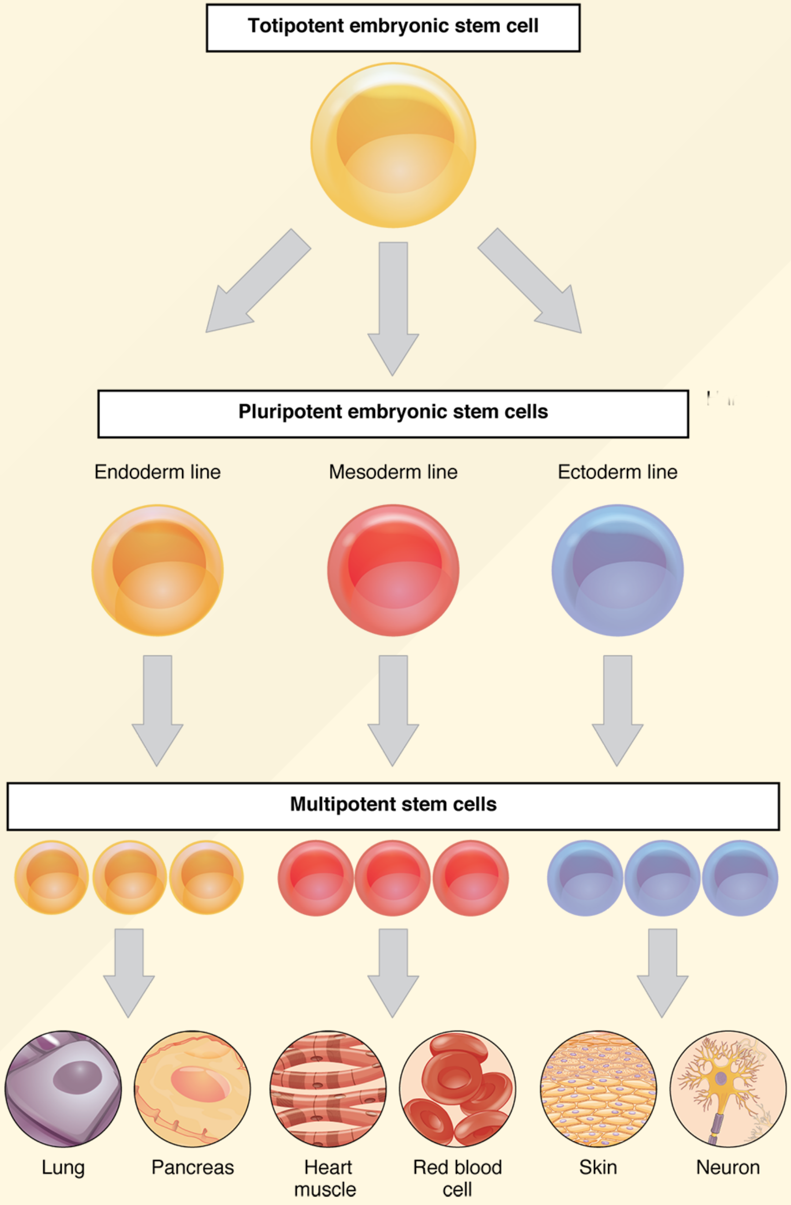
DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications

A. Sekhon, R. Singh, and Y. Qi, Bioinformatics 34 (17)

UnConference 2018

@_alesssia - *@AIClubGenderMinority* - *@H2Oai*

What? Why?



Totipotent embryonic stem cell

Pluripotent embryonic stem cells

Endoderm line

Mesoderm line

Ectoderm line

Multipotent stem cells

Lung

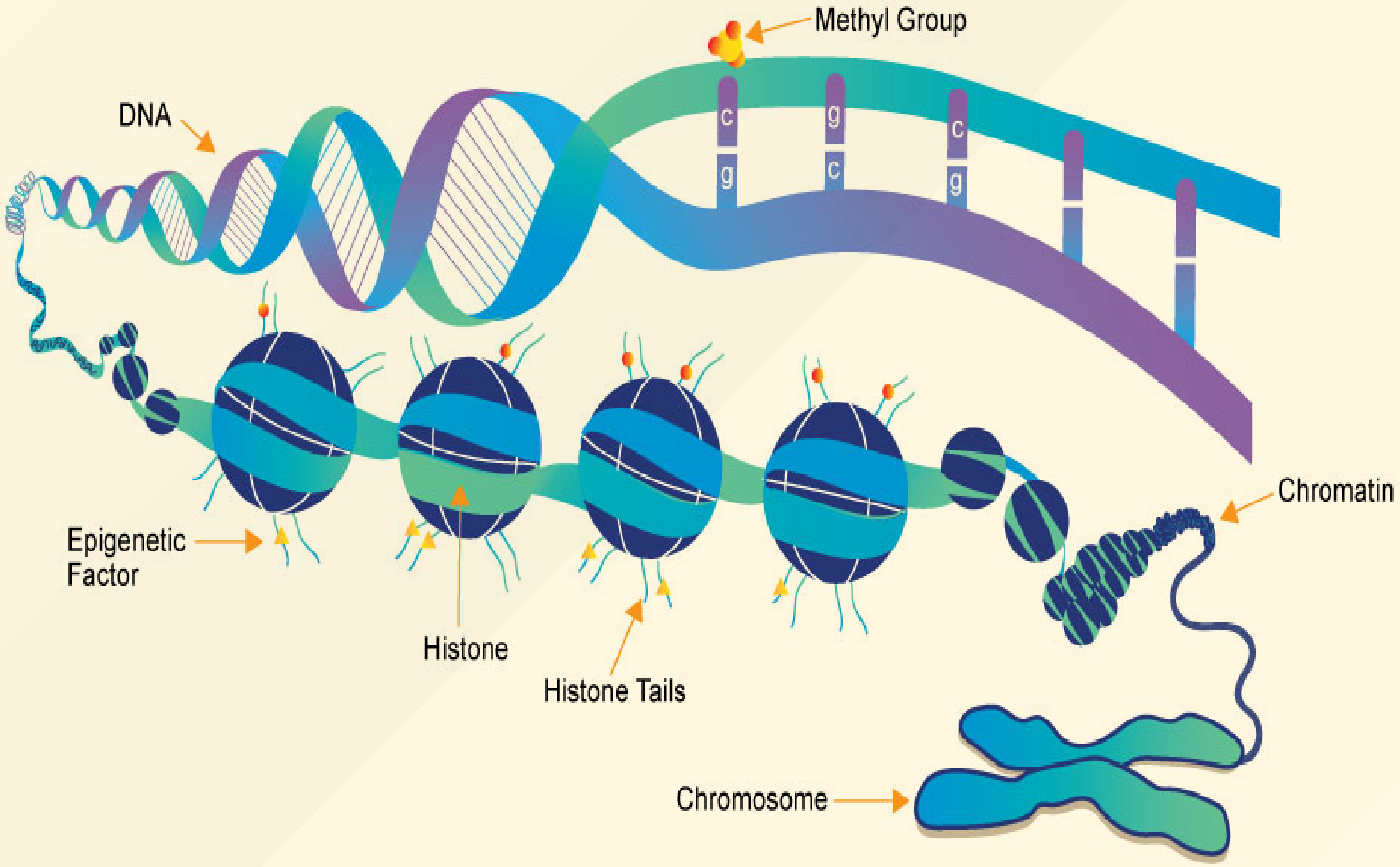
Pancreas

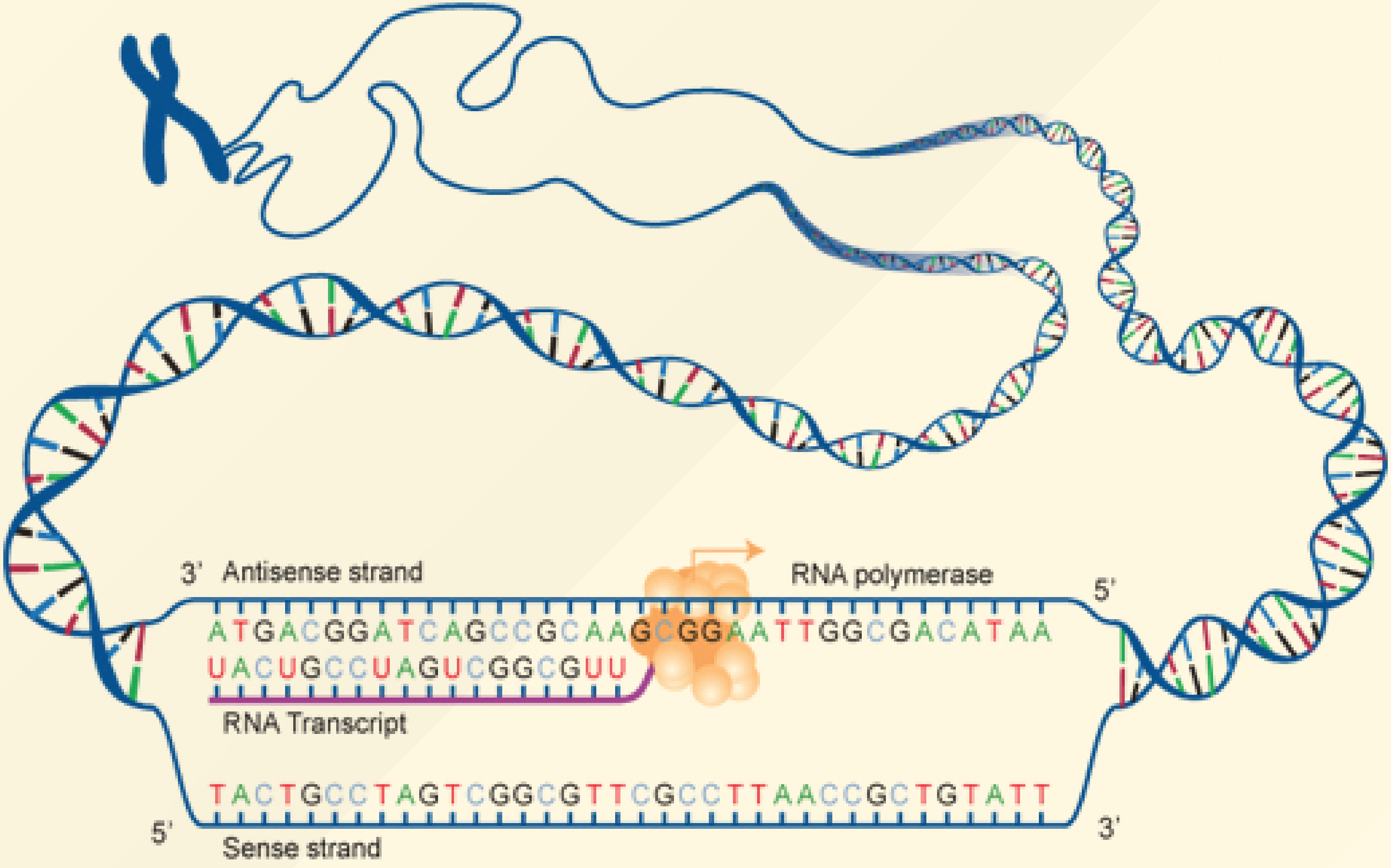
Heart muscle

Red blood cell

Skin

Neuron


Methyl Group
DNA
c
g
g
c
c
g
Chromatin
Epigenetic Factor
Histone
Histone Tails
Chromosome


3'
Antisense strand
RNA polymerase
5'
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
UACUGCCUAGUCGGCGUU
RNA Transcript
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT
5'
3'
Sense strand

Differential gene expression: cell differentiation and diseases

Challenges

1. **Genome-wide histone modification (HM) signals are spatially structured and may have long-range dependency**

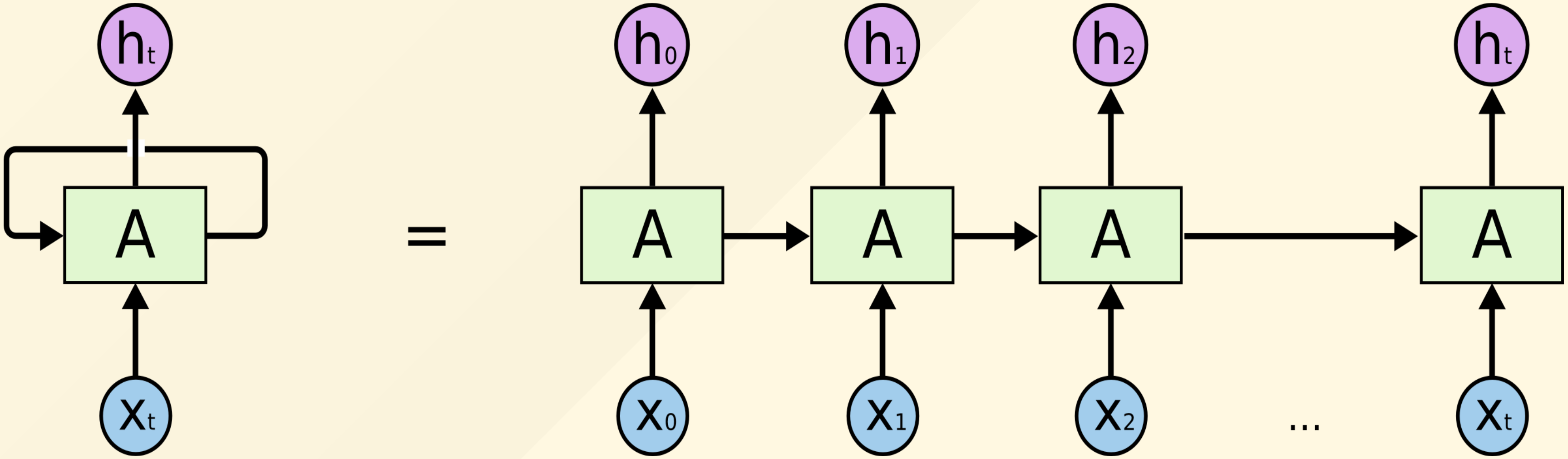
2. We want to understand what the relevant HM factors are, and how they work together

3. We want to understand how HMs affect gene regulation, therefore we require a model with a degree of interpretability

4. We are dealing with multiple cells

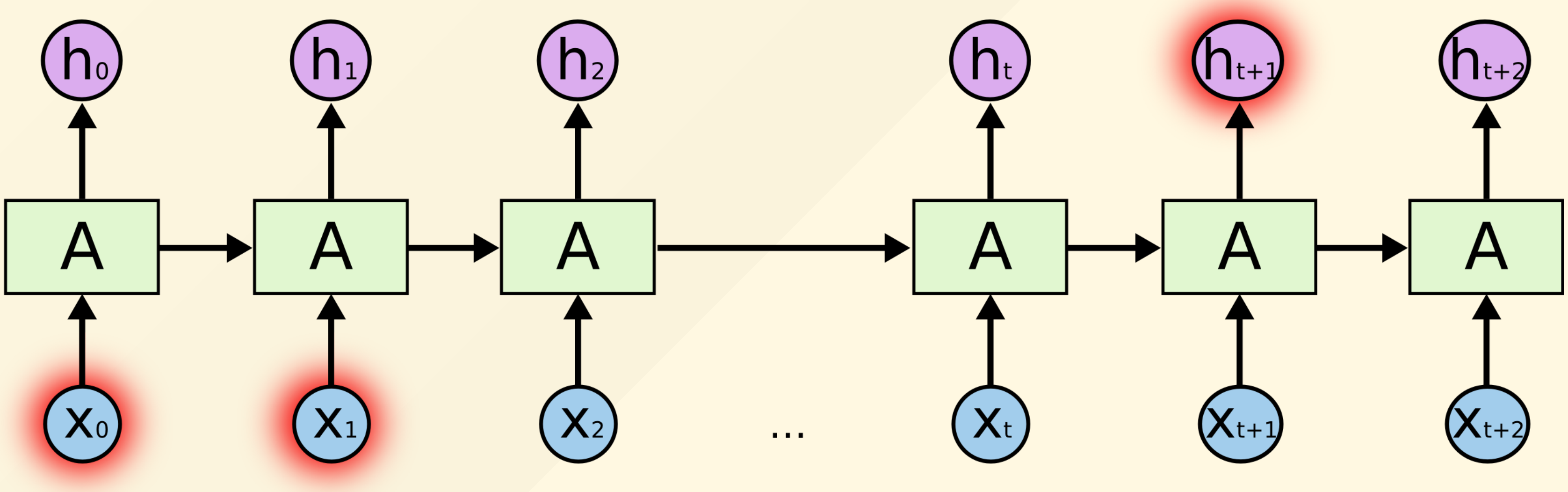
How?

Recurrent Neural Networks (RNNs)



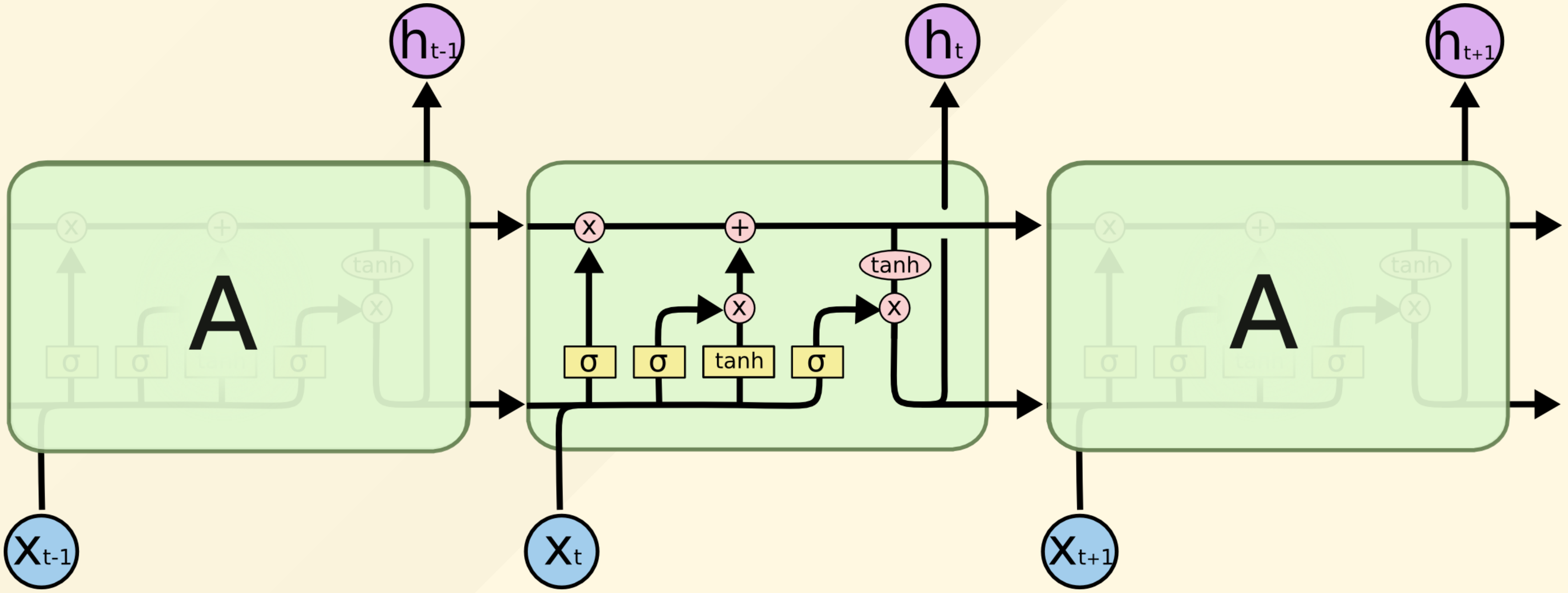
Understanding LSTM Networks

Recurrent Neural Networks (RNNs)



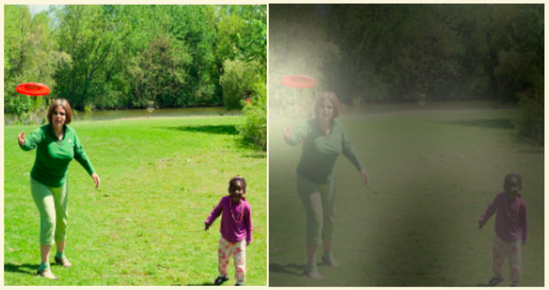
Understanding LSTM Networks

Long Short Term Memory networks (LSTMs)

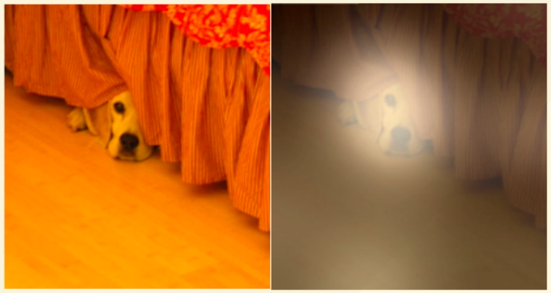


Understanding LSTM Networks

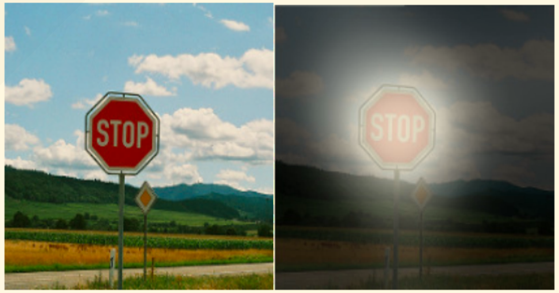
Attention-based deep-learning methods



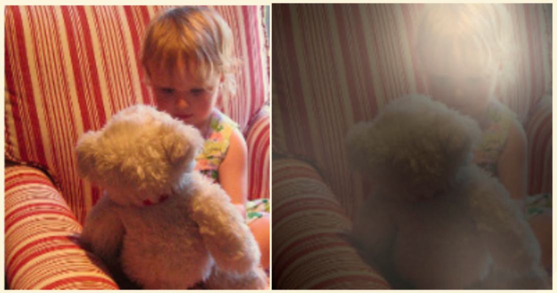
A woman is throwing a frisbee in a park.



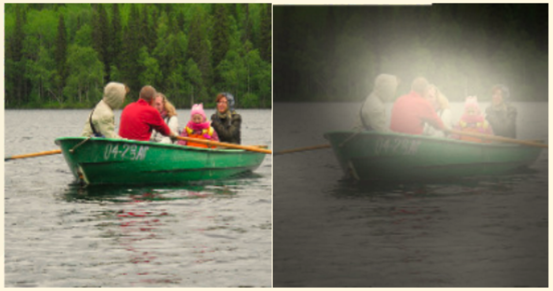
A dog is standing on a hardwood floor.



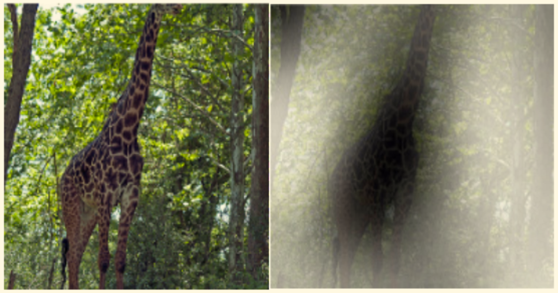
A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.

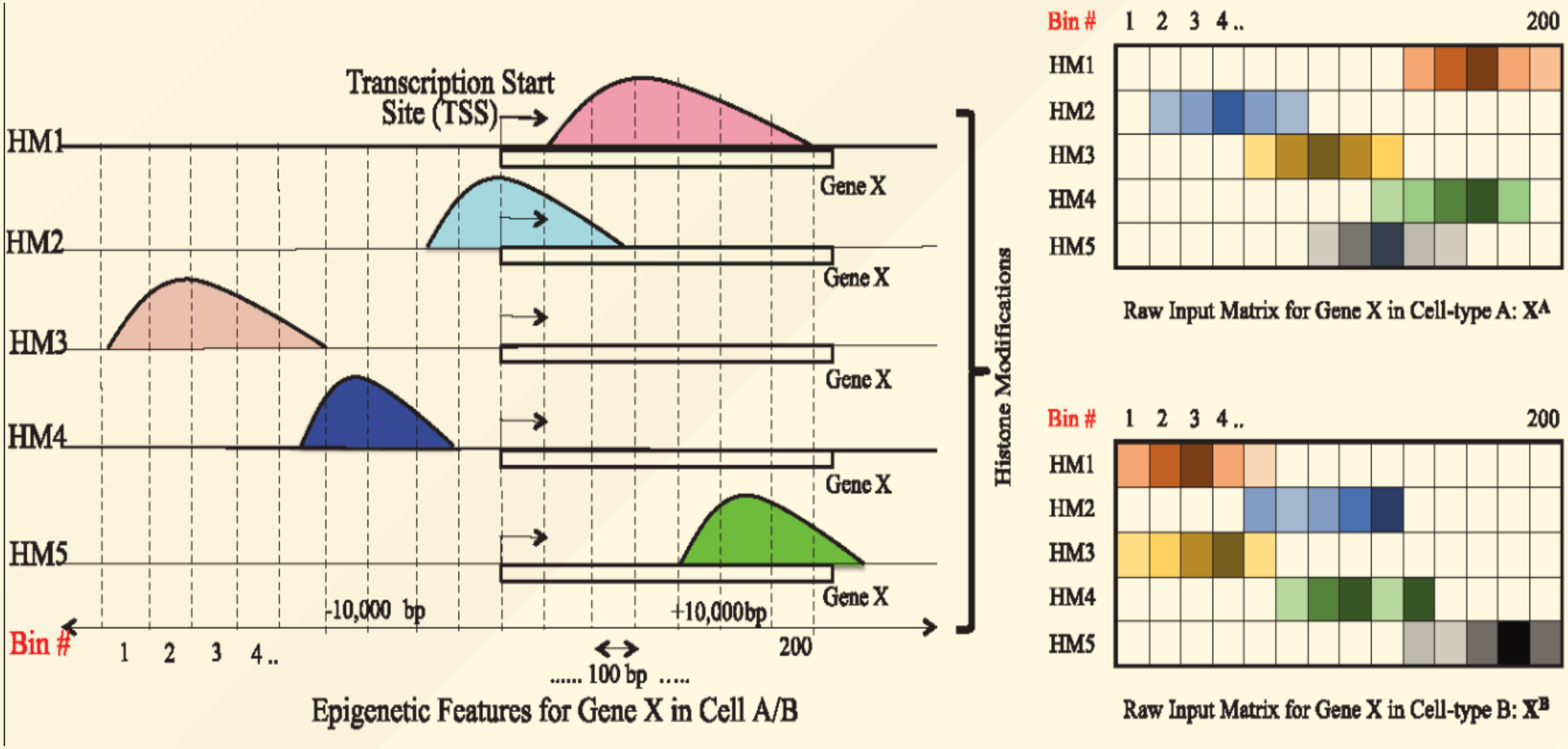


A giraffe standing in a forest with trees in the background.

How to Automatically Generate Textual Descriptions for Photographs with Deep Learning

DeepDiff

Input generation

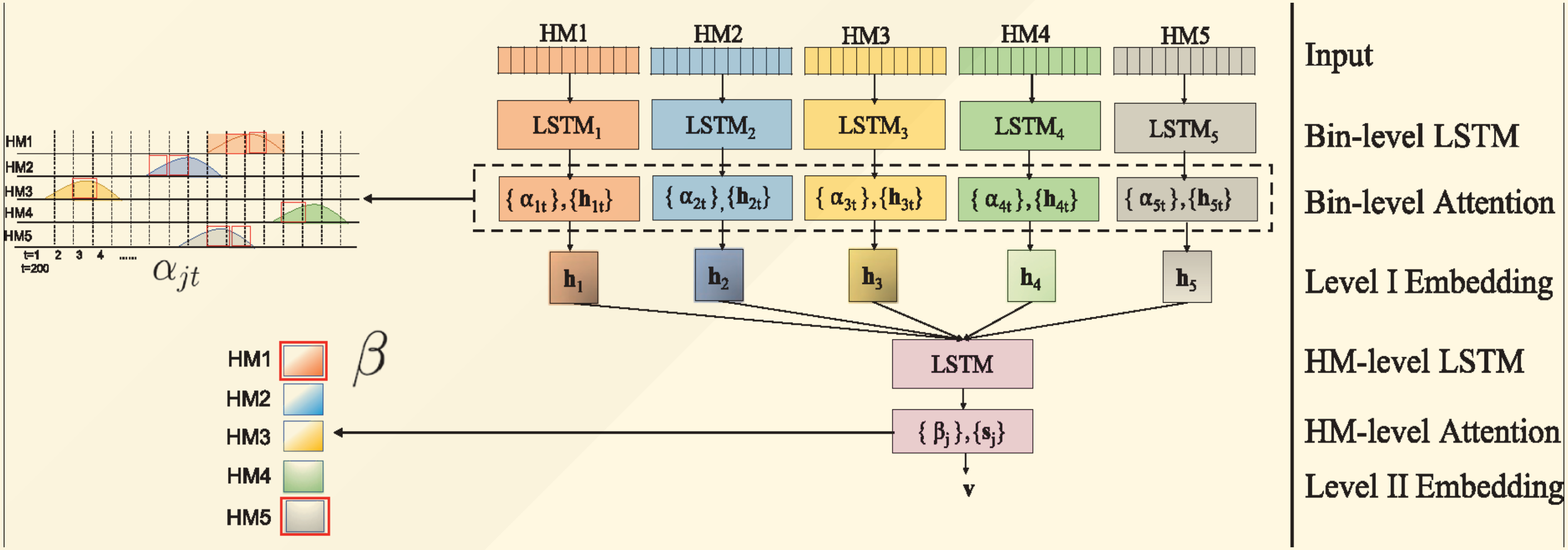


Epigenetic Features for Gene X in Cell A/B

Raw Input Matrix for Gene X in Cell-type A: $\mathbf{X}^A$

Raw Input Matrix for Gene X in Cell-type B: $\mathbf{X}^B$

DeepDiff network architecture



Multitasking with auxiliary tasks

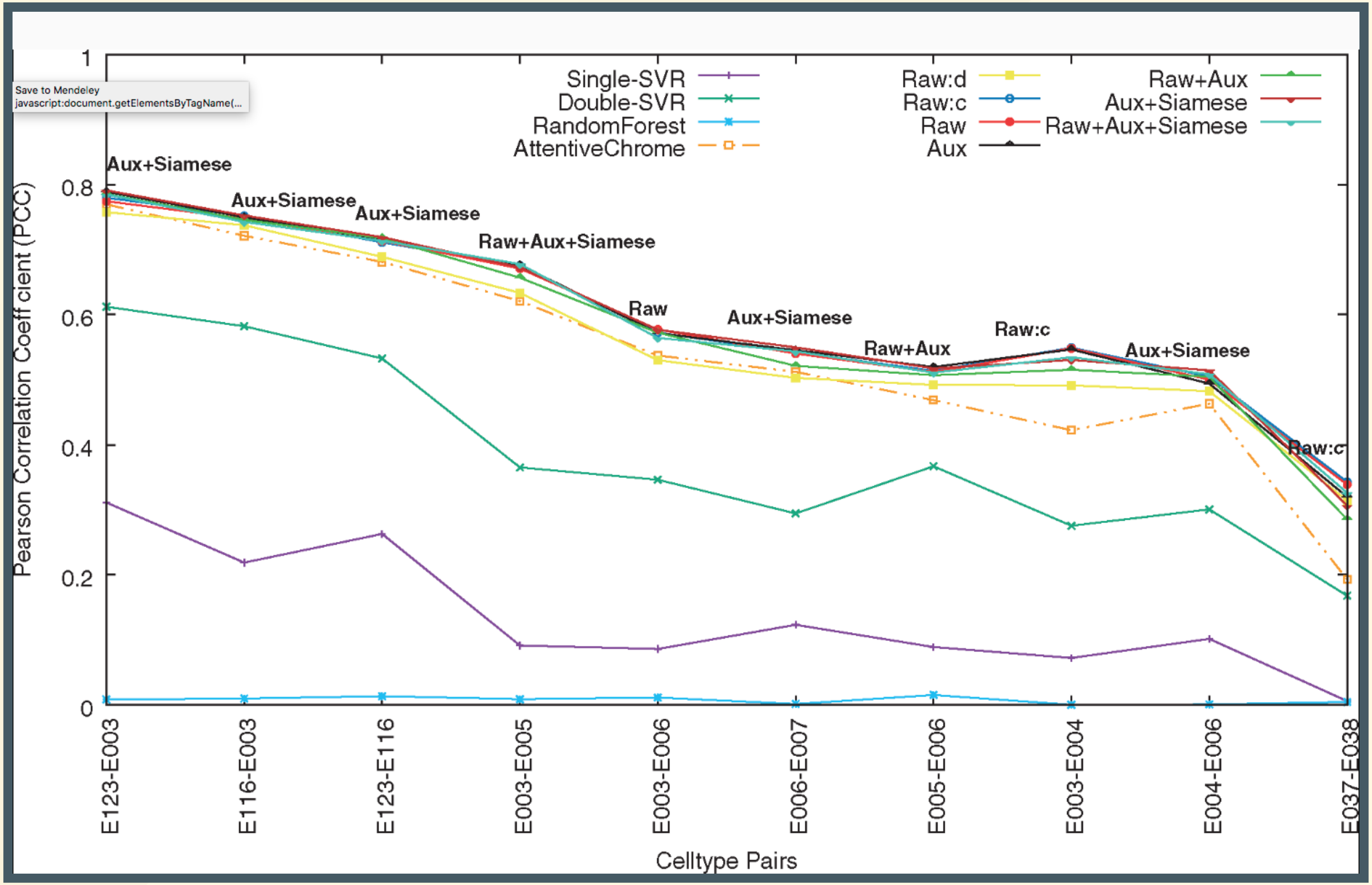
1. Cell-type specific prediction
2. Siamese architecture formulation

Results

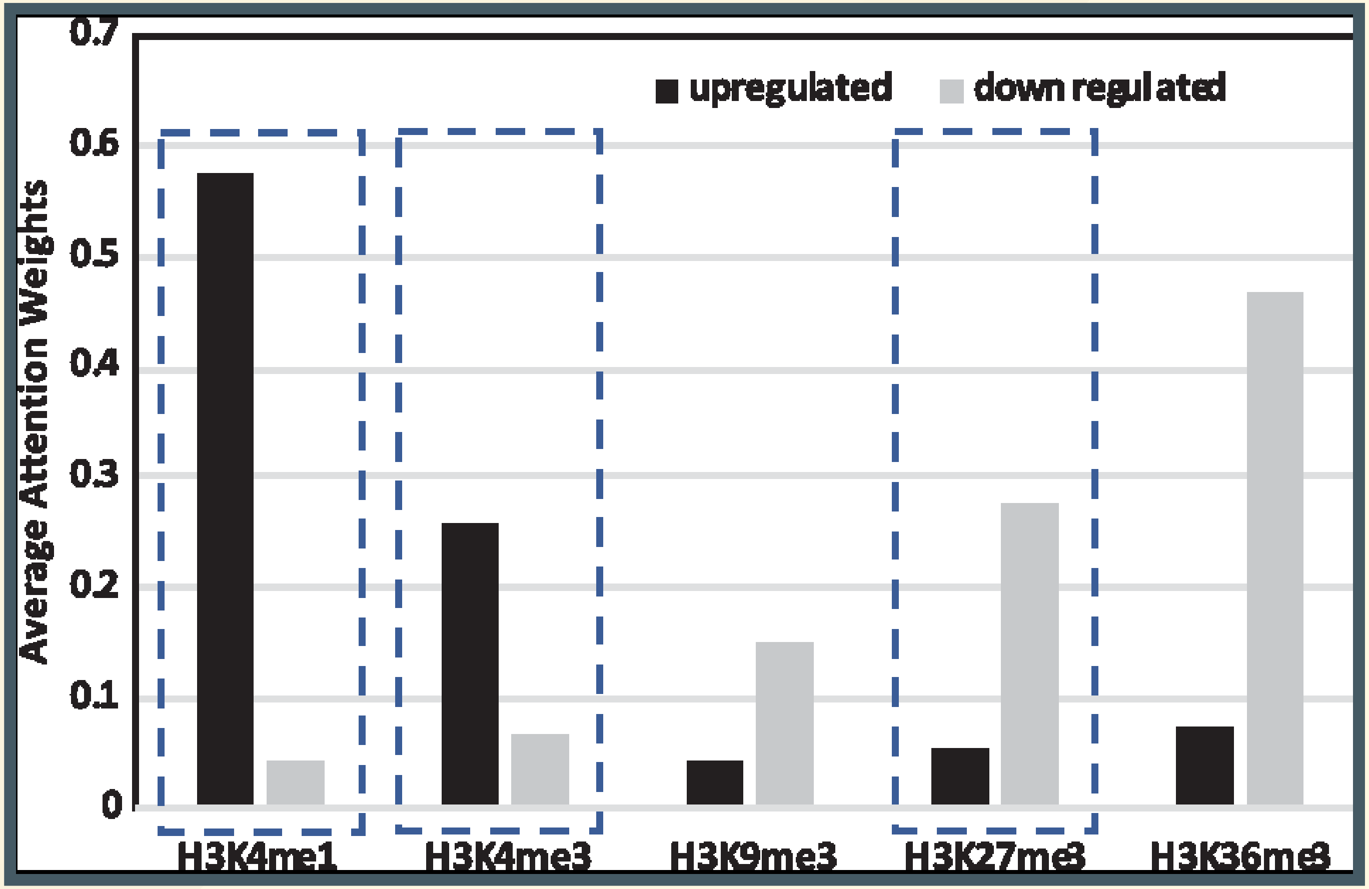
Dataset

- 10 cell lines from the RoadMap project (18,460 genes)
    - 10,000 genes used for training
    - 2,360 genes used for validation (and tuning)
    - 1,100 genes used for testing
- Performances as PCC between DeepDiff variants and *baselines* (aka, alternative approaches)

Performances


Save to Mendeley
javascript:document.getElementsByTagName(...

Interpretation via Attention



Discussion

Points for discussion

1. Any thoughts on DeepDiff architecture?
2. Will the DeepDiff generalise with new cell lines?
3. Is this the best way to evaluate DeepDiff?
4. Is attention helpful?

Thanks for surviving until the end of the last session 😎

You deserve 🍕

H_2O.ai

#H2OAIWorld London