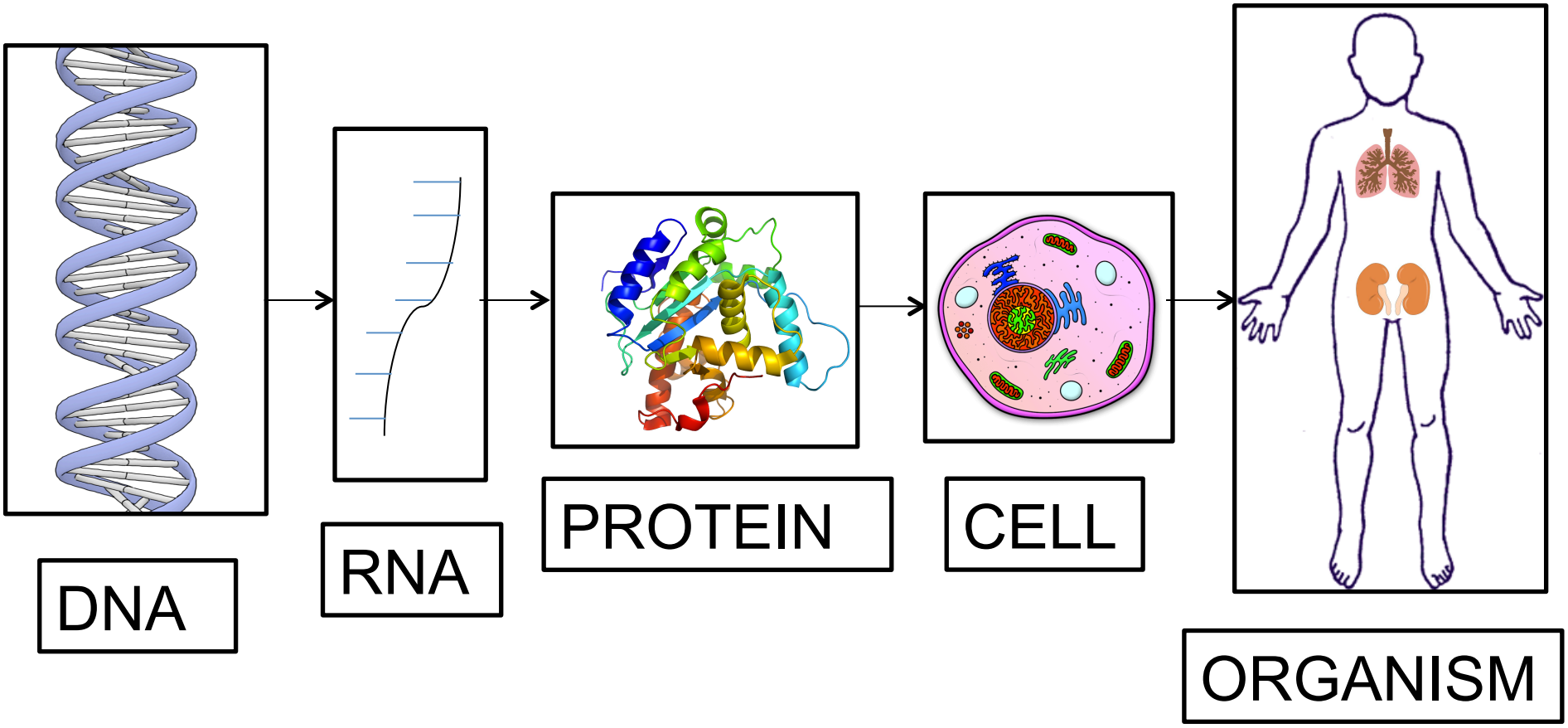
DeepChrome

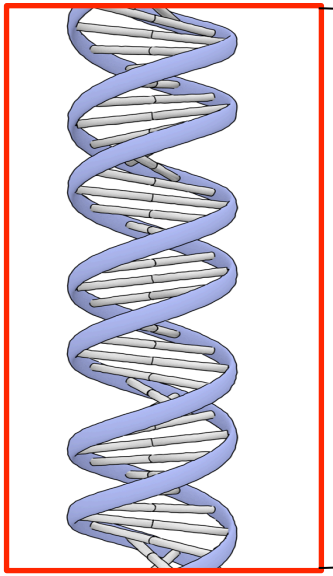
Deep-learning for predicting gene expression from histone modification

Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi

Biology in a Slide



DNA and Diseases



DNA

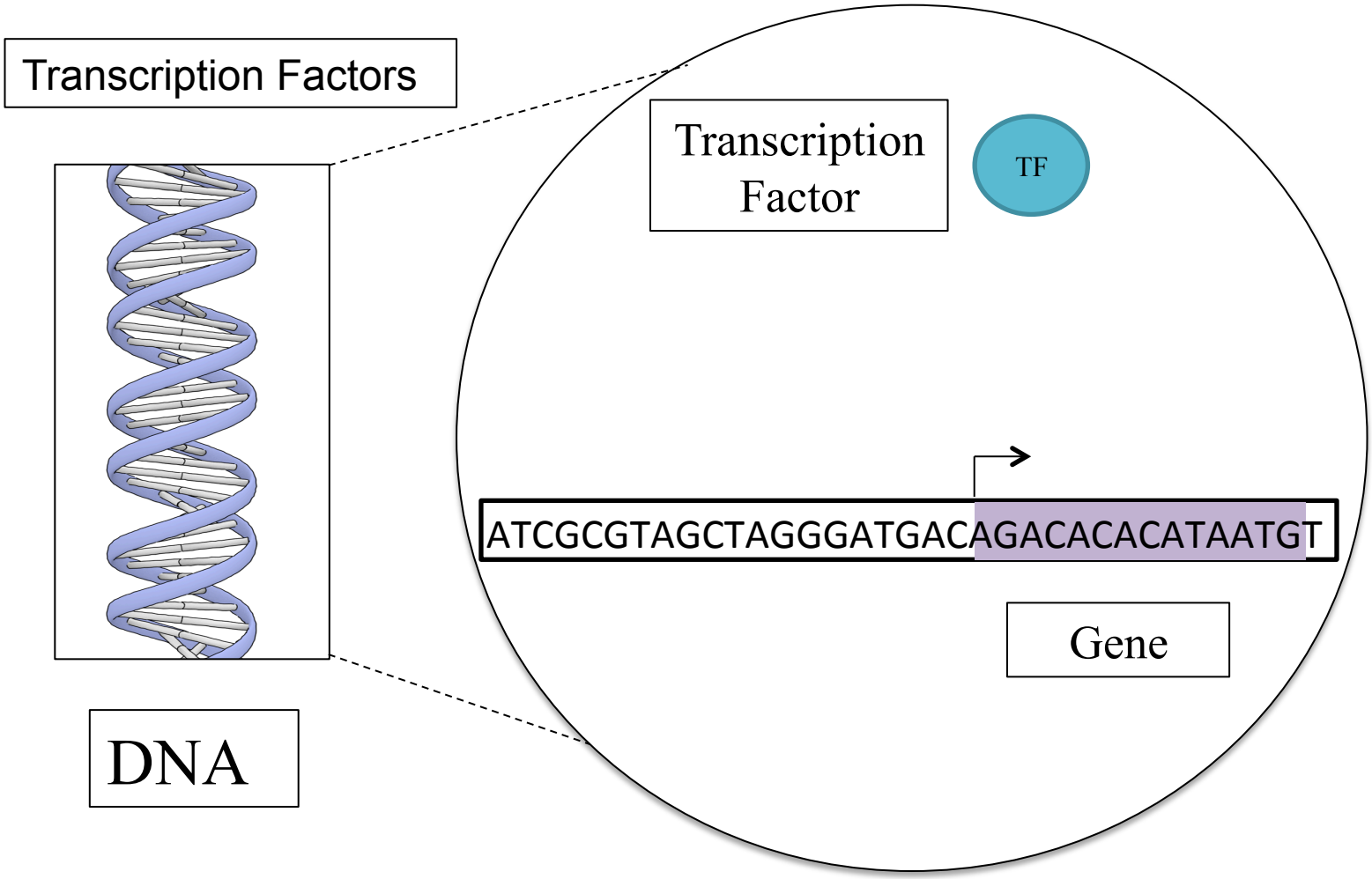
- Down Syndrome
- Parkinson's Disease
- Autism
- Muscular Atrophy
- Sickle Cell Disease

…….

……..

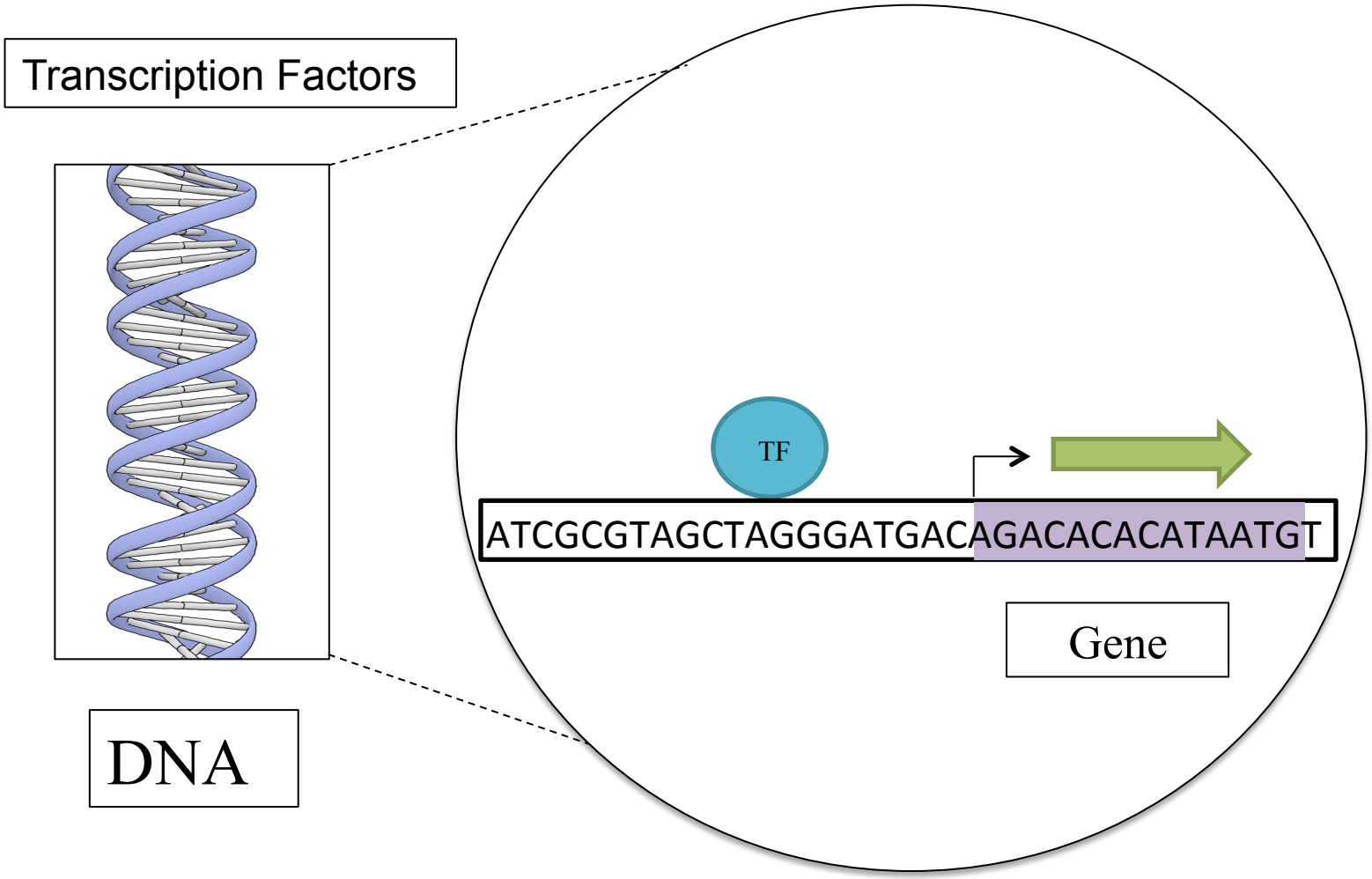
Epigenetic Factors

- Encyclopedia of DNA Elements (ENCODE)
- Roadmap Epigenetics Project (REMC)



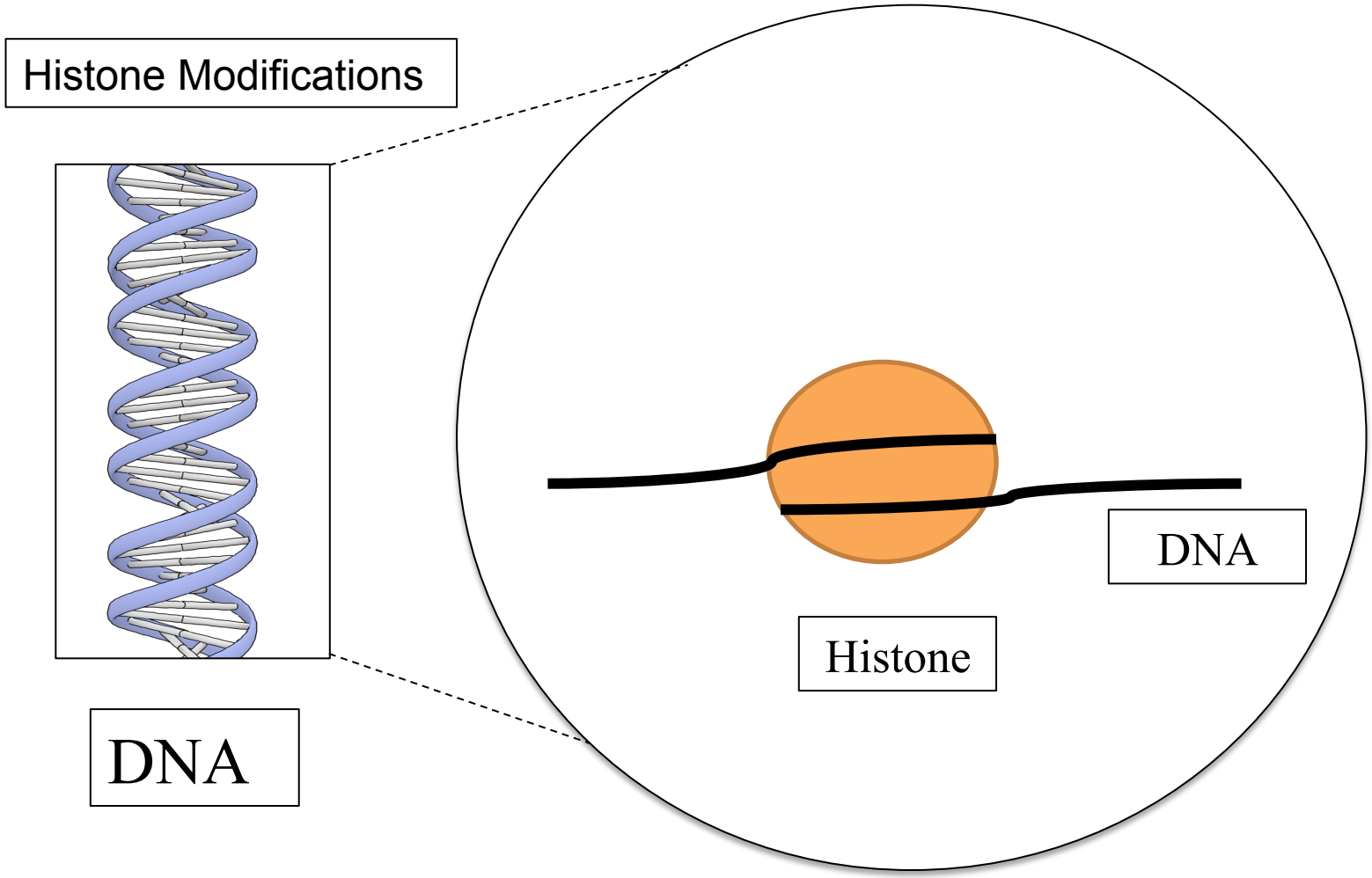
Epigenetic Factors

- Encyclopedia of DNA Elements (ENCODE)
- Roadmap Epigenetics Project (REMC)



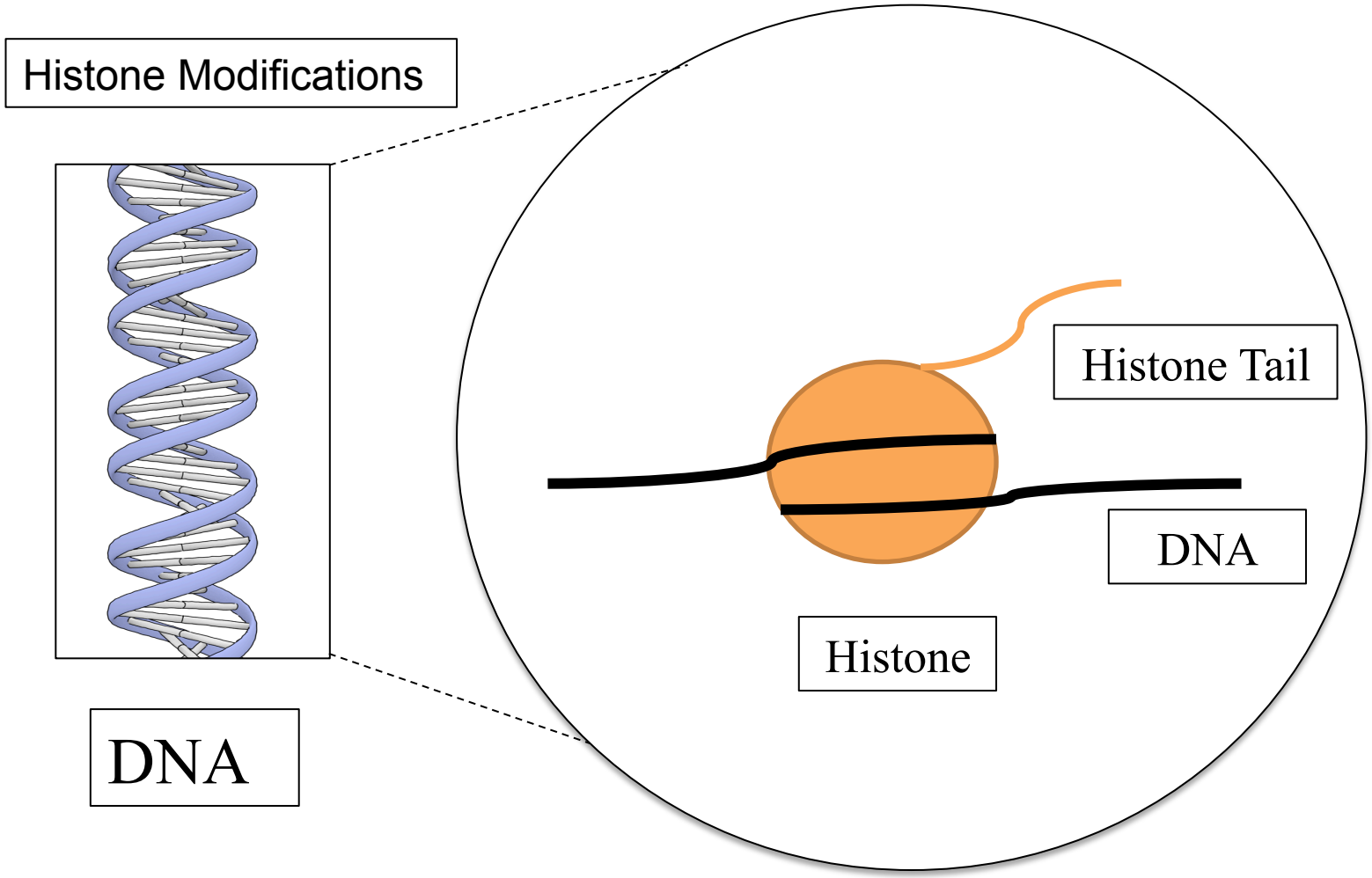
Epigenetic Factors

- Encyclopedia of DNA Elements (ENCODE)
- Roadmap Epigenetics Project (REMC)



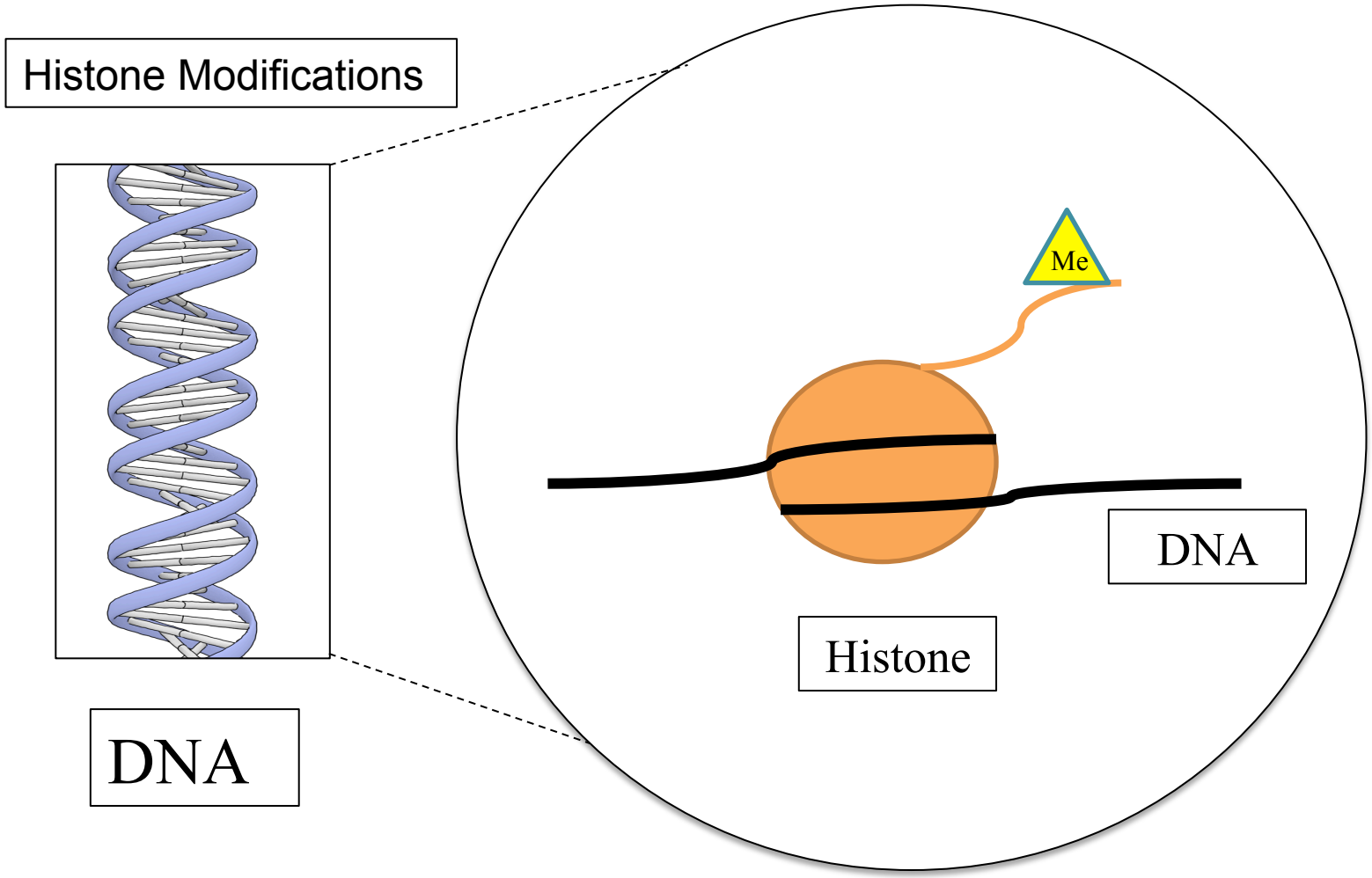
Epigenetic Factors

- Encyclopedia of DNA Elements (ENCODE)
- Roadmap Epigenetics Project (REMC)



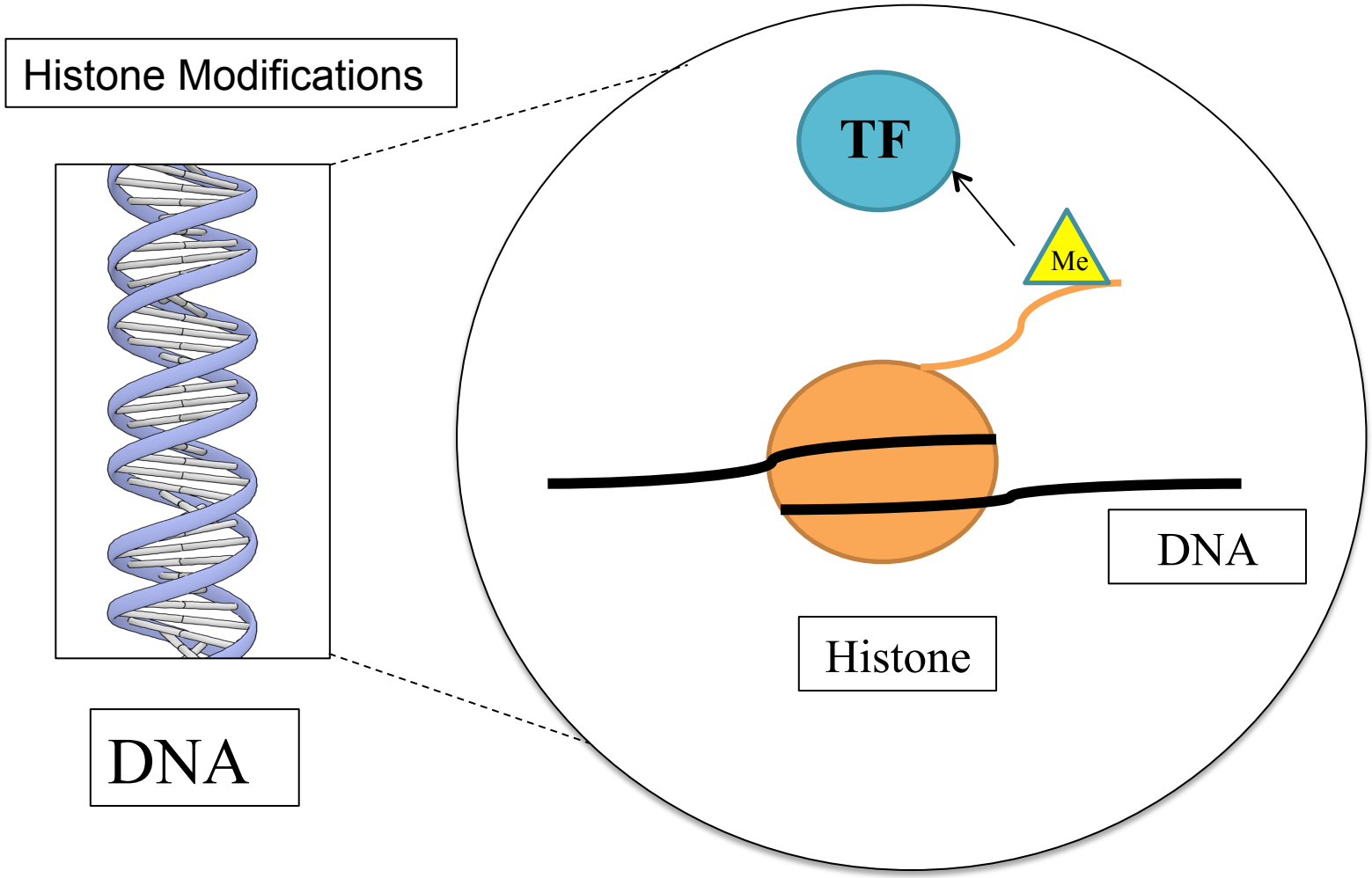
Epigenetic Factors

- Encyclopedia of DNA Elements (ENCODE)
- Roadmap Epigenetics Project (REMC)

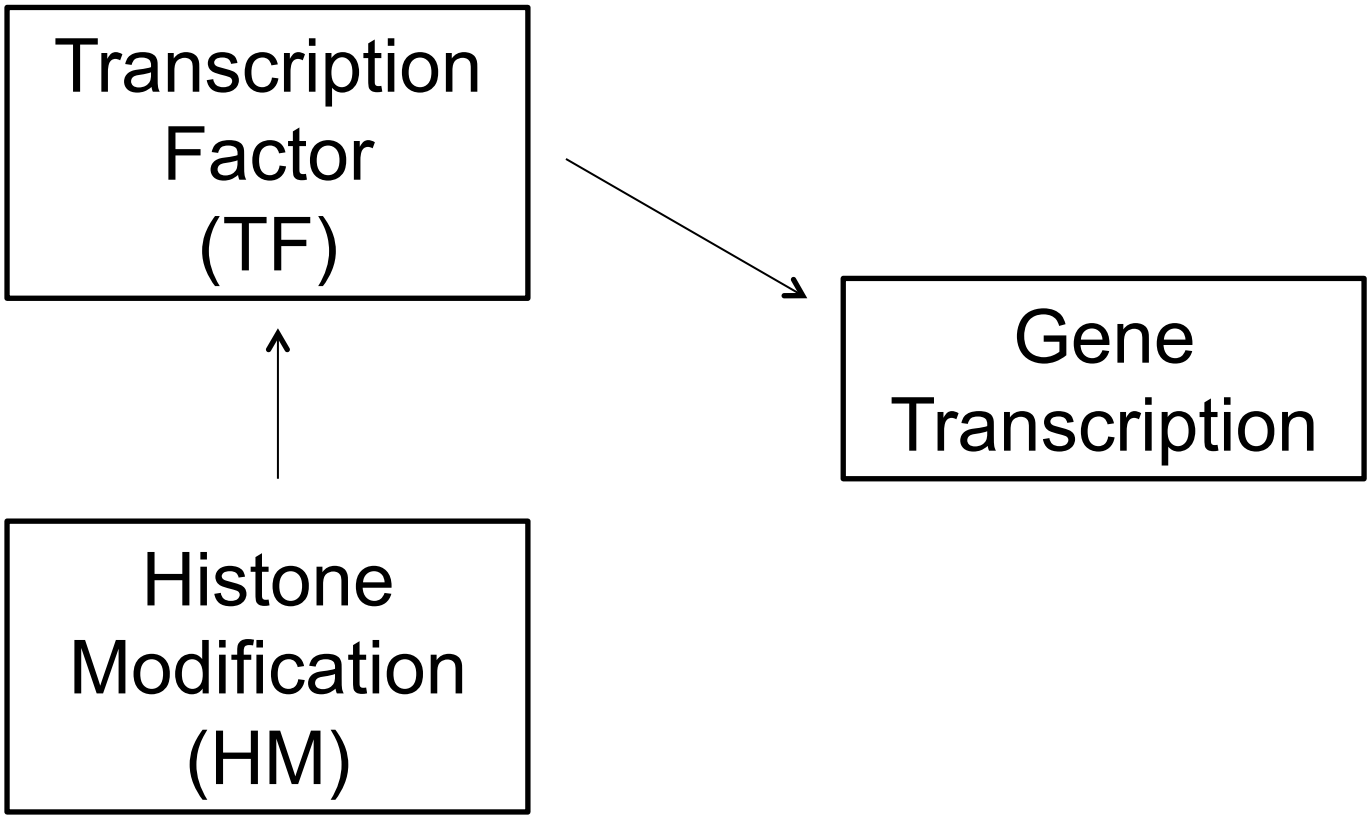


Epigenetic Factors

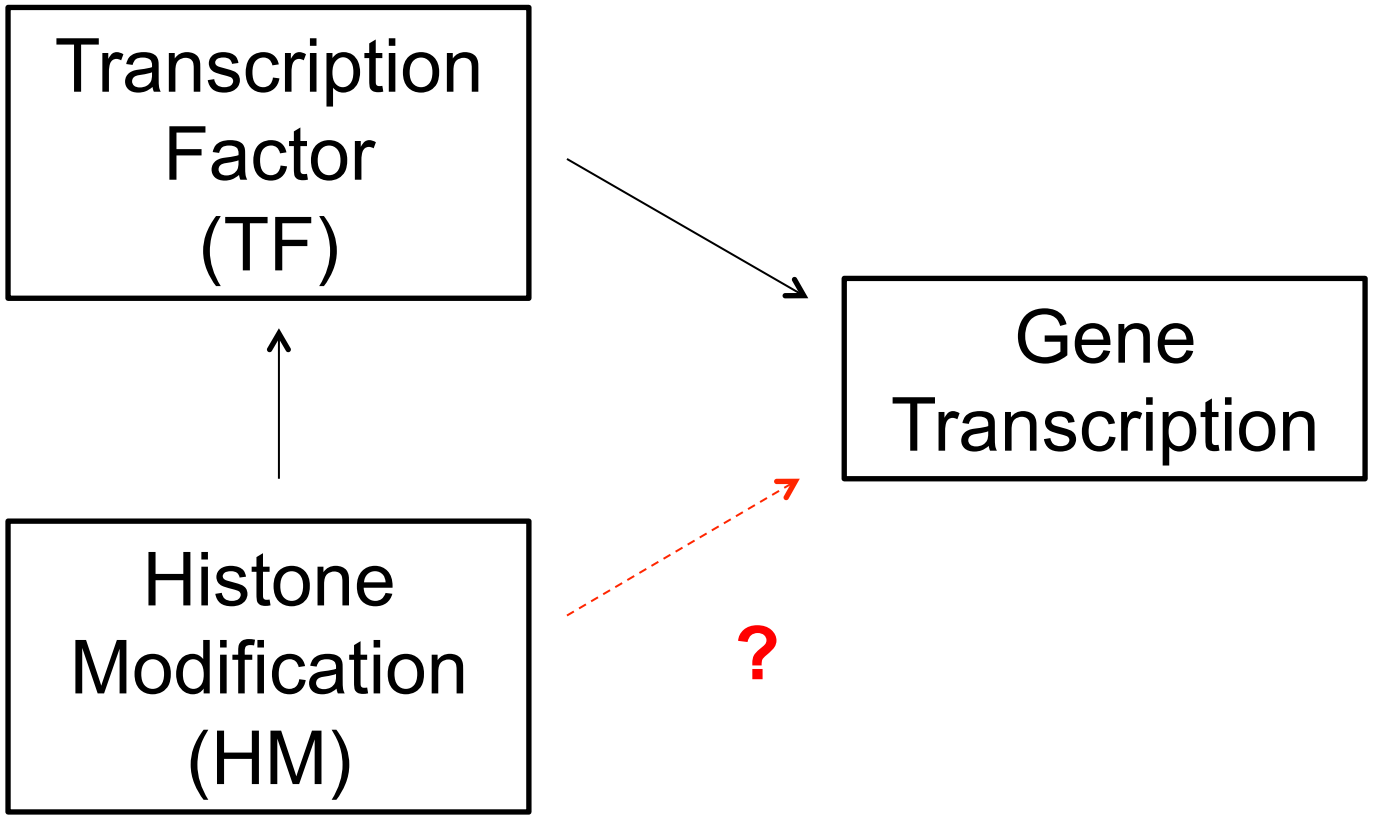
- Encyclopedia of DNA Elements (ENCODE)
- Roadmap Epigenetics Project (REMC)



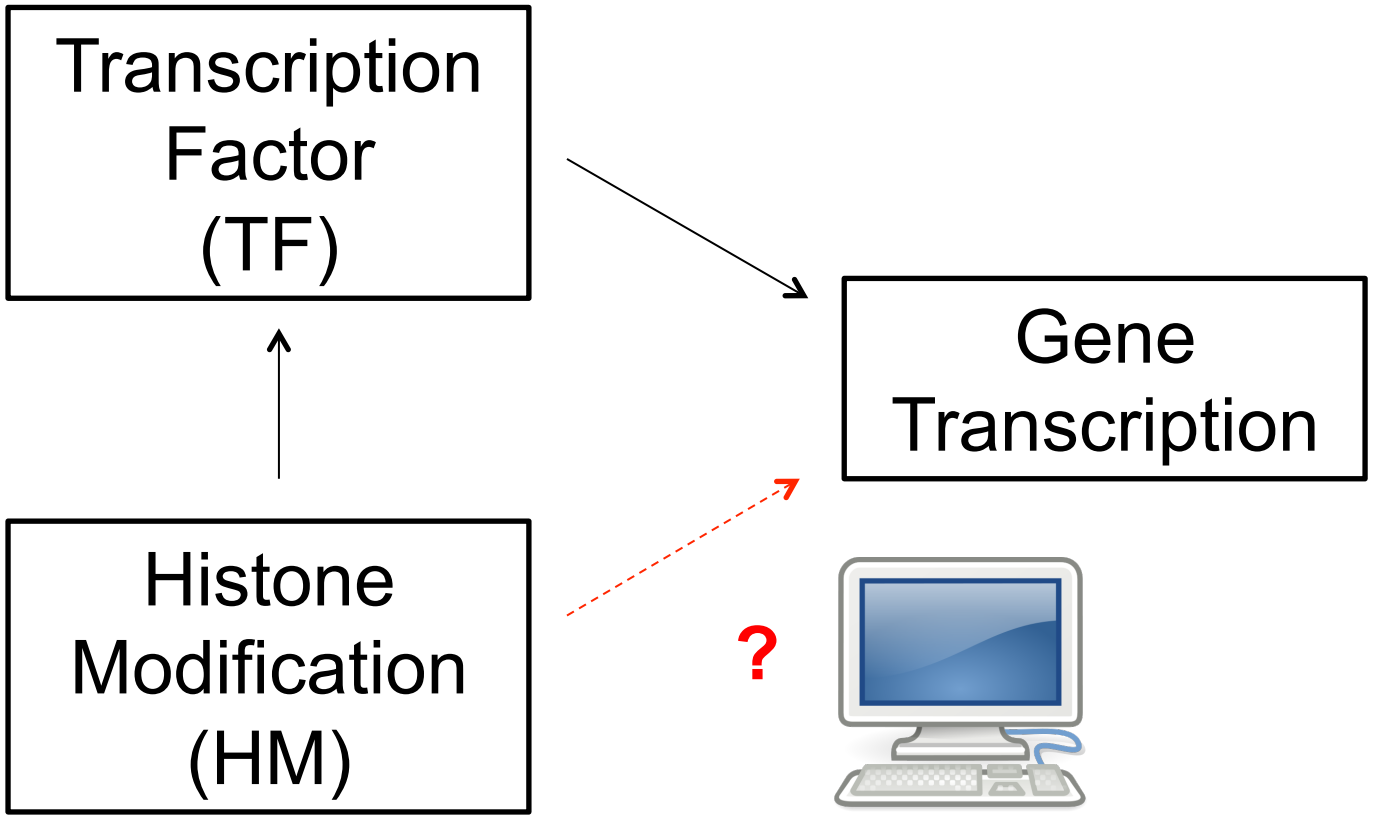
Histone Modification and Gene Transcription



Histone Modification and Gene Transcription

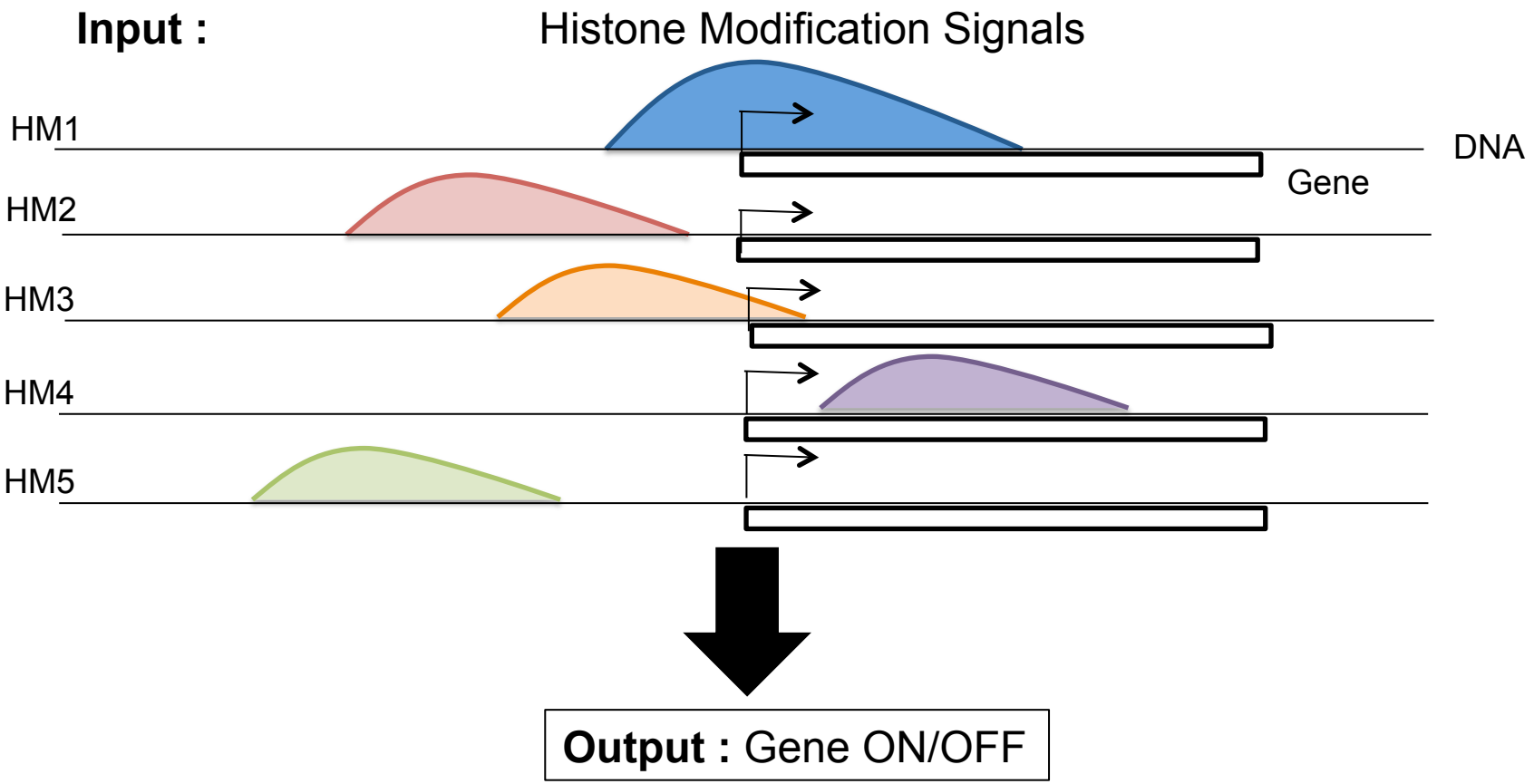


Histone Modification and Gene Transcription

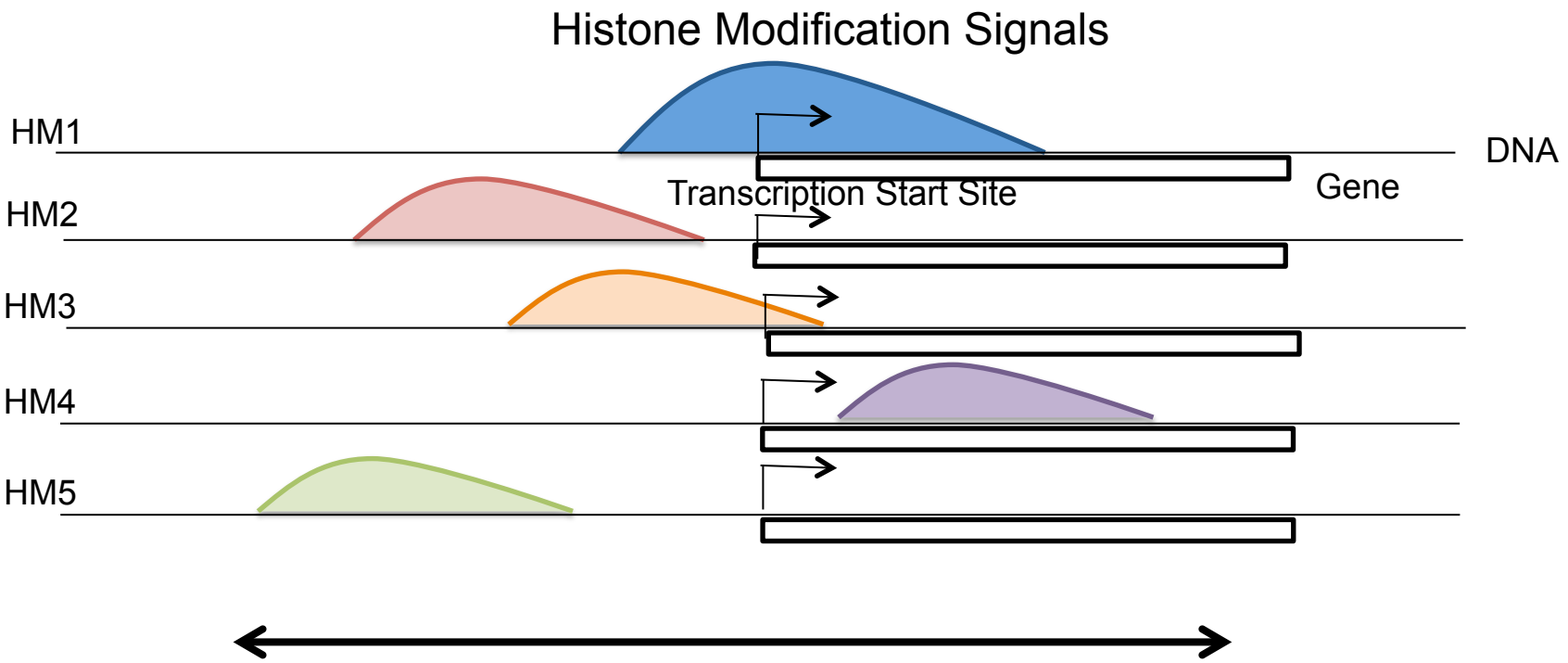


Task Formulation

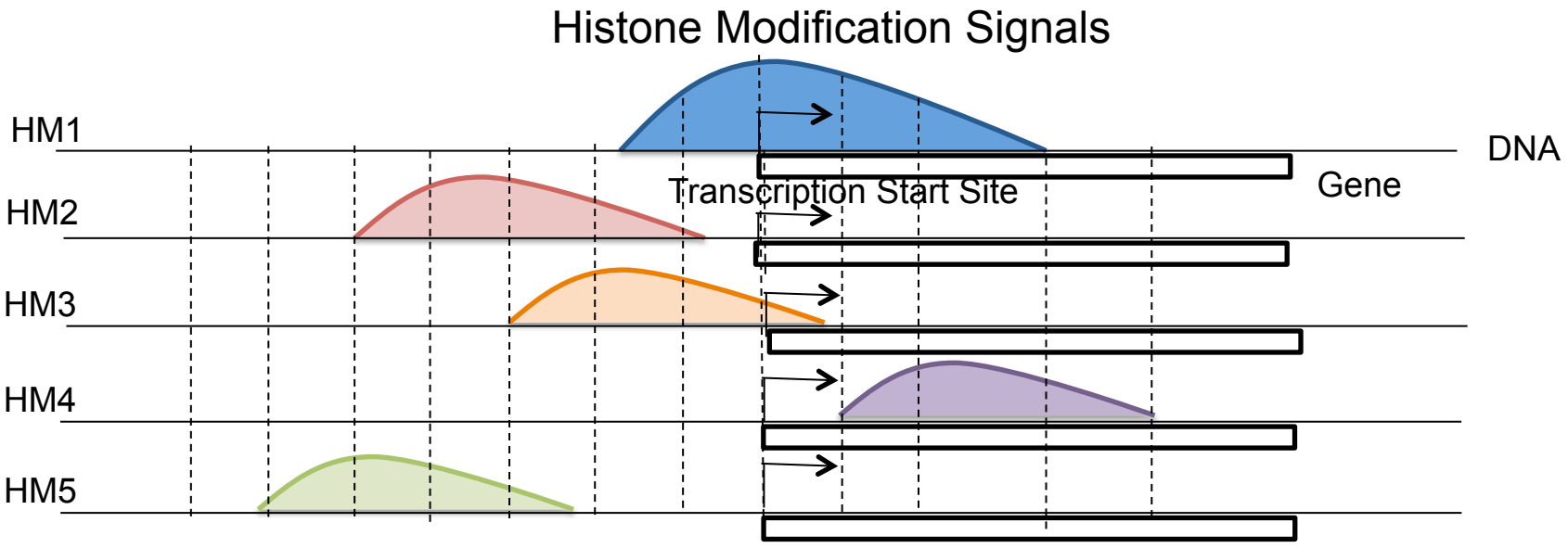
Prediction Task



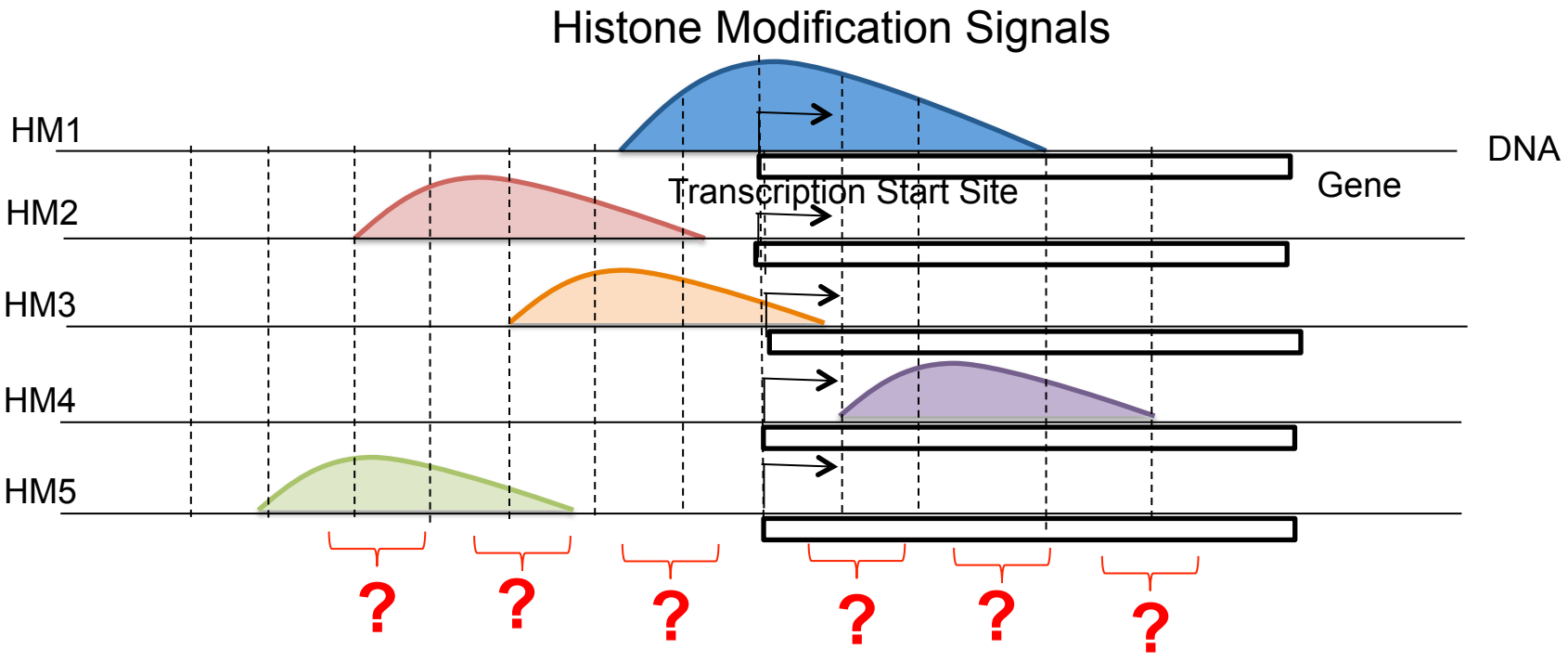
Input


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
DNA
Gene
Transcription Start Site

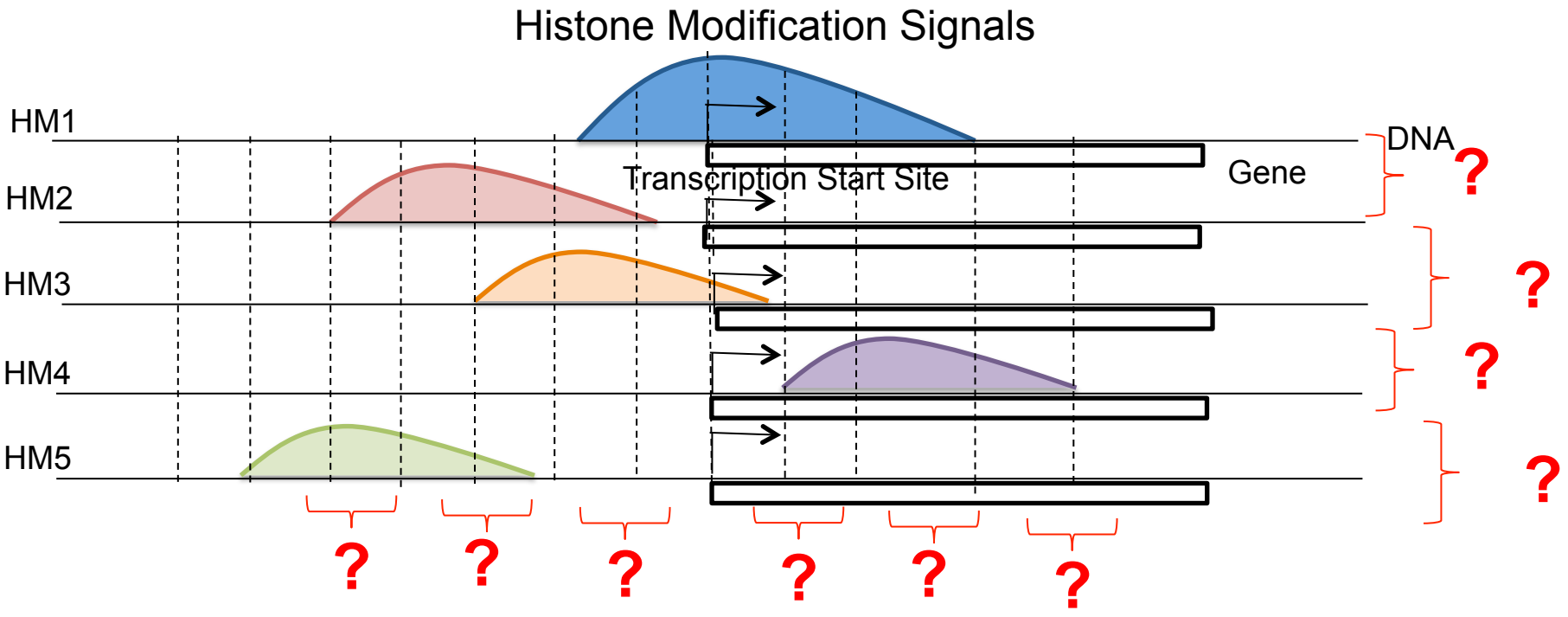
Input


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
DNA
Gene
Transcription Start Site

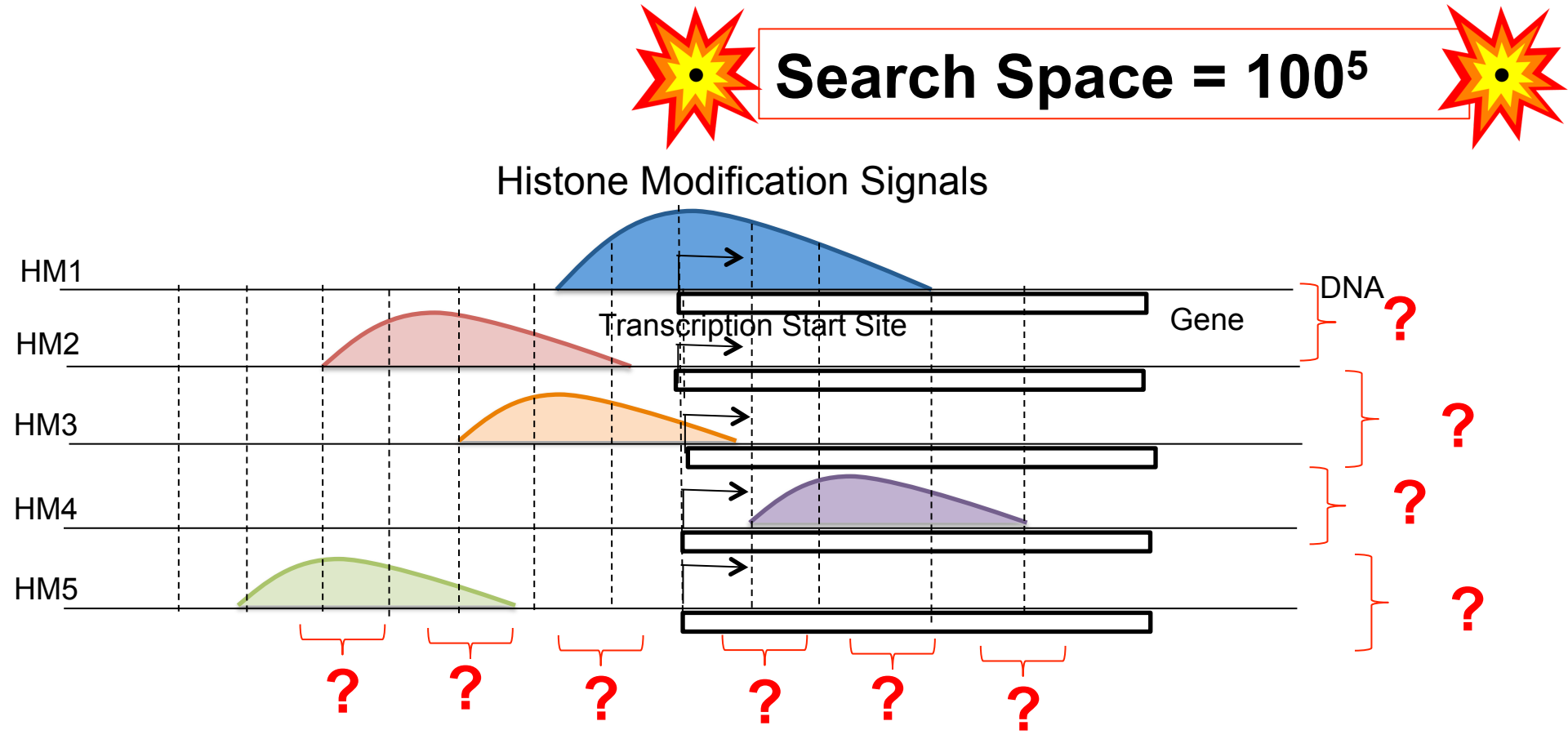
Challenge


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
DNA
Gene
Transcription Start Site
?
?
?
?
?
?

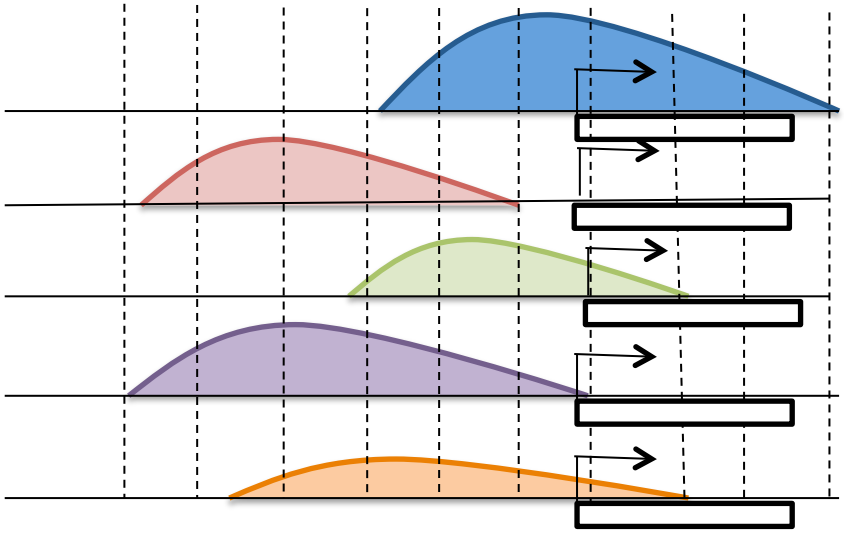
Challenge


Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
Transcription Start Site
Gene
DNA
?
?
?
?
?
?
?
?
?
?
?

Challenge


Search Space = 100^5
Histone Modification Signals
HM1
HM2
HM3
HM4
HM5
DNA
Gene
Transcription Start Site
?

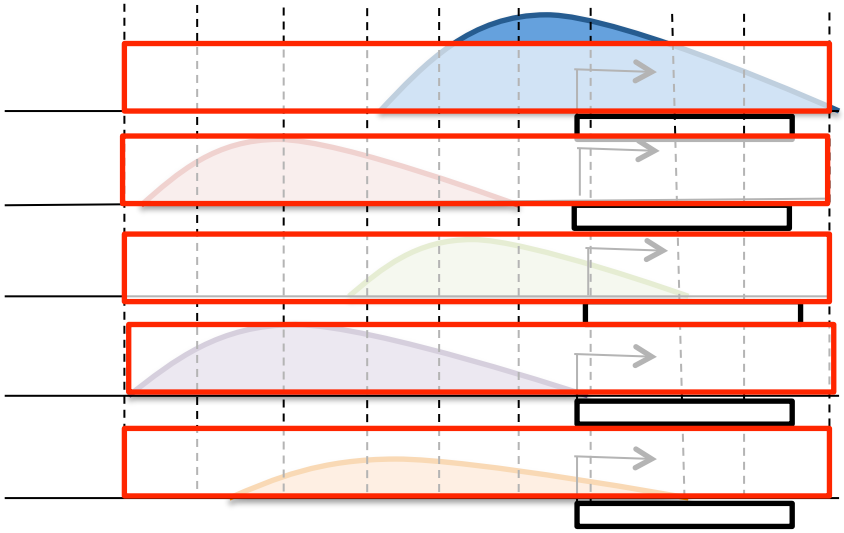
Related Work



Linear Regression,
SVM,
Random Forest

Gene ON/OFF

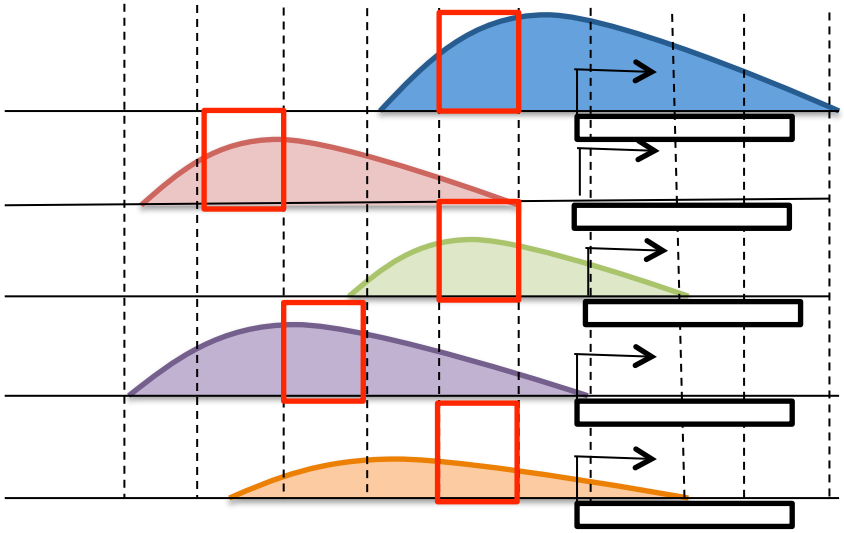
Related Work



Linear Regression,
SVM,
Random Forest

Gene ON/OFF

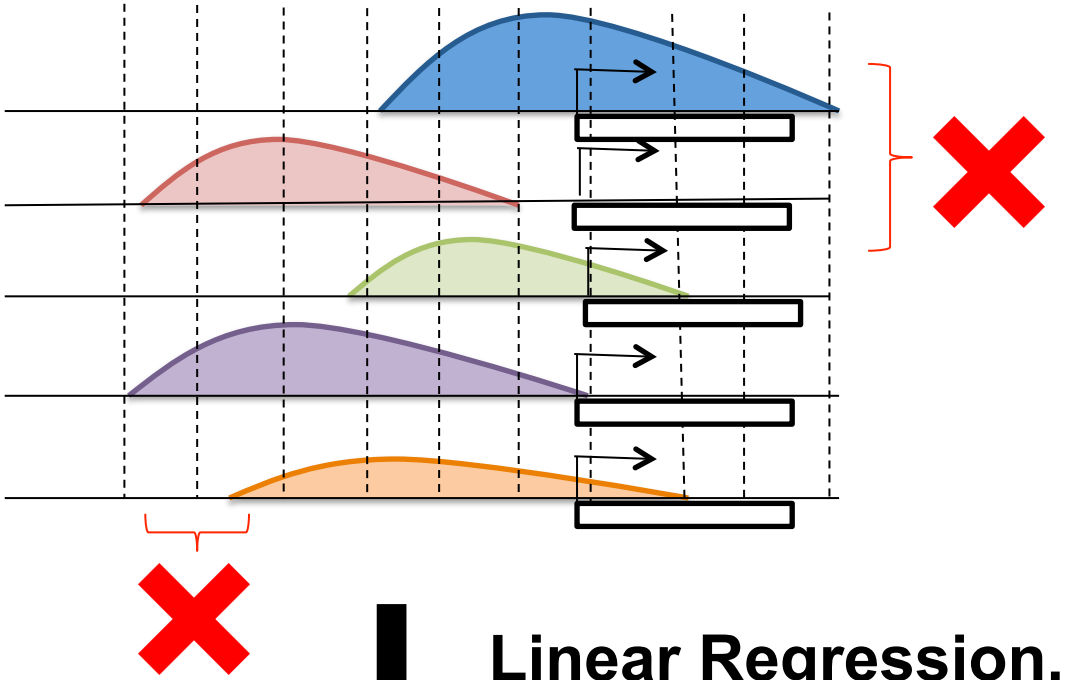
Related Work



Linear Regression,
SVM,
Random Forest

Gene ON/OFF

Drawback

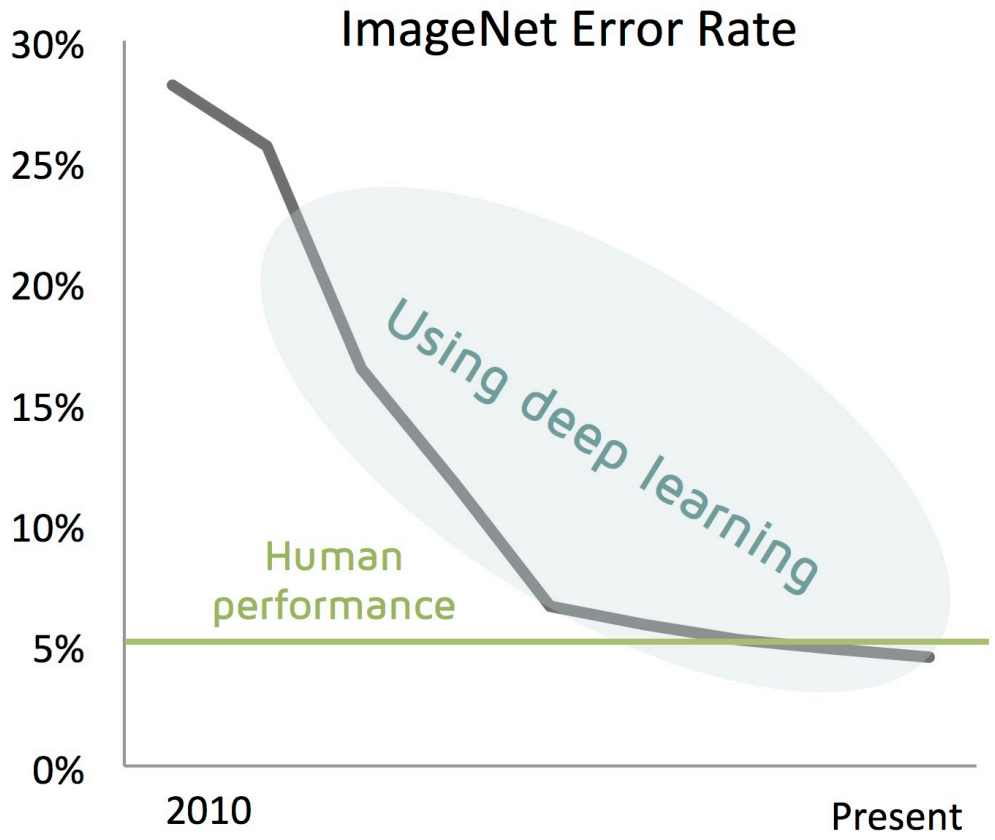


Linear Regression,
SVM,
Random Forest

Gene ON/OFF

Solution

Convolutional Neural Network (CNN)



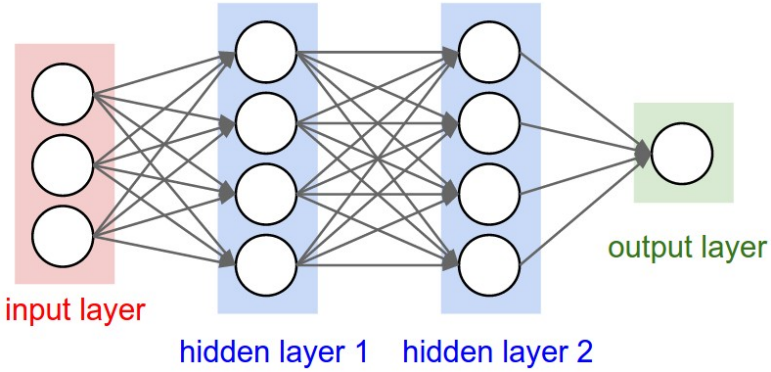
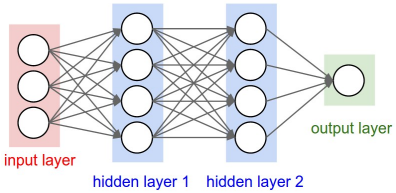


Image courtesy: (1) https://www.nervanasys.com/deep-learning-and-the-need-for-unified-tools / (2) http://cs231n.github.io/convolutional-networks/

Solution

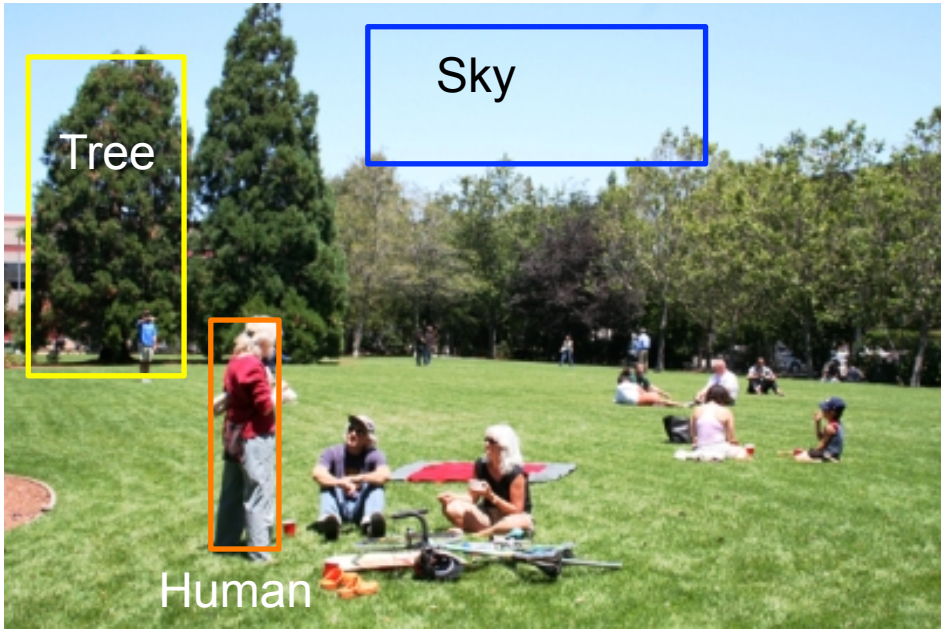
Convolutional Neural Network (CNN)

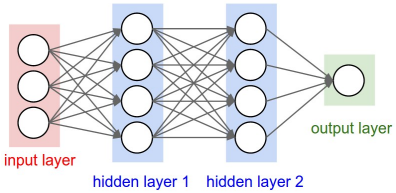


Park

Solution

Convolutional Neural Network (CNN)

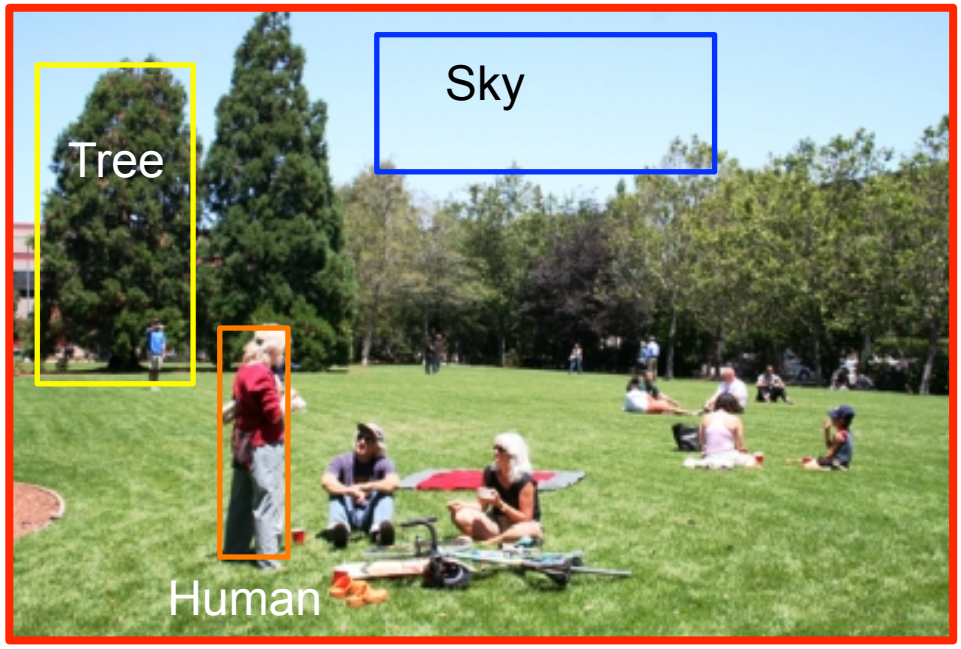


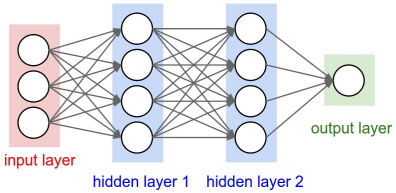


Park

Solution

Convolutional Neural Network (CNN)

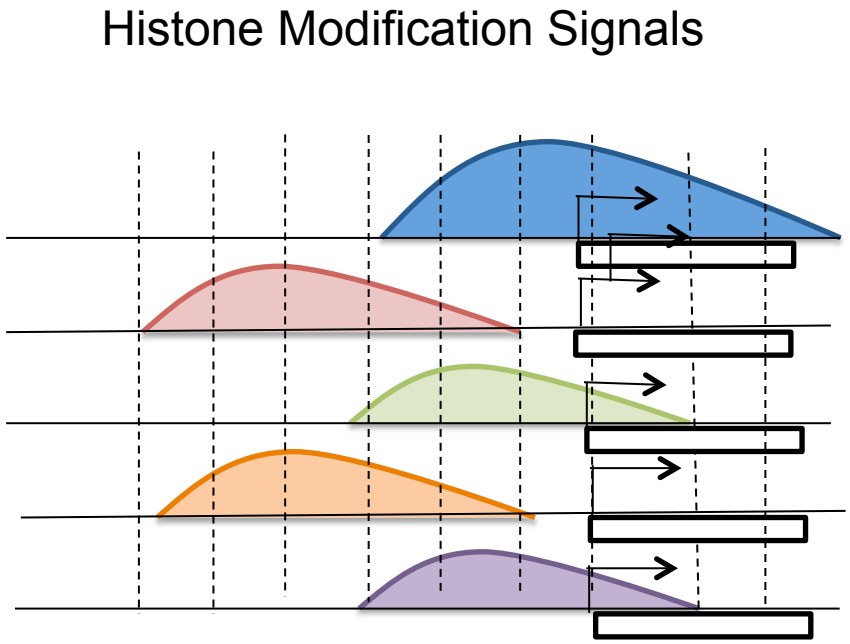




Park

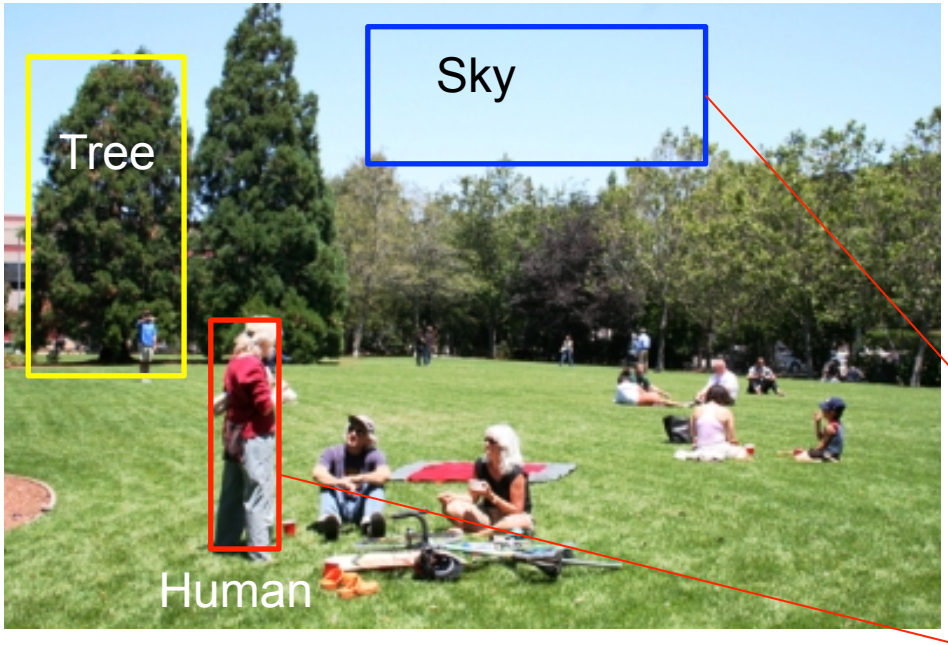
Solution

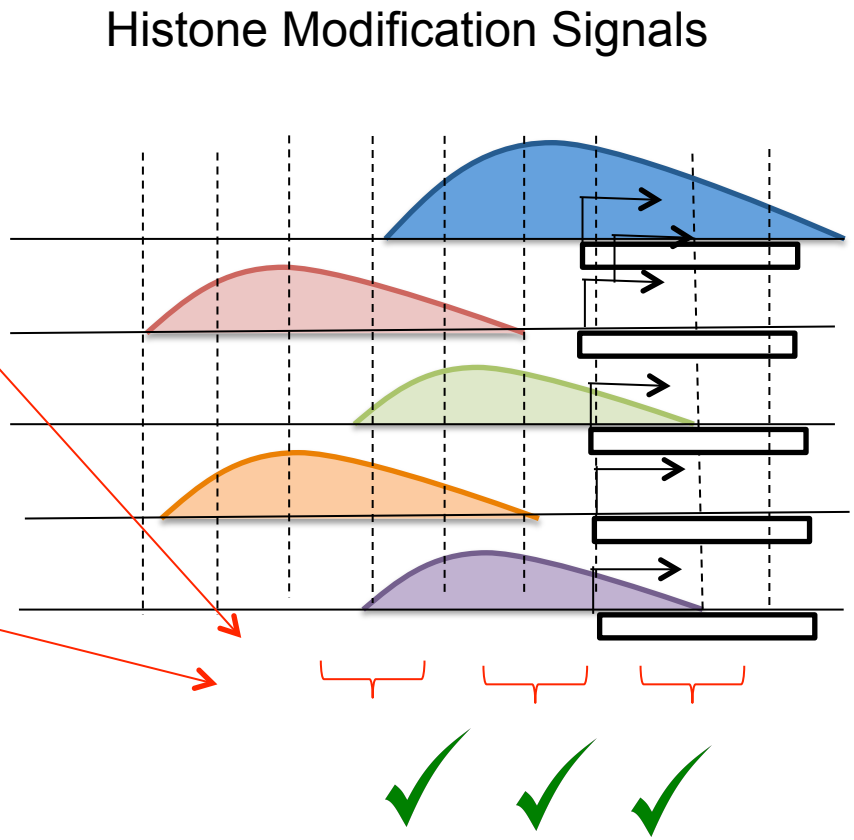
Analogy to our task



Solution

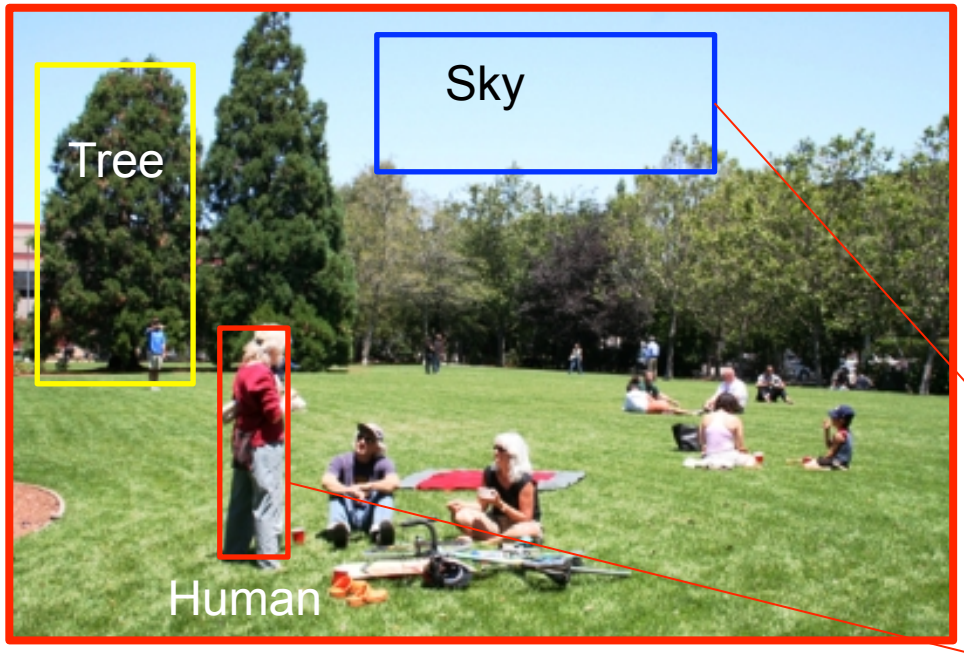
Analogy to our task



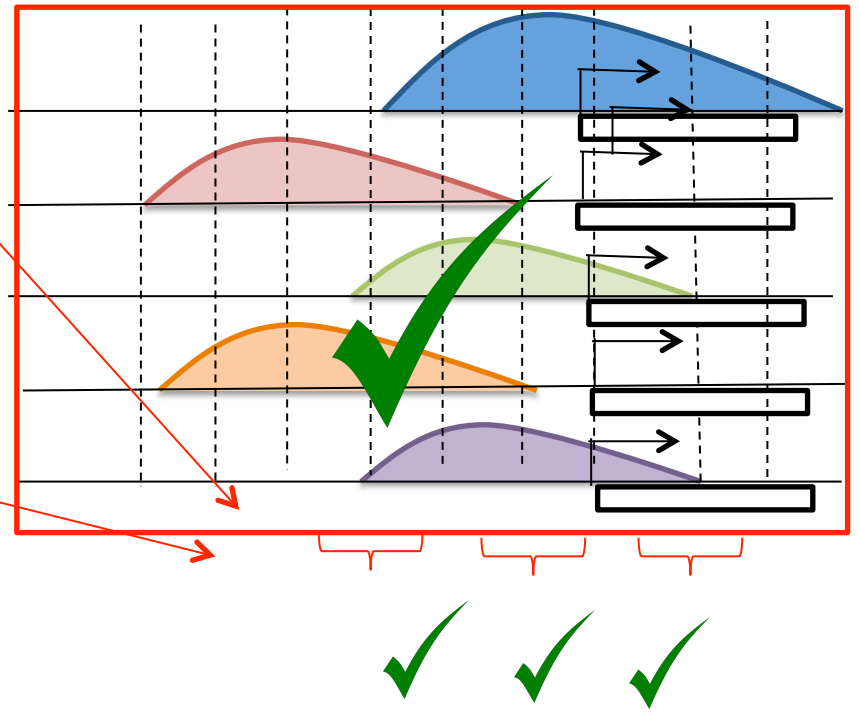


Solution

Analogy to our task

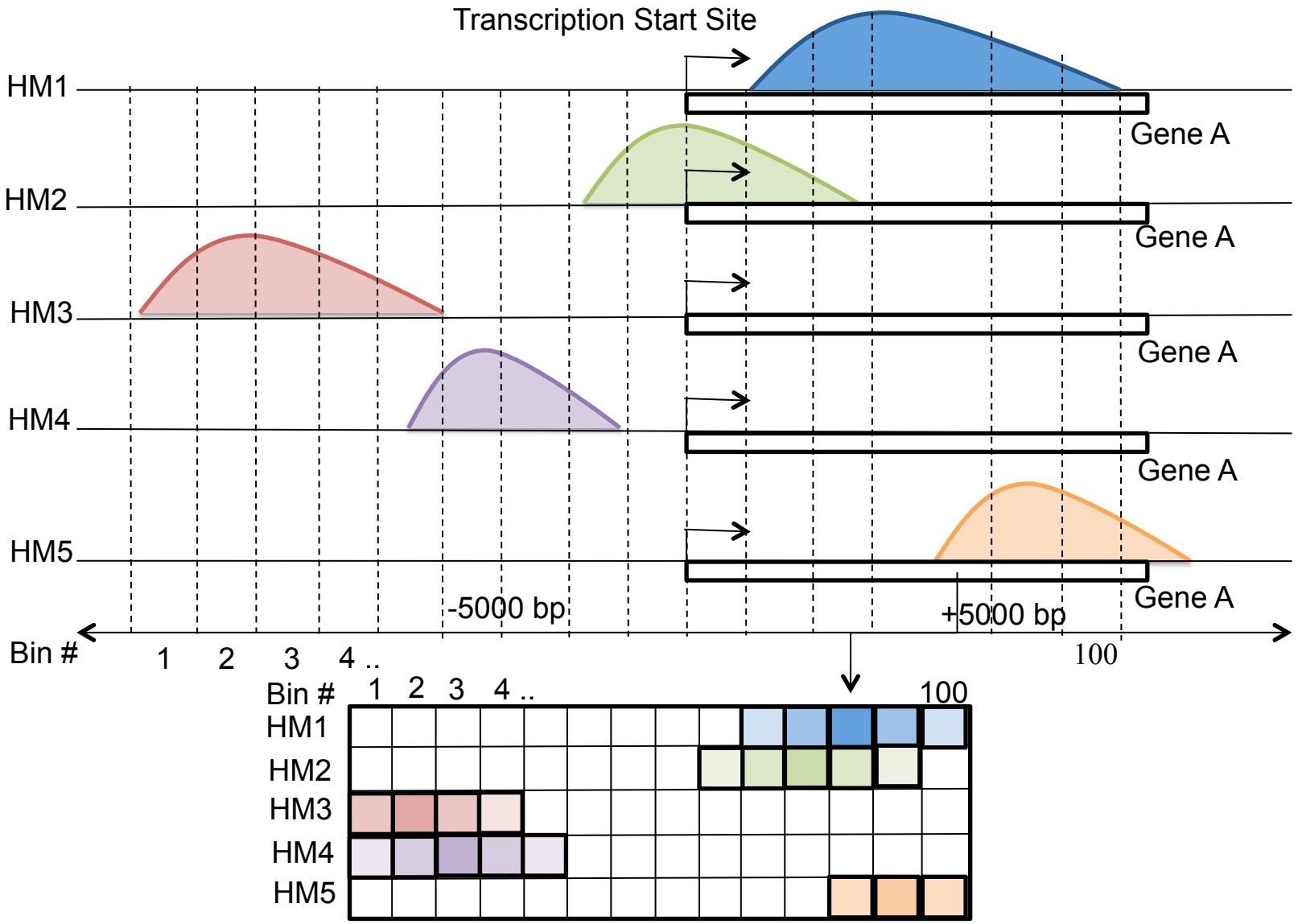


Histone Modification Signals

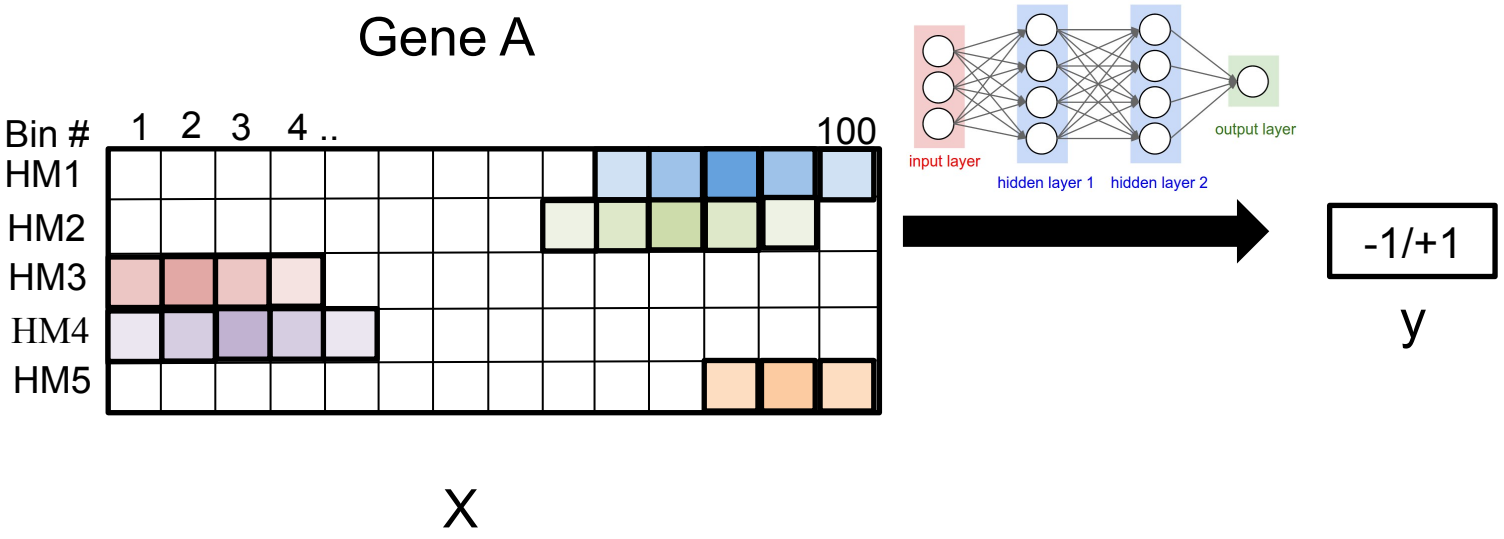


DeepChrome

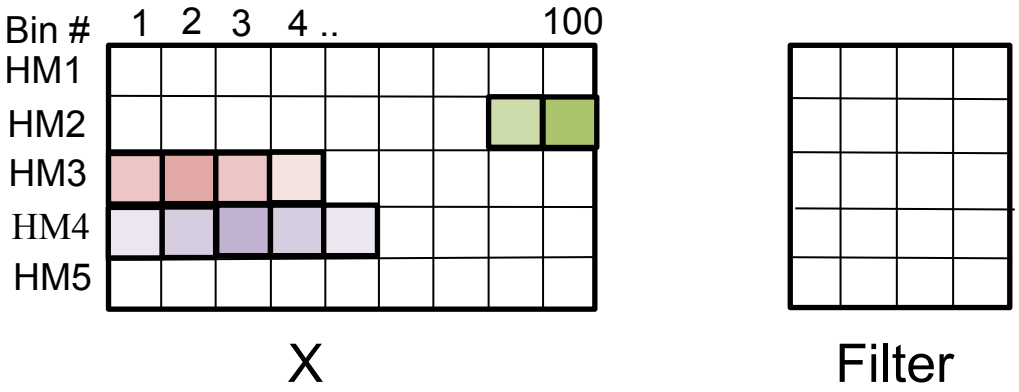
Data


Transcription Start Site
HM1
HM2
HM3
HM4
HM5
Gene A
Gene A
Gene A
Gene A
Gene A
-5000 bp
+5000 bp
Bin #
1
2
3
4 ..
100
Bin #
1
2
3
4 ..
100
HM1
HM2
HM3
HM4
HM5

Overview

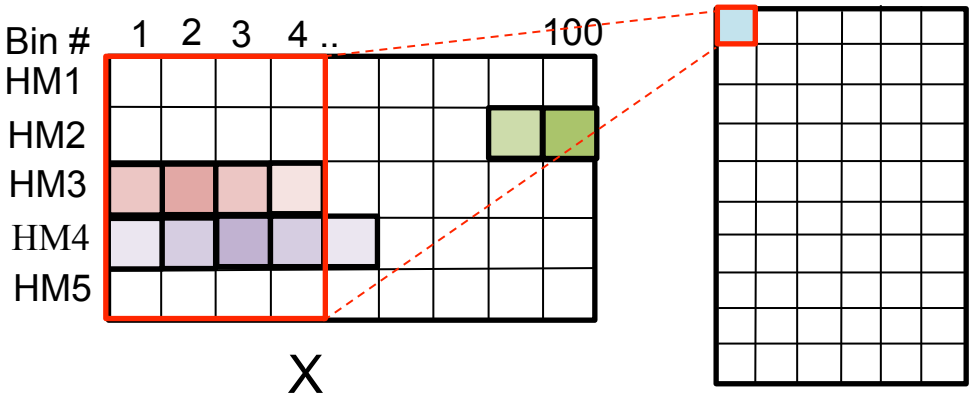

Gene A
Bin #
1
2
3
4 ..
100
HM1
HM2
HM3
HM4
HM5
X
input layer
hidden layer 1
hidden layer 2
output layer
-1/+1
y

CNN Model



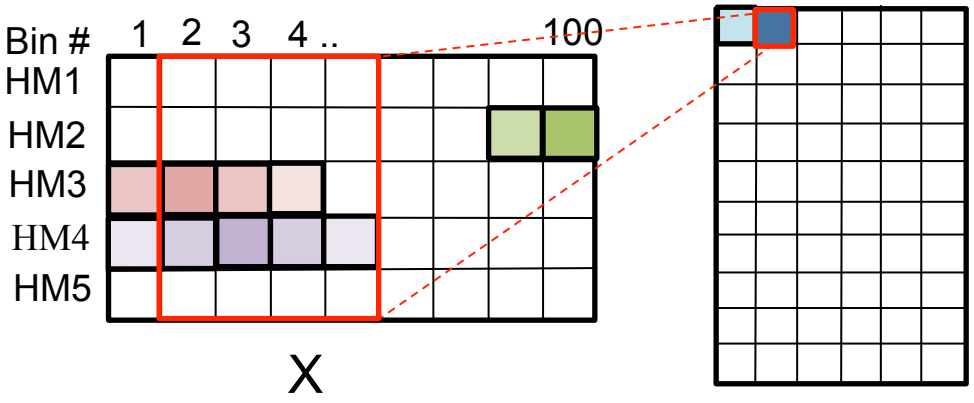
1. Convolution

CNN Model



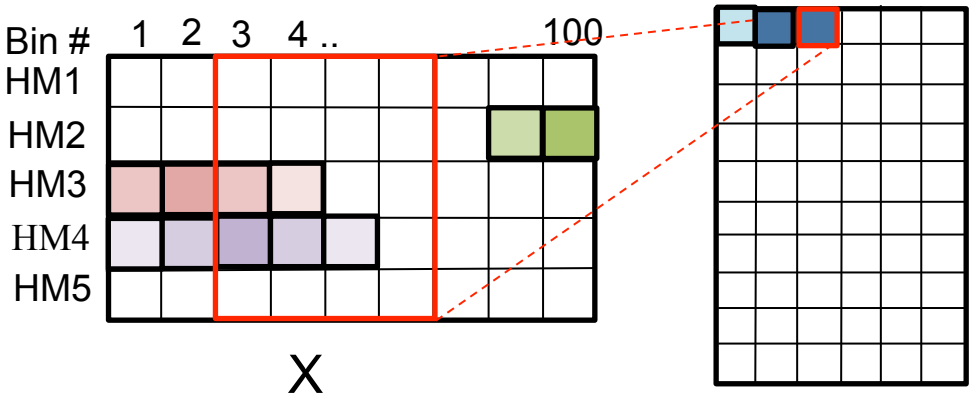
1. Convolution

CNN Model



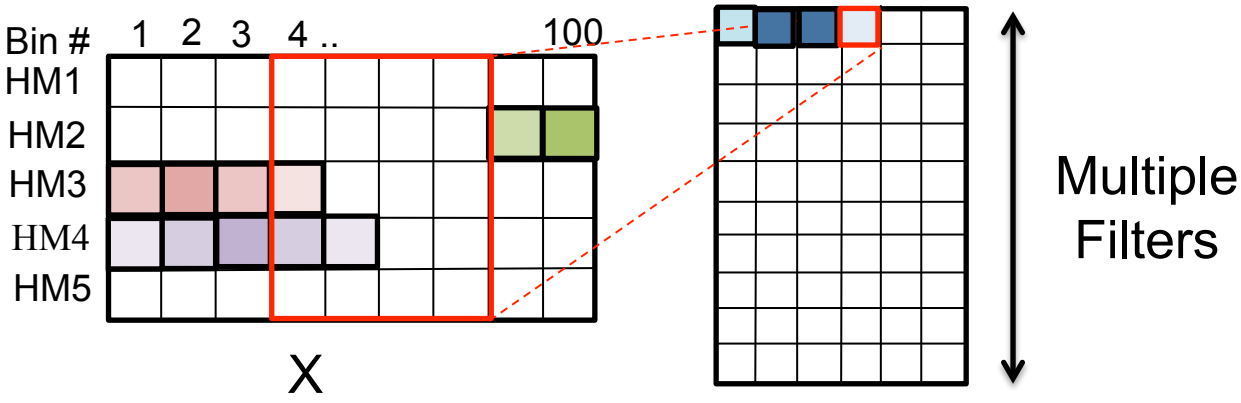
1. Convolution

CNN Model



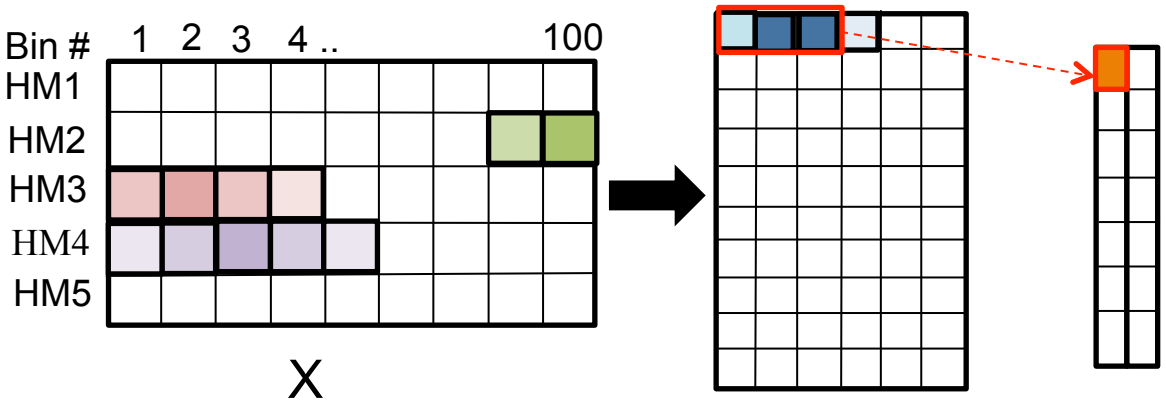
1. Convolution

CNN Model



1. Convolution

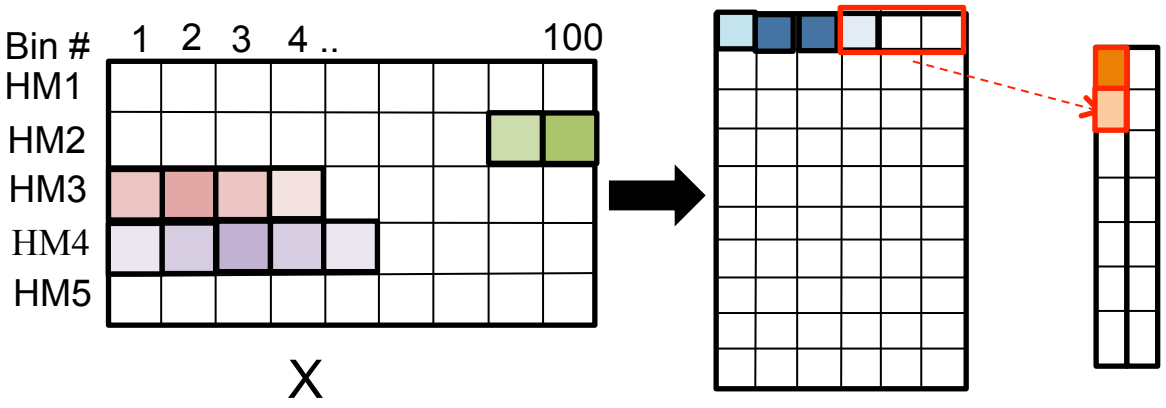
CNN Model



1. Convolution

2. Max Pooling

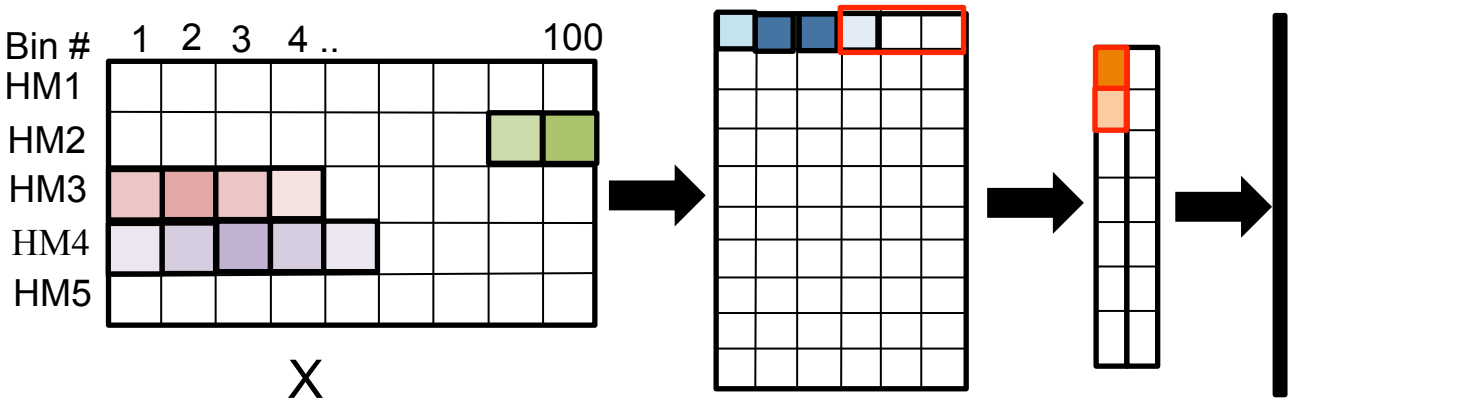
CNN Model



1. Convolution

2. Max Pooling

CNN Model

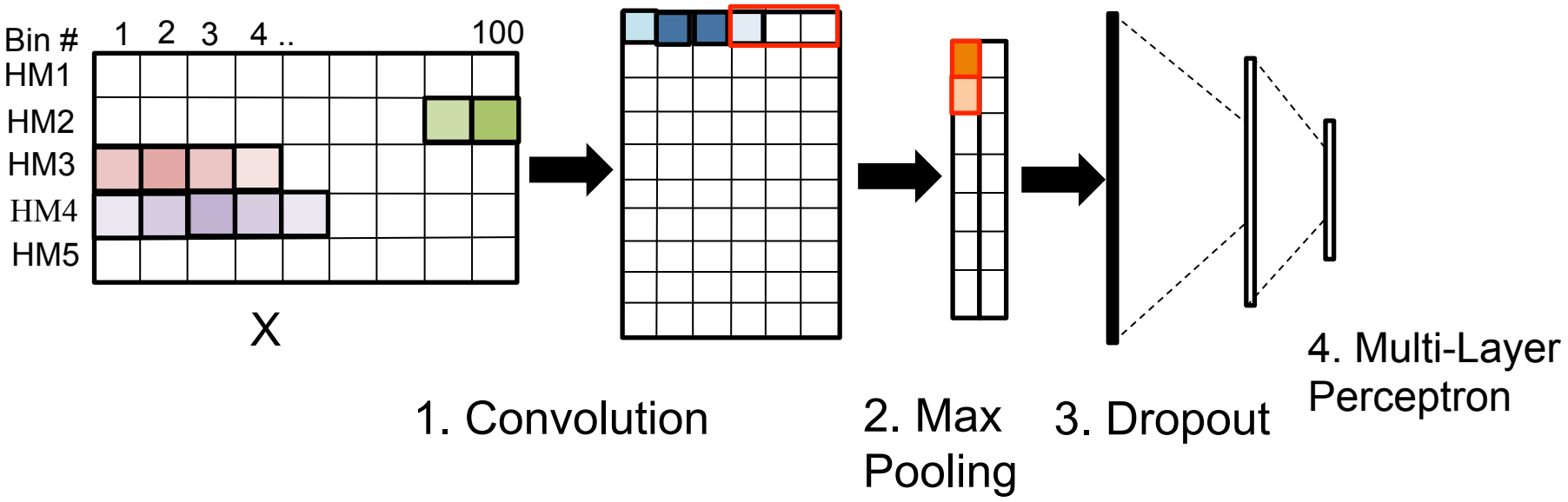


1. Convolution

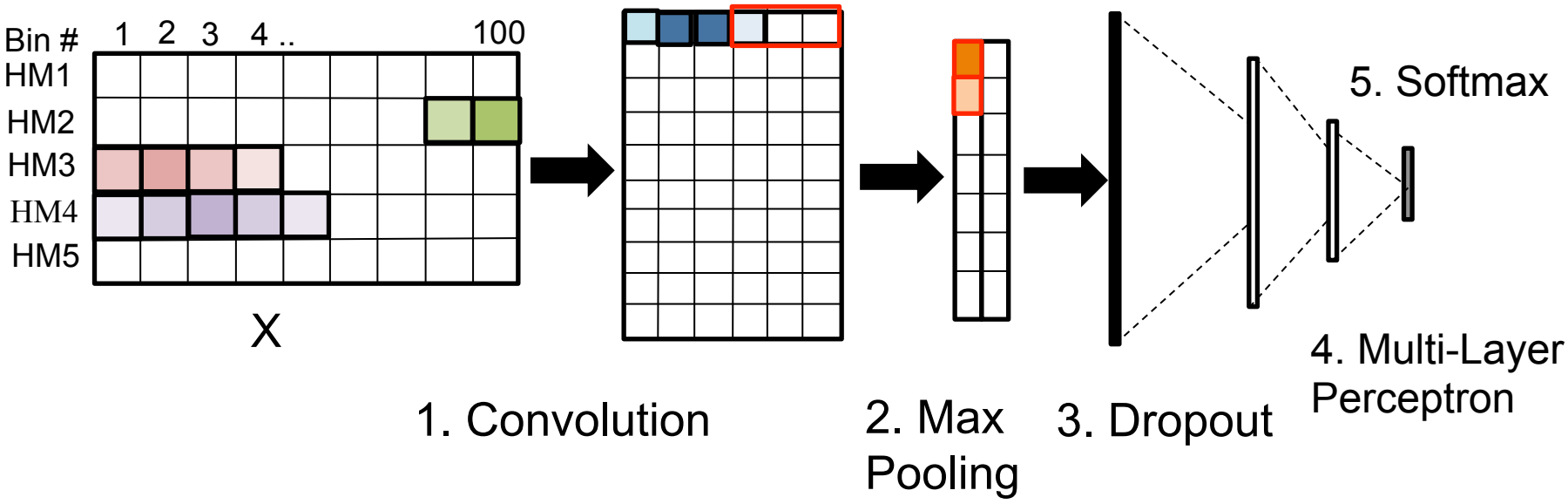
2. Max Pooling

3. Dropout

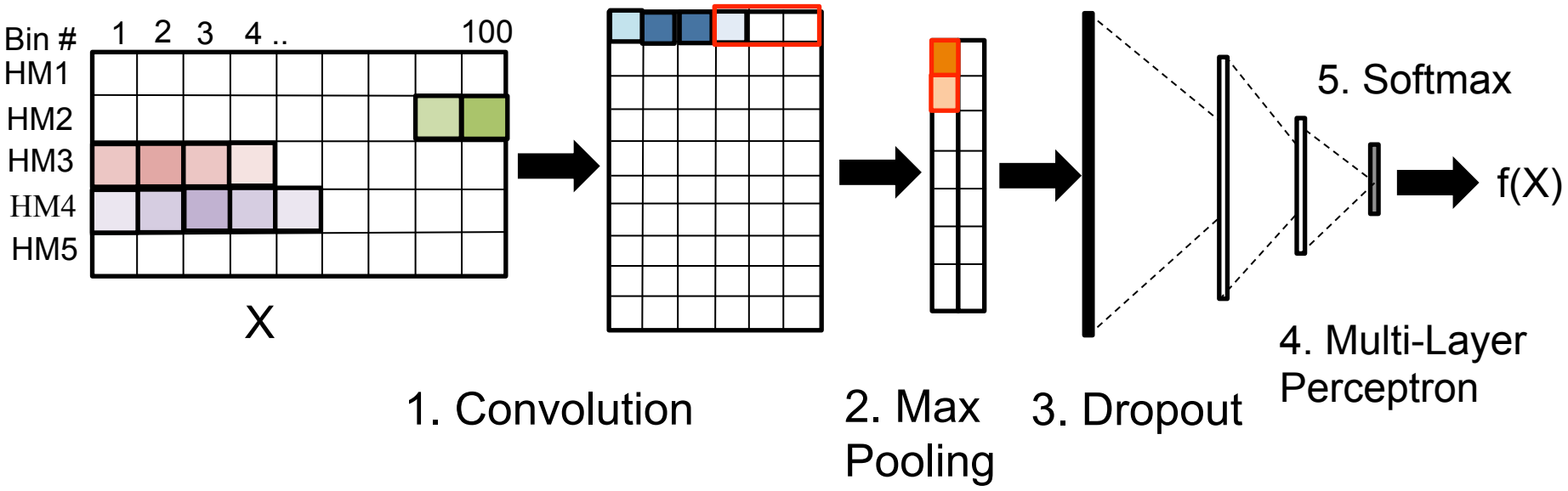
CNN Model


Bin #
1
2
3
4 ..
100
HM1
HM2
HM3
HM4
HM5
X
1. Convolution
2. Max Pooling
3. Dropout
4. Multi-Layer Perceptron

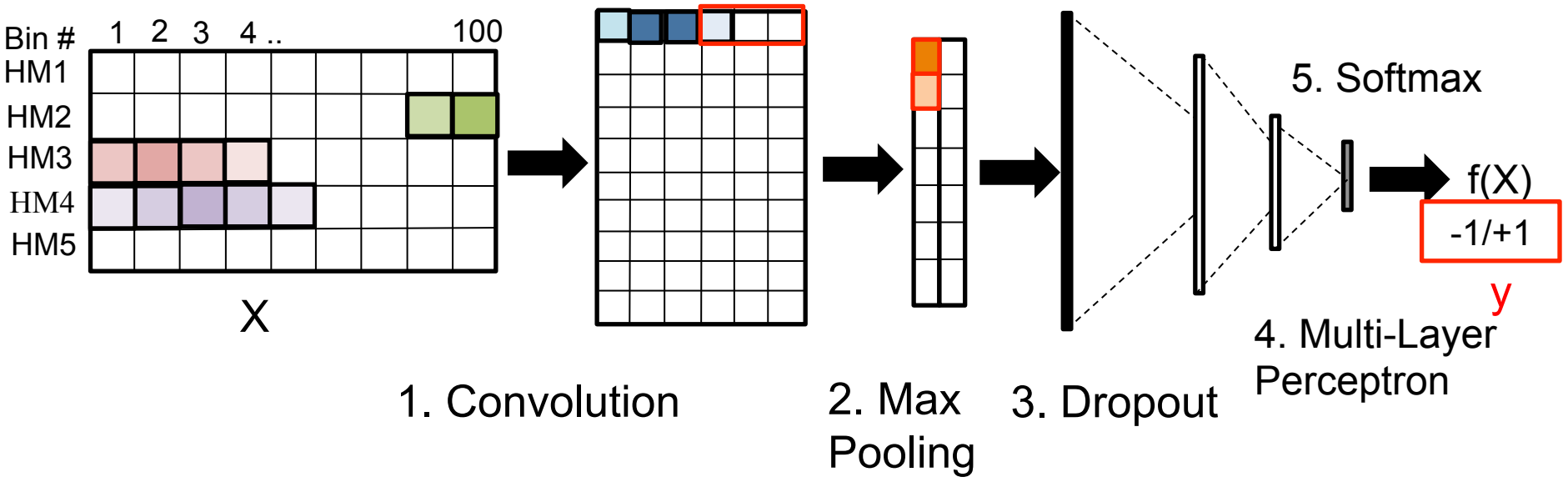
CNN Model


Bin #
1
2
3
4 ..
100
HM1
HM2
HM3
HM4
HM5
X
1. Convolution
2. Max Pooling
3. Dropout
4. Multi-Layer Perceptron
5. Softmax

CNN Model

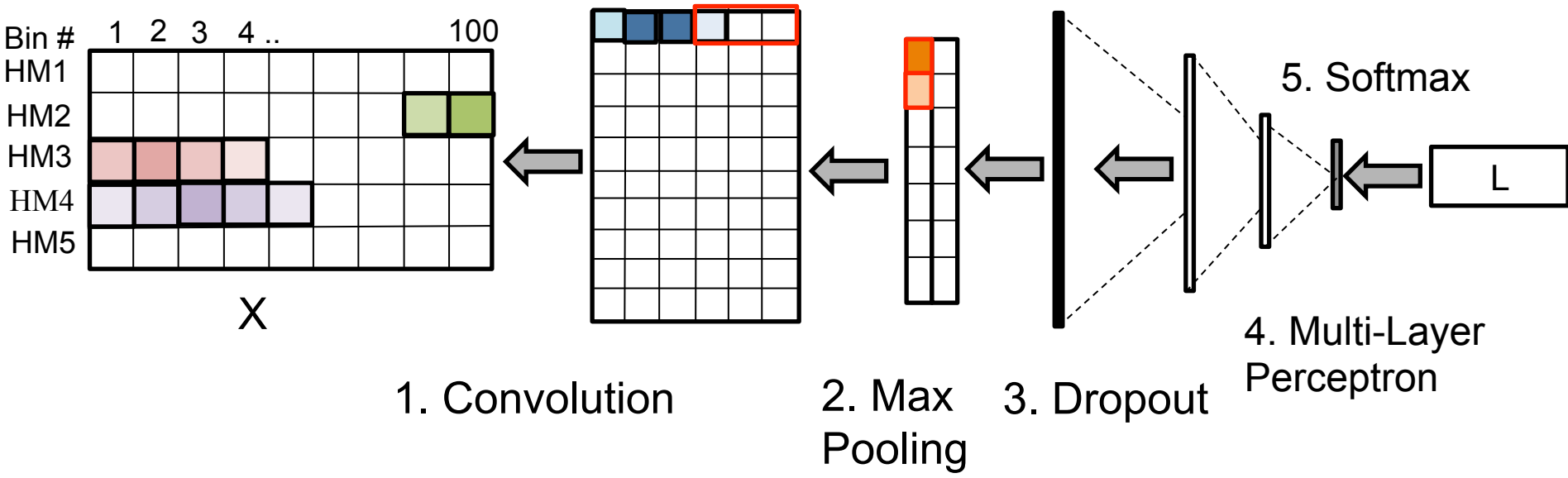

Bin #
1
2
3
4 ..
100
HM1
HM2
HM3
HM4
HM5
X
1. Convolution
2. Max Pooling
3. Dropout
4. Multi-Layer Perceptron
5. Softmax
f(X)

CNN Model



$$L = \sum_{n=1}^{N_{samp}} loss(f(X^{(n)}), y^{(n)})$$

CNN Model



Back-propagation:

$$\Theta \leftarrow \Theta - \eta \frac{\partial L}{\partial \Theta}$$

Experimental Setup

- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps (REMC)
- **Output (Gene Expression):** Discretized RNA-Seq (REMC)

Experimental Setup

- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps (REMC)
- **Output (Gene Expression):** Discretized RNA-Seq (REMC)

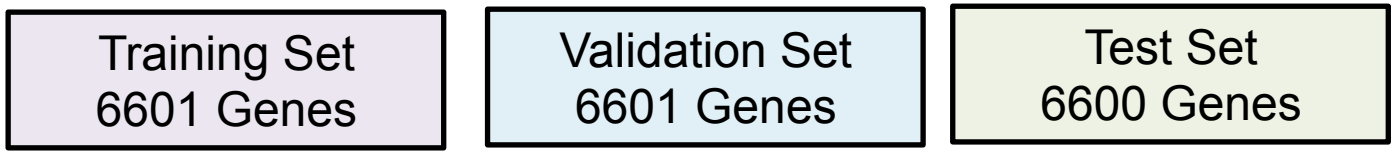
Histone Mark	Functional Category
H3K27me3	Repressor
H3K36me3	Promoter
H3K4me1	Distal Promoter
H3K4me3	Promoter
H3K9me3	Repressor

Experimental Setup

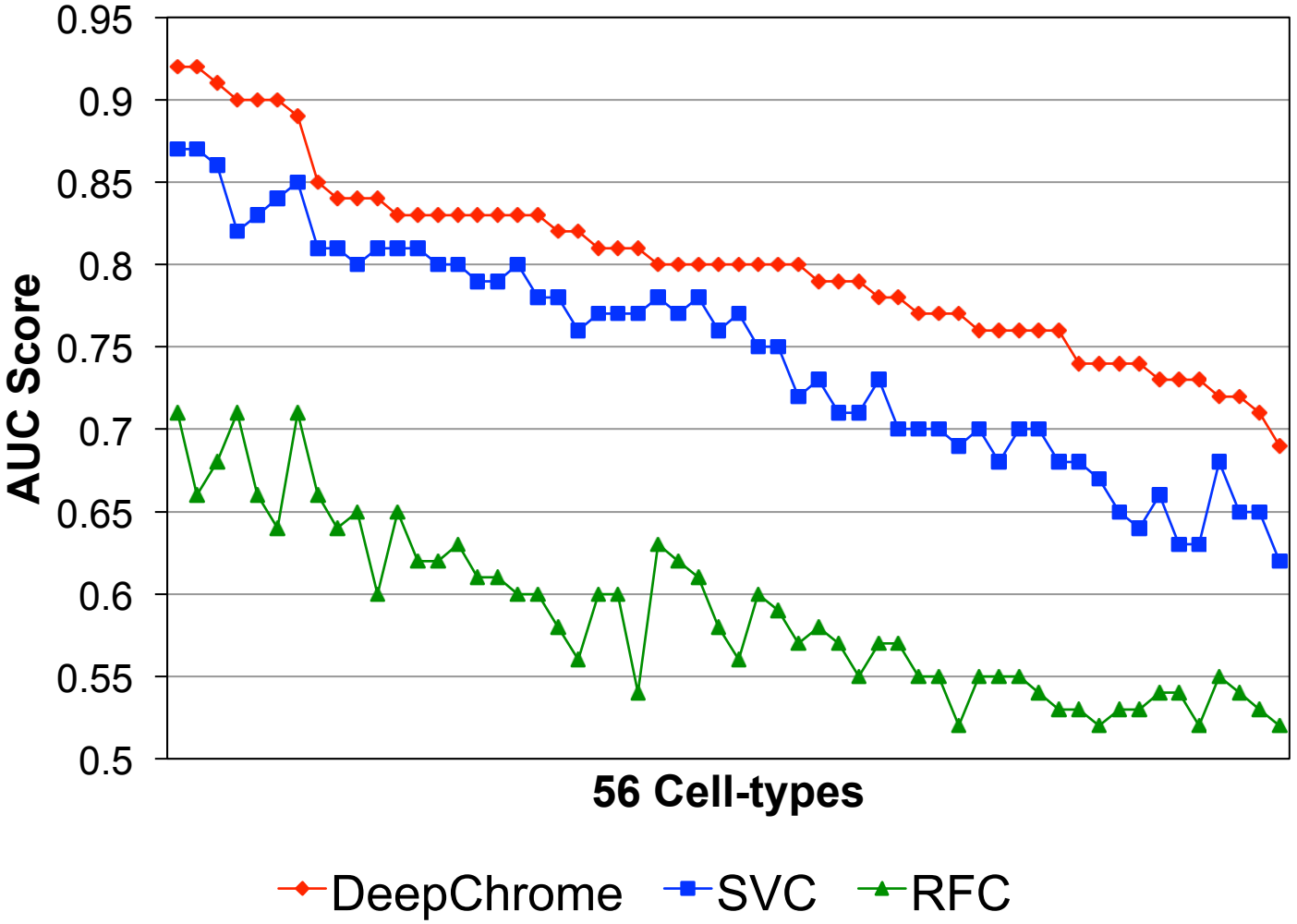
- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps (REMC)
- **Output (Gene Expression):** Discretized RNA-Seq (REMC)

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

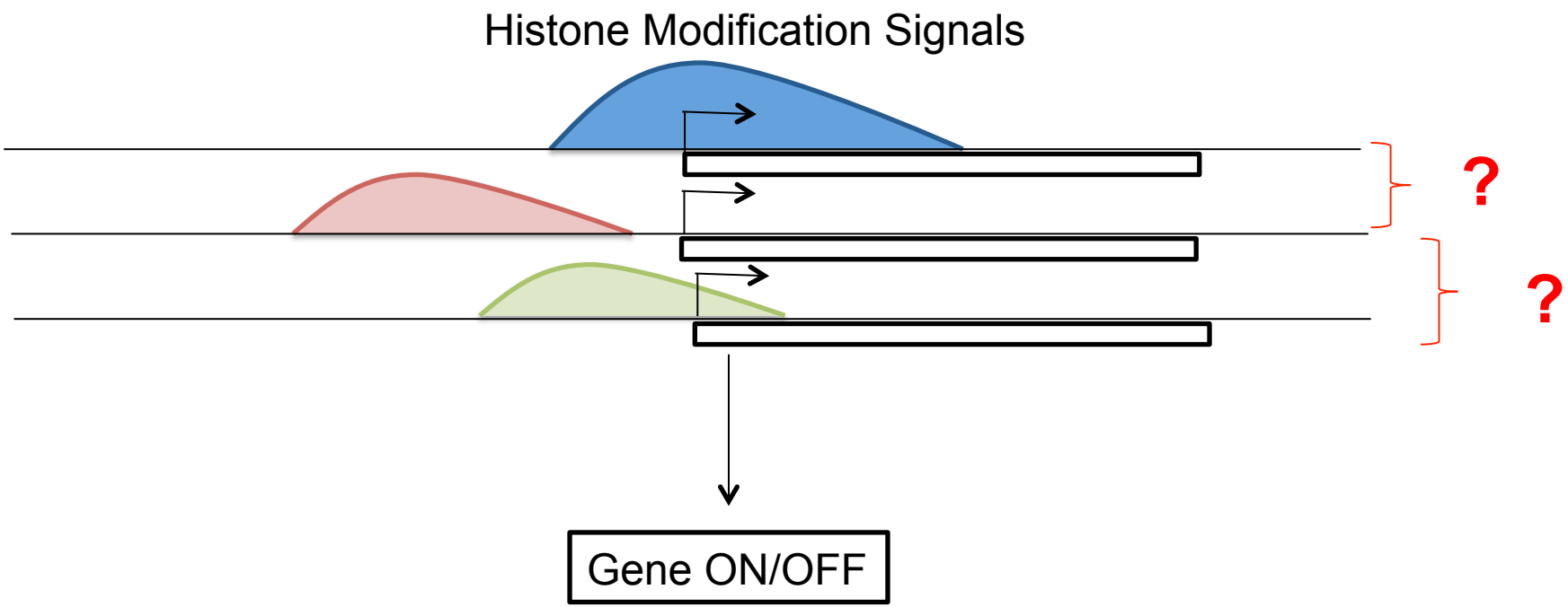
- **Baselines:** Support Vector Classifier (SVC) and Random Forest Classifier (RFC)



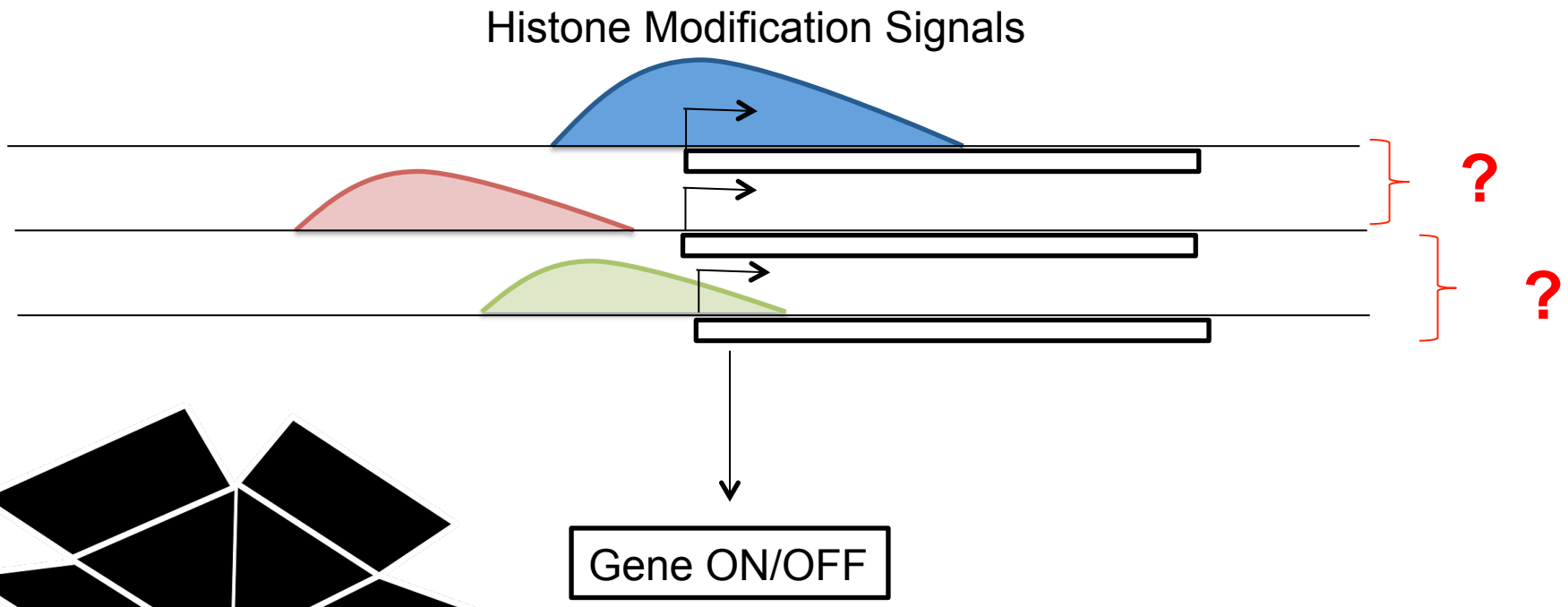
Results: Accuracy

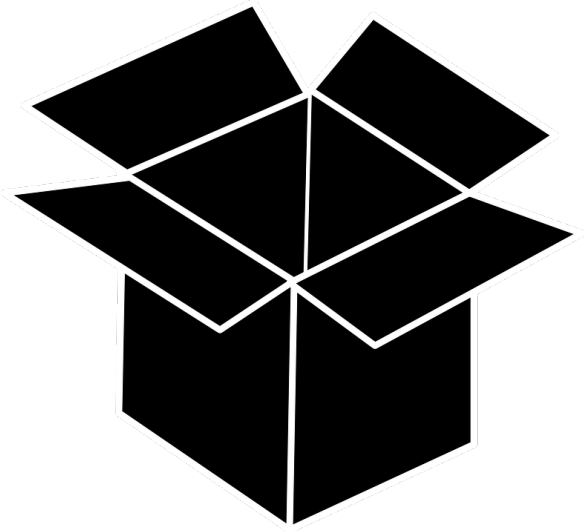


Visualization



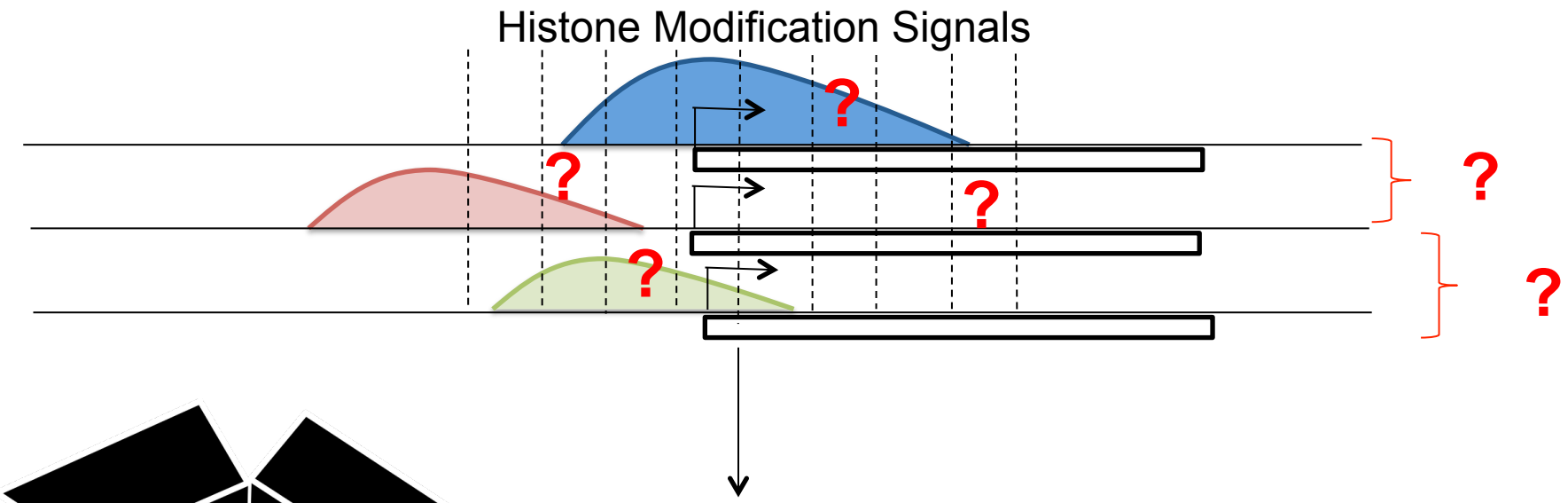
Visualization

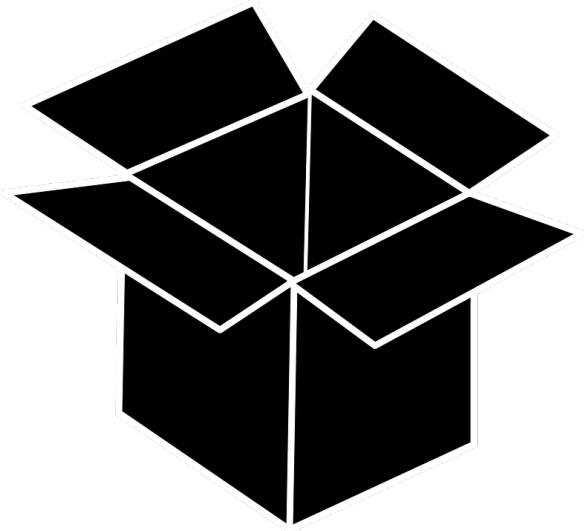




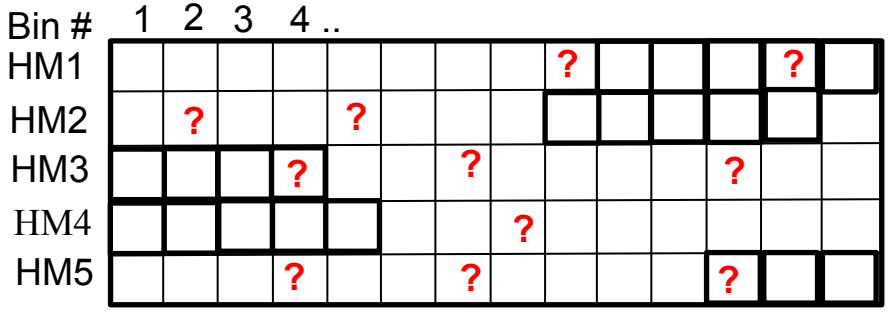
CNN Model

Visualization

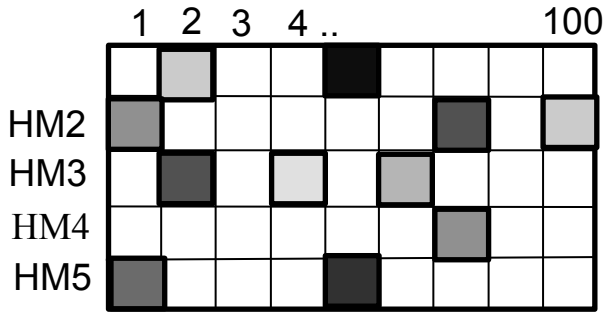




CNN Model

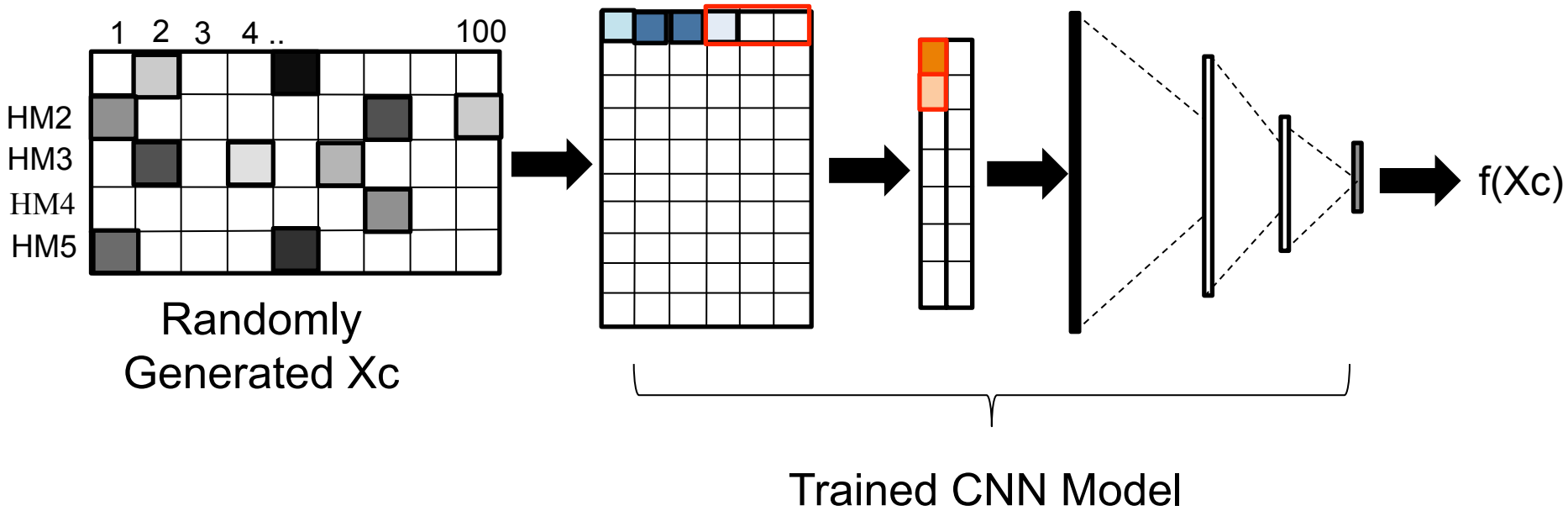


Iterative Most-likely Class Method

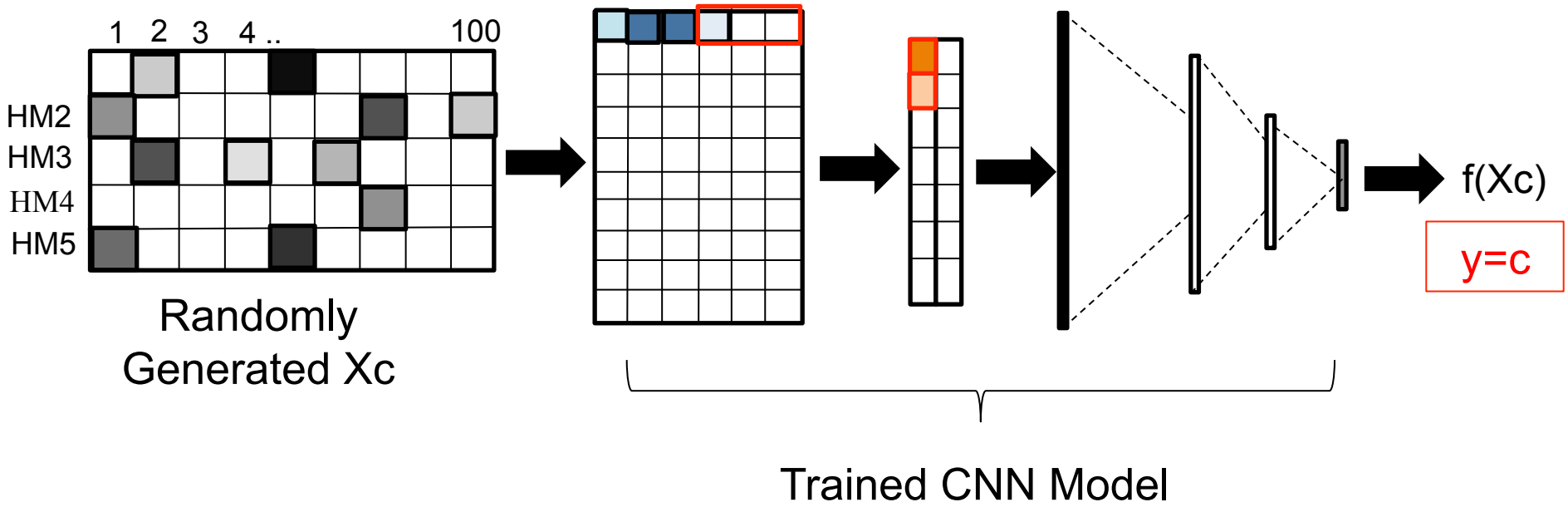


Randomly
Generated Xc

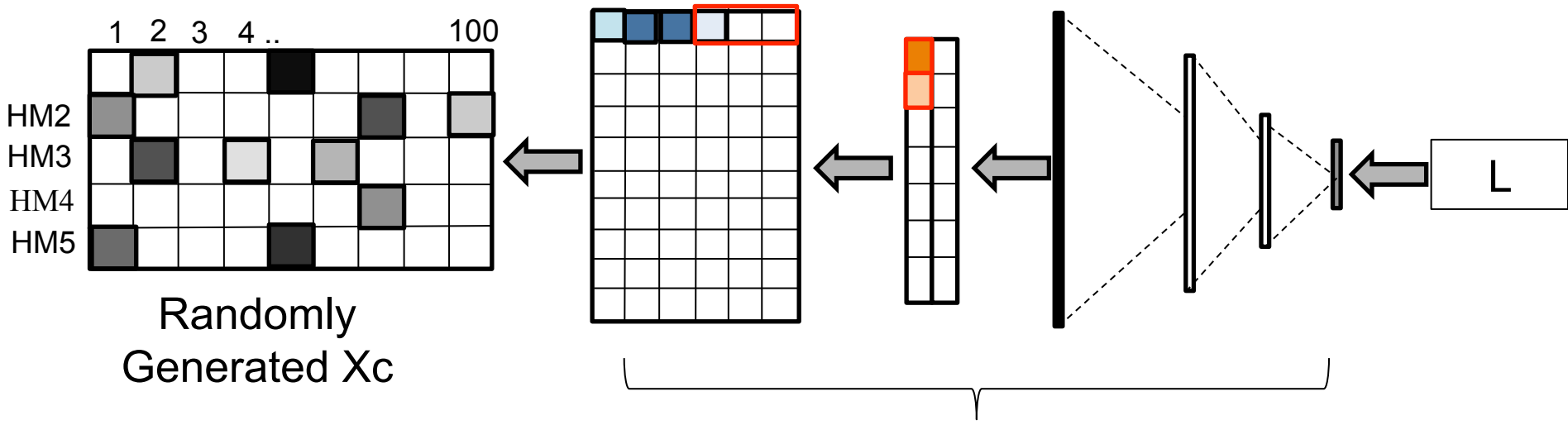
Iterative Most-likely Class Method



Iterative Most-likely Class Method

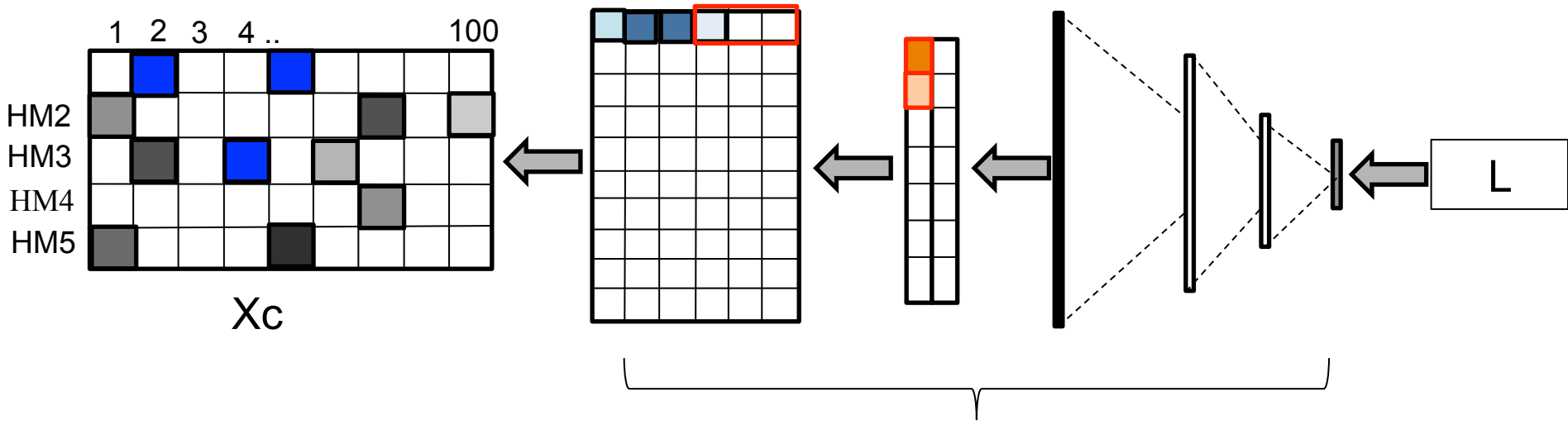


Iterative Most-likely Class Method



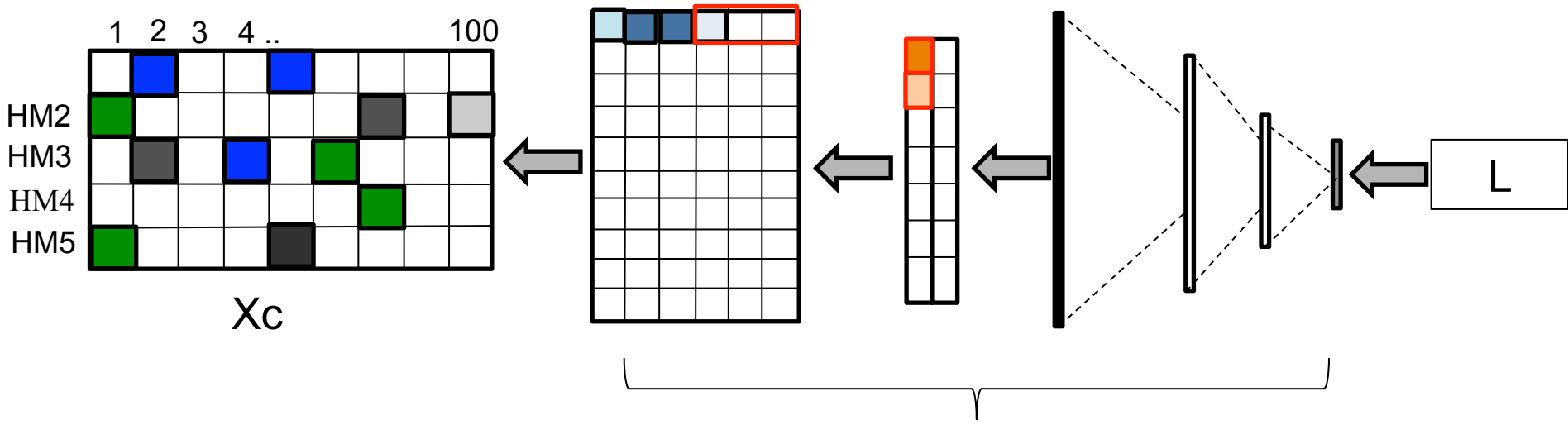
$$\arg\min_{X_c} \{L(f(X_c), y = c)\}$$

Iterative Most-likely Class Method



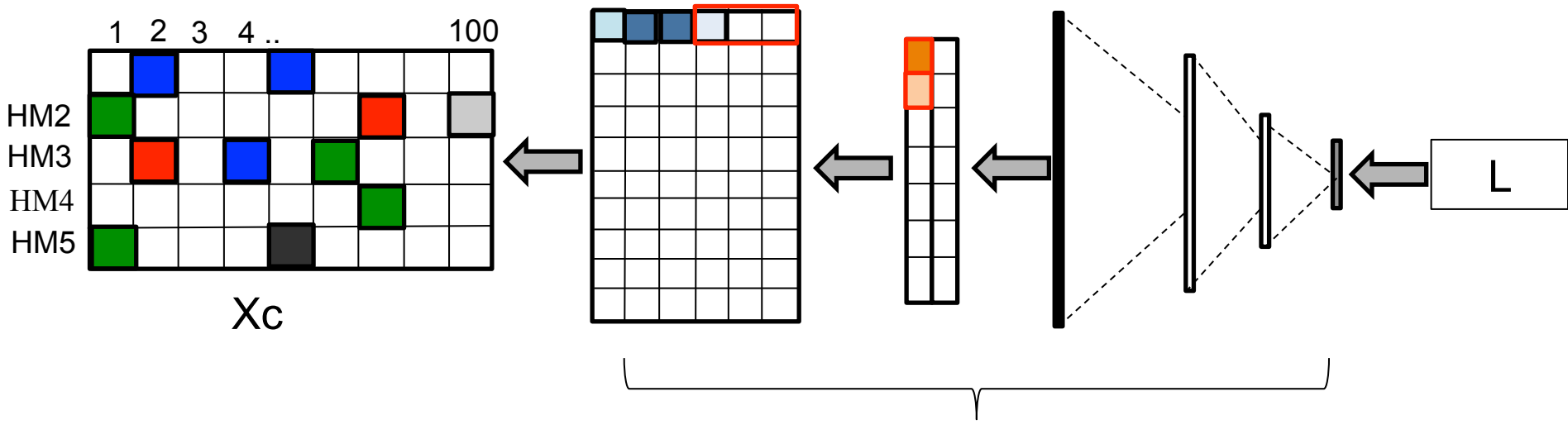
$$\arg\min_{X_c} \{L(f(X_c), y = c)\}$$

Iterative Most-likely Class Method



$$\arg\min_{X_c} \{L(f(X_c), y = c)\}$$

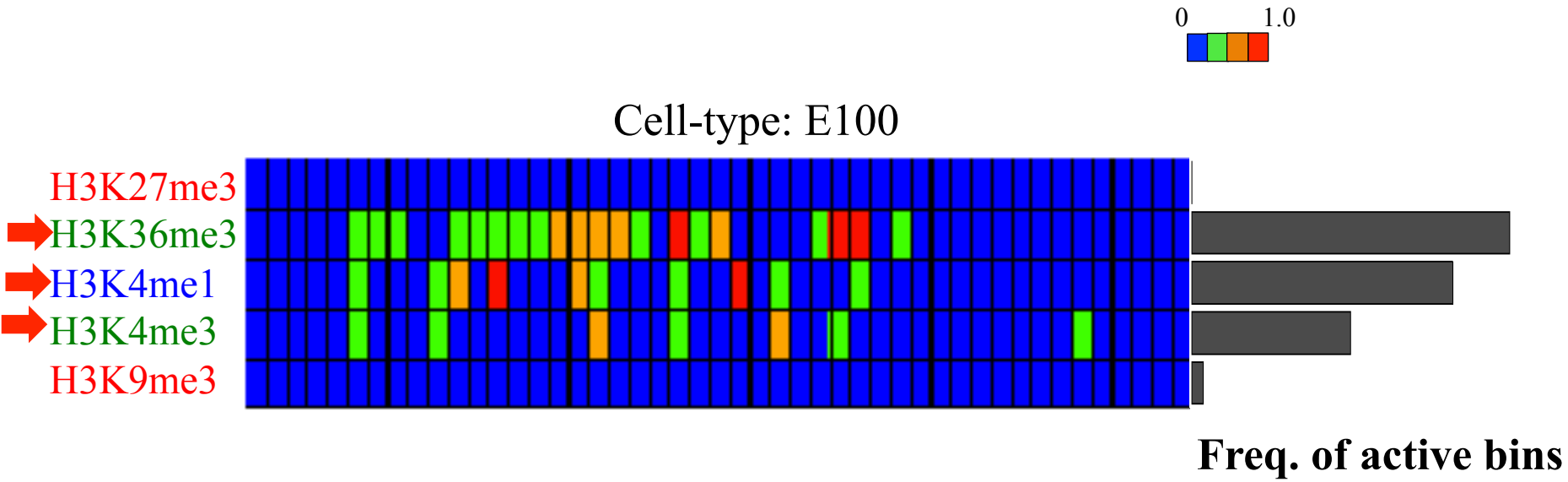
Iterative Most-likely Class Method



$$\arg\min_{X_c} \{L(f(X_c), y = c)\}$$

Results: Visualization

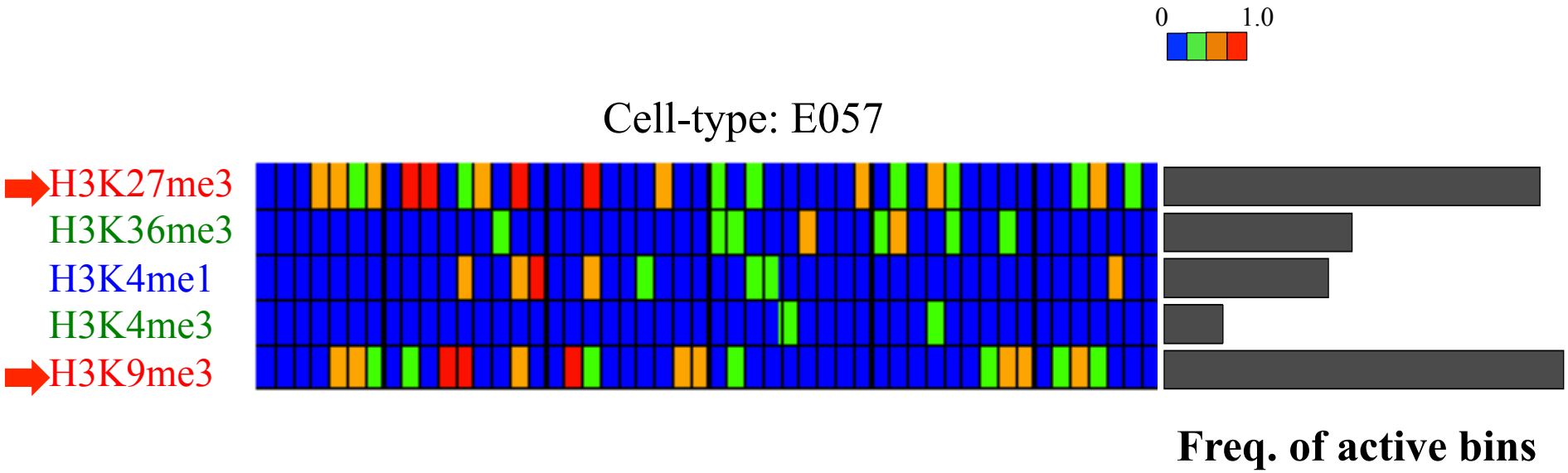
Gene : ON (y=+1)



PROMOTER
DISTAL PROMOTER
REPRESSOR

Results: Visualization

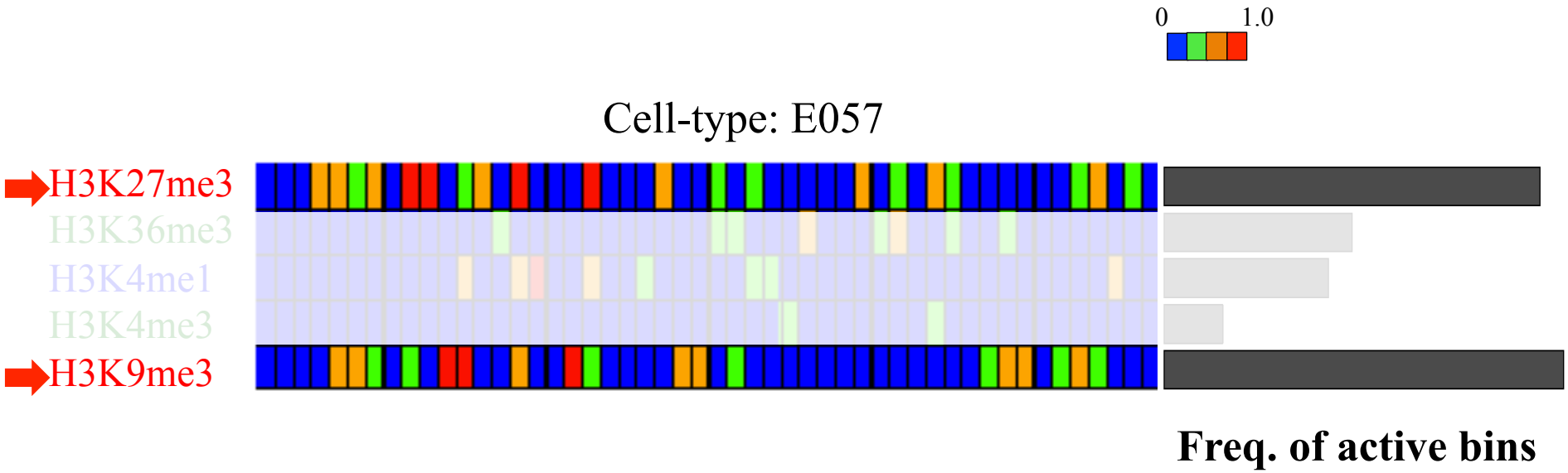
Gene : OFF (y= -1)



PROMOTER
DISTAL PROMOTER
REPRESSOR

Results: Visualization

Gene : OFF (y= -1)



PROMOTER
DISTAL PROMOTER
REPRESSOR

Conclusion

1. First deep learning implementation for gene expression prediction

2. Unified Framework

   a. Outperforms state-of-the-art implementations

   b. Visualization of high-order combinatorial relationships

Available @ www.deepchrome.org

Acknowledgements

Computer Science at the University of Virginia

Machine Learning and Bioinformatics Lab

Dr. Yanjun Qi

Jack Lanchantin

Dr. Gabriel Robins

Marina Sanusi

Beilun Wang

Weilin Xu

Ji Gao

Kamran Kowsari

Department of Biochemistry and Molecular Genetics: Dr. Mazhar Adli

Travel Fund

ECCB 2016

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY


Thank you