Published at NIPS 2017

Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin

Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi
Department of Computer Science
University of Virginia

Presented by
Johra Muhammad Moosa
Bioinformatics Lab

What is Gene Regulation?

Process of controlling **Gene Expression**

What is Gene Expression?

Abundance of a gene or a protein in a cell

DNA to Protein

- DNA: coded message
- Gene: part (from TSS to TTS) of the message to be decoded
- Ribosome decoder
- Protein decoded output

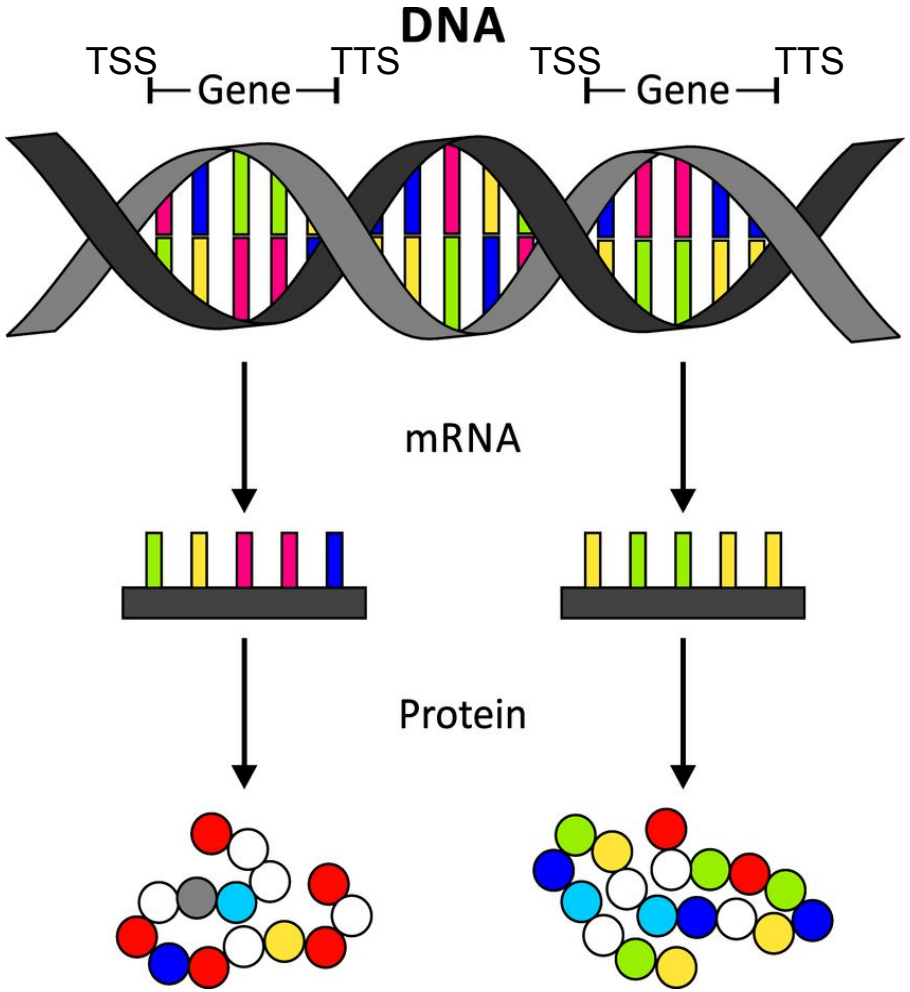


Image source:
https://www.researchgate.net/publication/318420329_Health_technology_assessment_of_next-generation_sequencing/figures?lo=1

Gene Expression

- Each cell has different functionality
  - needs different proteins
  - gene expression is different for each cell
- Detect abnormal cells

Histone & Chromatin

1. **Histones** lock and compact DNA around it
2. Form a structure called **Chromatin**
3. To protect DNA during cell division and replication

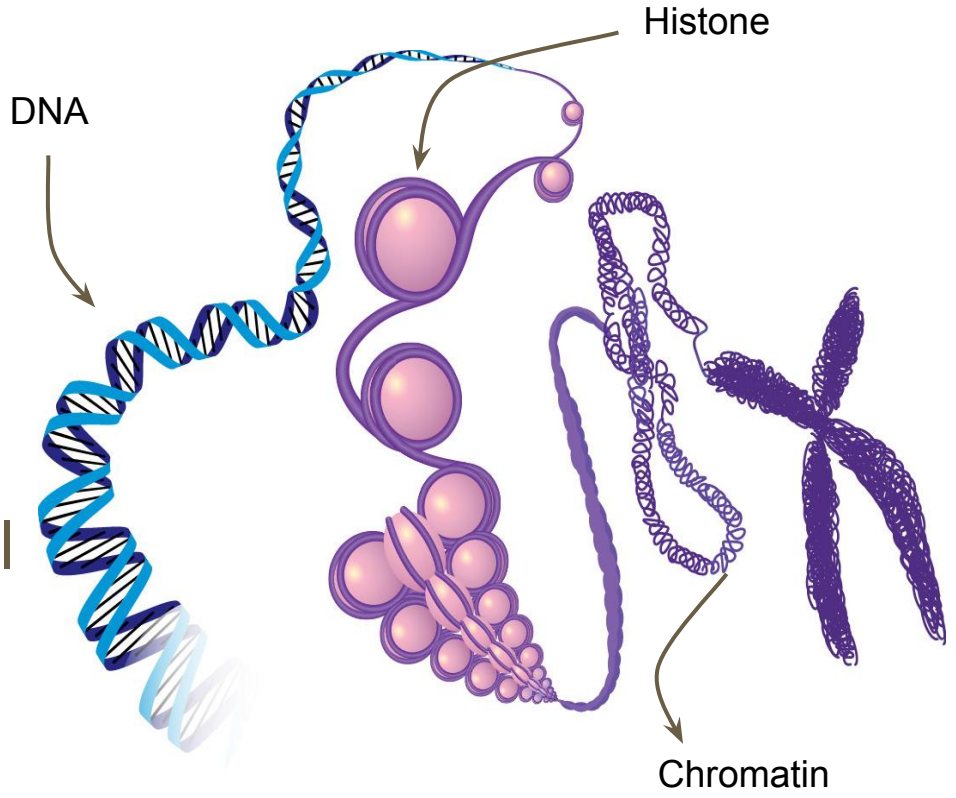
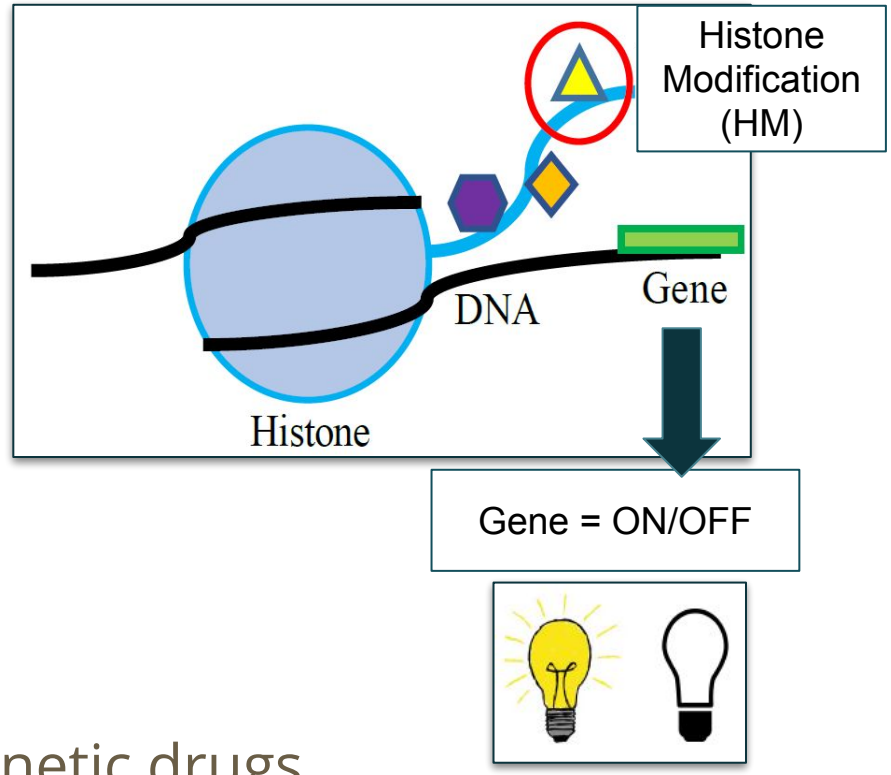


Image source: http://sites.tufts.edu/insight/2016/05/16/adenine-methylation-in-mammals-n6-ma-is-the-new-5mc/

Histone Code Hypothesis

1. Histone Modification Marks
    a. chemically modified
2. Genes: accessible or restricted
3. Neighboring region
    a. Gene expression affected

4. HM: can be reversed
    a. unlike genetic mutation
    b. this can help develop epigenetic drugs



Objective

- Biologists now can measure gene expressions and HM mark signals easily

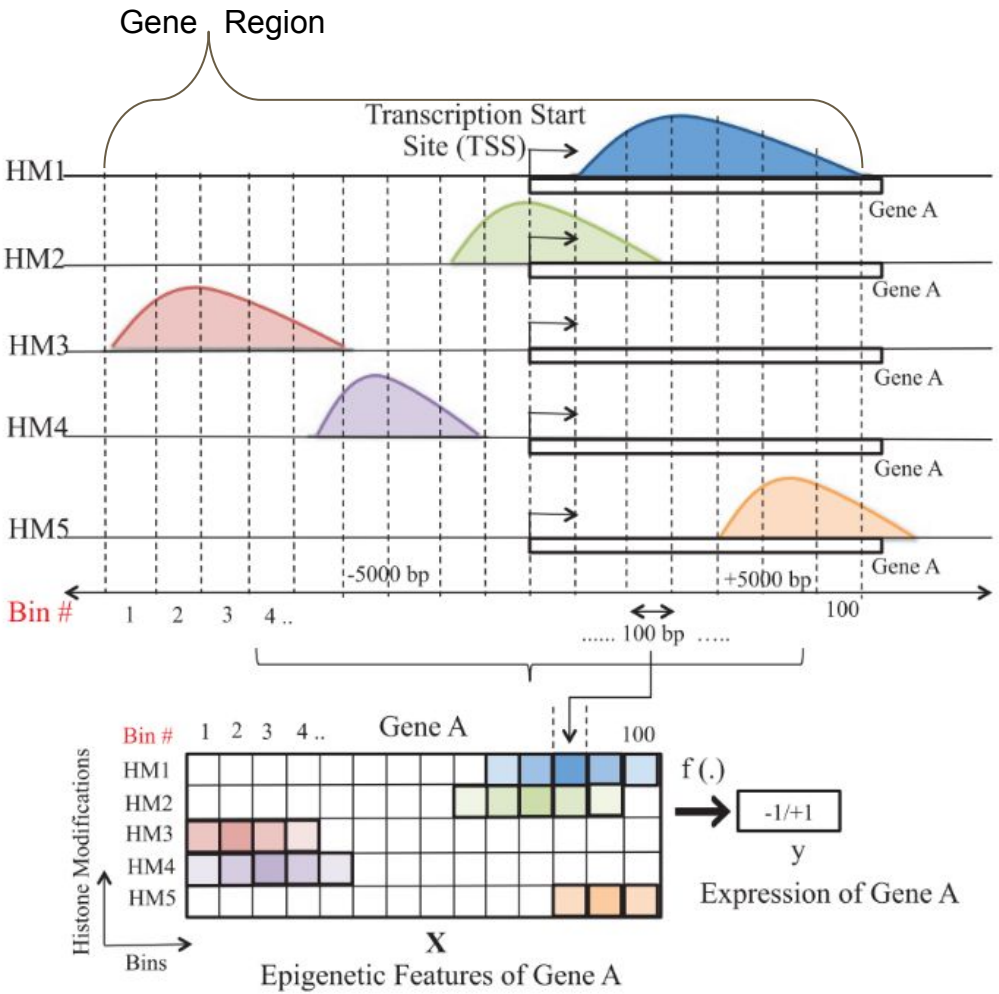
- For each cell type, we want to find the important HM marks and how they interact to control gene expression

*How **HM** MARKS control GENE REGULATION in a cell*

Histone Modification Signal

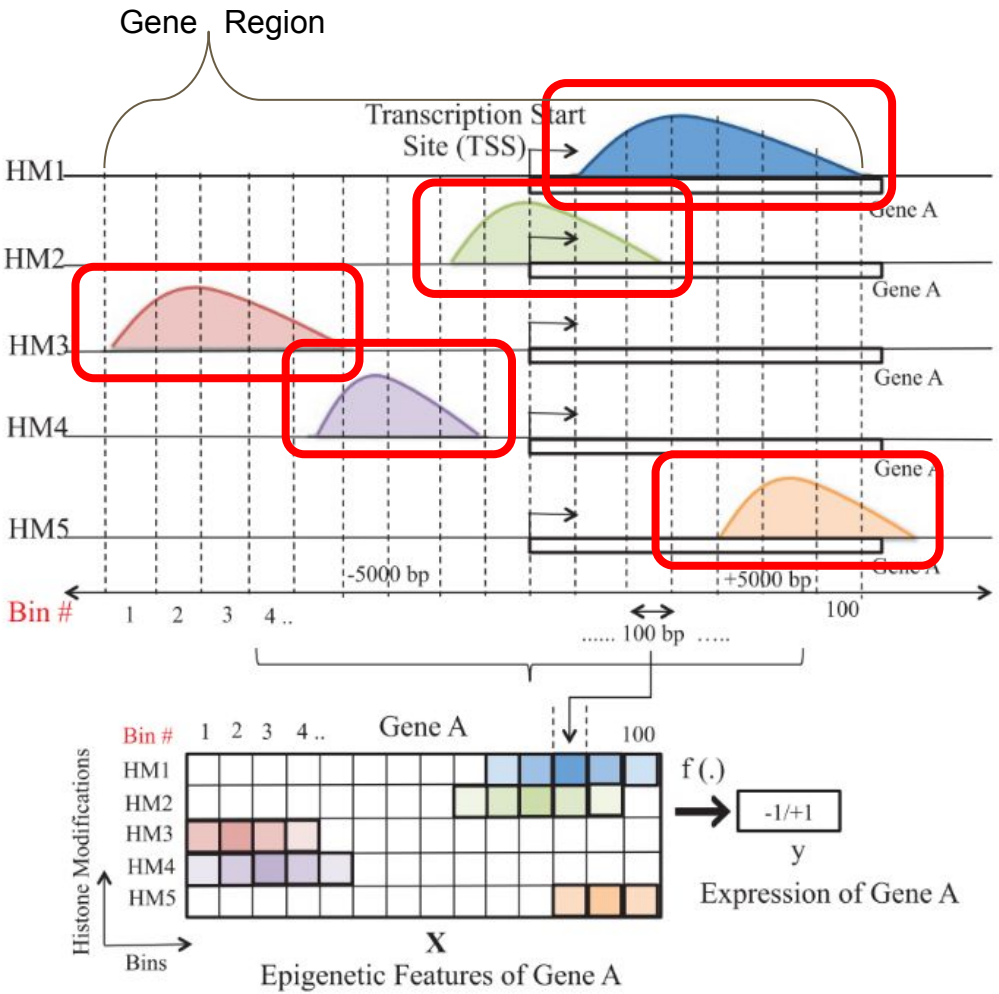
- Might appear anywhere in the gene region

- Gene region
  - 10k base pairs (bp)
  - Centered at TSS



Histone Modification Signal

- HM signals can span across **multiple neighboring bins**



Data Modelling/ Factors

- Large number of histone modification marks
- The five standard histones can be modified simultaneously
- Different sites/ locations
- Different chemical modifications

Learn dependencies among different marks

Input data

REMC database: 56 different cell types

- Histone modification signals covering the **gene region**

Feature vector for each mark: signals surrounding a gene's TSS position

- Flanking region: ± 5k bp from the TSS, total 10k bp
- Divide into 100 bins, with bins of 100 bp length each

Five (5) core histone modifications

- Uniformly profiled across all cell types

Core Histone Marks

1. Five core histone marks
2. Uniformly profiled in the REMC dataset

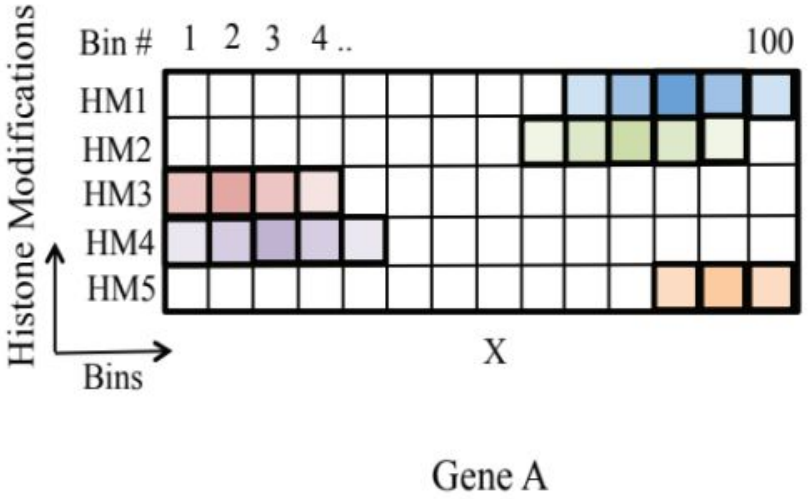
Histone Mark	Associated with	Functional category	Tag
H3K4me3	Promoter regions	Promoter mark	H_{prom}
H3k4me1	Enhancer regions	Distal mark	H_{enhc}
H3K36me3	Transcribed regions	Structural mark	H_{struct}
H3K9me3	Heterochromatin regions	Repressor mark	H_{reprA}
H2K27me3	Polycomb repression	Repressor mark	H_{reprB}

Express

Supress

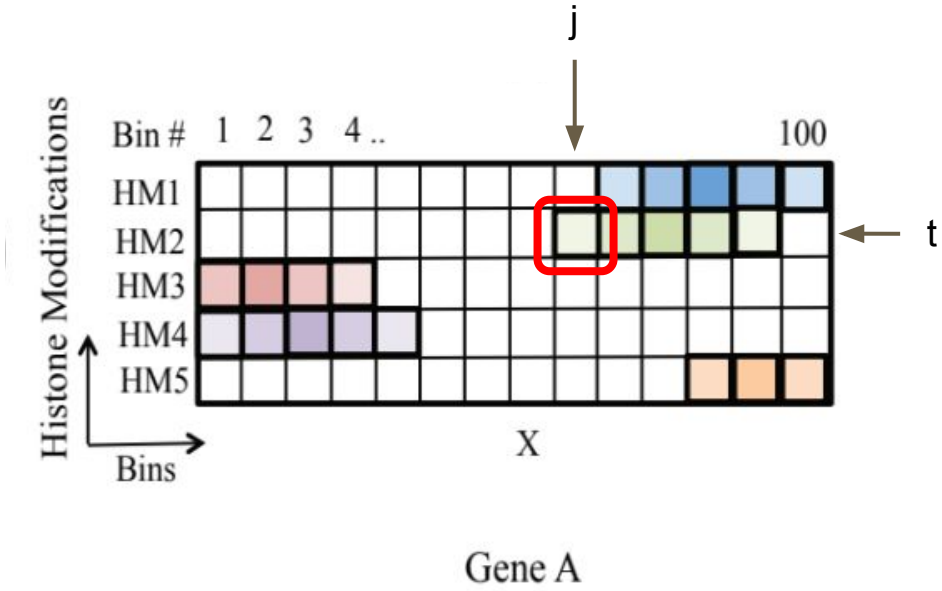
Input Format

- Columns: bins, T = 100 bins
- Rows: HMs, M = 5
- For each gene,
  - Input, **X**: 5 × 100
  - Label, y, binary:
    - Discretized
    - High: +1
    - Low: -1



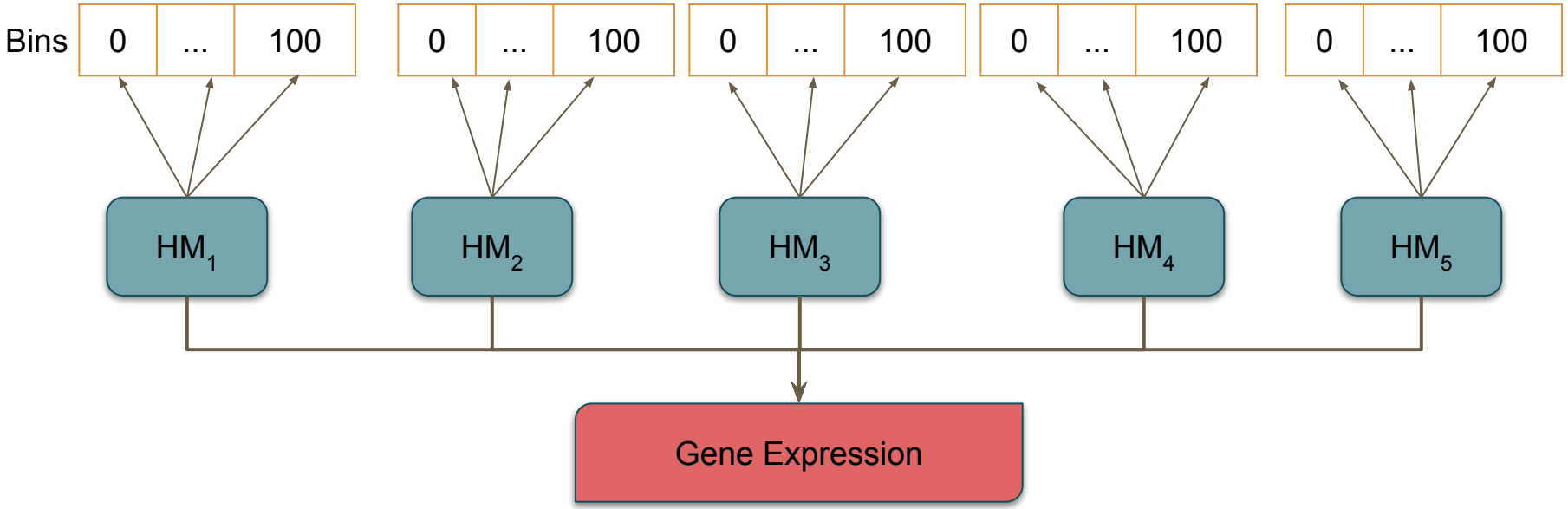
Input Format

- N: number of genes
- For the n^{th} gene, the sample is (X^n, y^n)



x_t^j t^{th} bin, from j^{th} HM

Dependencies



Model Formulation

- RNN: to capture the spatial dependencies

- **One** LSTM for the five HMs
  - Combinatorial features: To model how they interact
- **One** LSTM for **each** HMs
  - Local features: To model the spatial dependencies among the bins

Model Formulation

- RNN: to capture the spatial dependencies

- Two levels of soft attention
  1. Attends the important regions of a HM mark
  2. Attends the important HM marks

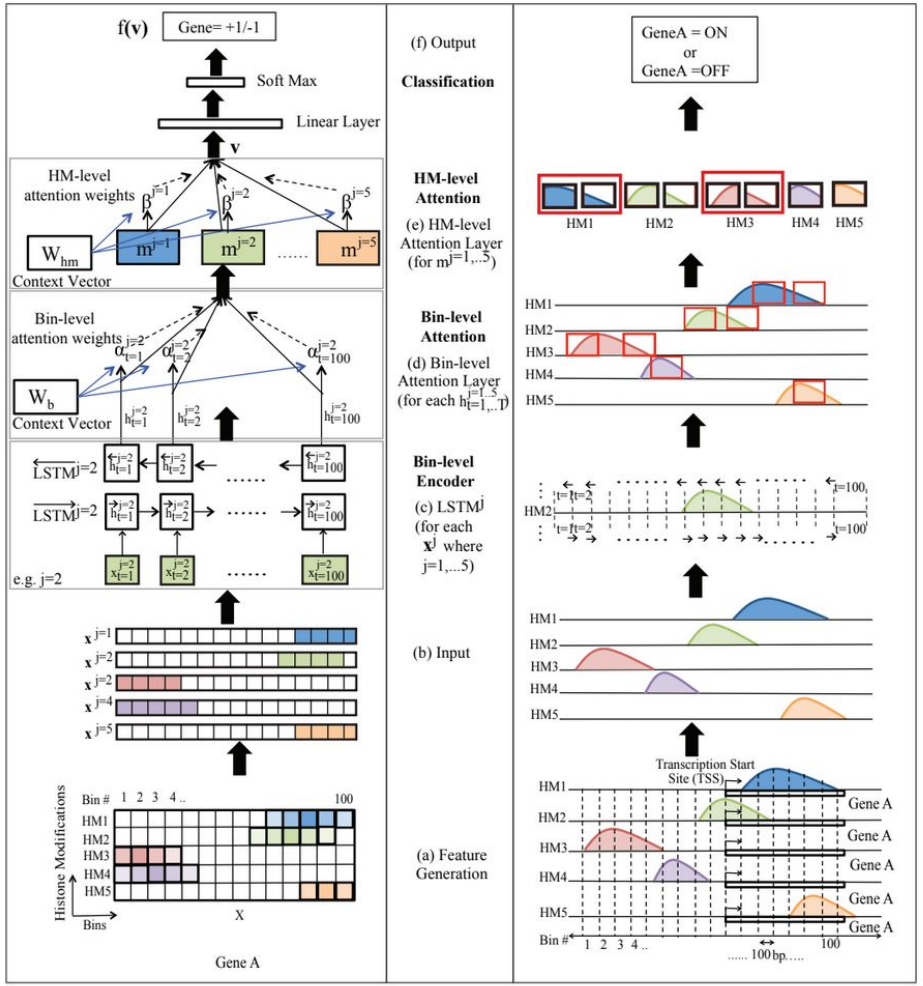
Model Objective

For each gene,

- Local feature
    - Attend most relevant bin positions of an HM mark
- Combinatorial features
    - Recognize and attend the relevant HM marks

Model Formulation

1. Bin-level LSTM encoder
    a. For each HM mark
2. Bin-level Attention
    a. For each HM mark
3. HM-level LSTM encoder
    a. encoding all HM marks
4. HM-level Attention
    a. One over all the HM marks
5. Final classification module



Bin-Level Encoder

One for each HM

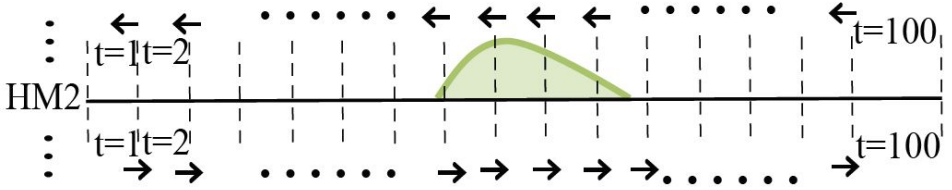
t = 1 to 100 & j = 1 to 5

Bin-Level Encoder

5 bin-level encoder

Bidirectional LSTM$_j$ for each HM$_j$

Backward: from x^T_j to x^1_j $\overrightarrow{\mathbf{h}_t^j} = \overrightarrow{LSTM^j}(x_t^j)$



concat

Final embedding vector

$$\mathbf{h}_t^j = [\overrightarrow{\mathbf{h}_t^j}, \overleftarrow{\mathbf{h}_t^j}]$$

Forward: from x^1_j to x^T_j $\overleftarrow{\mathbf{h}_t^j} = \overleftarrow{LSTM^j}(x_t^j)$.

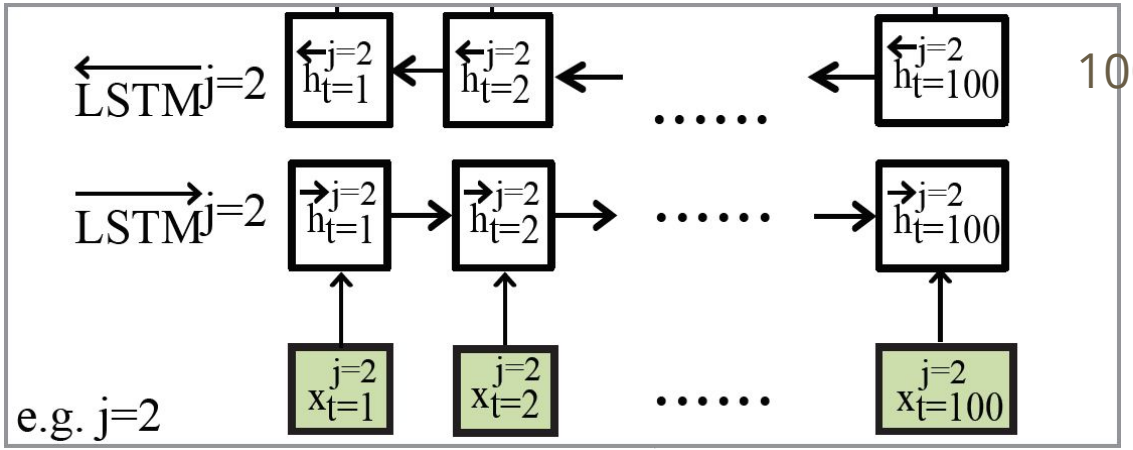
Bin-Level Encoder

t = 1 to 100 & j = 1 to 5

5 bin-level encoder

Bidirectional $LSTM_j$ for each HM_j

for each HMs



100 embedding vectors, of size d

Example: HM_2

Bin-Level α-Attention

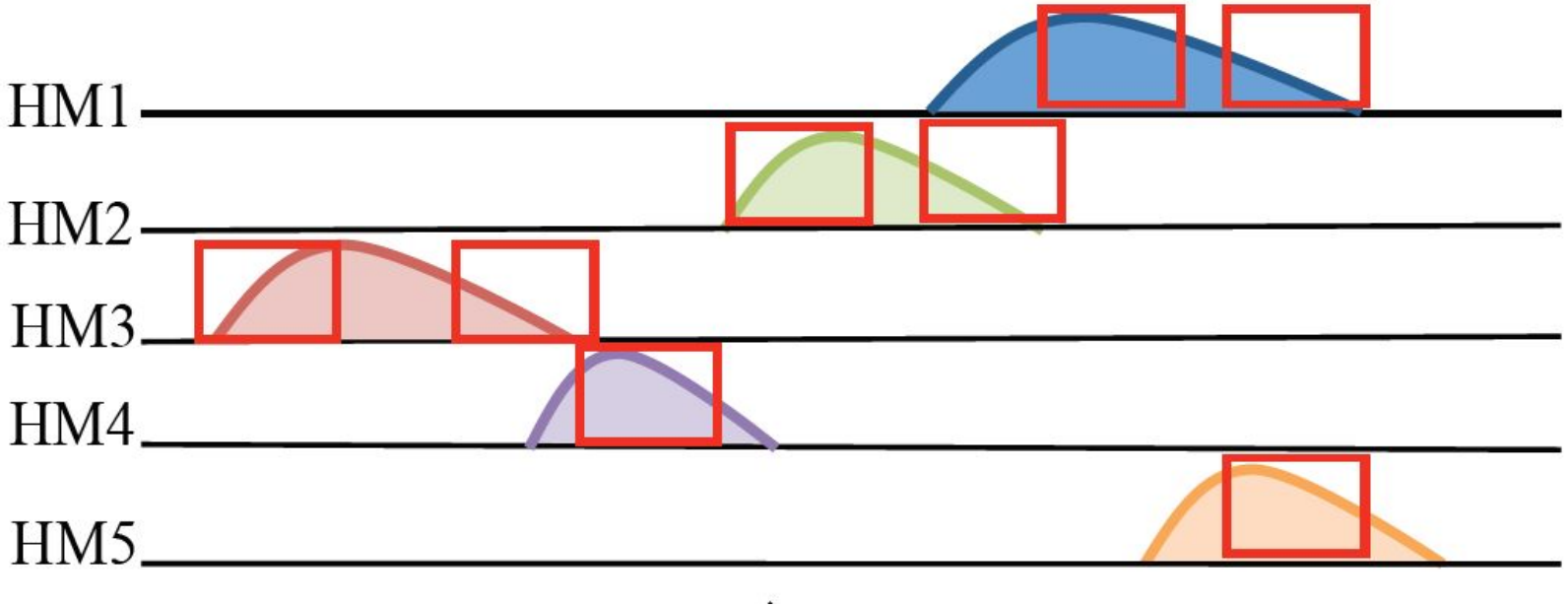
One for each HM

t = 1 to 100 & j = 1 to 5

Bin-Level α-Attention

5 bin-level attention

Finds the bins important for the encoding of j^{th} HM mark



Bin-Level α-Attention

Learnable weight vector: α^j, of size T

Bin-level attention weight, $\boldsymbol{\alpha}_t^j = \frac{exp(\mathbf{W}_b \mathbf{h}_t^j)}{\sum_{i=1}^{T} exp(\mathbf{W}_b \mathbf{h}_i^j)}$

T embedding vectors

Intuition: W_b learns the positional relevance

HM Encoding

Representation of the entire j^{th} HM mark, for this gene X

$$\mathbf{m}^j = \sum_{t=1}^{T} \alpha_t^j \times \mathbf{h}_t^j$$

HM-Level Encoder

One to encode all HMs

HM-Level Encoder

Bidirectional LSTM

On imagined sequence of HMs

Input: m^j representation of j^{th} HM

Output: of size d',

$$\mathbf{s}^j = [\overrightarrow{LSTM}_s(\mathbf{m}^j), \overleftarrow{LSTM}_s(\mathbf{m}^j)]$$

The Dependencies

Bin-Level Encoding, m^j

Dependency among the bins for j^{th} HM

HM-Level Encoding, s^j

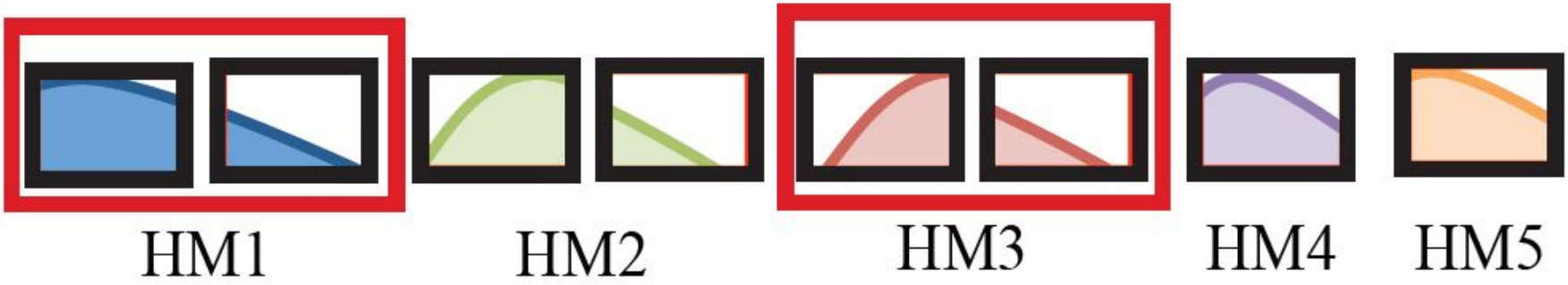
Dependency between the j^{th} HM and the other HMs

HM-Level β-Attention

One to attend the HMs

HM-Level β-Attention

Finds the important HM marks for classifying gene expression



HM-Level β-Attention

Finds the important HM markers for classifying gene expression

Learnable weights: β^j for j ∈ {1, ..., M}

Represents the relative importance of HM^j

HM-level attention weight, $$\beta^j = \frac{exp(\mathbf{W}_s \mathbf{s}^j)}{\sum_{i=1}^{M} exp(\mathbf{W}_s \mathbf{s}^i)}$$

Gene Region Encoding

Entire “**gene region**”, for the current sample **X**, encoded into a hidden representation,

$$\mathbf{v} = \sum_{j=1}^{M} \beta^j \mathbf{s}^j$$

Weighted sum of the embeddings from all HM marks

Summarizes the information of all HMs to represent the entire gene region

Classification of Gene Expression

Predict the gene expression from the encoding vector, **v**

$$f(\mathbf{v}) = \text{softmax}(\mathbf{W}_c\mathbf{v} + b_c)$$

Learnable parameters: W_c and b_c

Loss function: argmin(-log(likelihood))

Model Hyperparameters

Description	Hyperparameter	Size/Value
Bin level embedding size	d	32
HM level embedding size	d’	16
Bin level context vector size	Wb	64
HM level context vector size	Ws	32

Experimental Results

- Performance evaluation using AUC scores
  - DeepChrome: CNN
  - LSTM: without attention
  - Five variants of AttentiveChrome
    - CNN-Attn
    - CNN-α, β
    - LSTM-α
    - LSTM-Attn
    - LSTM-α, β (AttentiveChrome)

AUC Score

Table 2: AUC score performances for different variations of AttentiveChrome and baselines

	Baselines		AttentiveChrome Variations				
Model	DeepChrome (CNN) [29]	LSTM	CNN-Attn	CNN-α, β	LSTM-Attn	LSTM-α	LSTM-α, β
Mean	0.8008	0.8052	0.7622	0.7936	0.8100	**0.8133**	0.8115
Median	0.8009	0.8036	0.7617	0.7914	0.8118	**0.8143**	0.8123
Max	**0.9225**	0.9185	0.8707	0.9059	0.9155	0.9218	0.9177
Min	0.6854	0.7073	0.6469	0.7001	**0.7237**	0.7250	0.7215
Improvement over DeepChrome [29] (out of 56 cell types)		36	0	16	49	**50**	49

Experimental Results

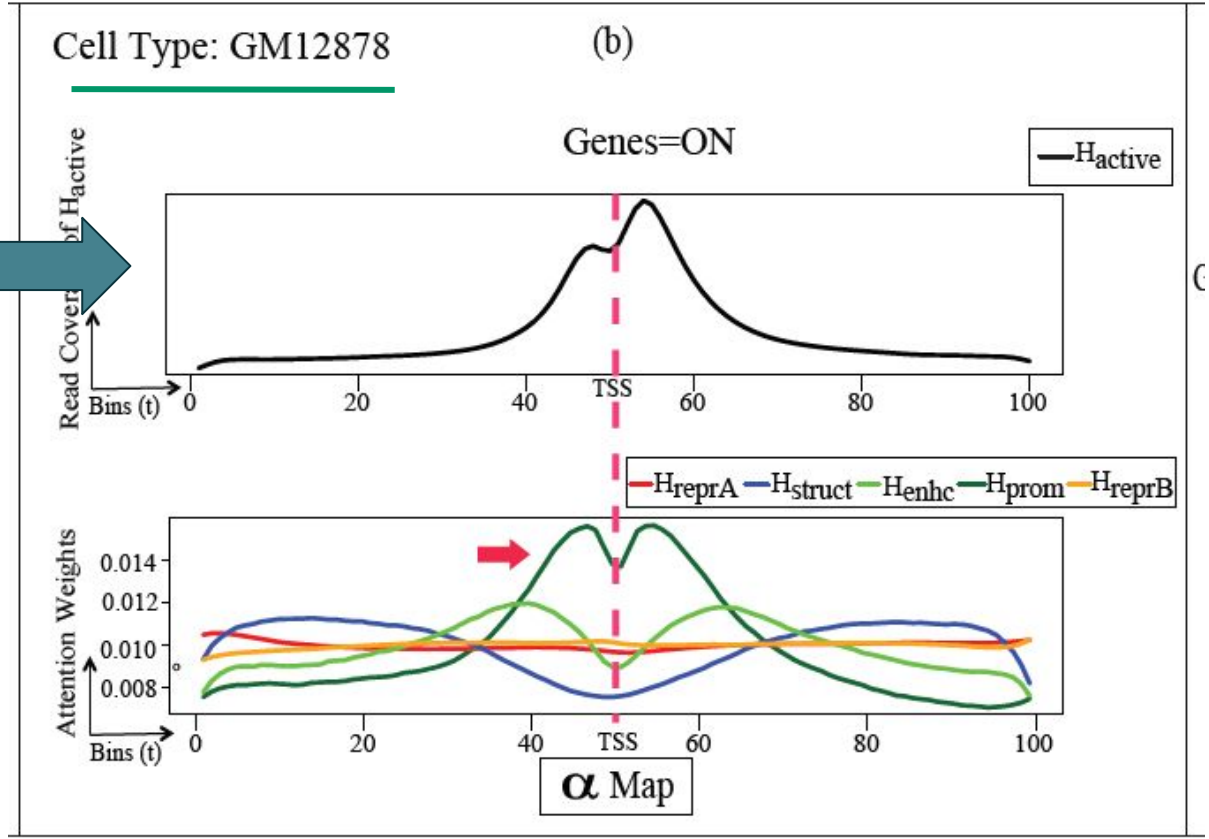
- Evaluation of Interpretation
  - Correlation of HM β-Attention Weight
  - Visualization of bin α-Attention Weights for each HM for cell GM12878
  - Attention Weight of bins with H_{active}
  - Visualization of HM-level Attention Weight for Gene PAX5

Interpretability Evaluation

- A new HM signal H_{active} - H3K27ac from REMC database
  - Active when gene is known
  - Not included in training
- For all the active genes:
  - average read counts of H_{active} across all 100 bins
- For all predicted ON genes:
  - Importance weights calculated by all visualization methods for our active input mark, H_{prom}

HM Profile

AttentiveChrome



Pearson Correlation

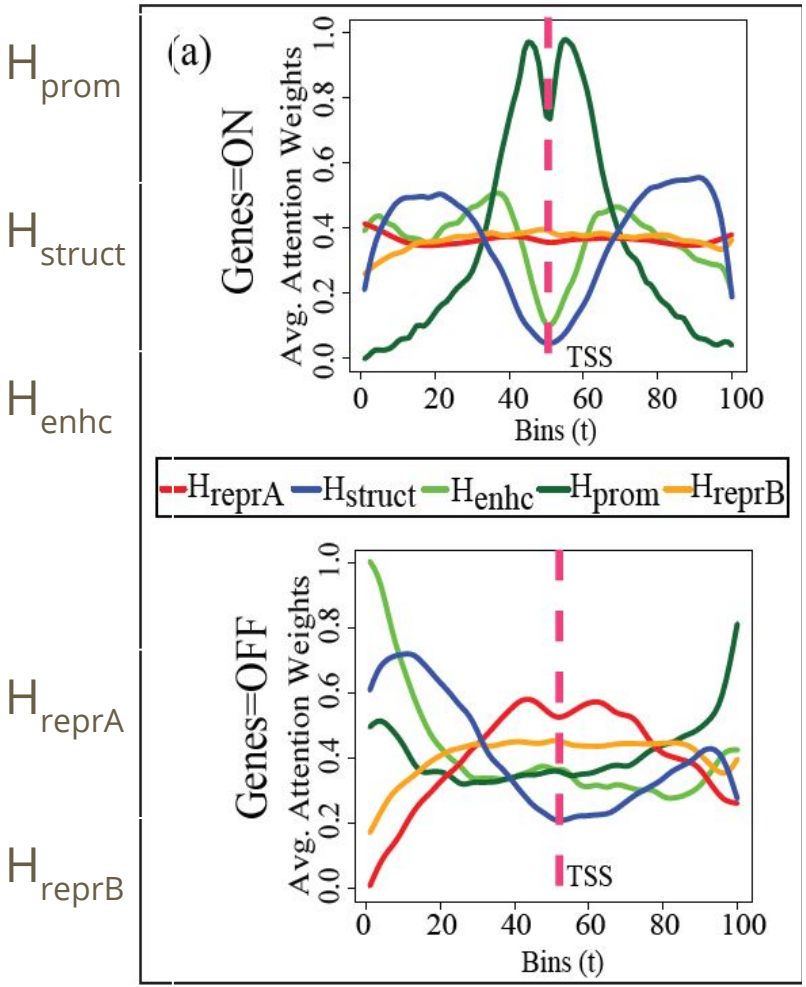
Between H_{active} and H_{prom}

	Stem cell	Blood cell	Leukemia cell
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

Visualization of α-attention

For GM12878 (blood cell)

Average α weights for all the predicted genes (ON-OFF)



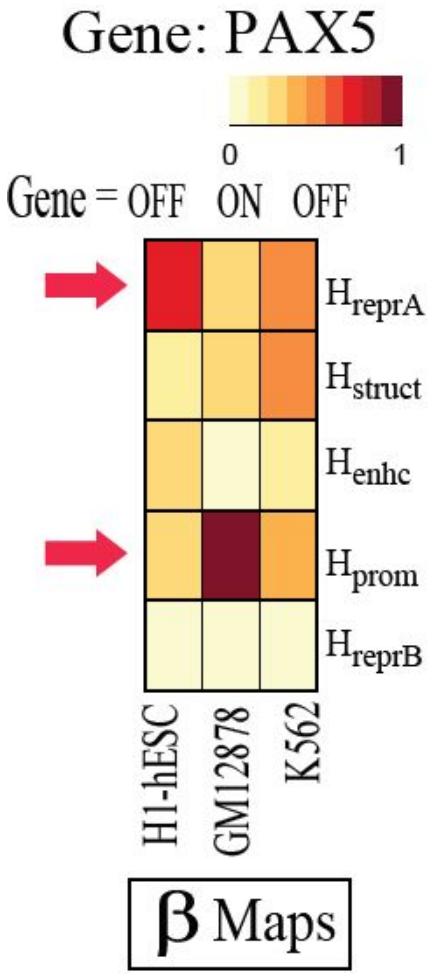
β Heatmap

Advantage of AttentiveChrome over LSTM-α

Gene PAX5: when stem cell converts to blood

- Stem cell: OFF
- Blood cell: ON

Heatmaps visualizing the HM-level weights



Previous Works

1. Linear regression
2. Support vector machines
3. Random forests
4. DeepChrome (CNN)
   a. automatically learns combinatorial interactions

Signal average as input feature

Multiple models

Requires additional feature analysis

Computational Study	Unified	Non-linear	Bin-Info	Representation Learning		Prediction	Feature Inter.	Interpretable
				Neighbor Bins	Whole Region			
Linear Regression ([14])	×	×	×	×	✓	✓	×	✓
Support Vector Machine ([7])	×	✓	Bin-specific	×	✓	✓	✓	×
Random Forest ([10])	×	✓	Best-bin	×	✓	✓	×	×
Rule Learning ([12])	×	✓	×	×	✓	×	✓	✓
DeepChrome-CNN [29]	✓	✓	Automatic	✓	✓	✓	✓	×
AttentiveChrome	✓	✓	Automatic	✓	✓	✓	✓	✓

Previous Works

One model each bin → Multiple models

Select best bins only → Ignores neighboring bins

Contributions

Classification + Visualization

Better prediction

- More accurate than the state of the art

Better interpretation

- Validation of interpretation score
- With a new mark signal (which is not included in the modeling)

First attention-based deep learning method

- Molecular biology data modeling

Critique

Pros

Direction of how to incorporate deep models with biological dataset

Cons

Numbers do not suggest better performance

No specific HM ordering: why LSTM in HM-level encoding ?

Functional aspects of models are defined and evaluated by the authors

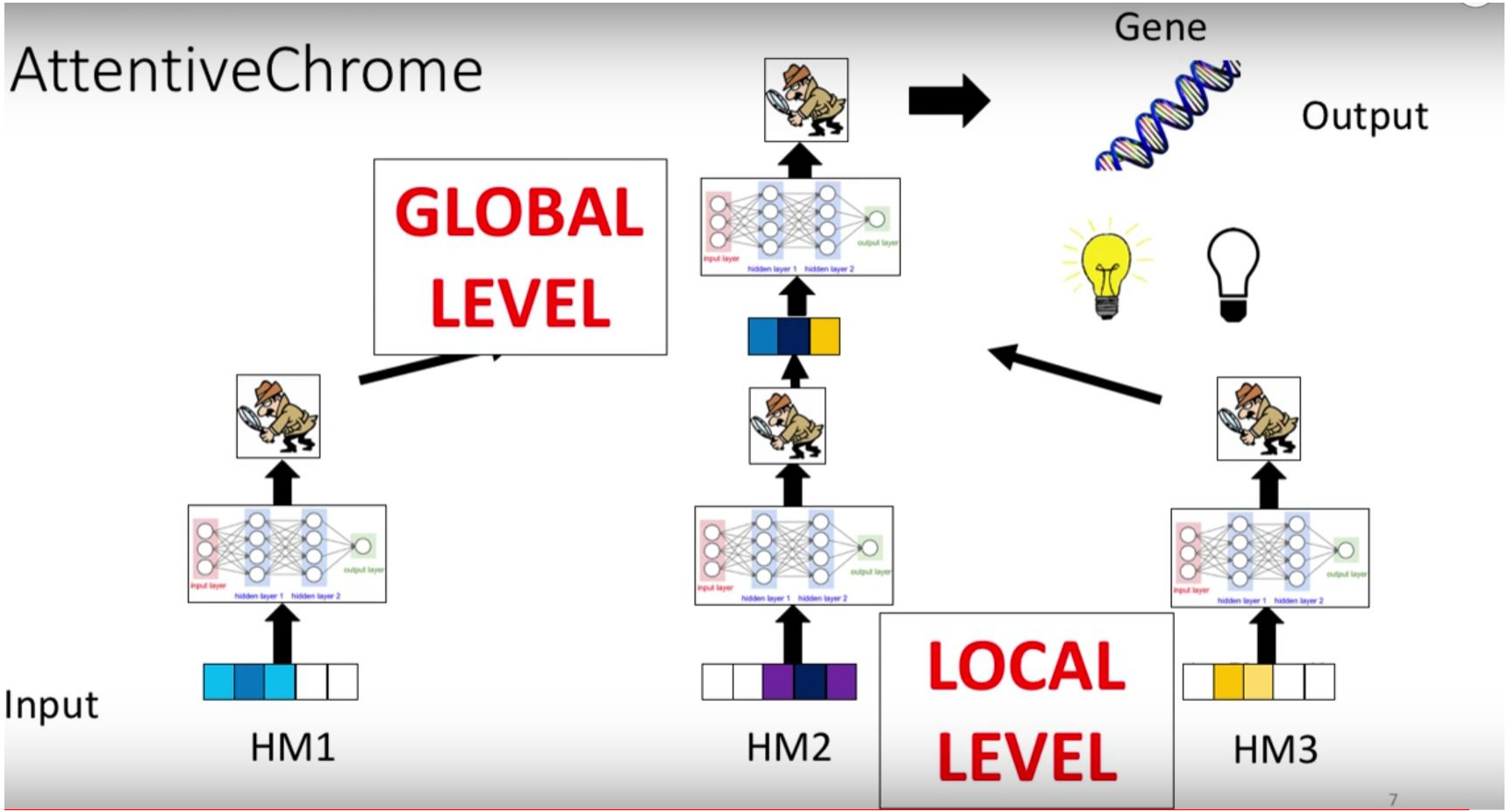
Extension of DeepChrome: however no comparative discussion provided

Future Direction

Try Regression!

Questions?

Thank You


AttentiveChrome
Gene
Output
GLOBAL LEVEL
LOCAL LEVEL
Input
HM1
HM2
HM3
7