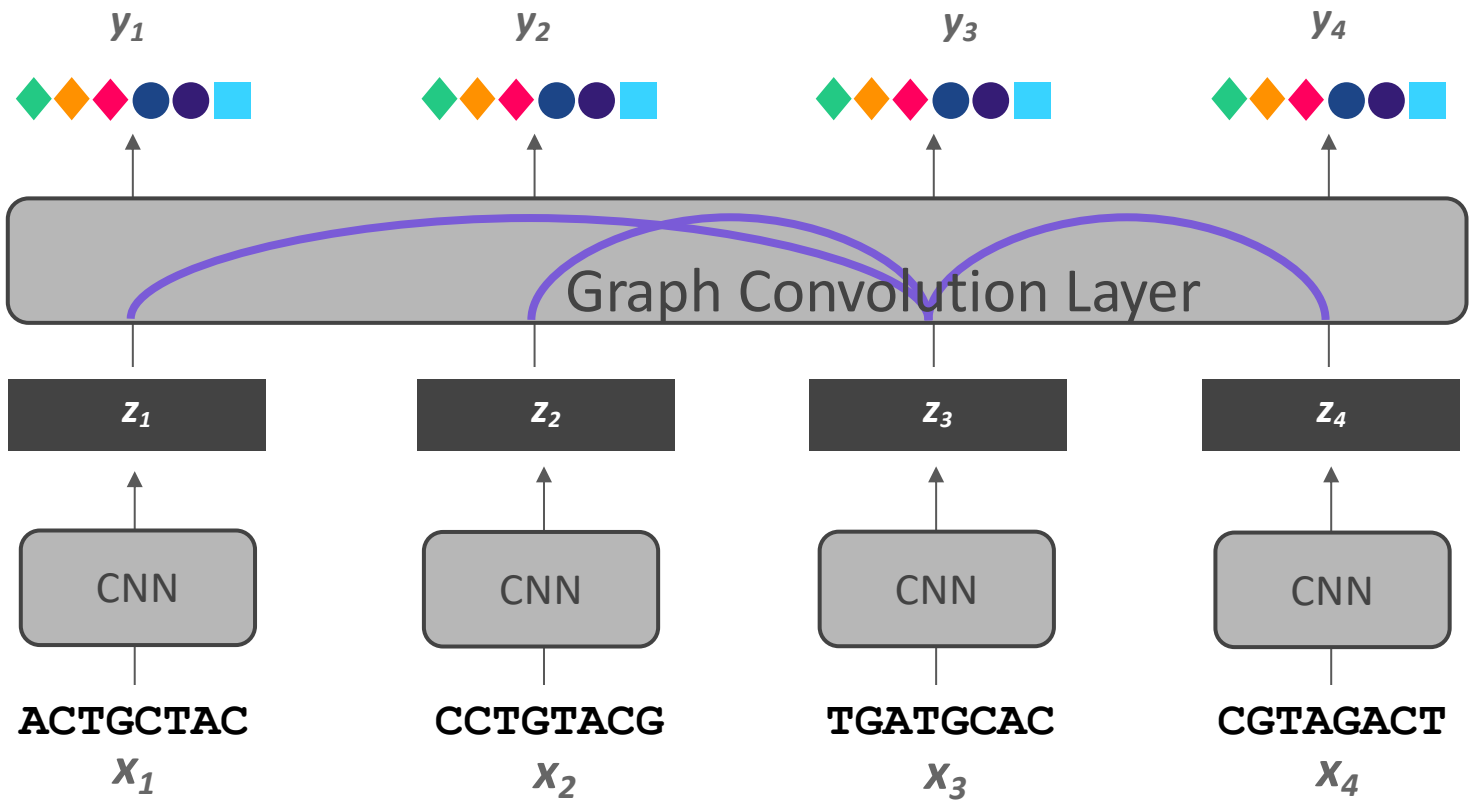
High-throughput Chromosome Conformation Capture (Hi-C)

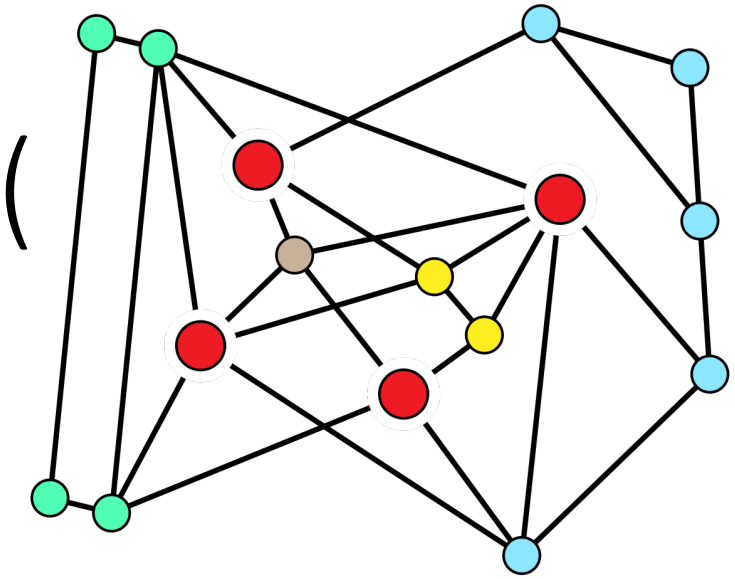


"structural blueprint" indicating interactions that may matter for regulation

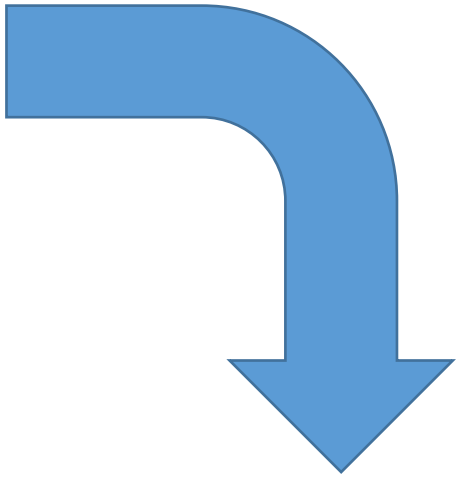
Lieberman et al. - Science 2009

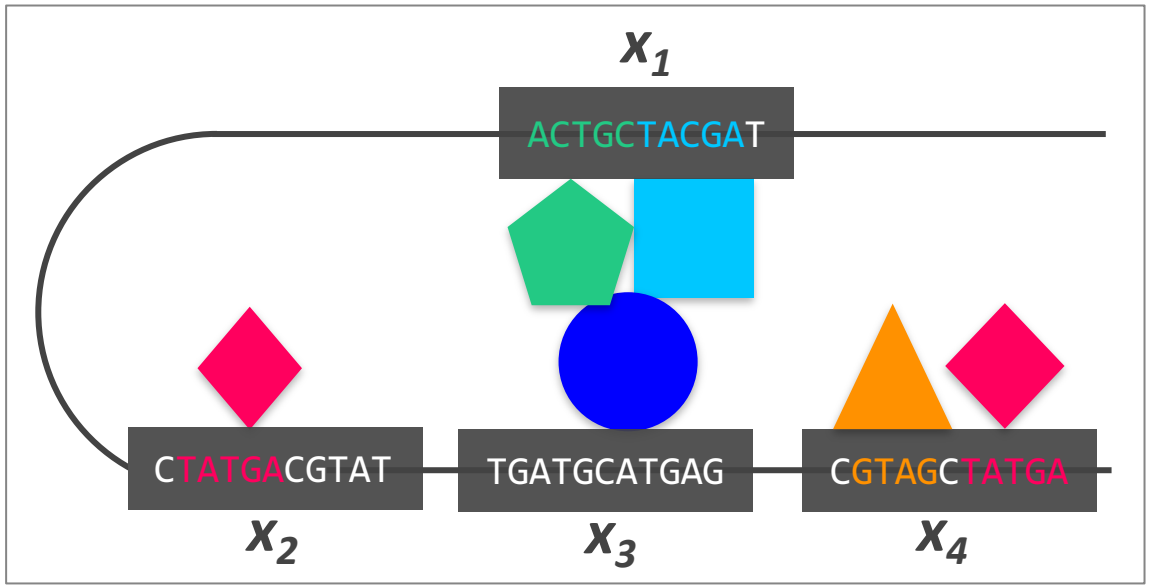
ChromeGCN: Combining Sequence and Graph Learning for Chromatin Profile Prediction



GCN()

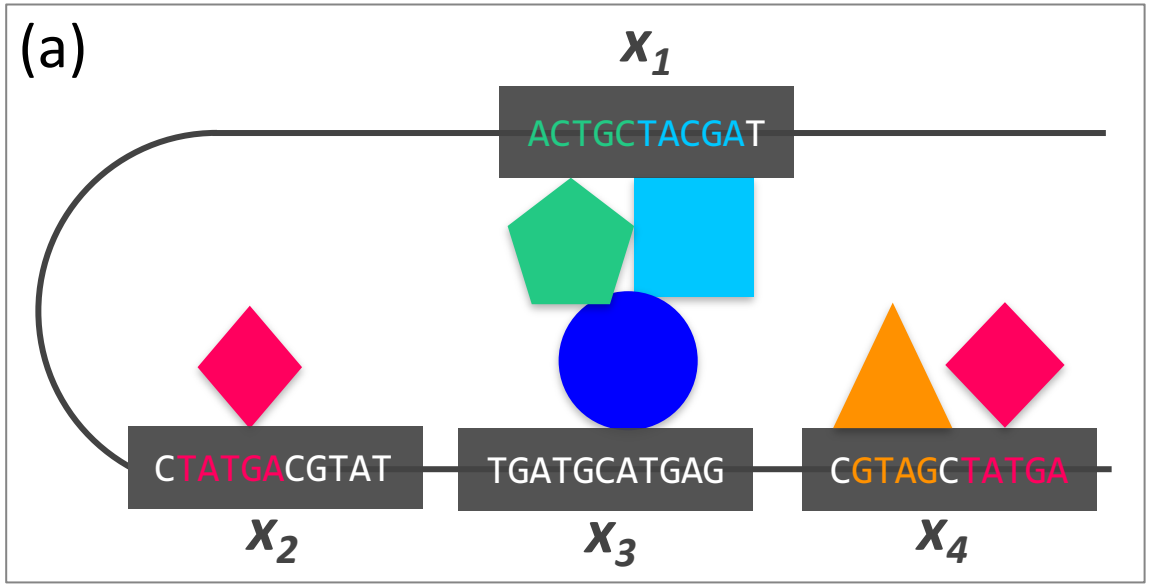
(X,A)

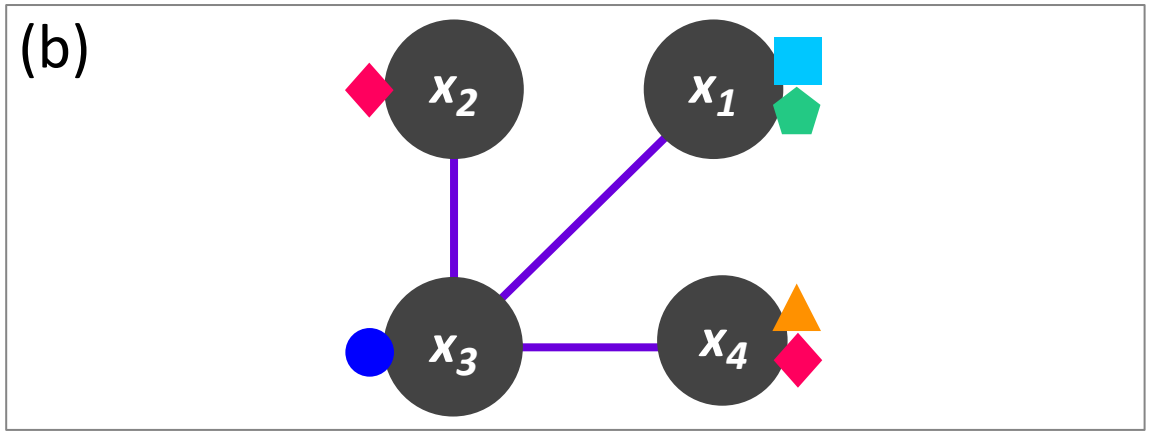


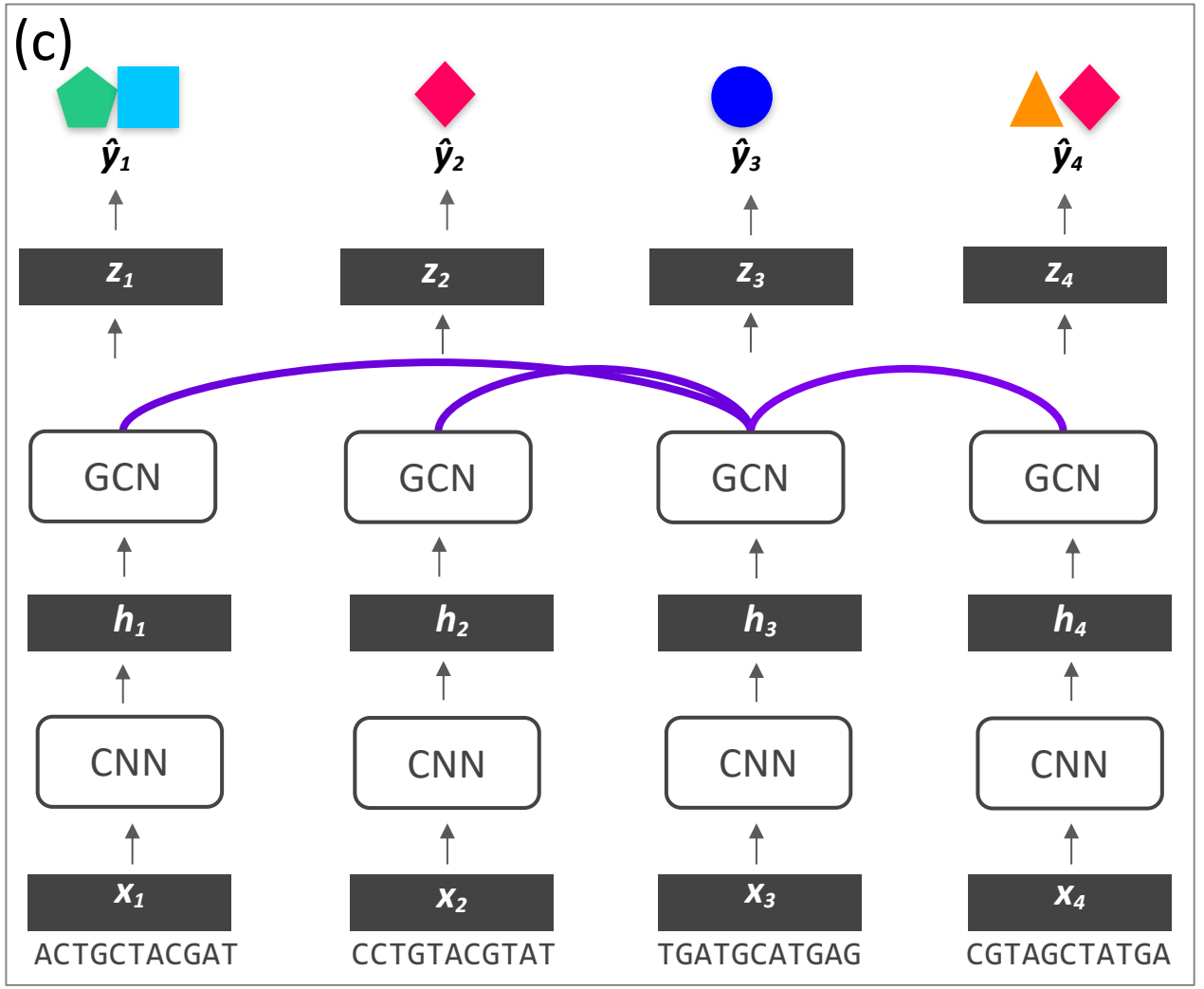

x_1
ACTGCTACGAT
CTATGACGTAT
TGATGCATGAG
CGTAGCTATGA
x_2
x_3
x_4

ChromeGCN: Combining Sequence and Graph Learning for Chromatin Profile Prediction

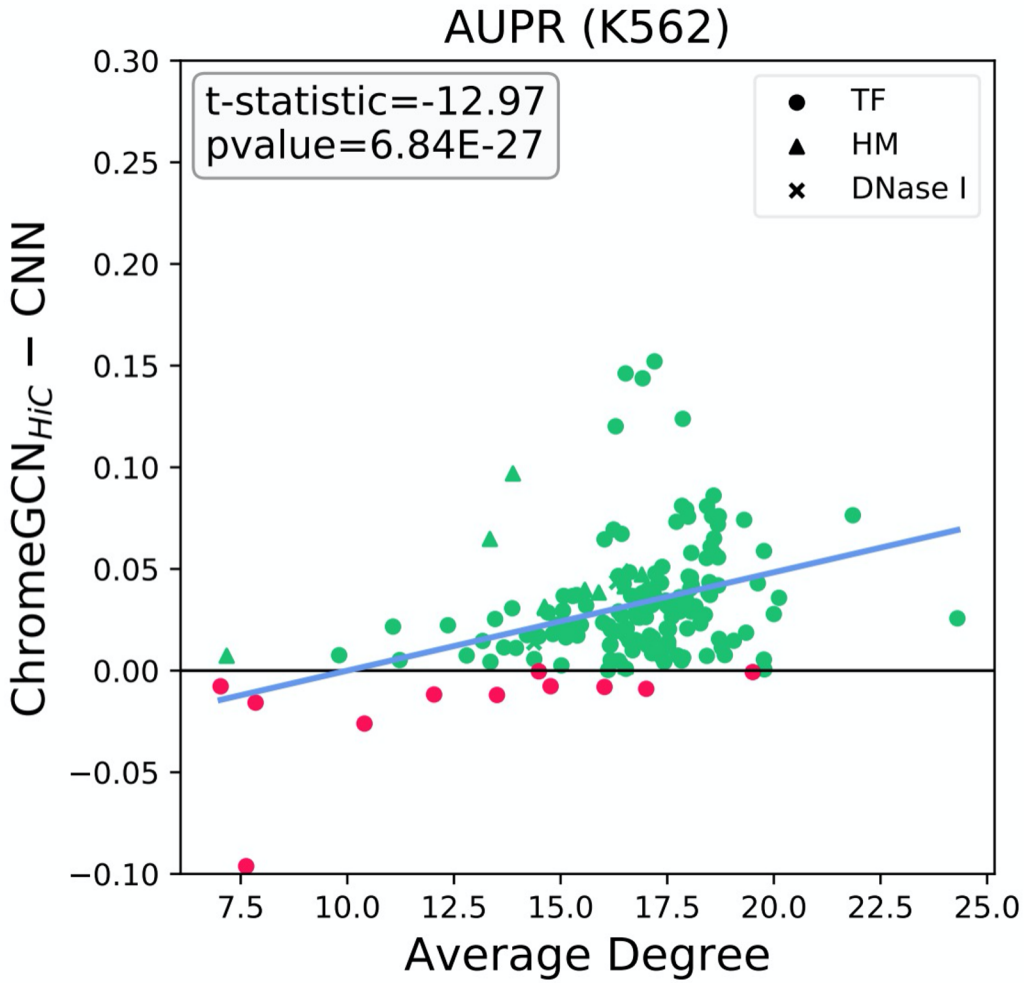
Graph Convolutional Networks for Epigenetic Activity Prediction Using Both Sequence and 3D Genome







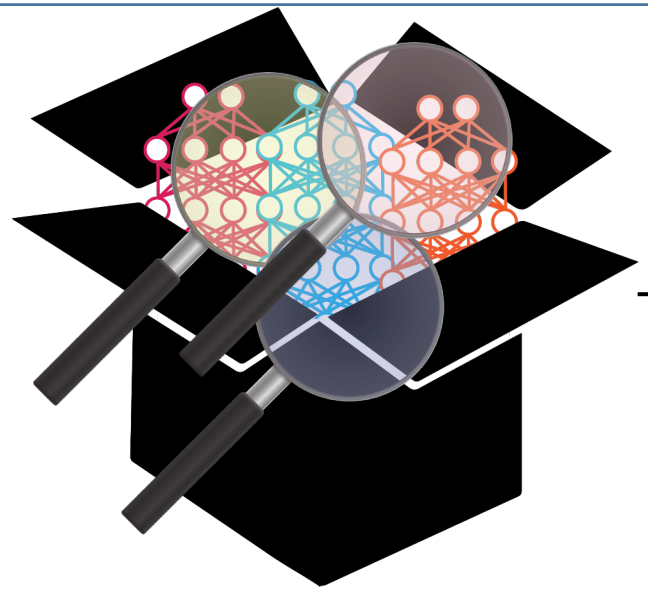
ChromeGCN: Combining Sequence and Graph Learning for Chromatin Profile Prediction

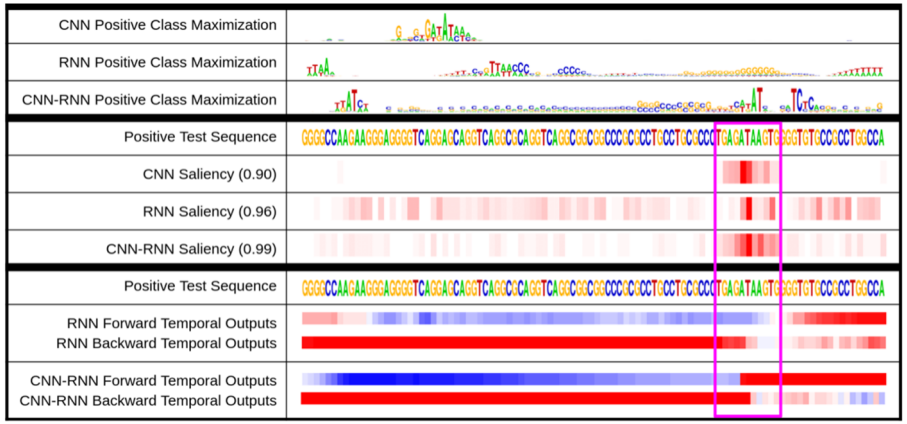


ChromeGCN - **Lanchantin** and Qi. ECCB 2020, Bioinformatics 2020

Deep Motif Dashboard: Understand DNNs by Post Analysis

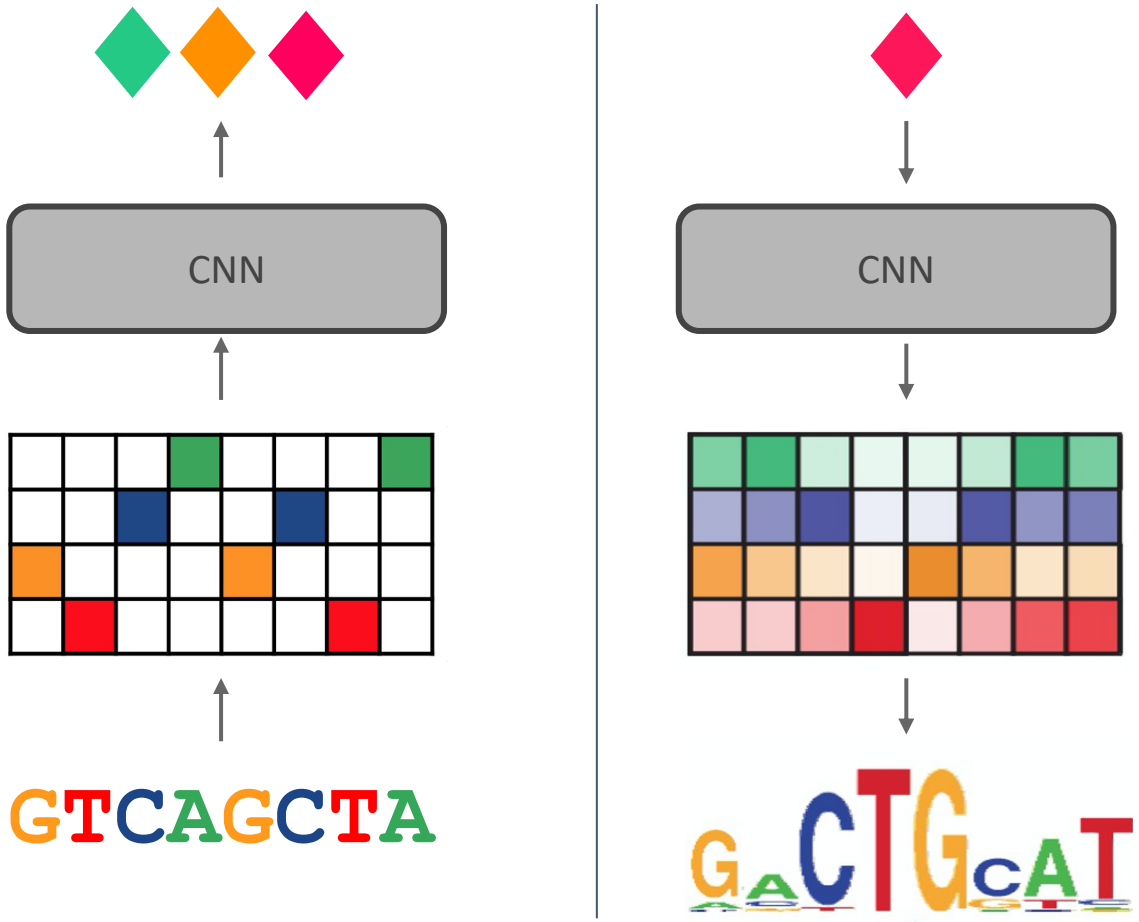
Lanchantin, Singh, Wang & Qi - Pacific Symposium on Biocomputing, 2017





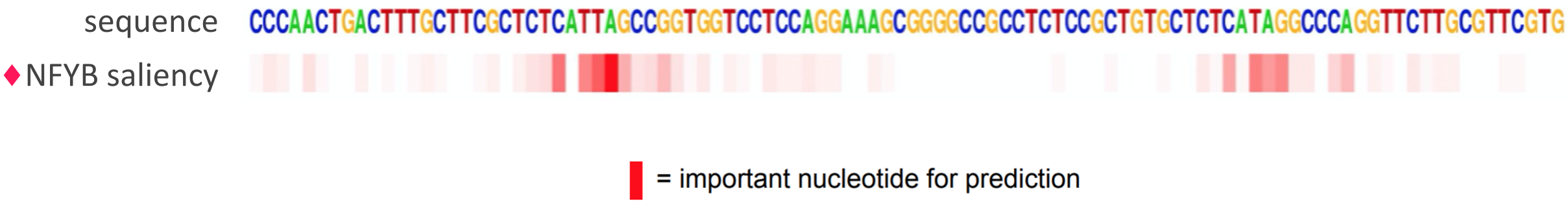
1. Saliency Maps - recommending on CNN kind
2. Class Optimization - recommending on CNN kind
3. Temporal Output Values - recommending on RNN kind

Interpreting Sequence Syntax with Class Optimization

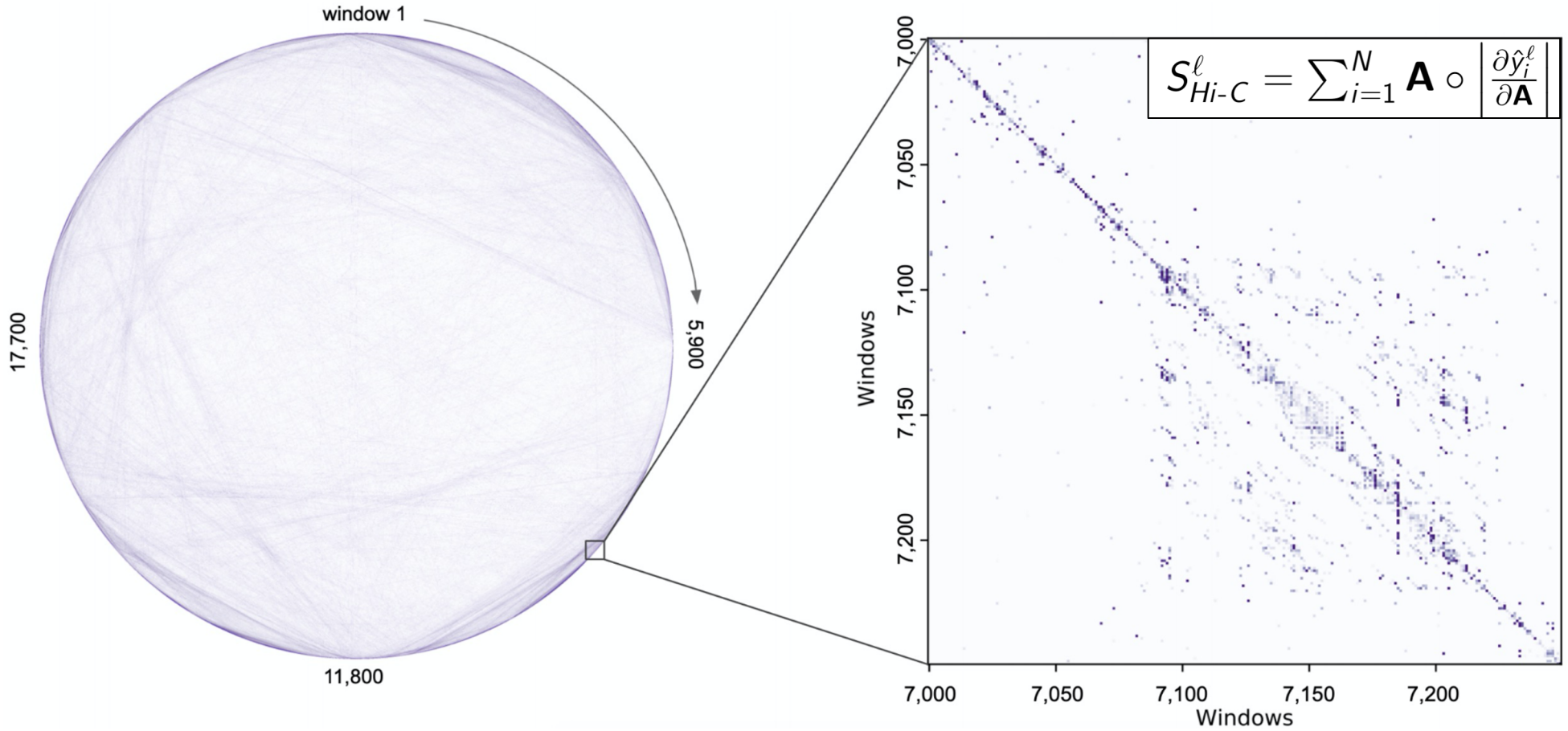


Deep Motif Dashboard - **Lanchantin**, Singh, Wang, Qi. ICLR Workshop Track 2016, PSB 2017

Interpreting Sequence Syntax with Saliency Maps



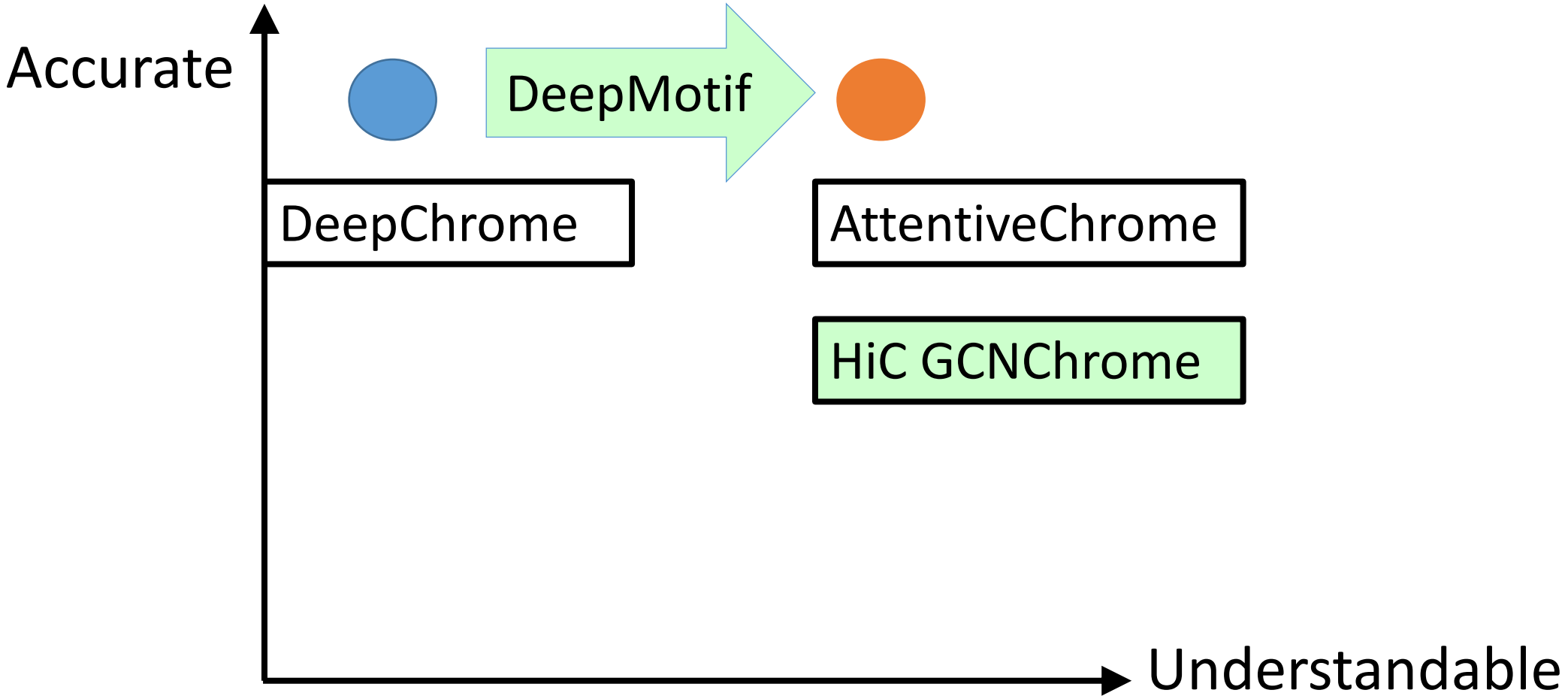
Interpreting Long Range Interactions with Hi-C Saliency Maps



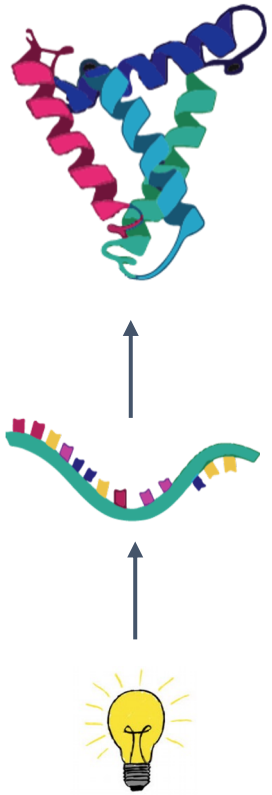
YY1 Hi-C saliency map for all 500k edges in **A**

https://www.deepchrome.net

Summary of tools



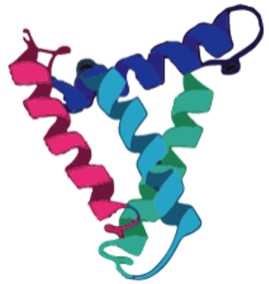
Third Task:



gene expressed

ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

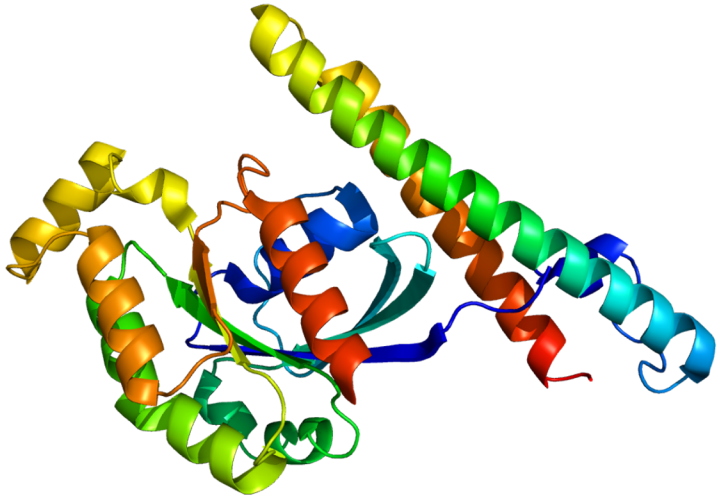
Third Task:



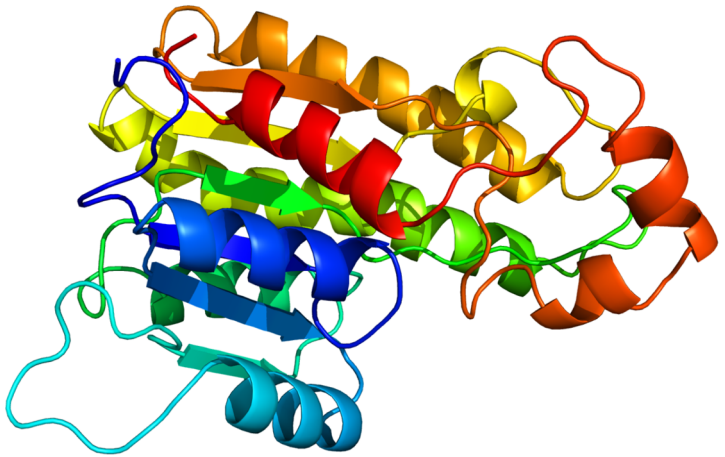
gene expressed

ATGCTCGATGCTAATACGACTGAGATTACTGAGACTGAGACTCTAGAT

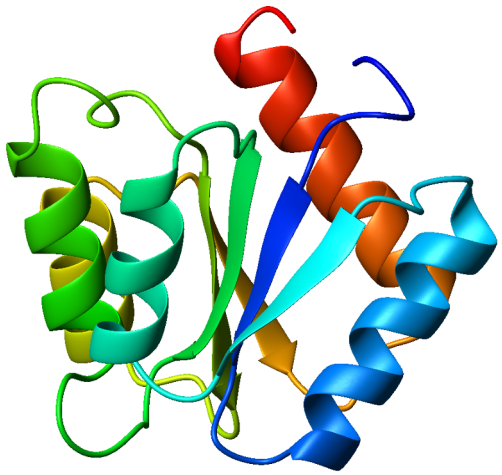
Proteins: the building blocks of life



oxygen transportation



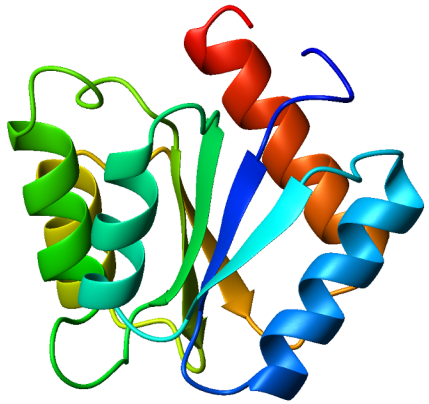
antibodies

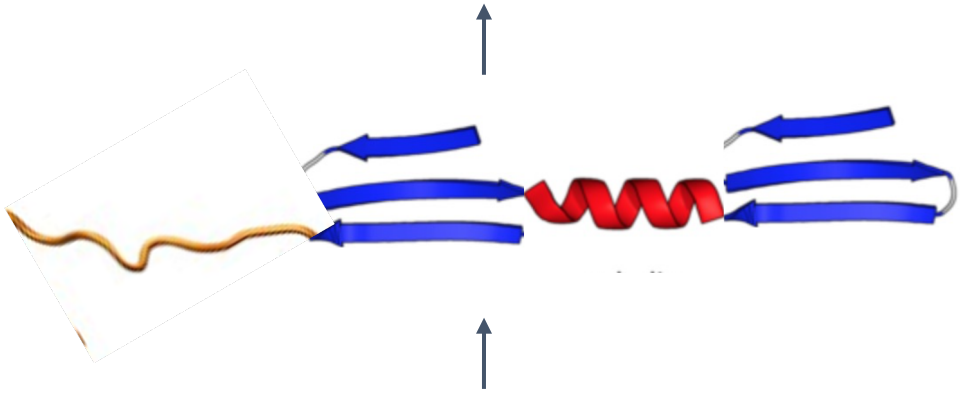


digestive enzymes

Protein Structures

Tertiary Structure



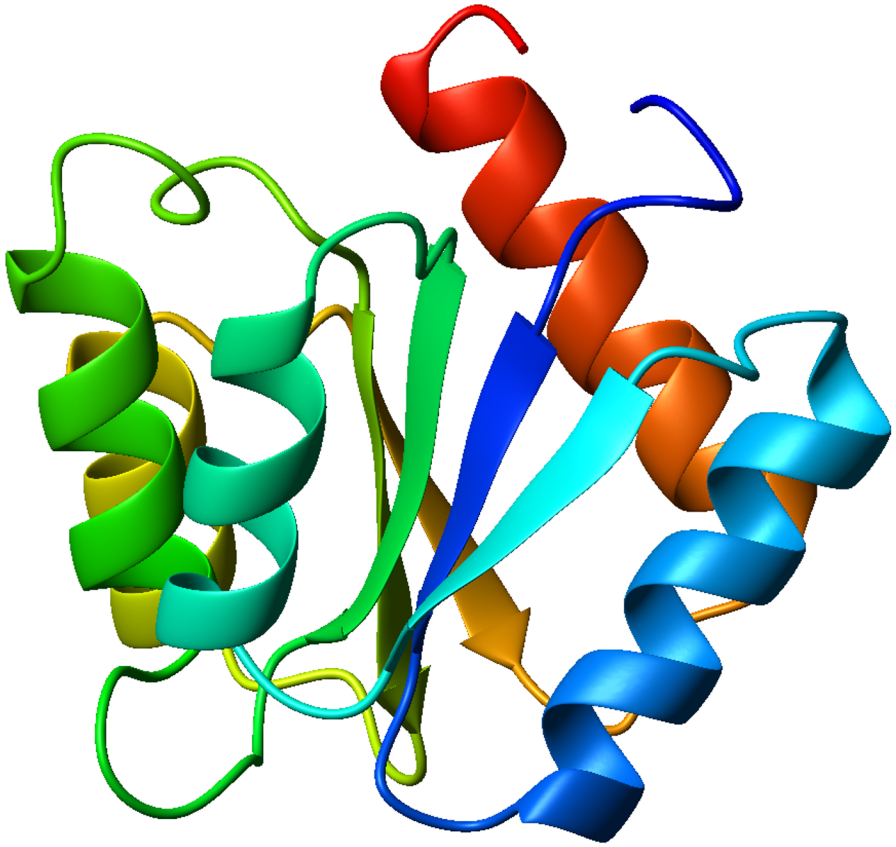


Secondary Structure

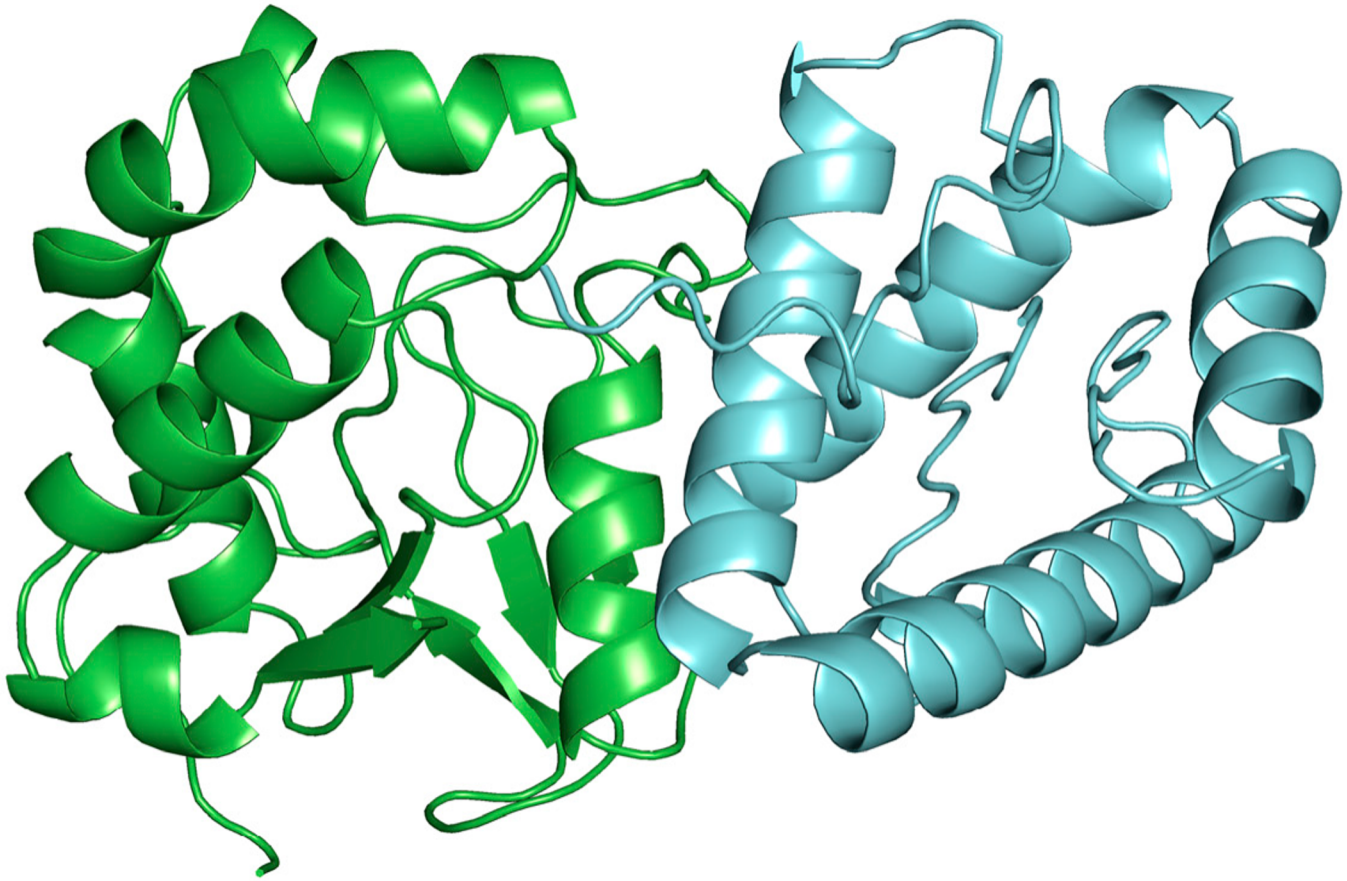
Primary Sequence

MHFTEDKATILWGKVNVEGETLGRVYPWQ

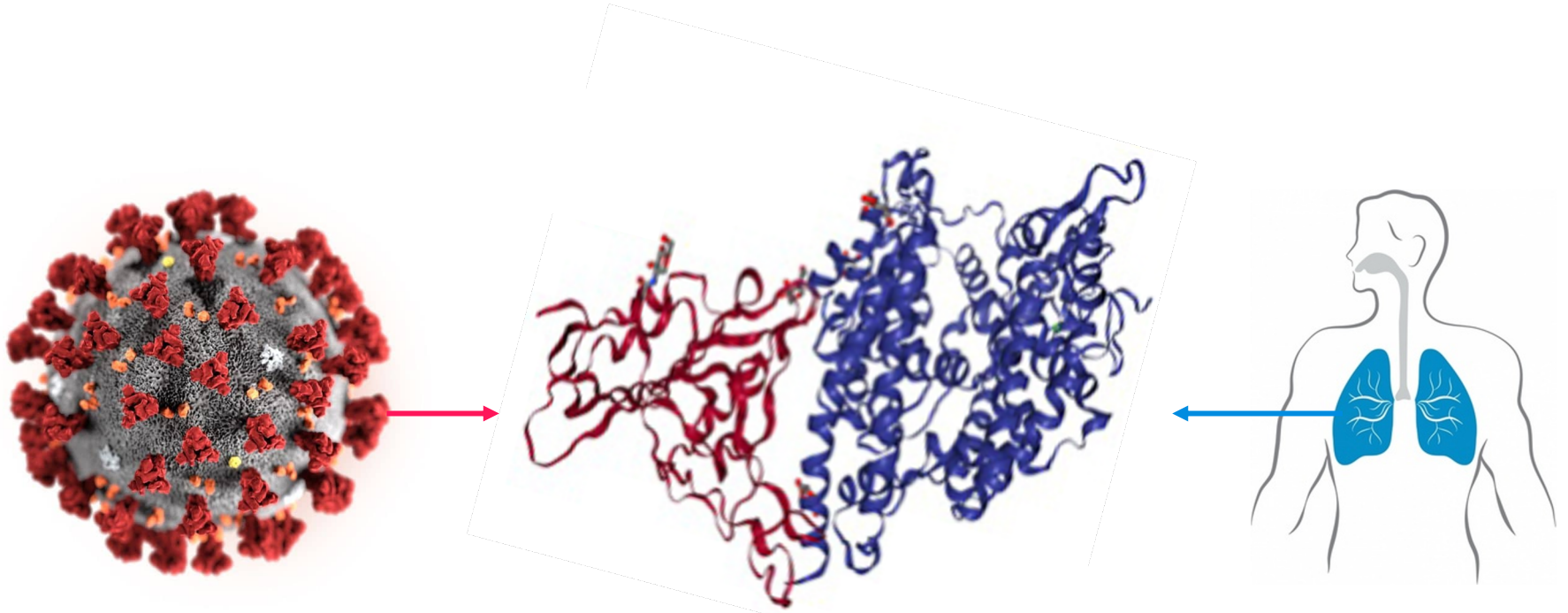
Structure Determines Function



One primary function: Protein-Protein Interactions (PPIs)



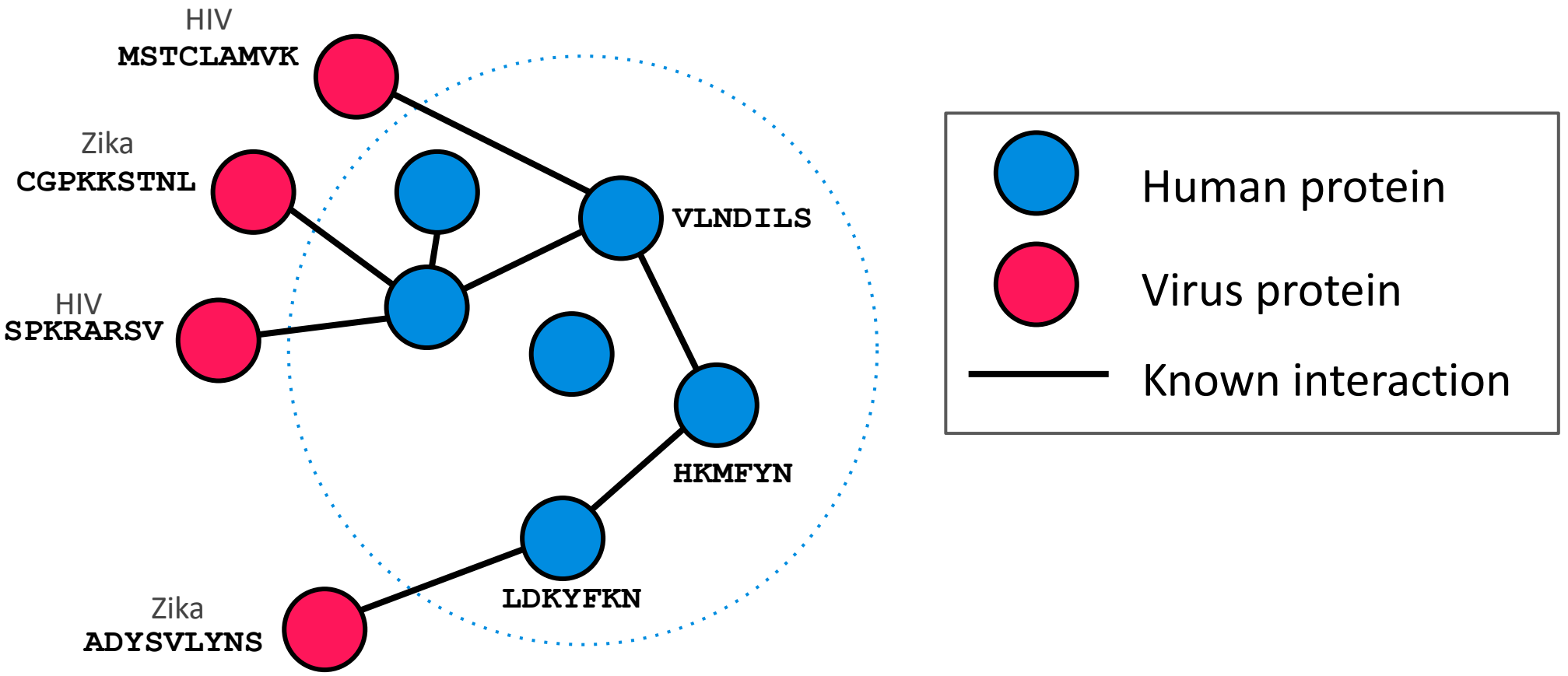
Our Task: To Discover Human-Virus Protein-Protein Interactions



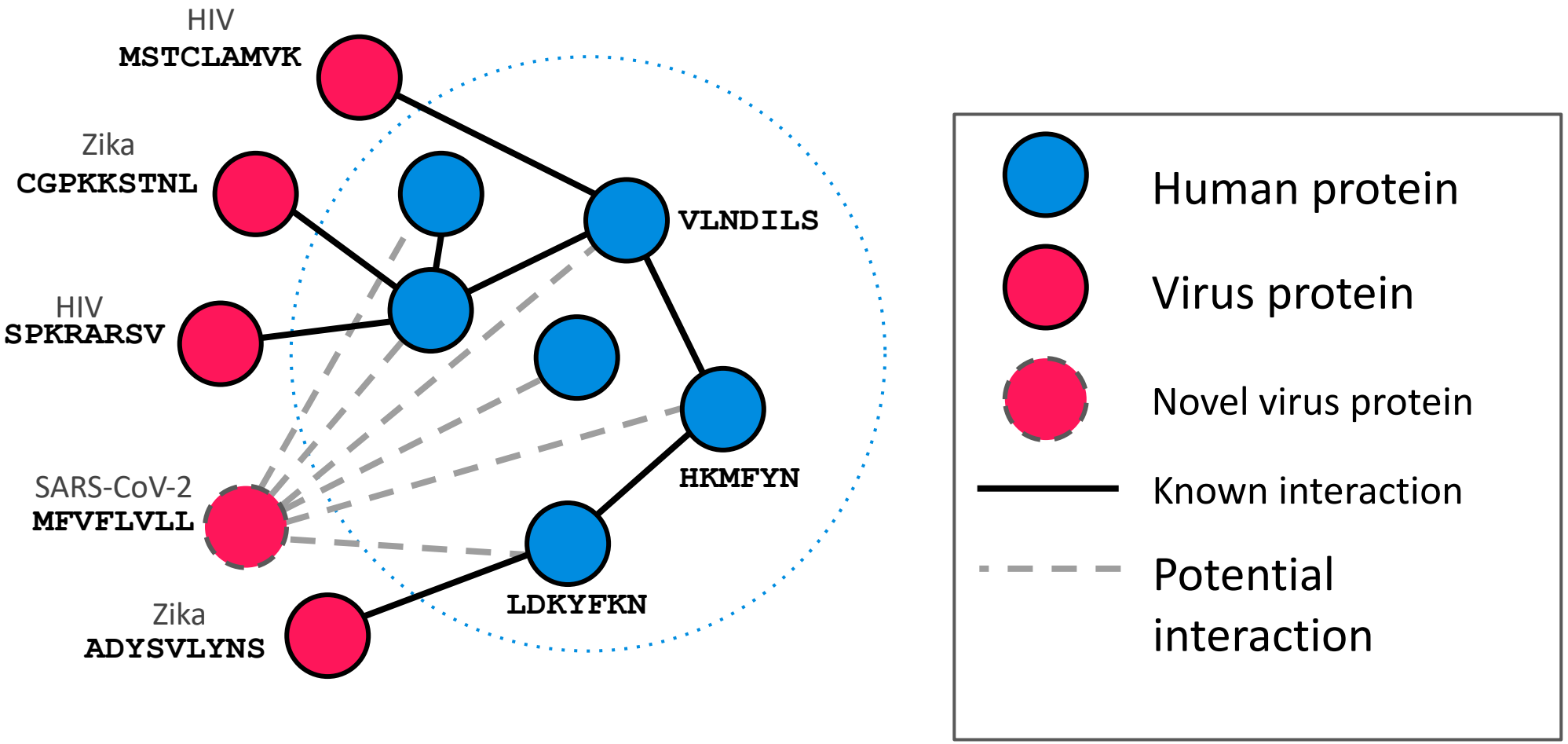
SARS-CoV-2 Spike Protein

Human ACE2 Protein

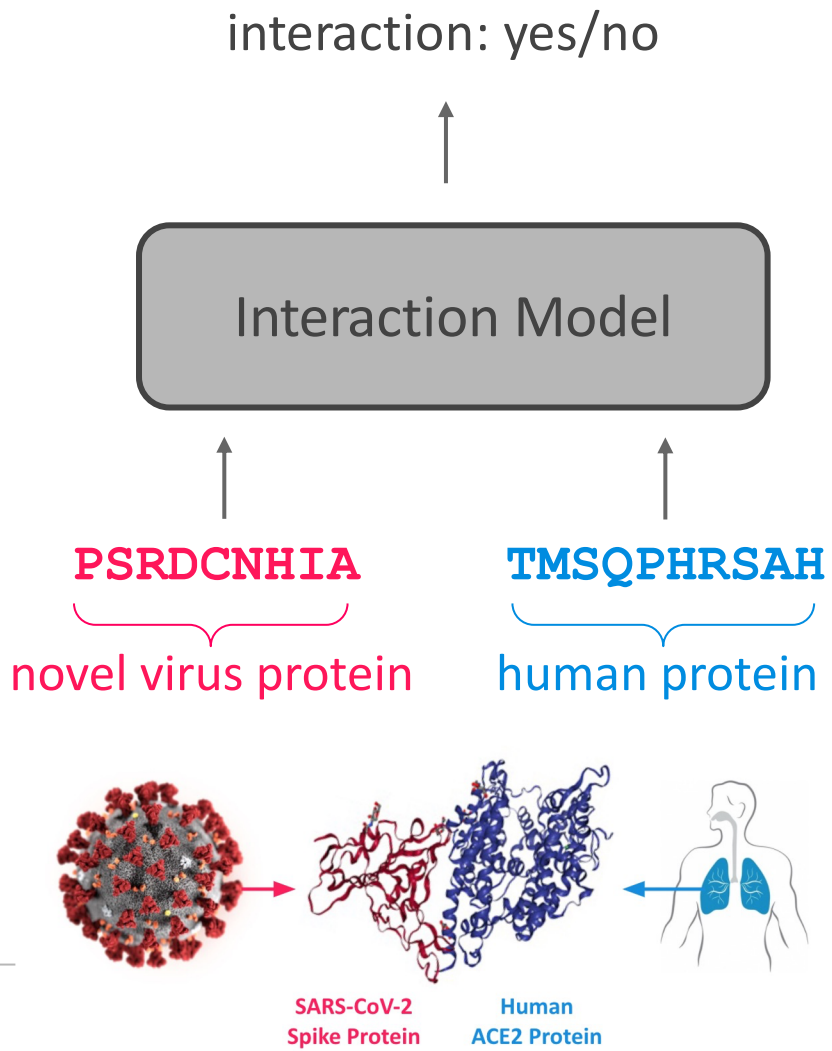
Human-Virus Protein-Protein Interactions



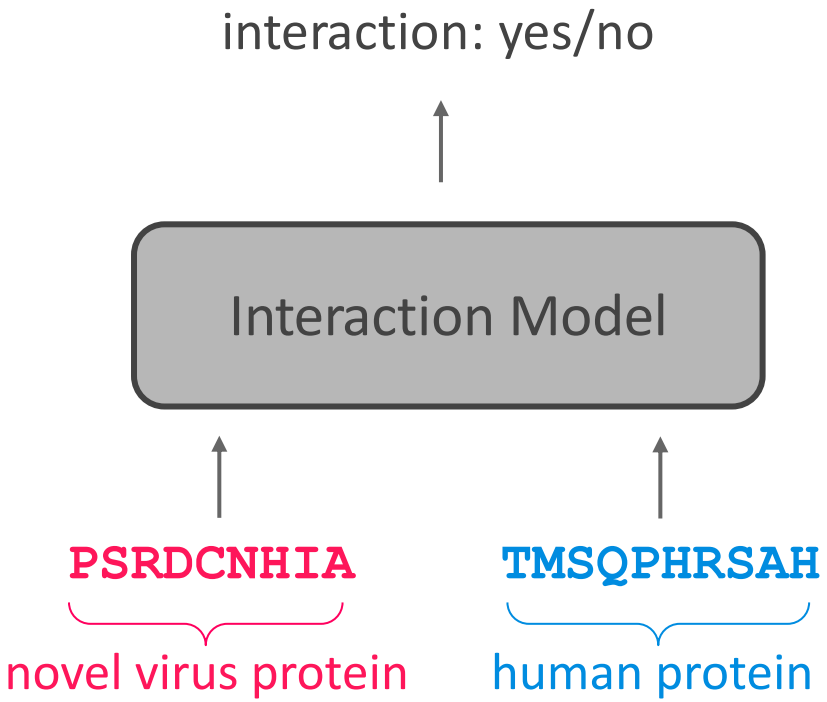
Human-Virus Protein-Protein Interactions



Novel Virus-Human Protein Interaction Prediction from Sequence

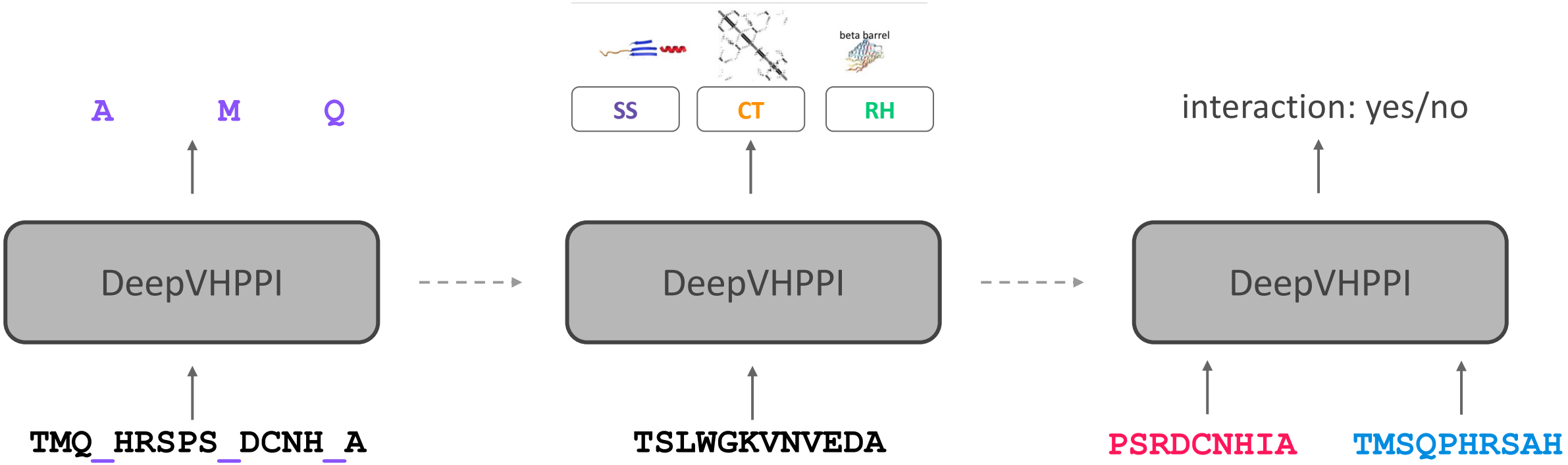


Novel Virus-Human Protein Interaction Prediction from Sequence

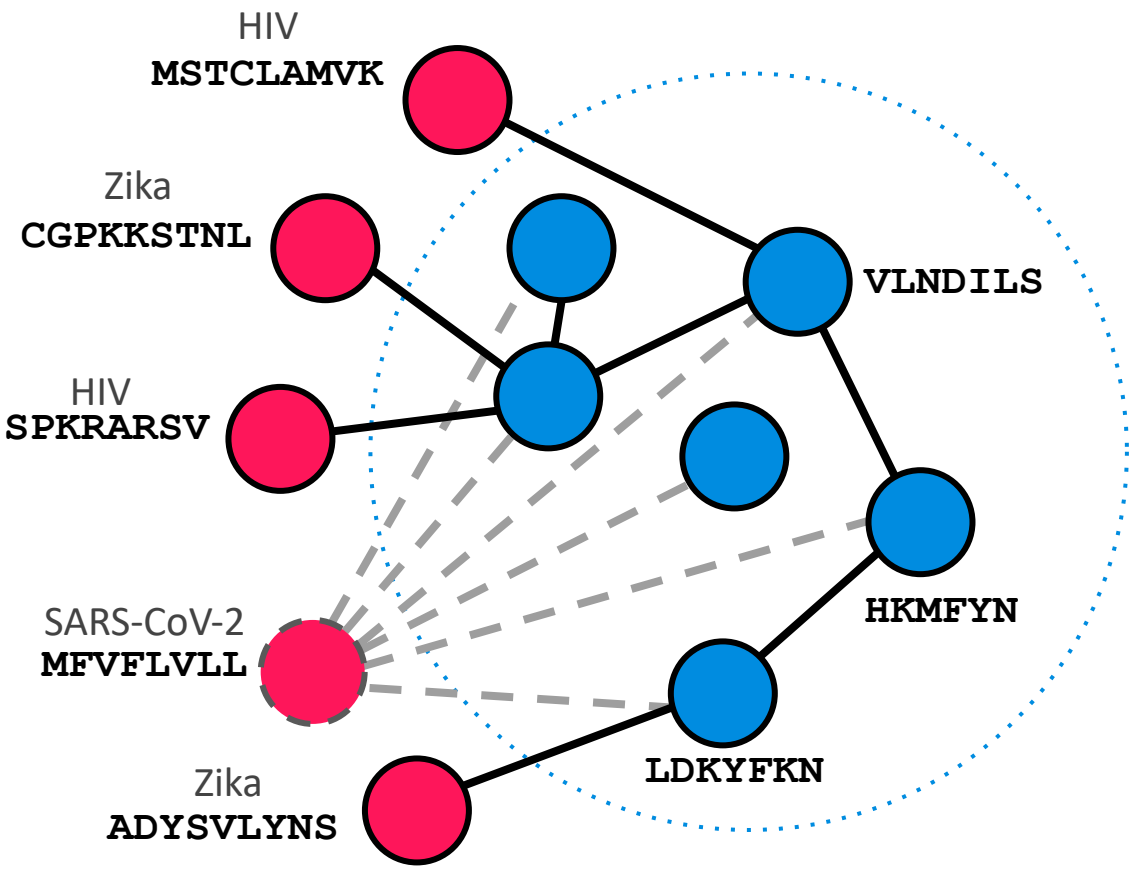


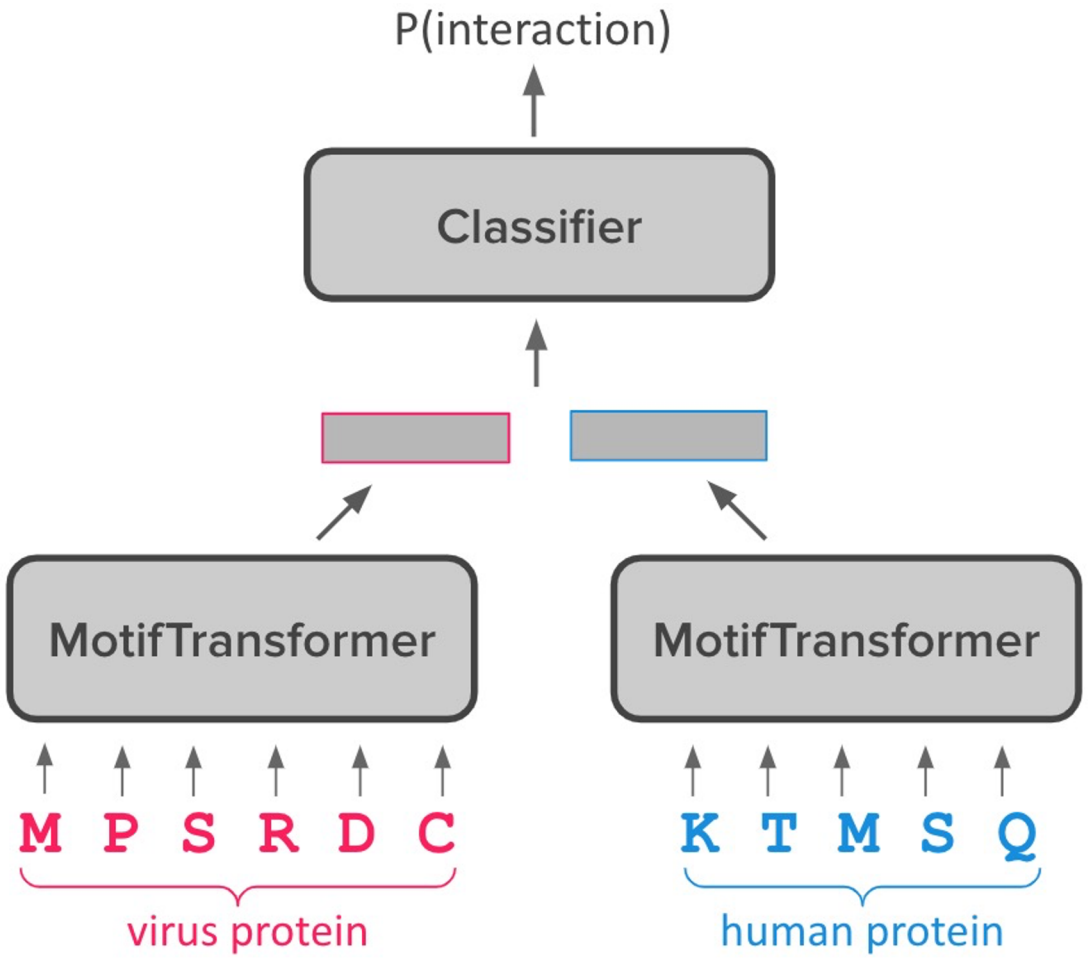
1. Limited interaction data available
2. Interactions are largely determined by structure

Transfer Learning for Sequence-Based Interaction Prediction

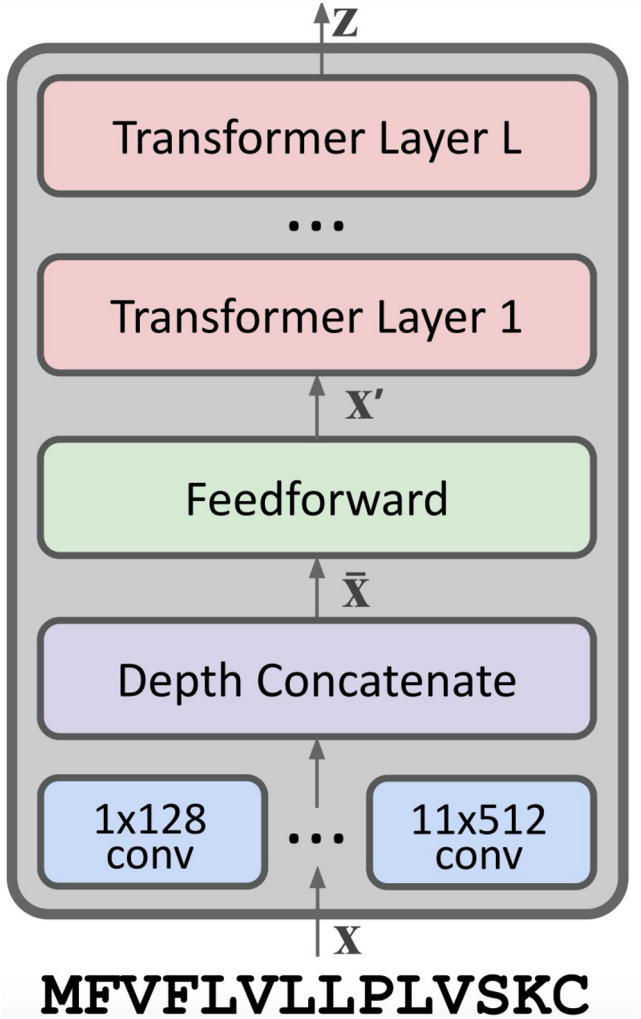


Interaction Prediction

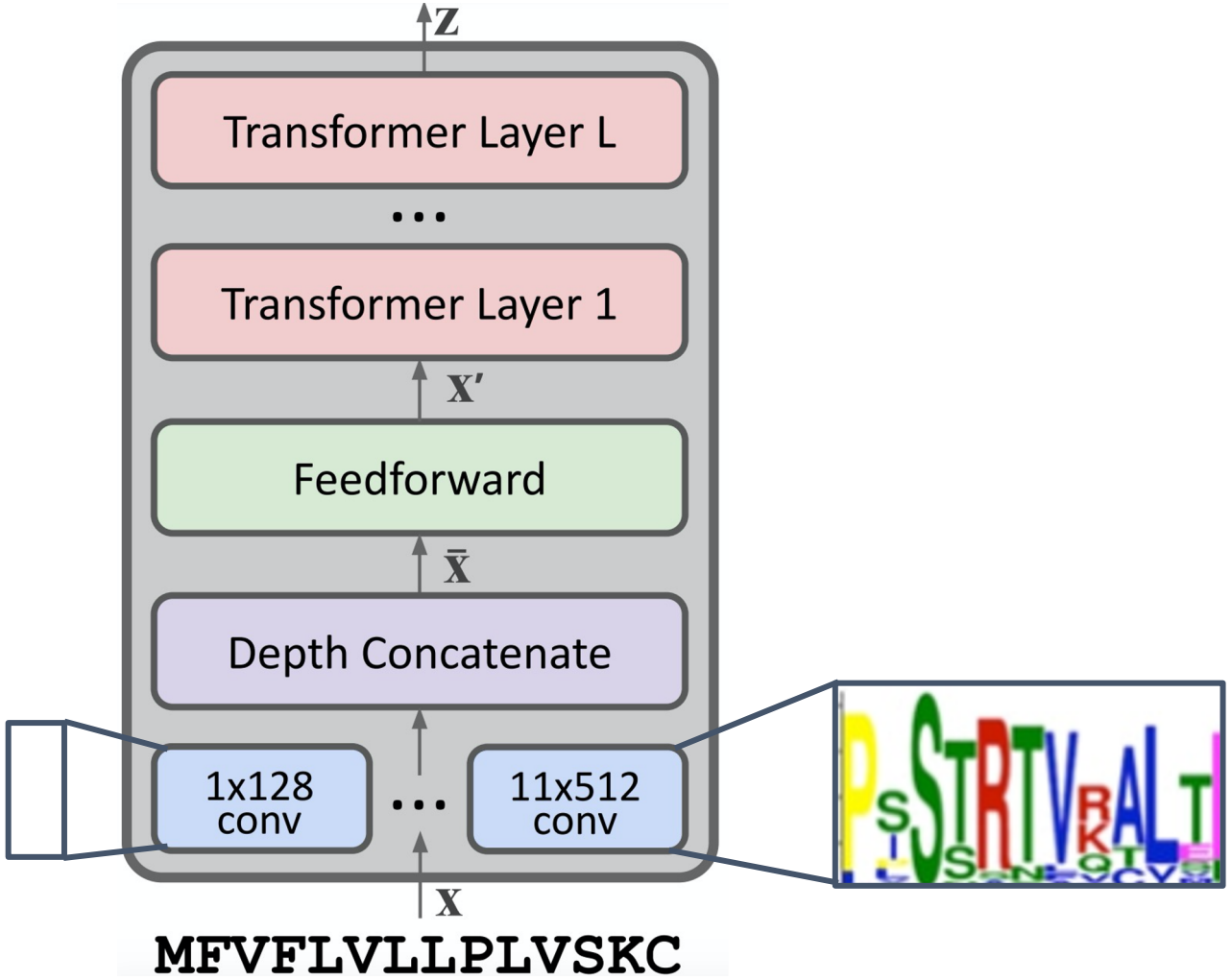




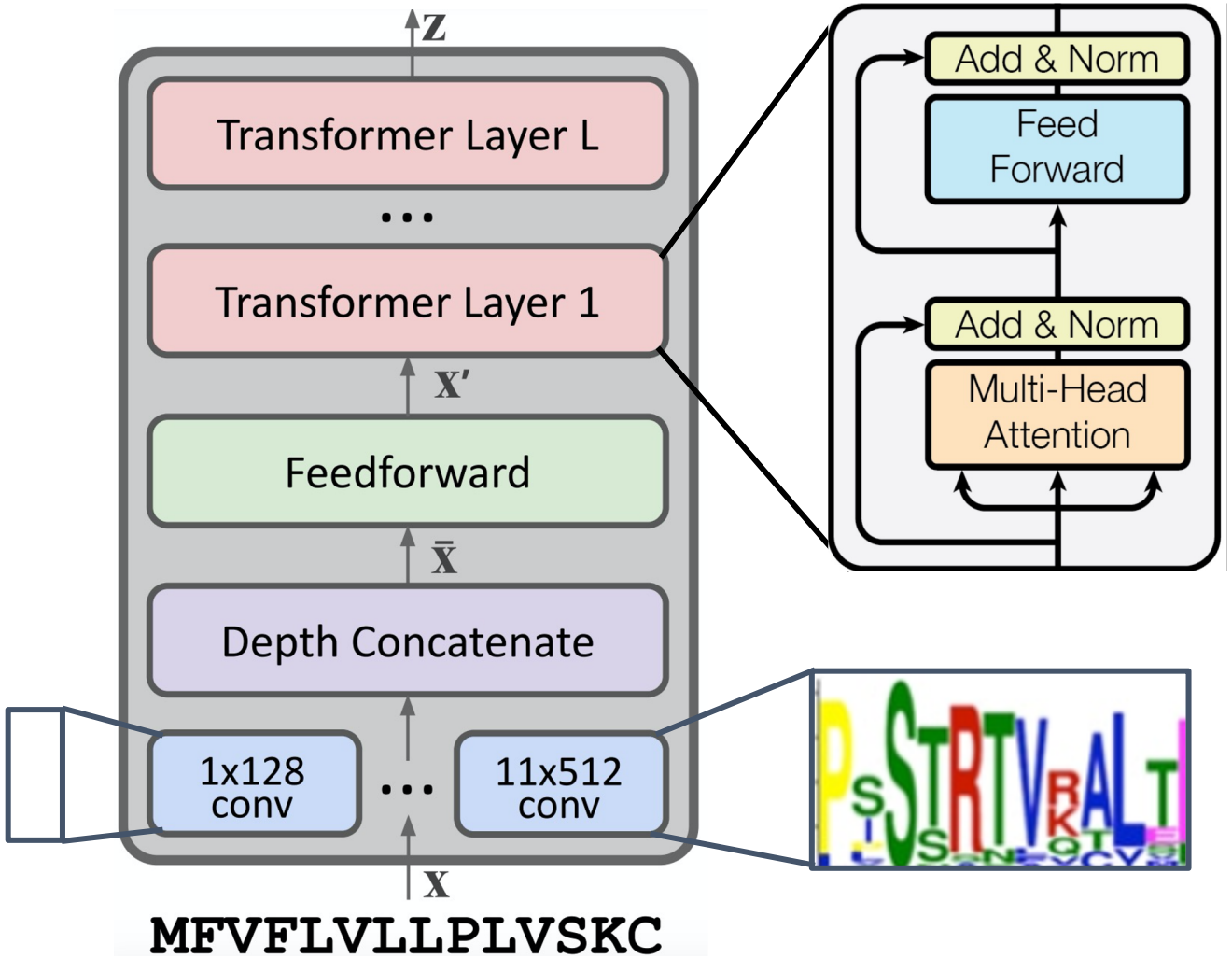
Motif Transformer



Motif Transformer



Motif Transformer



Experimental Setup

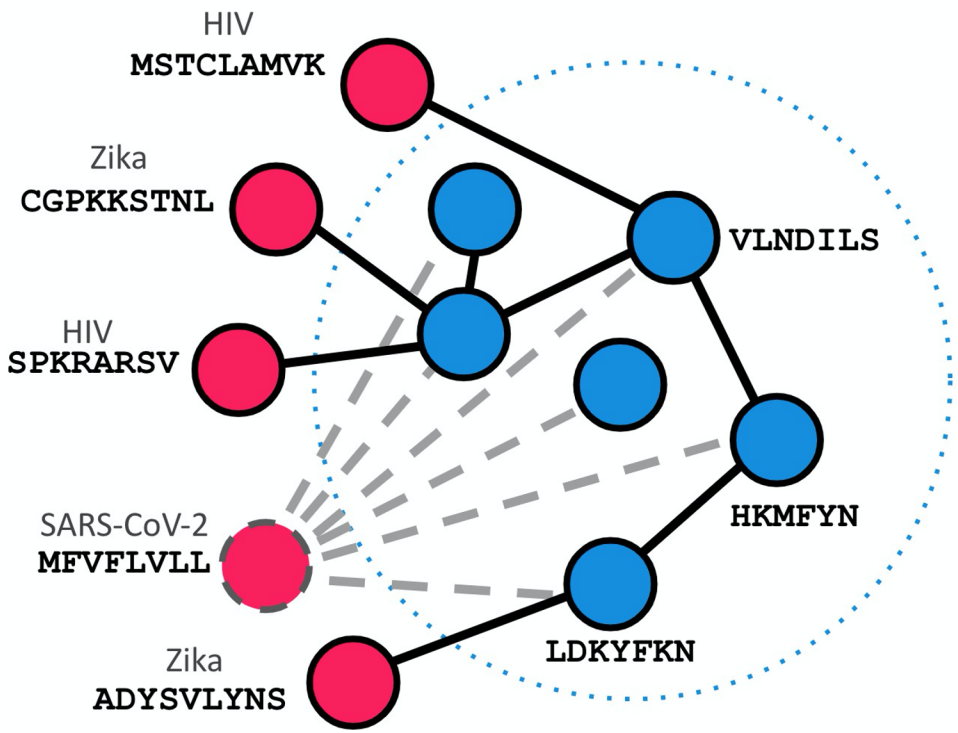
- **Training Data: HPIDB 3.0 Dataset**
  - 22,000 positive interactions, 226,000 negative interactions
  - 1,100k virus proteins, 20,000 host (human) proteins
- **Testing Data:**
  - HIV, Ebola interactions from Zhou et al.
  - Our own curated SARS-CoV-2 interactions collected from BioGrid

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

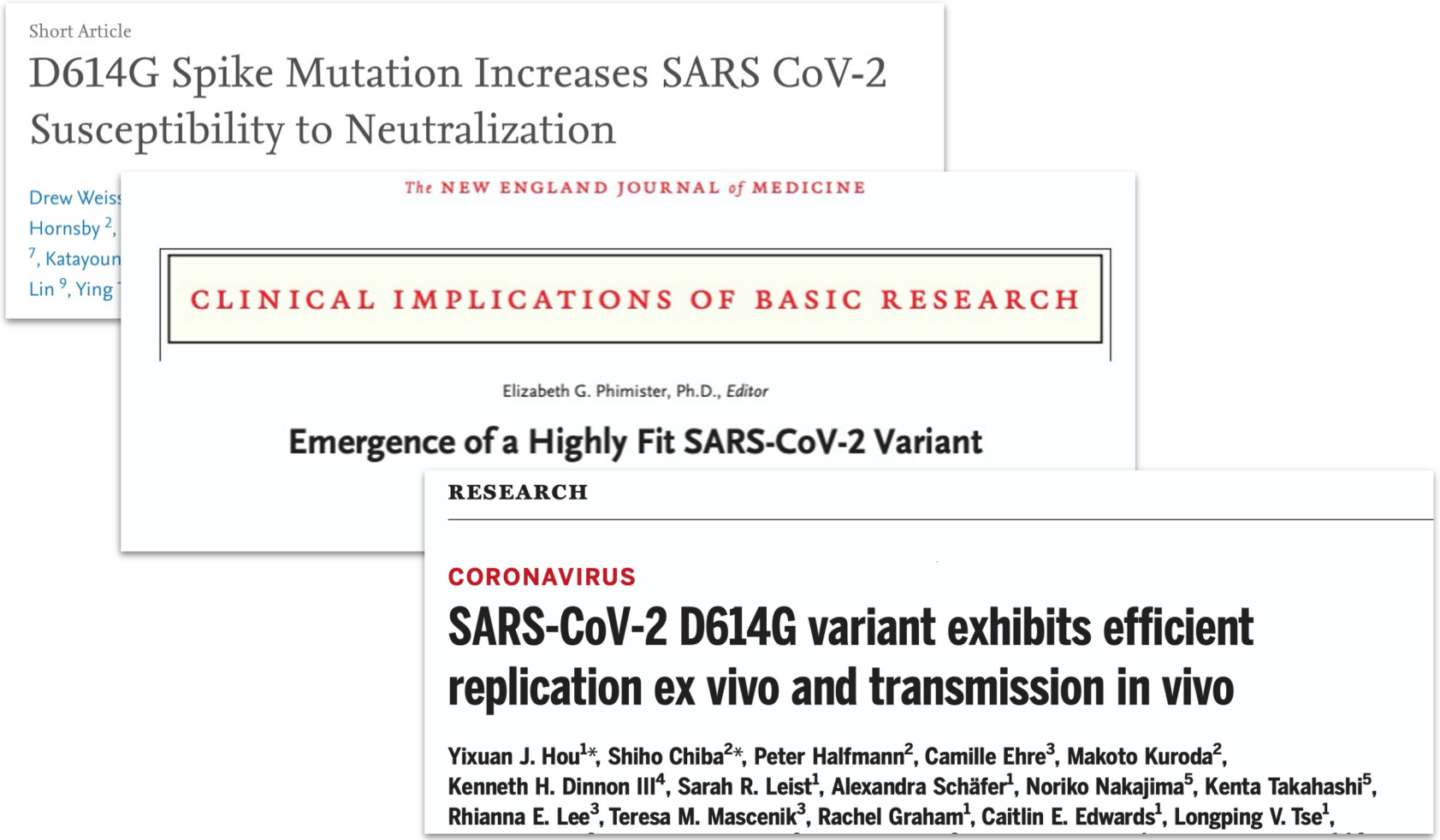
Use Cases of Sequence Based Interaction Predictors

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

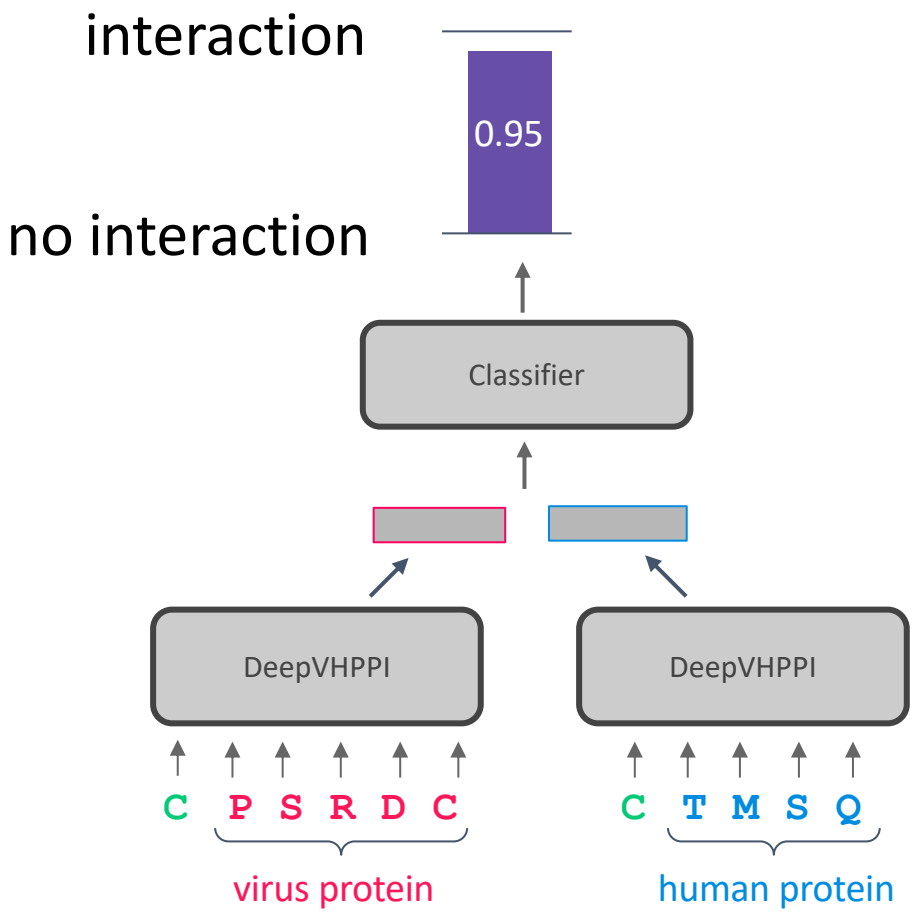
RESEARCH

CORONAVIRUS

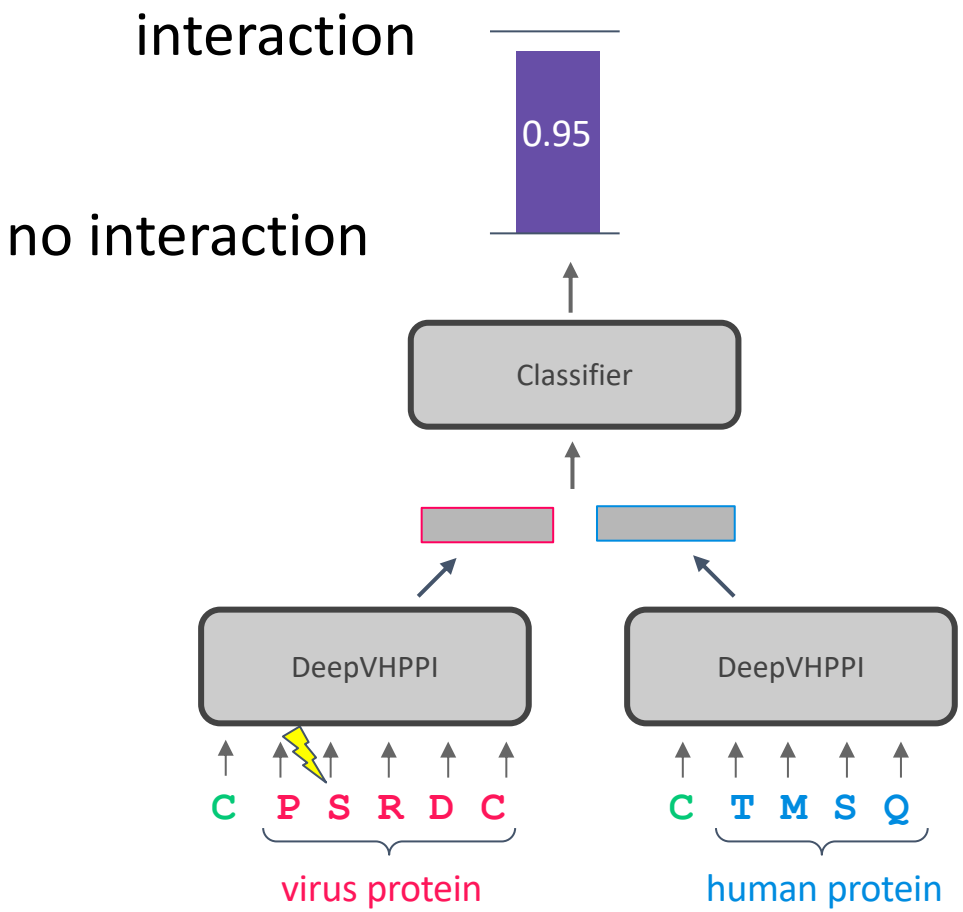
SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo

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],

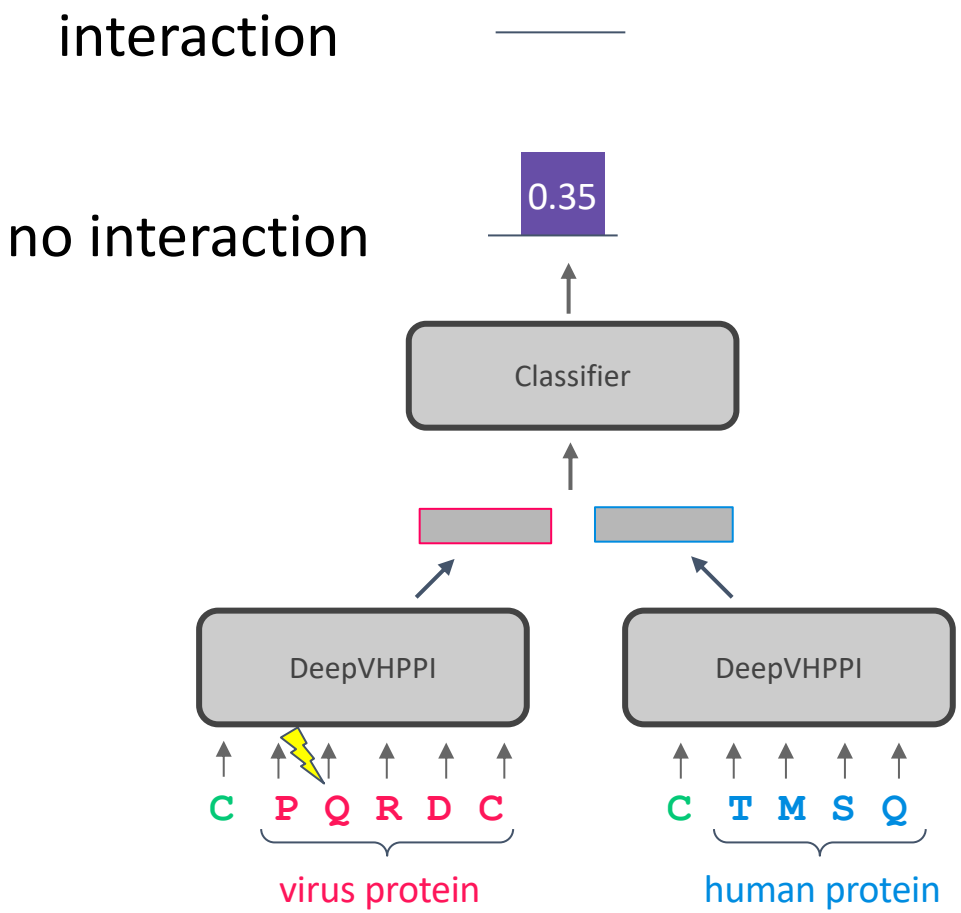
Perturbation Analysis



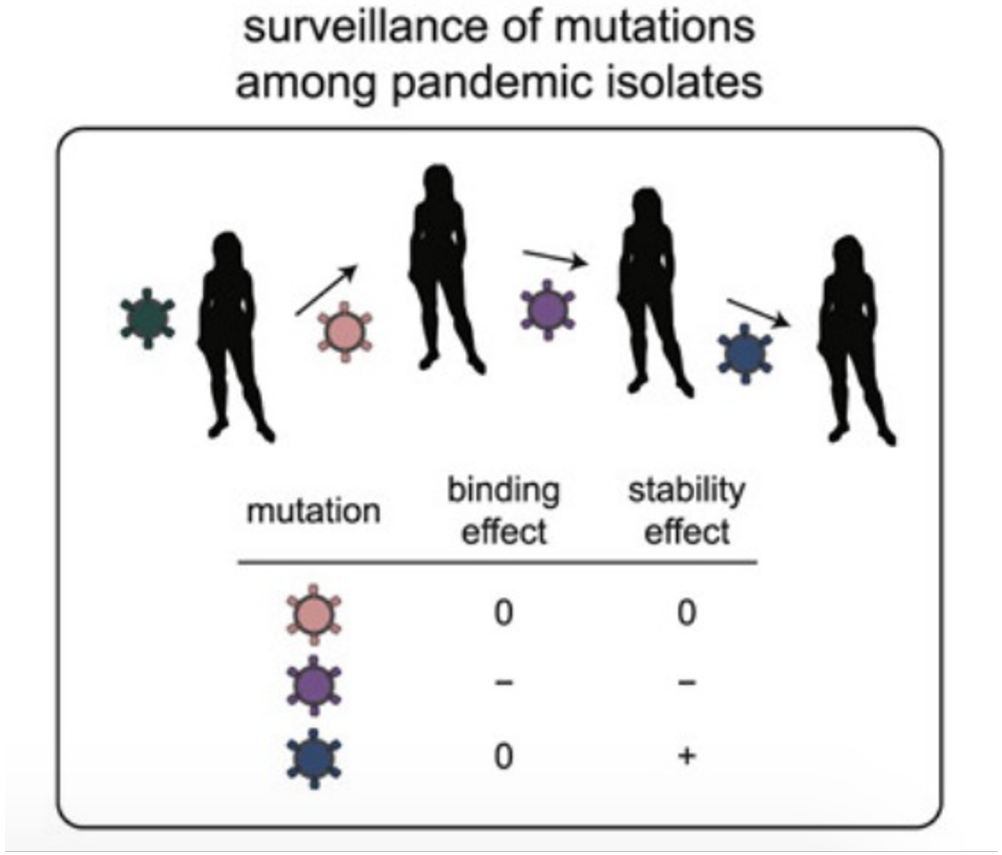
Perturbation Analysis

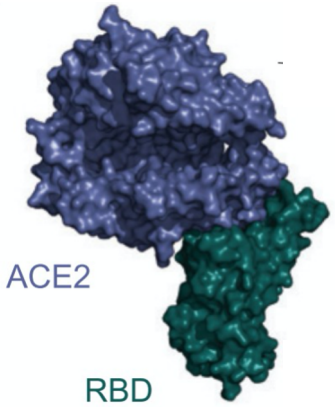


Perturbation Analysis



Perturbation Analysis: Investigating Mutations



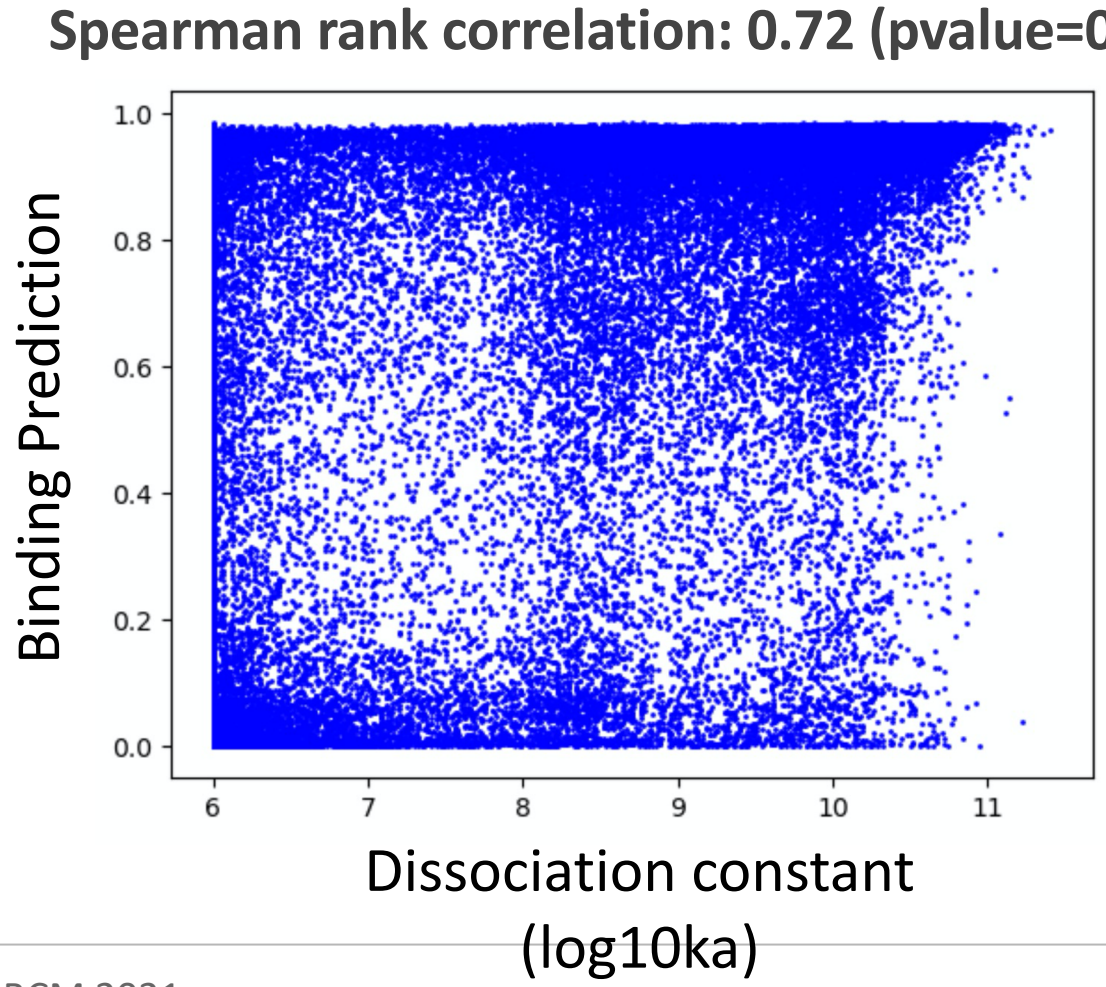


Starr et al. Cell 2020

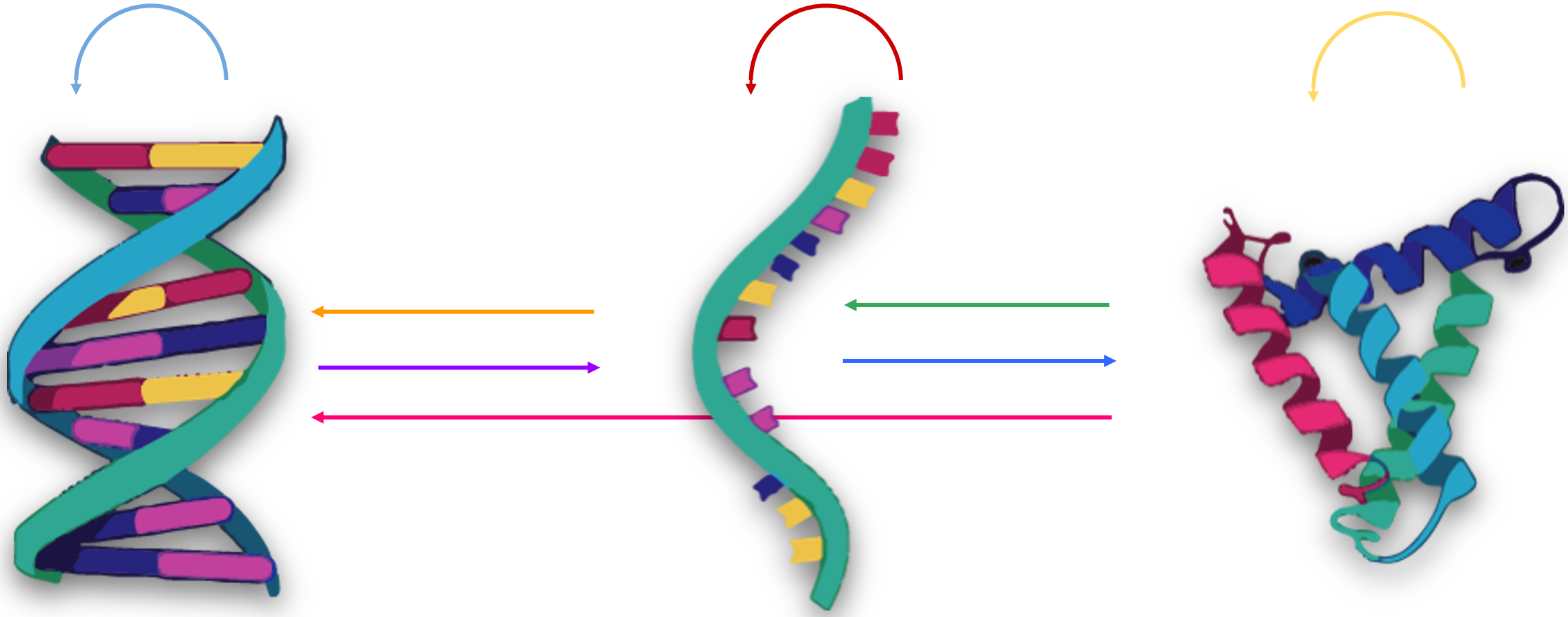
Perturbation Analysis: Mutated Spike and ACE2 Interactions

Training: 1,000 mutated seqs

Testing: 94,000 mutated seqs



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



DNA ***RNA*** ***PROTEIN***

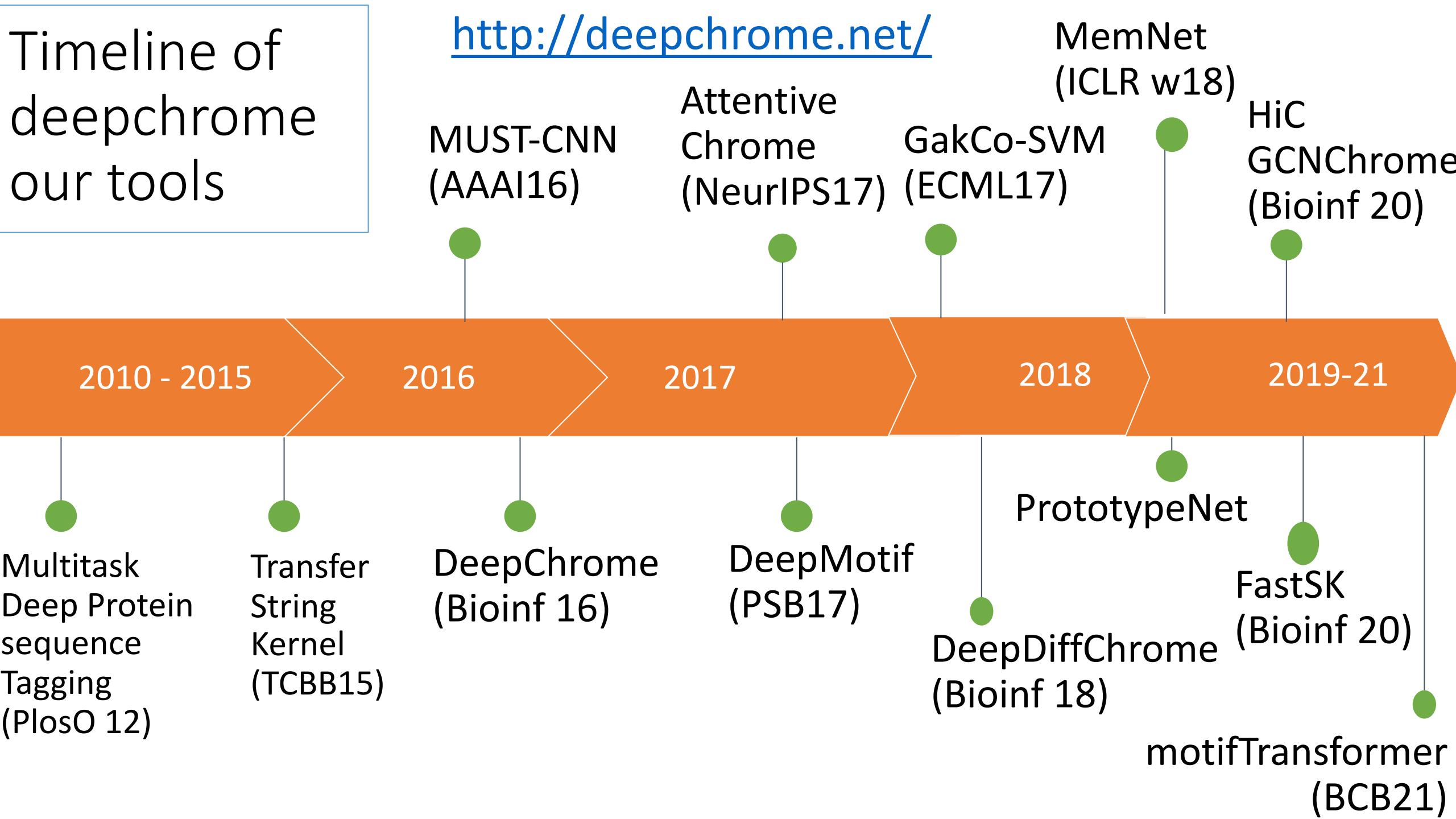
Journey Ahead

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

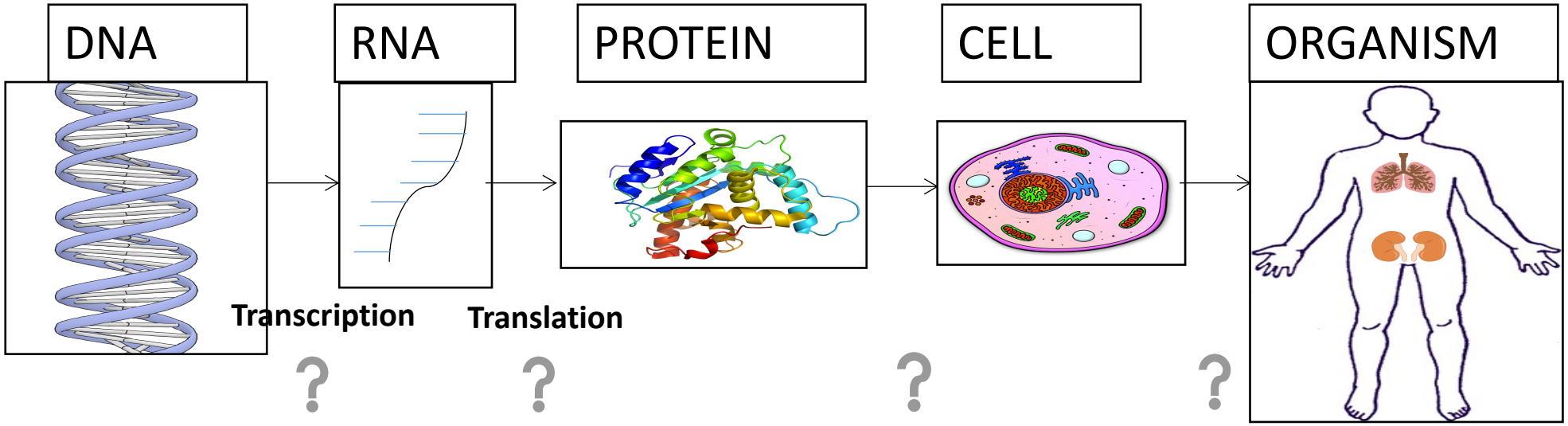
- Need help in biology
- Need help in medicine
- Need help in figuring out NIH grant applications

Timeline of deepchrome our tools

http://deepchrome.net/



Biological Modules

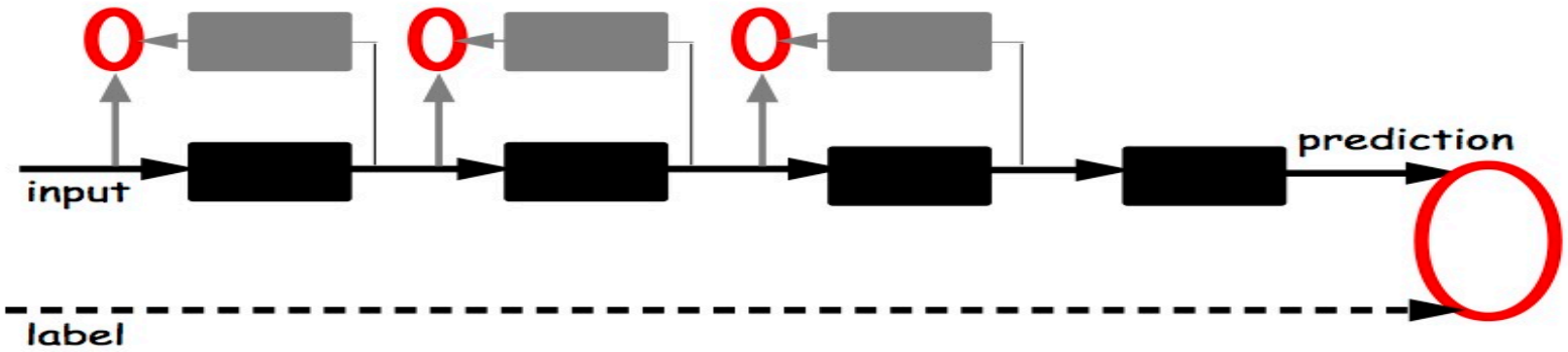


CATGACTG
CATGCCTG

Genetic Variant

Disease

Deep Learning Modules (composable)



Some Recent Trends

https://qdata.github.io/deep2Read/

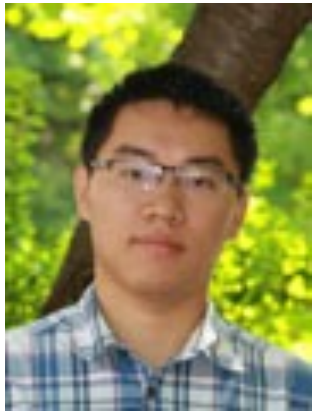
- 1. Autoencoder / layer-wise training
- 2. CNN / Residual / Dynamic parameter
- 3. RNN / Attention / Seq2Seq, …
- 4. Neural Architecture with explicit Memory
- 5. NTM 4program induction / sequential decisions
- 6. Learning to optimize / Learning DNN architectures
- 7. Learning to learn / meta-learning/ few-shots
- 8. DNN on graphs / trees / sets
- 9. Deep Generative models, e.g., autoregressive
- 10. Generative Adversarial Networks (GAN)
- 11. Deep reinforcement learning
- 12. Validate / Evade / Test / Understand / Verify DNNs

Acknowledgements

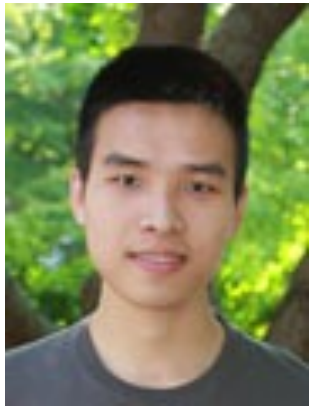
Ritambhara Singh
Now Assistant Professor @Brown

Jack Lanchantin

Arshdeep Sekhon



Beilun Wang
Now Associate Professor @ Southeastern Univ.

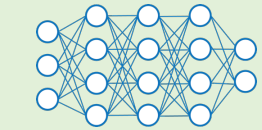


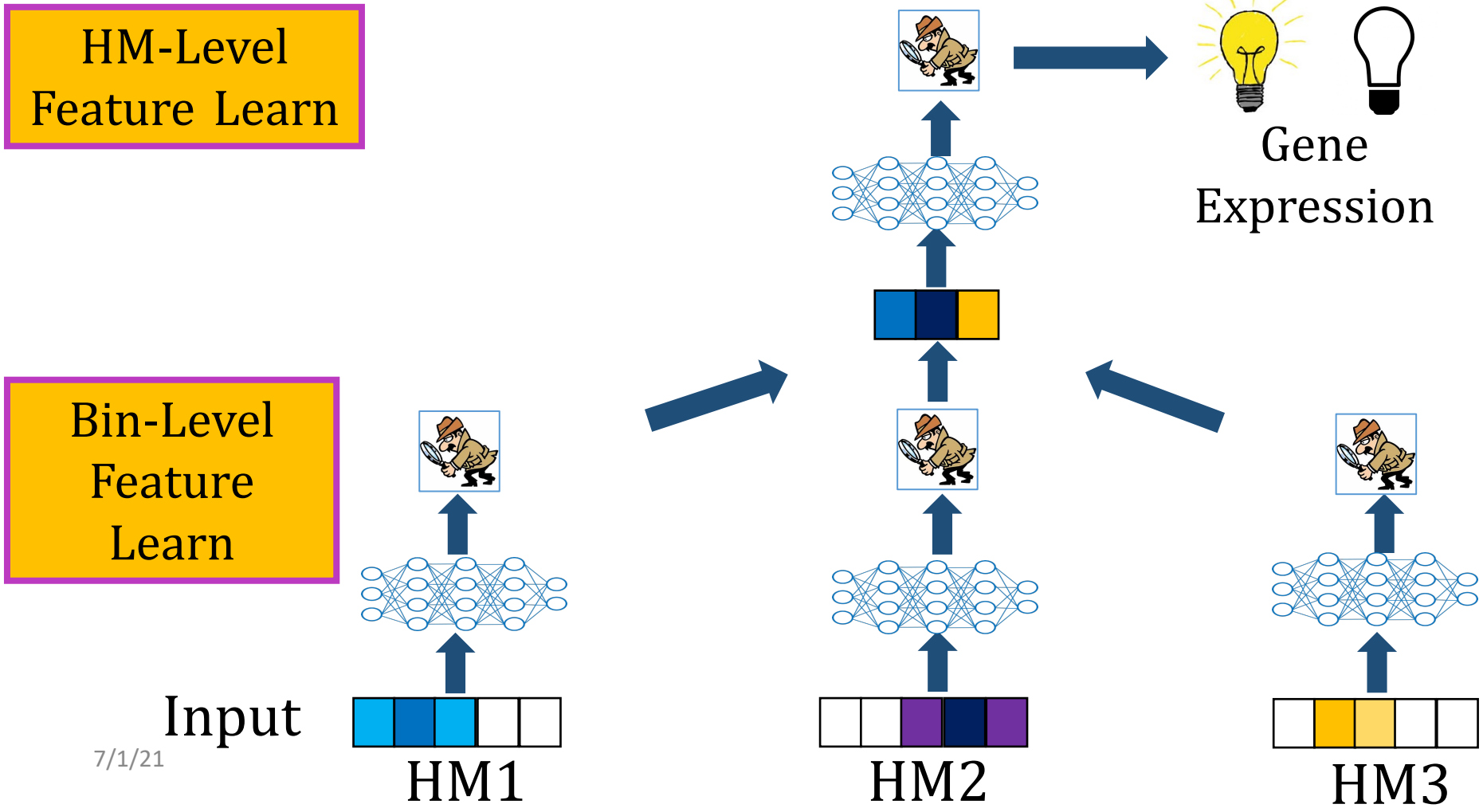
Weilin Xu, Now Research Staff @ Intel Labs

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

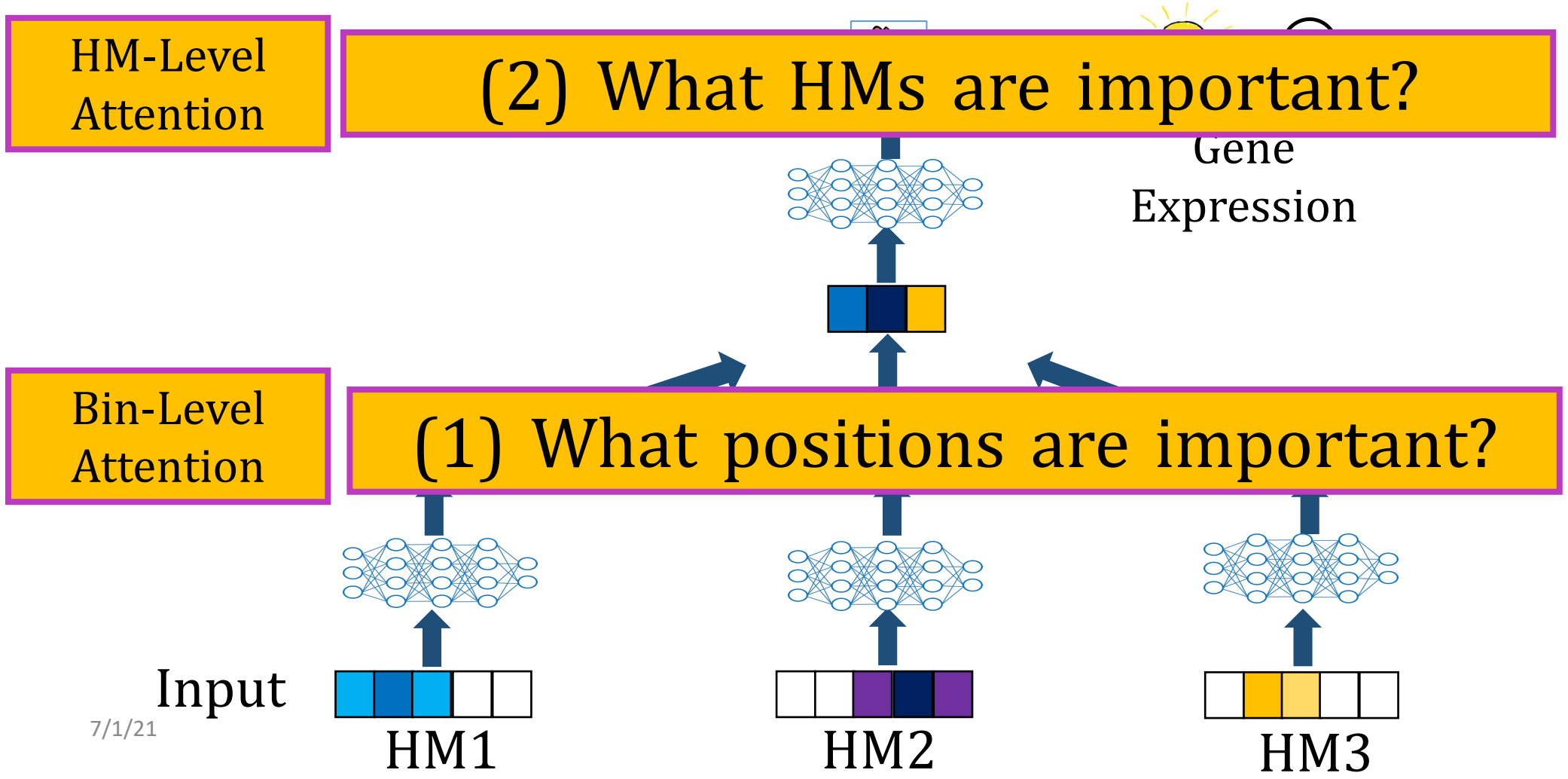

Thank you

AttentiveChrome

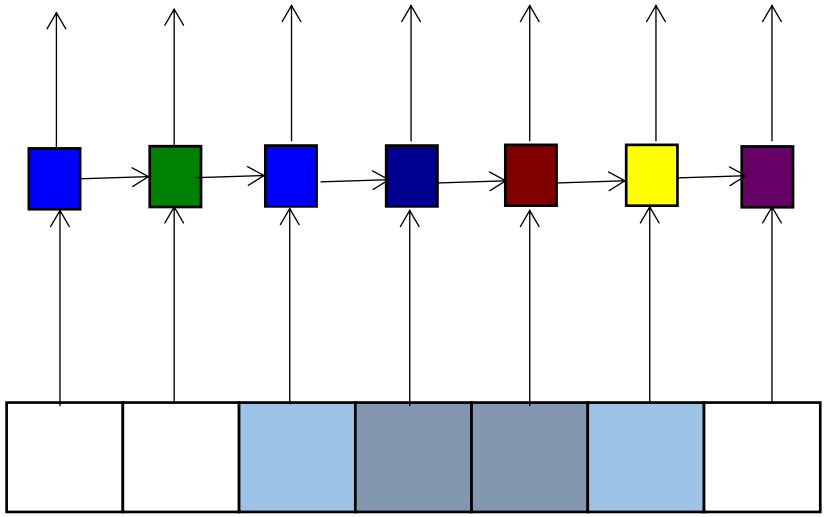




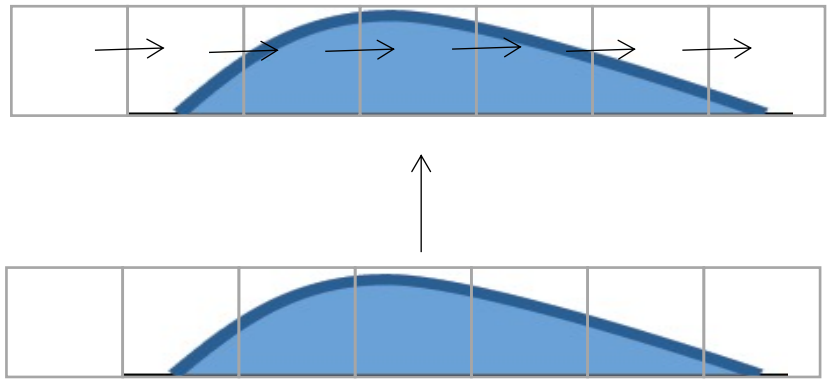
AttentiveChrome



Multiple Recurrent Neural Networks (Hierarchical RNNs) to model each HM and the Combination of all HMs : for example on HM1

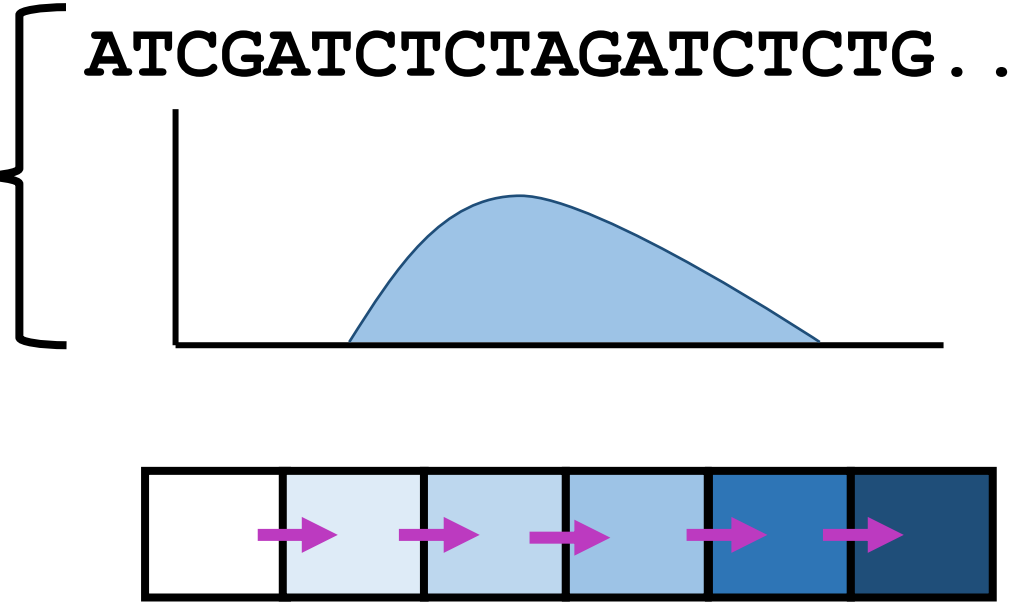


HM1



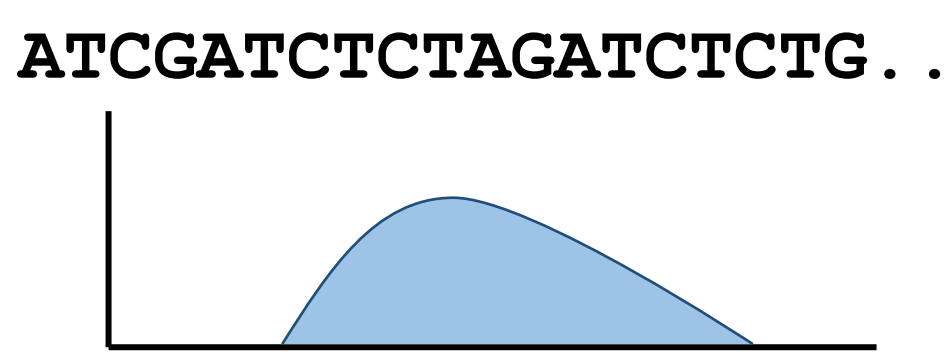
Signals on Genome are Sequential Input (X)

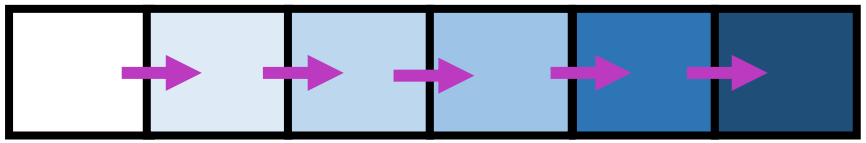
Strings, signals etc.



Sequential Input (X)

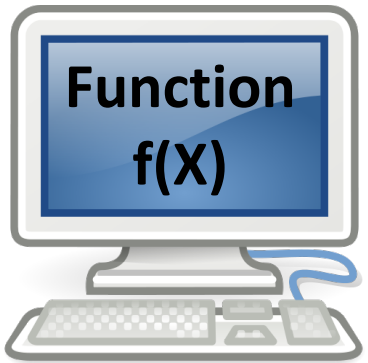
Strings, signals etc.





X

This Food is not good.



NO
(-1)

RNN models sequential inputs well

- Make fully-connected layer model each unit recurrently
- Units form a directed chain graph along a sequence
- Each unit uses recent history and current input in modeling

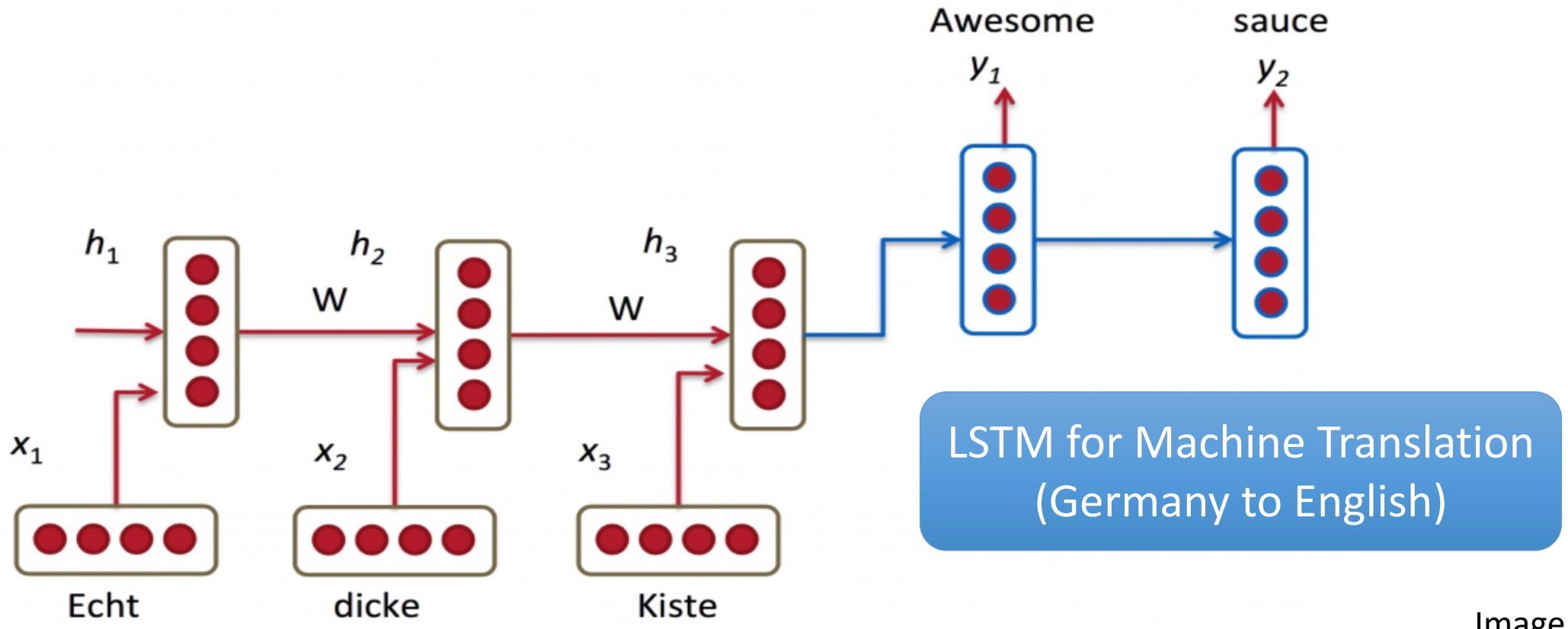
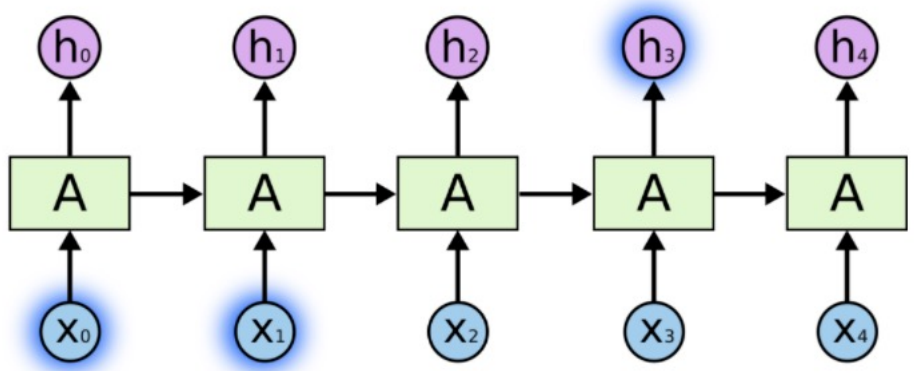


Image credit : wildML

Using Attention to Select RNN per-unit outputs



$$h_t = f_W(h_{t-1}, x_t)$$

new state — old state — input vector at some time step

some function with parameters W

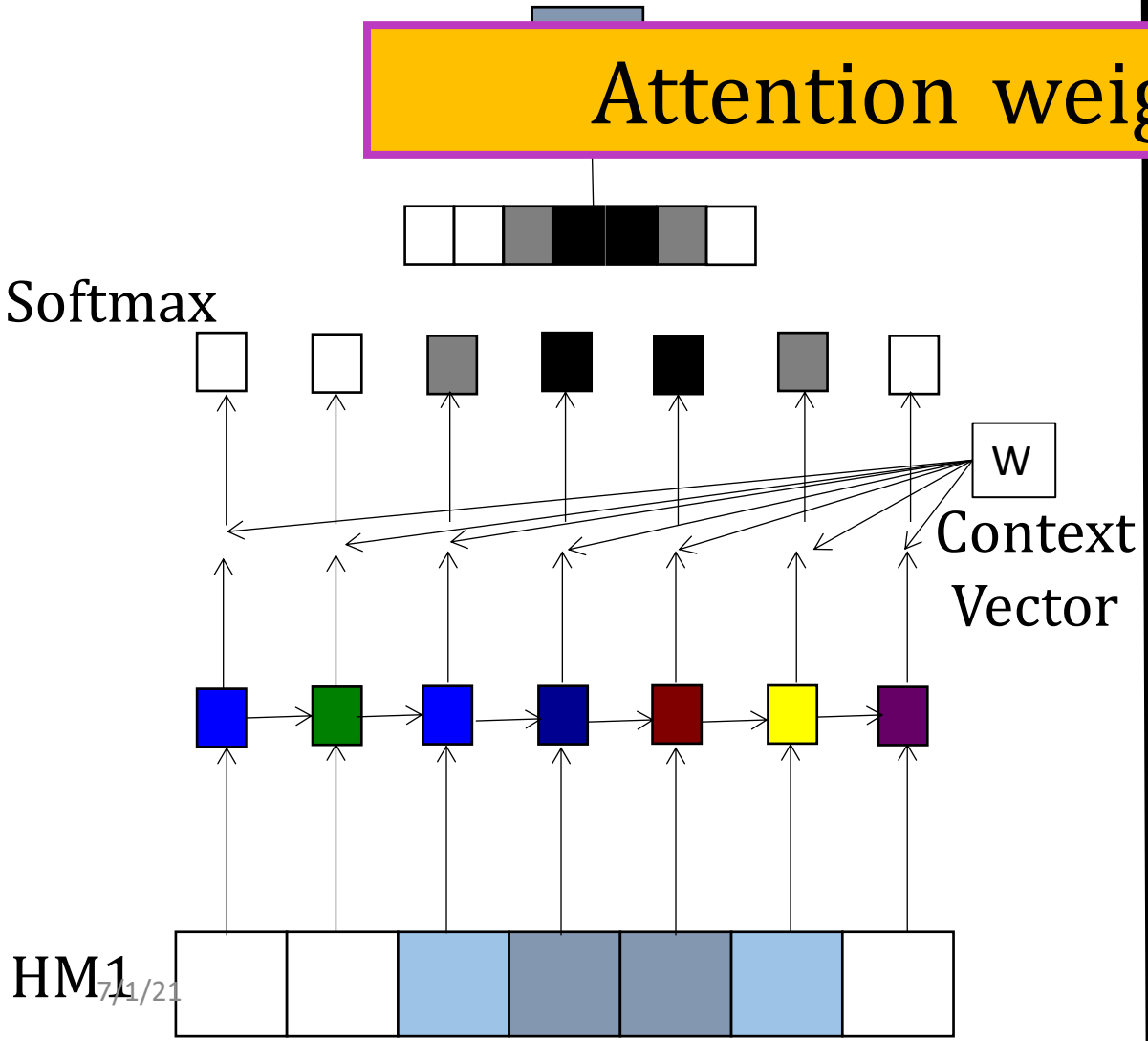
$$\alpha_t^j = \frac{exp(\mathbf{W}_b \mathbf{h}_t^j)}{\sum_{i=1}^{T} exp(\mathbf{W}_b \mathbf{h}_i^j)}$$

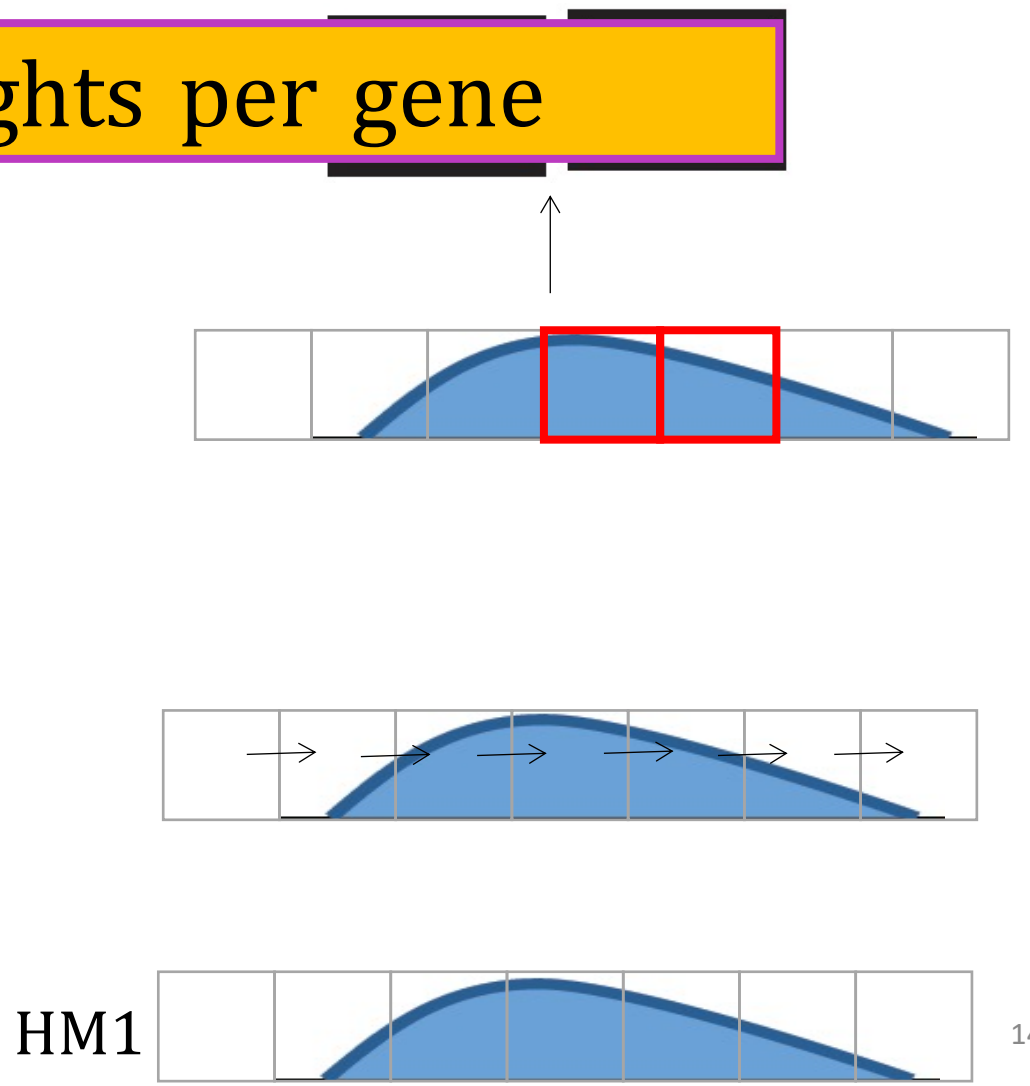
W_b is learned

Image Credits from Christopher Olah

Attention Mechanism

Attention weights per gene

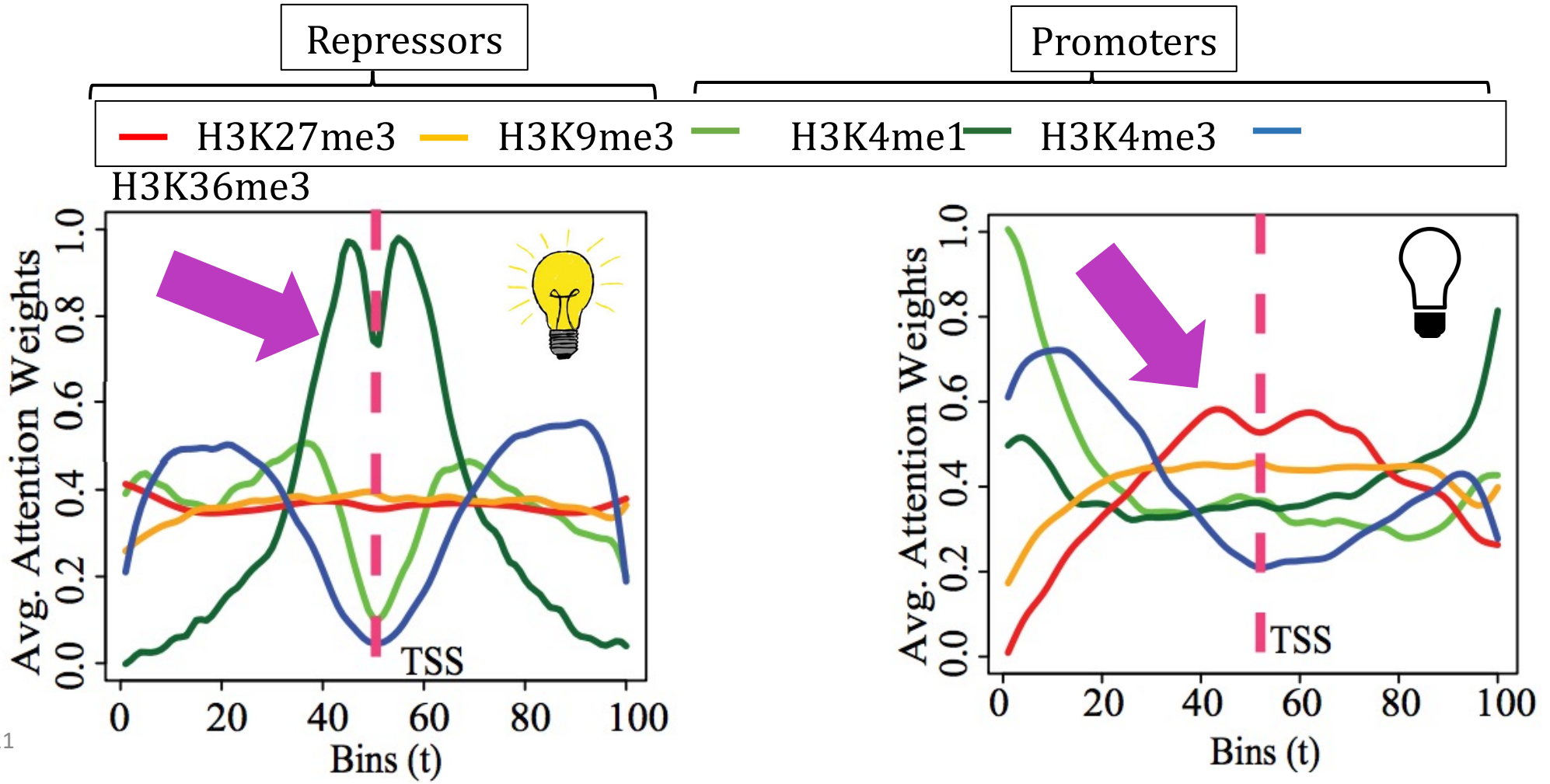




Bin-Level Visualization

(1) What positions are important?

CELL TYPE: GM12878 (Blood Cell)

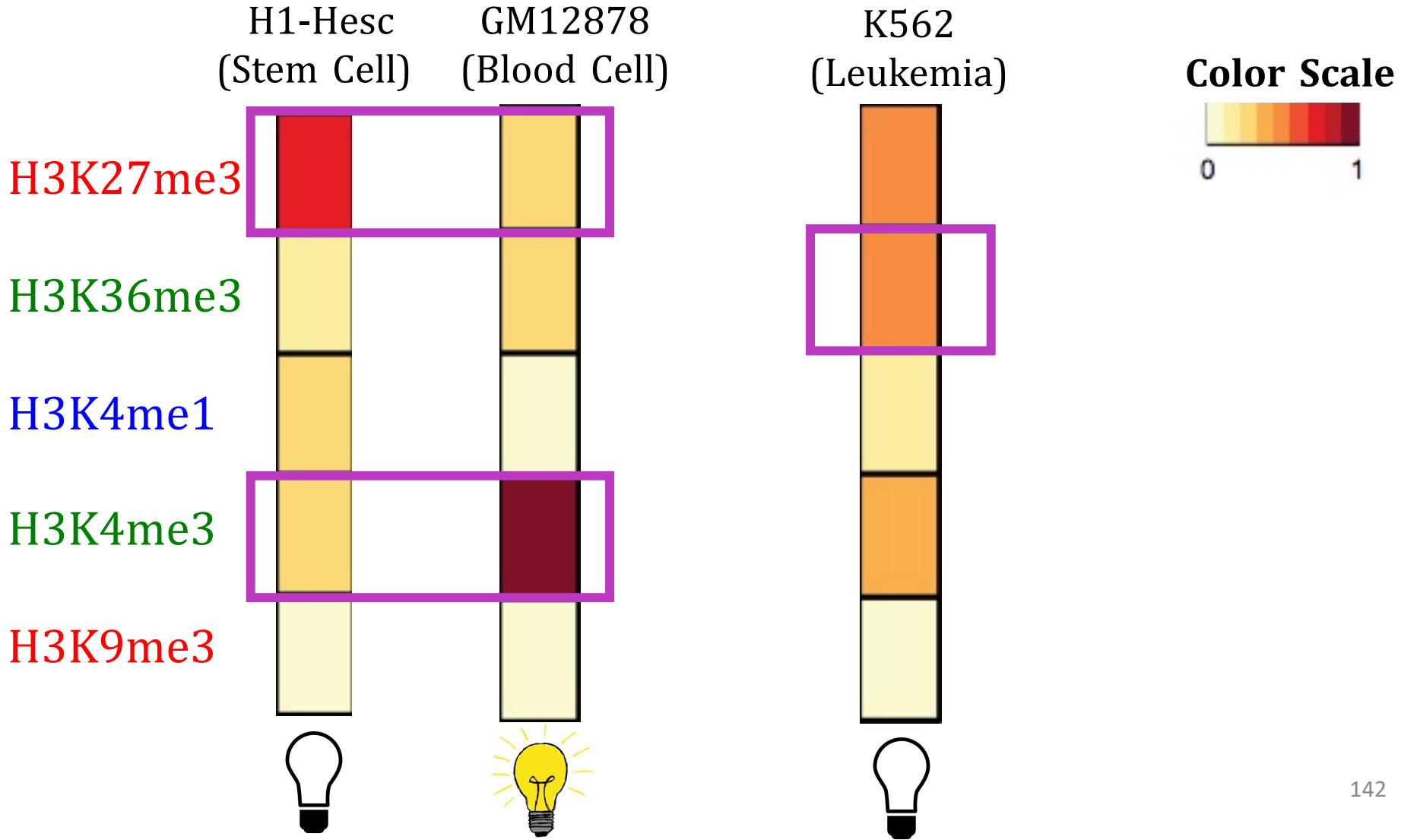


HM-Level Visualization

(2) What HMs are important?

Cell Types:

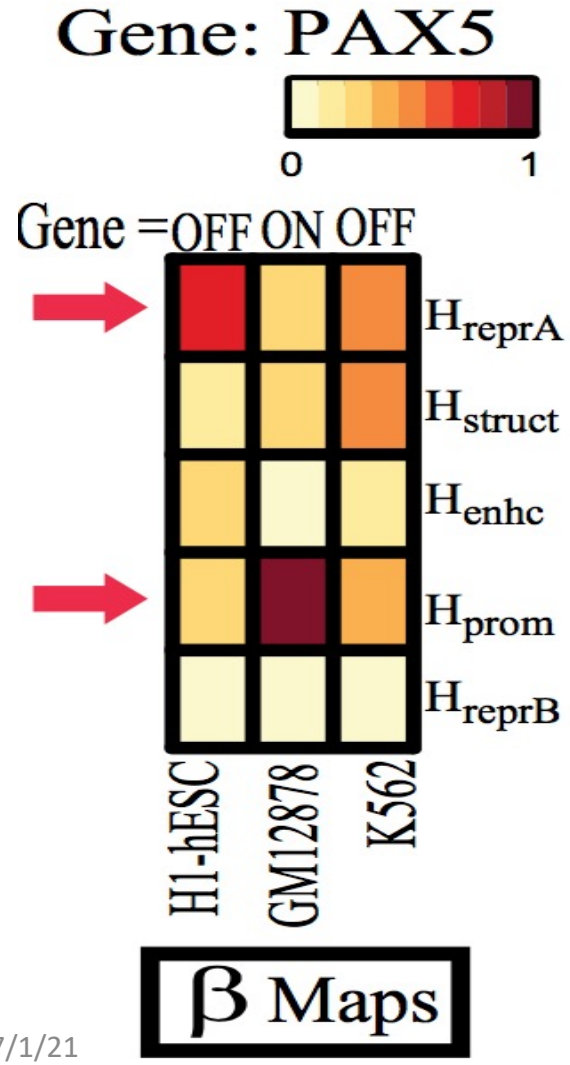
Gene: PAX5



PROMOTER
DISTAL PROMOTER
REPRESSOR

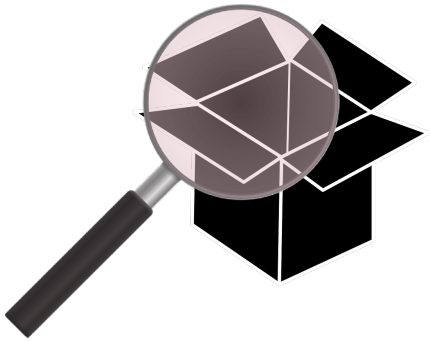
Results: HM level attention

(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.

Connecting to Related Work to Post-Understand DNN



- Deconvolution — Activation
- Perturbation-based — Temporal Output Values
- Backpropagation-based — Saliency Map, Class Optimization
- Difference to Reference — DeepLift
- Influence based — Influential Function / ICML27 Best Paper