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

Arshdeep Sekhon[1] Ritambhara Singh[1] Yanjun Qi[1]

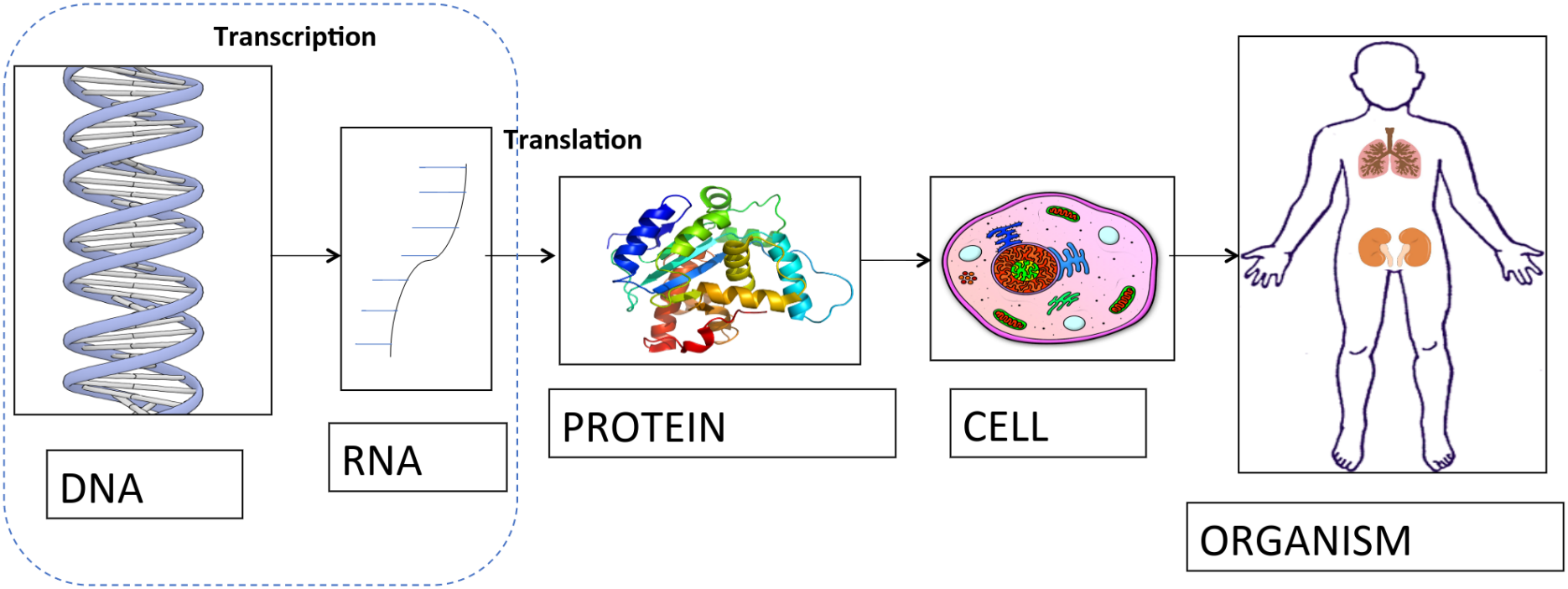
[1]Department of Computer Science
University of Virginia
http://deepchrome.org/

September 11, 2018

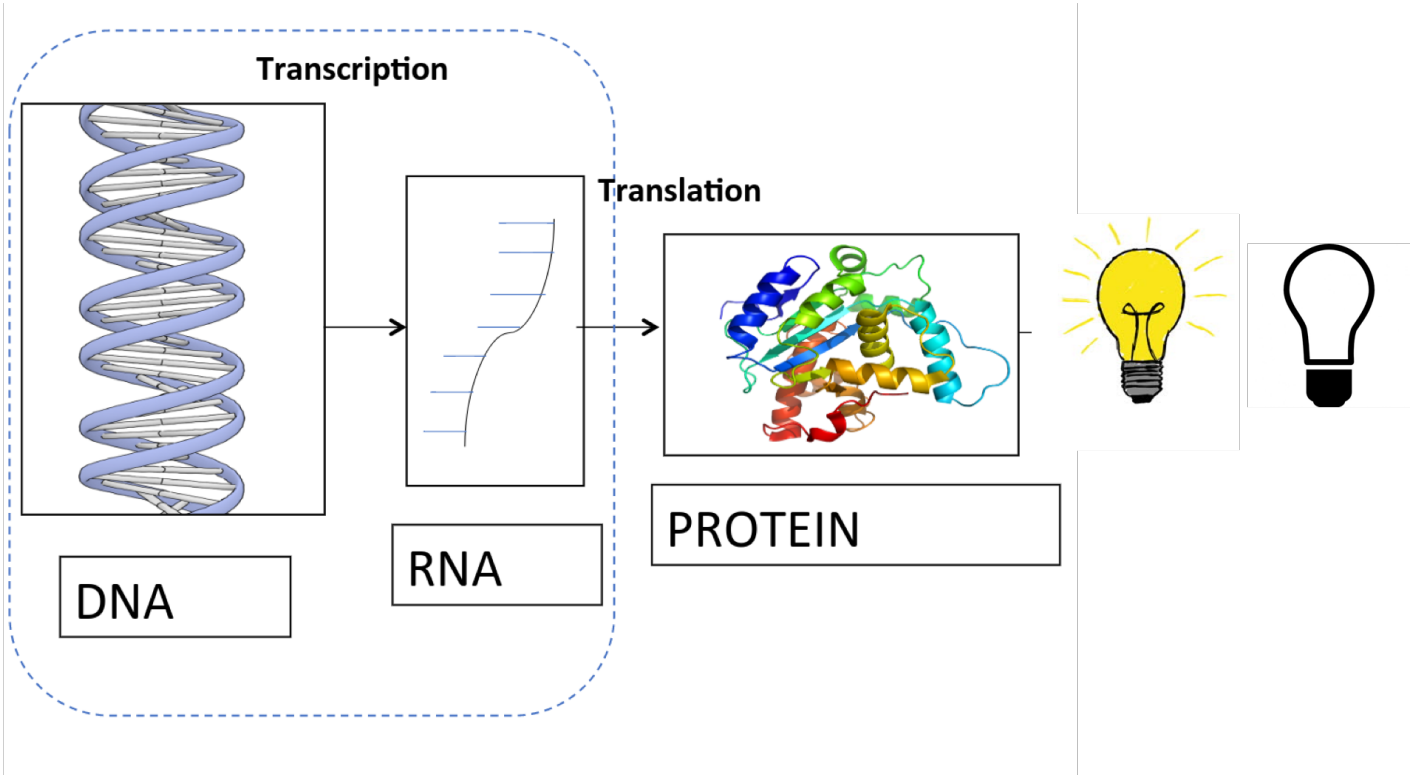
Outline

Outline

Biology in a slide

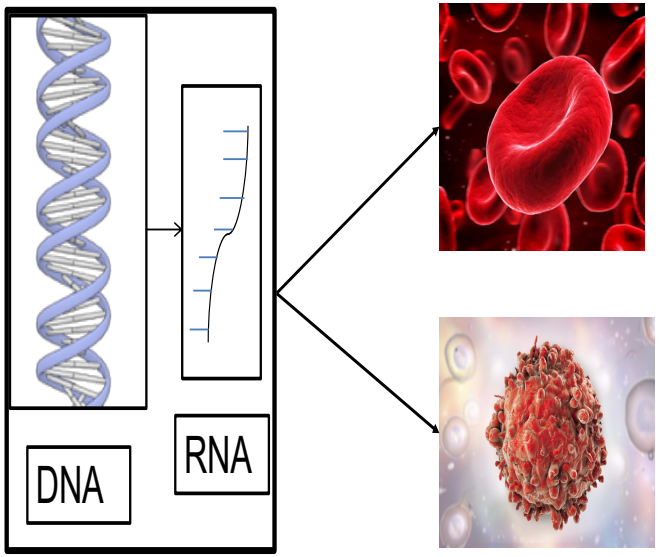


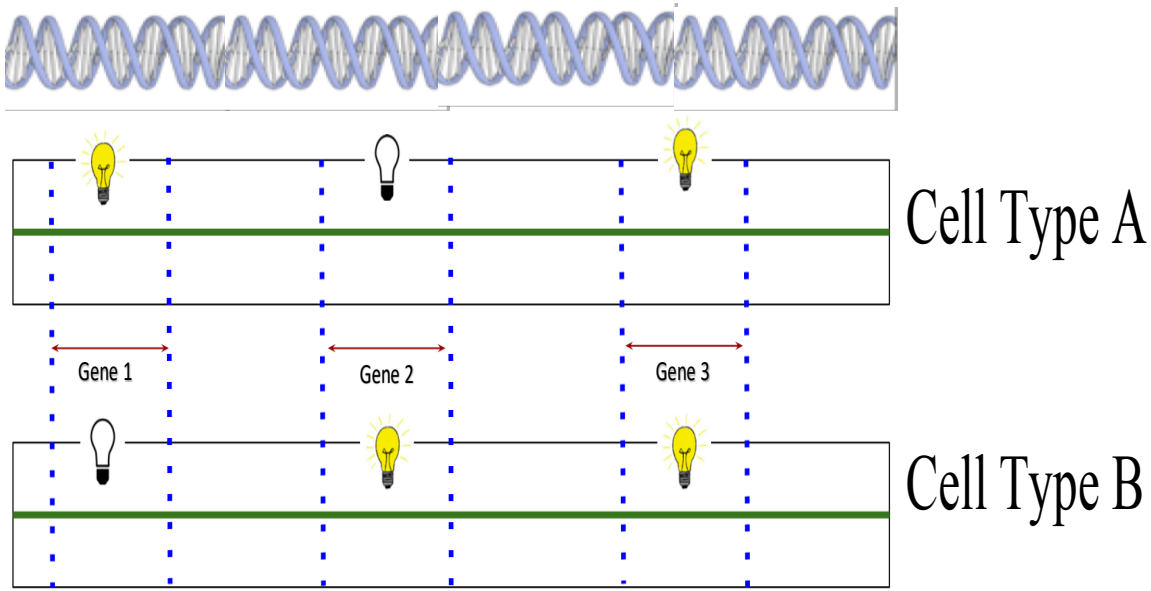
Gene Expression



Differential Gene Expression

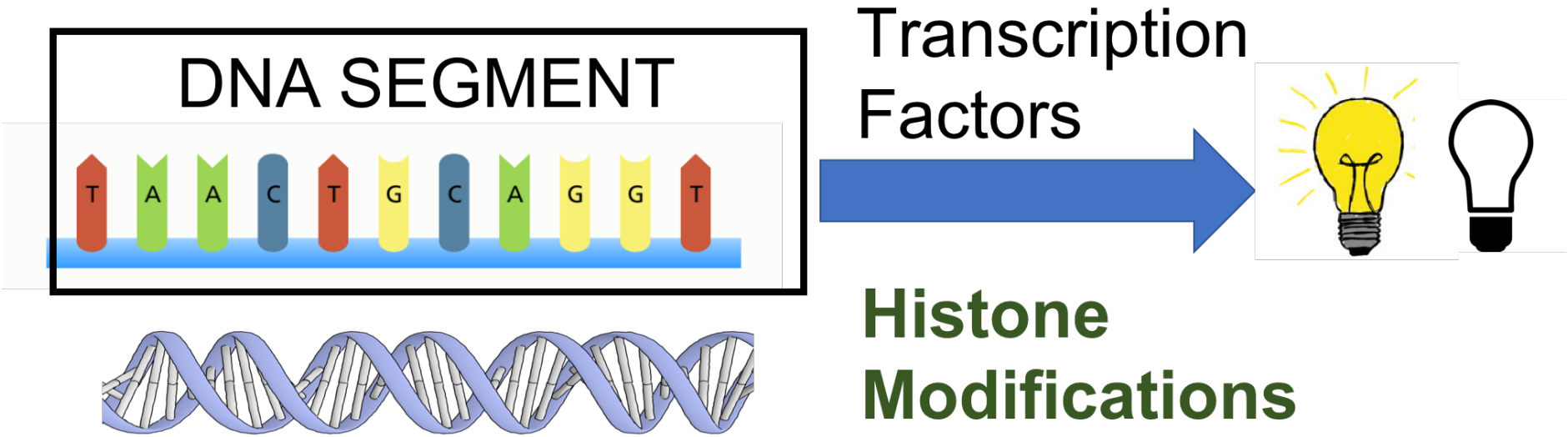
The cell arrives at its function through differential gene expression, the activation of the same gene differently in two cell conditions.



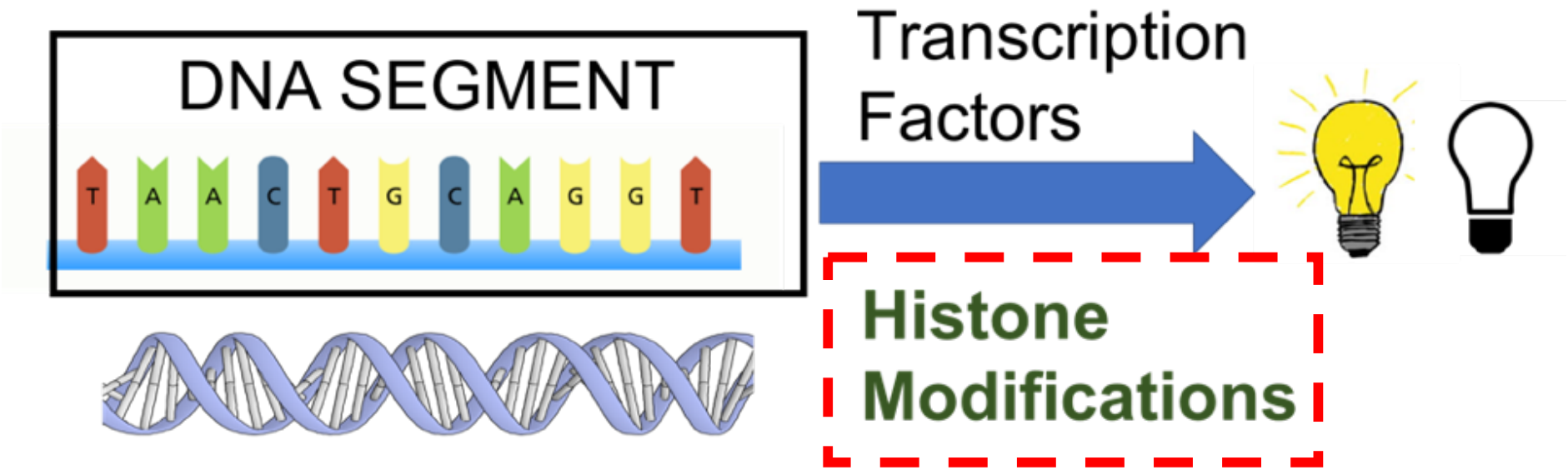


Gene Regulation

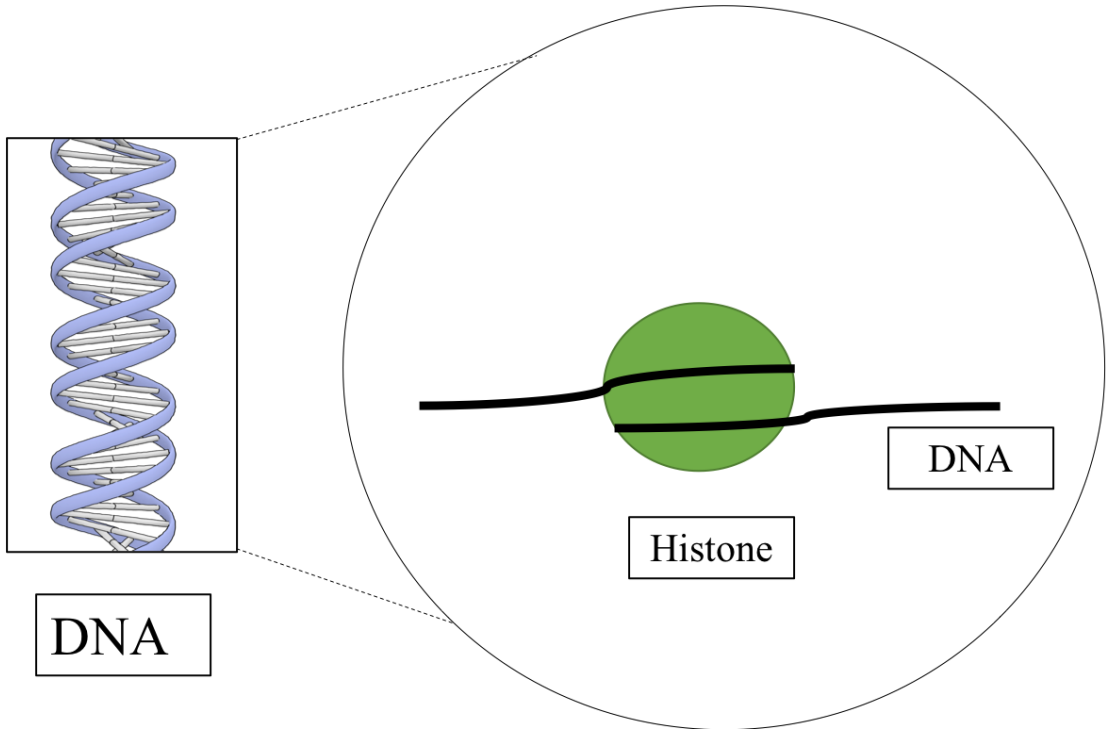
Wide range of mechanisms that are used by cells to increase or decrease gene expression



Gene Regulation

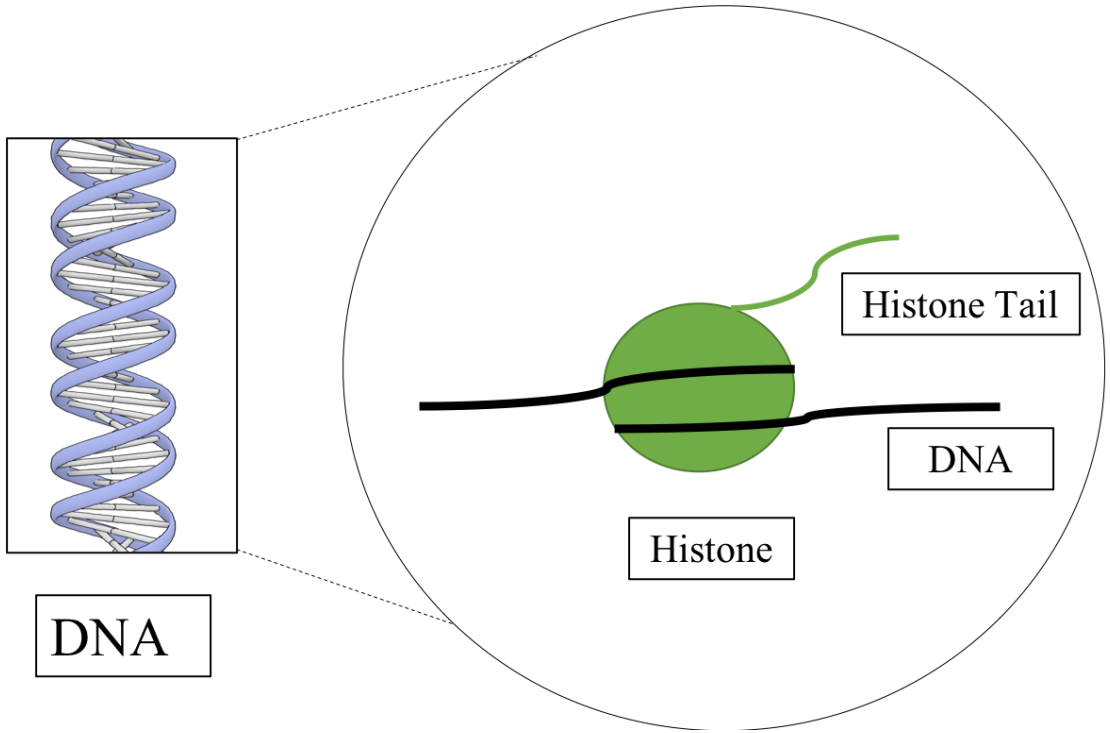


Histone Modifications



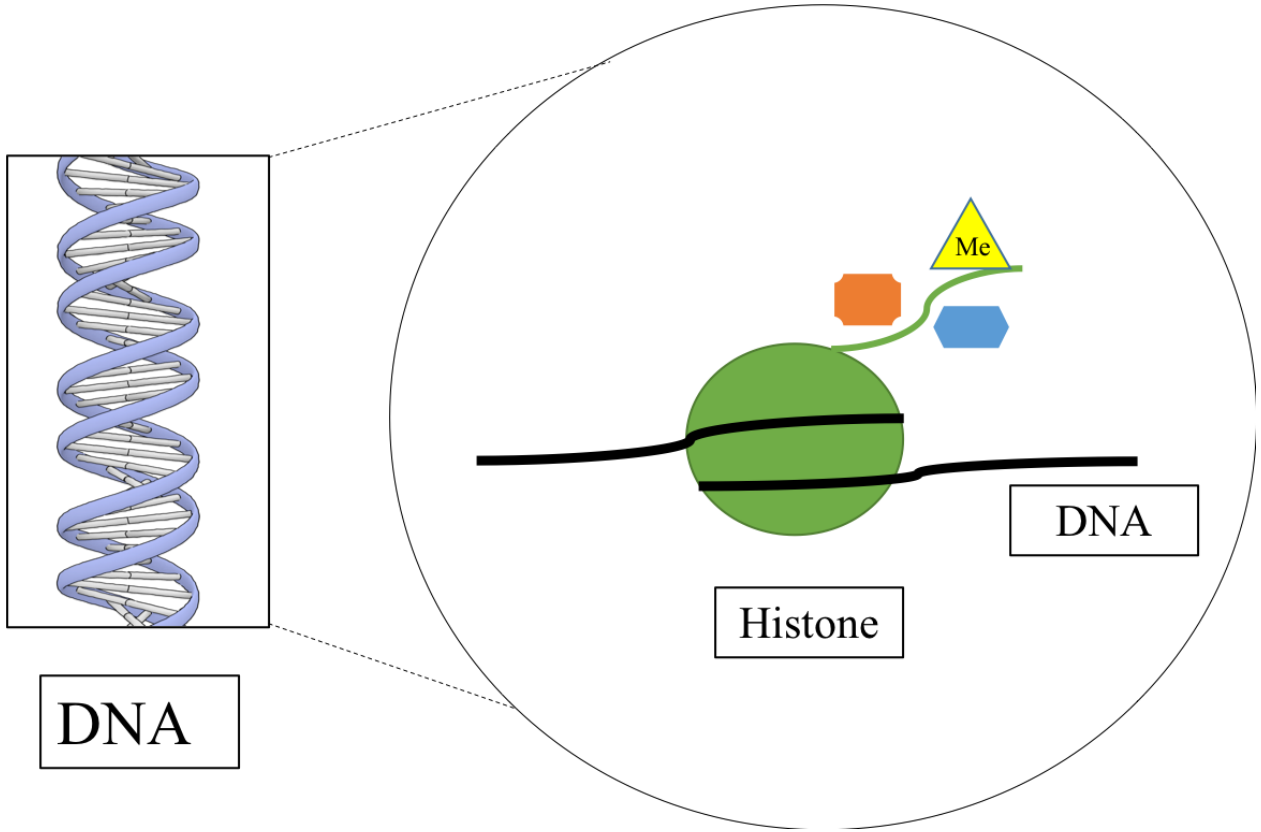
DNA is wrapped around nucleosomes, made of histone proteins.

Histone Modifications



24

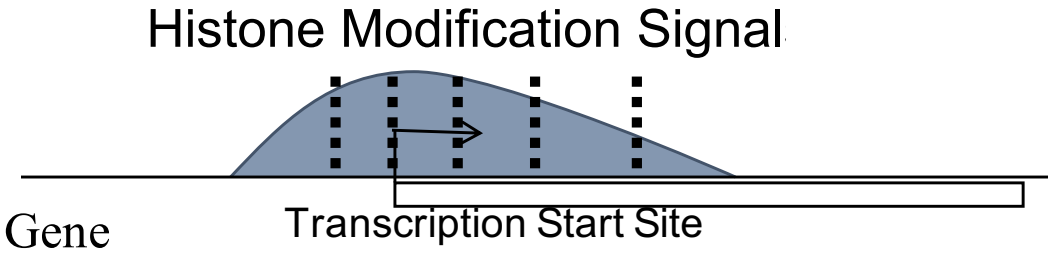
Histone Modifications



Gene Expression Regulation by:

- transcriptional activation/inactivation
- chromosome packaging

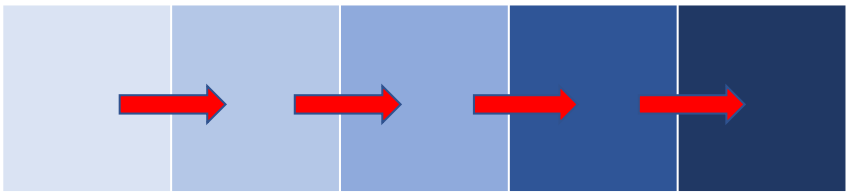
Data Properties: Histone Modifications



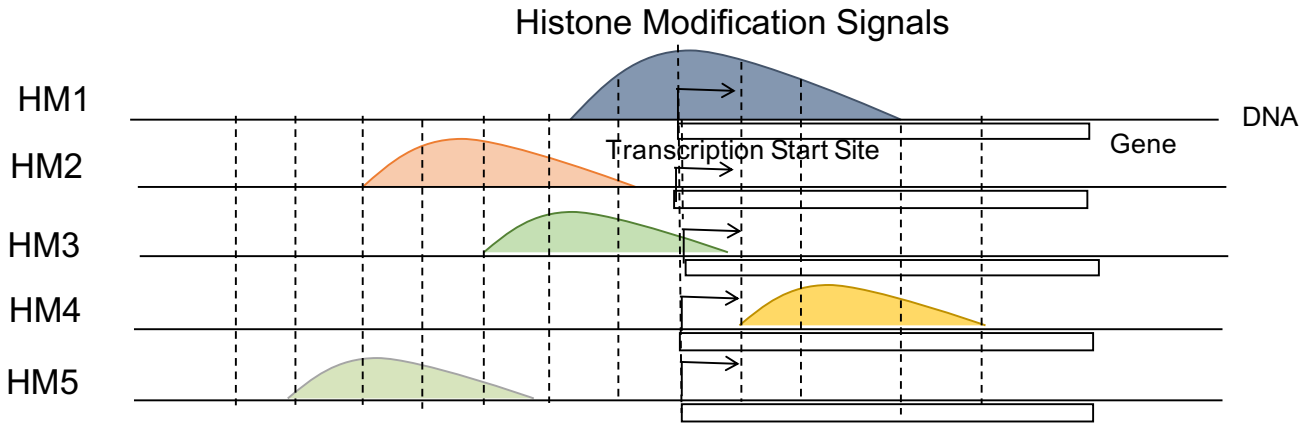
→**Sequential**

Modular

Combinatorial Interactions



Data Properties: Histone Modifications

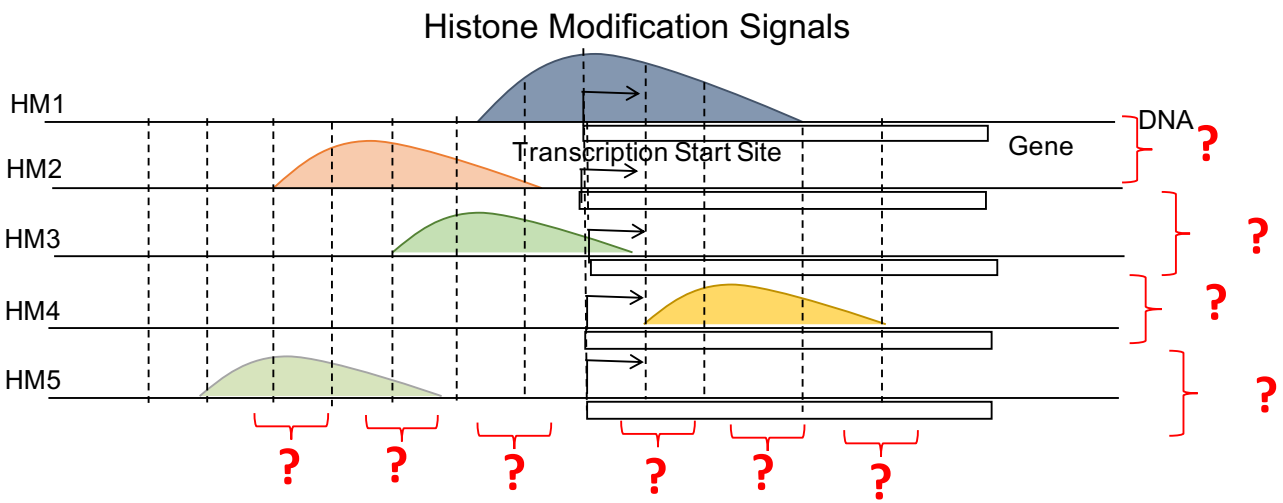


Sequential

$\rightarrow$ **Modular**

Combinatorial Interactions

Data Properties: Histone Modifications



4

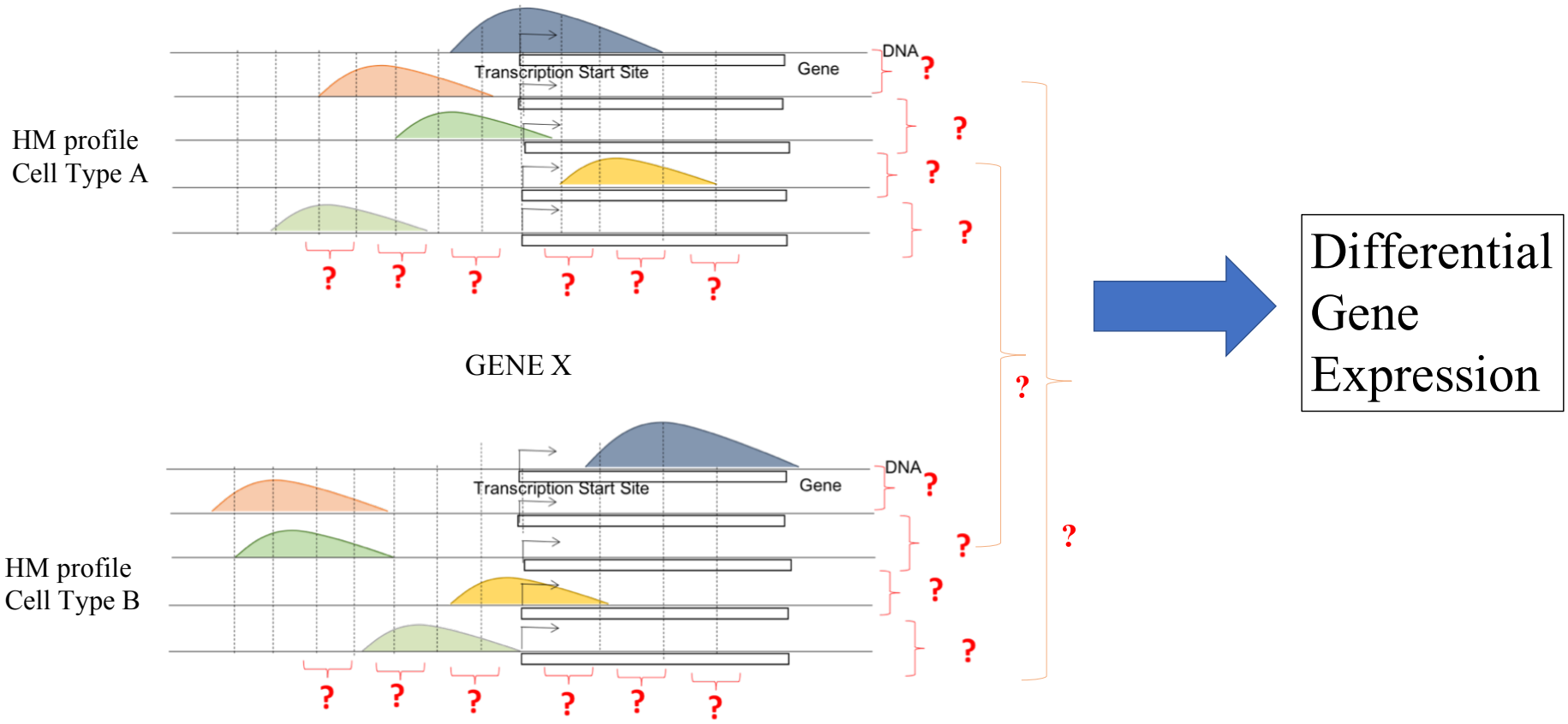
Sequential

Modular

$\rightarrow$ **Combinatorial Interactions**

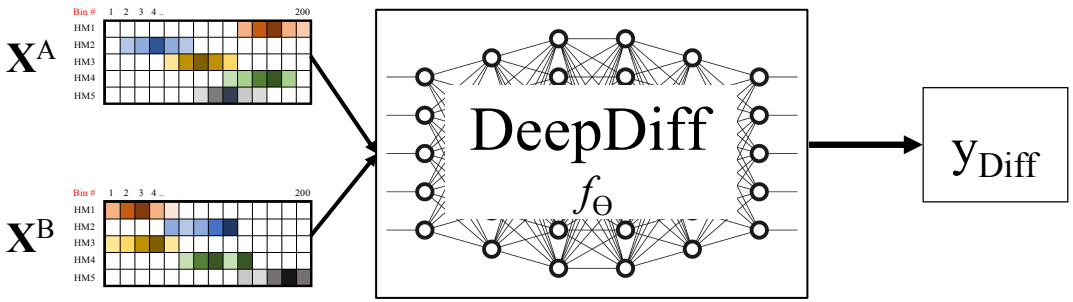
Histone Modifications and Differential Gene Expression

Which HMs at **what** positions across both of the two cell conditions determine differential expression of a gene?



Outline

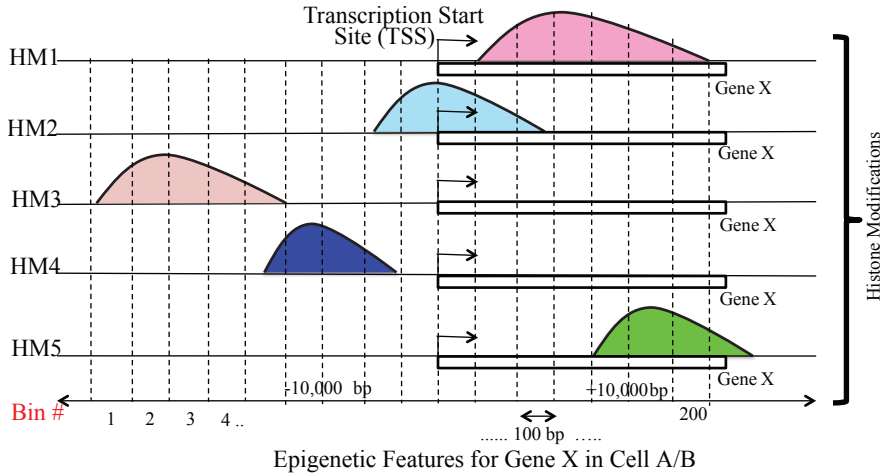
Predictive Modeling of Differential Gene Expression



- Given histone modification profiles $\mathbf{X}^A$ and $\mathbf{X}^B$
- Predict real target value:

$$y_{Diff} = log2\frac{Exp(B)}{Exp(A)}$$

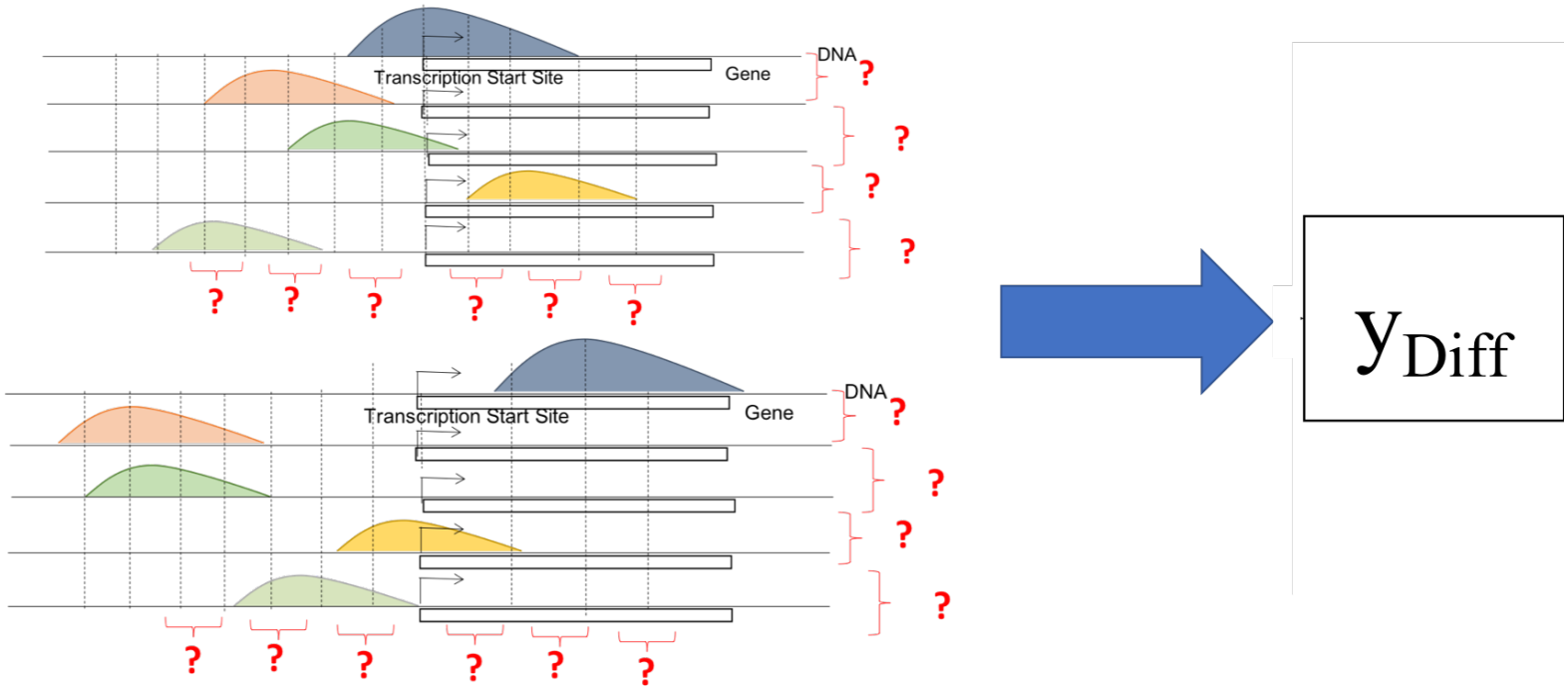
Data Challenge 1: HM signals are spatially structured and long range dependency



- learning methods typically need to use as input features all of the signals covering a DNA region of length 10, 000 base pair (bp) for each HM.
- multiple HMs that are sequentially ordered along the genome direction.

Data Challenge 2: Interpretability

The main goal is to understand what the relevant HM factors are and how they work together to control differential expression in two cell types.

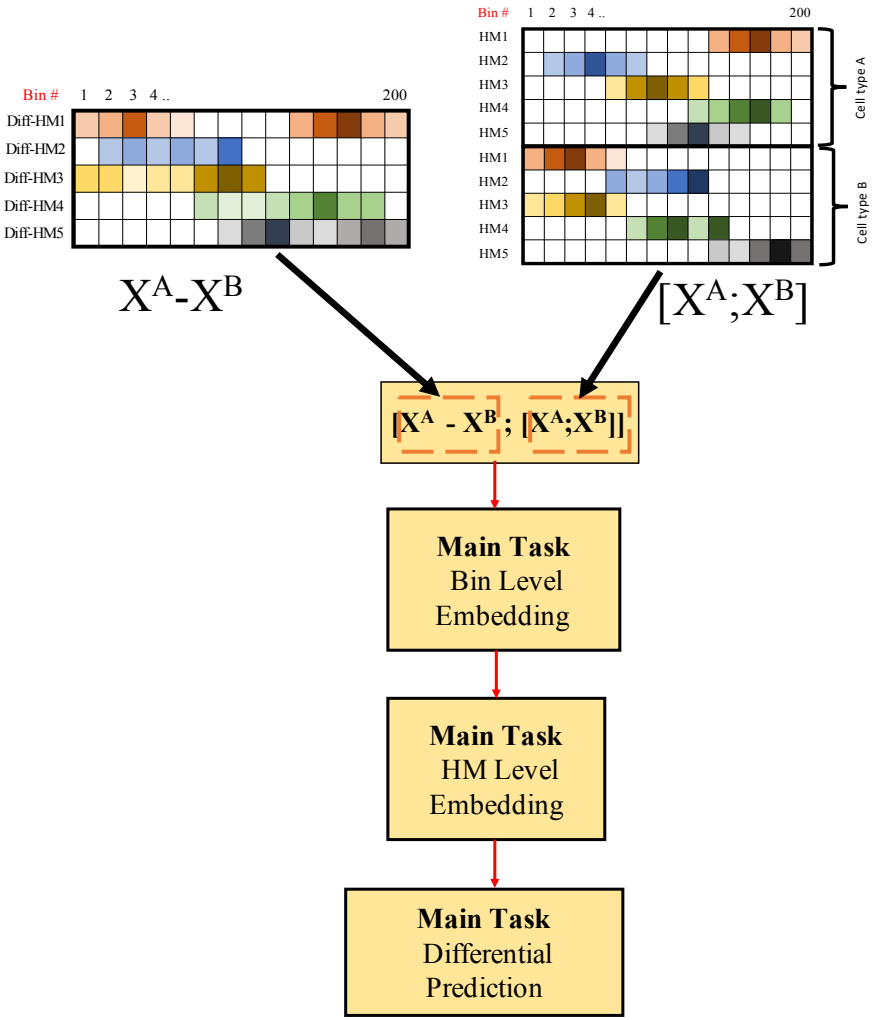


Related Work : Expression Prediction using HMs

Computational Study	Differential	Unified	Bin-Info	Feature Inter.	Interpretable
SVR (single layer) (Cheng and Gerstein [2011])	✓	×	Bin-specific	✓	×
SVR (two layer) (Cheng and Gerstein [2011])	✓	×	×	✓	×
ReliefF+Random Forest (Li et al. [2015])	✓	×	×	×	×
AttentiveChrome(Singh et al. [2017])	×	✓	Automatic	✓	✓

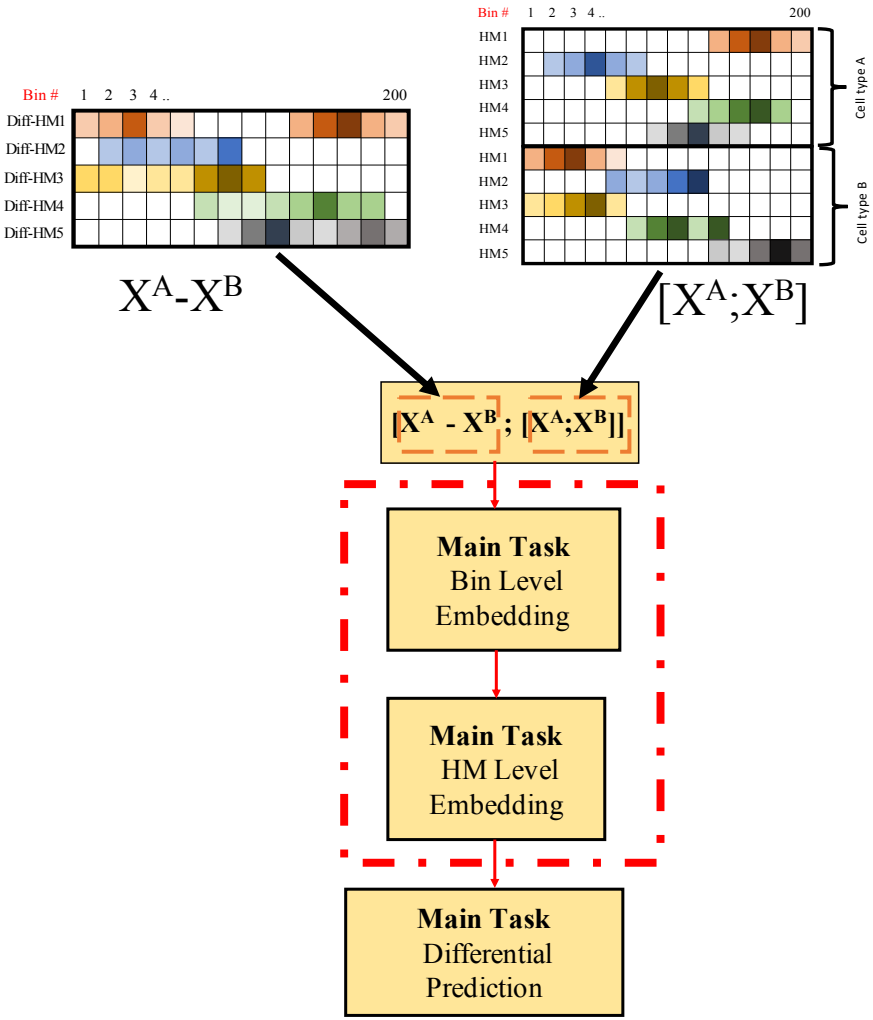
Outline

DeepDiff: Overview



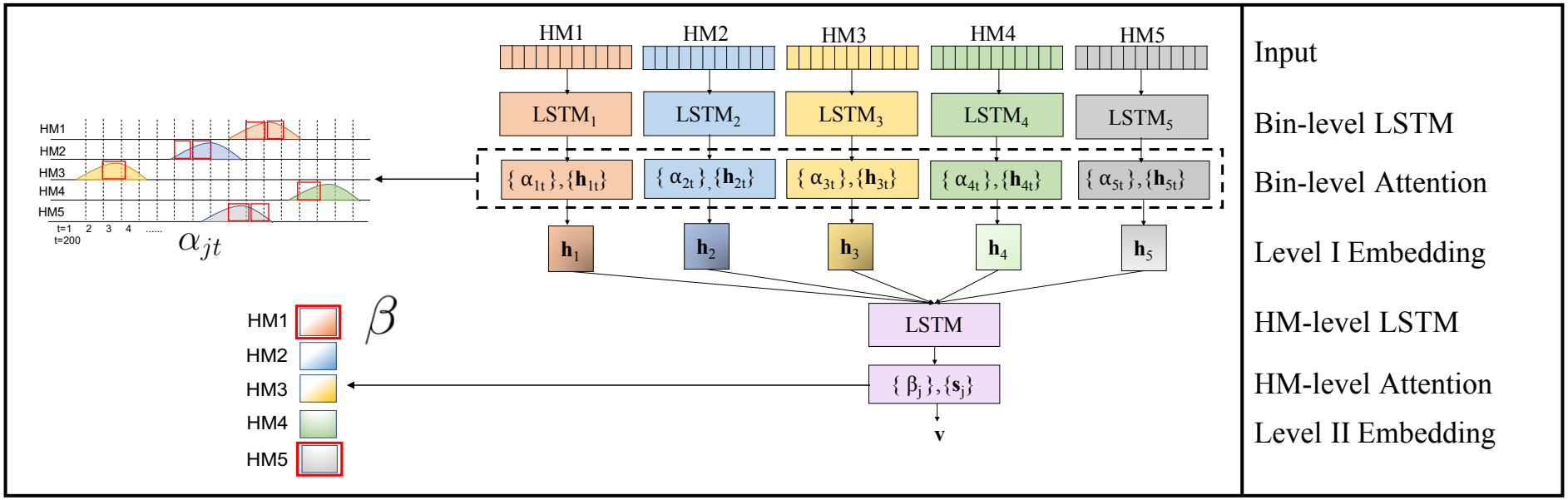
- **DEEP**-learning for predicting **DIFF**erential gene expression from histone modifications.
- Modular, interpretable and multitasking framework.

DeepDiff: Overview



- **DEEP**-learning for predicting **DIFF**erential gene expression from histone modifications.
- Modular, interpretable and multitasking framework.

DeepDiff basic module: Modularity and Interpretability



- Bin level attention(α) : relative importance of each bin or genome coordinate position for prediction
- HM-level attention(β): relative importance of each HM for prediction

Multitasking with DeepDiff

Improve representations using information from related tasks:

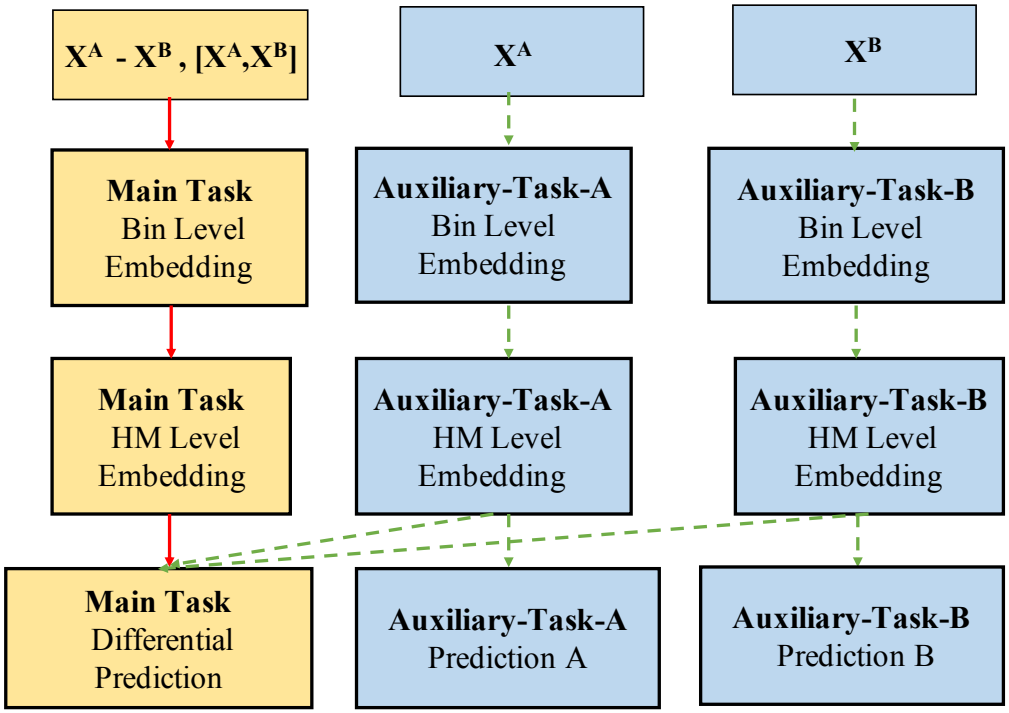
- Cell-specific prediction auxiliary task
- Siamese auxiliary loss term

Multitasking with DeepDiff

Improve representations using information from related tasks:

- $\rightarrow$ Cell-specific prediction auxiliary task
- Siamese auxiliary loss term

Auxiliary Task I: Cell-specific Prediction



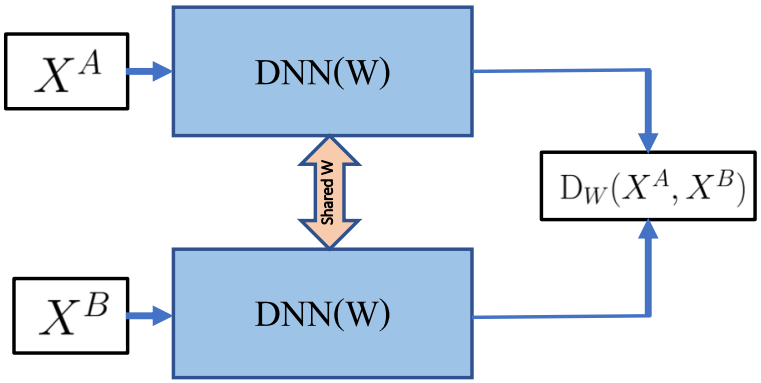
- Leverage Cell type specific gene expression
- In each cell type A and B, also predict Cell-Specific gene expression
- $Loss = \ell_{Diff} + \ell_{CellAux}$

Mutlitasking with DeepDiff

Improve representations using information from related tasks:

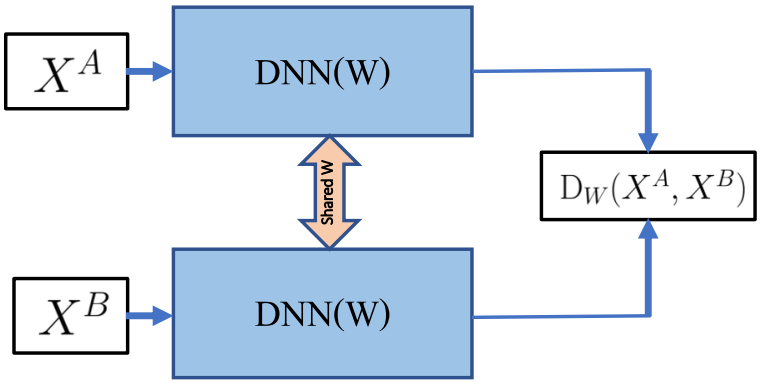
- Cell-specific prediction auxiliary task
- $\rightarrow$ Siamese auxiliary loss term

Auxiliary Task II: Siamese Auxiliary with Contrastive Siamese loss



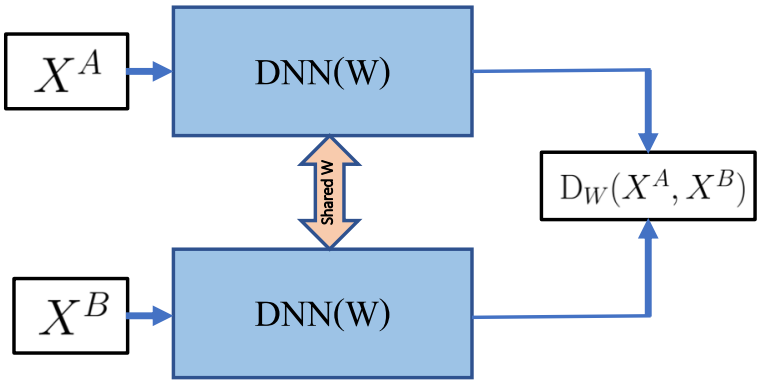
- Siamese Architecture: Twin DNNs with shared weights

Auxiliary Task II: Siamese Auxiliary with Contrastive Siamese loss



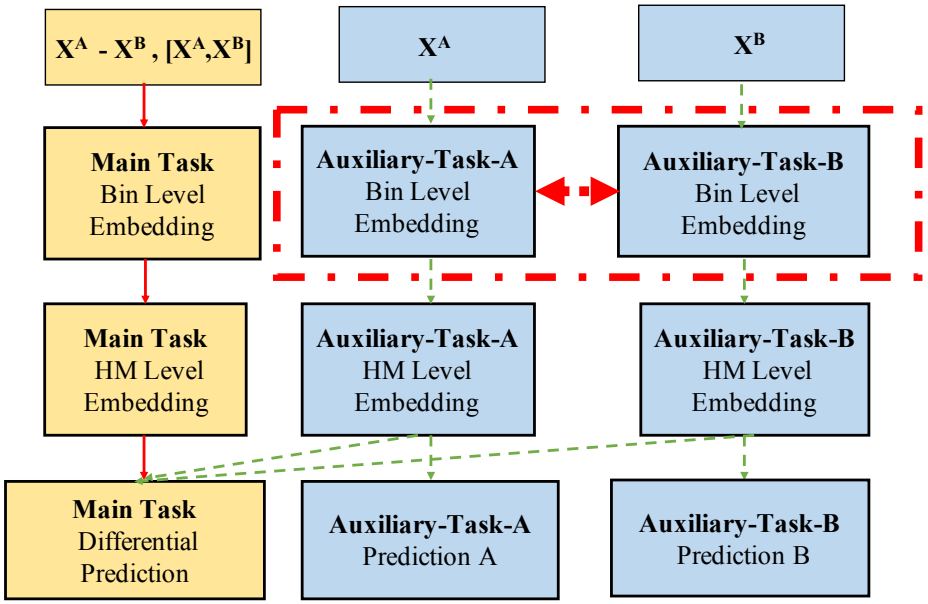
- Siamese Architecture: Twin DNNs with shared weights
- Encourage embeddings to map similar input vectors to nearby points on the output manifold.
- dissimilar input vectors to distant points

Auxiliary Task II: Siamese Auxiliary with Contrastive Siamese loss



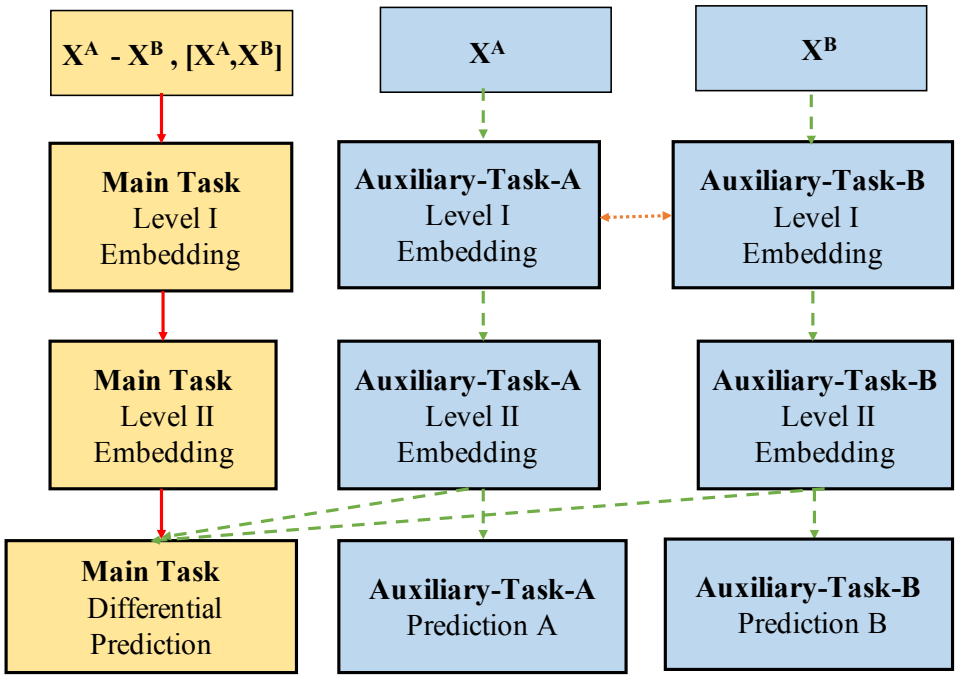
- Siamese Architecture: Twin DNNs with shared weights
- Encourage embeddings to map similar input vectors to nearby points on the output manifold.
- dissimilar input vectors to distant points
- $\ell_{Siamese} = \frac{1}{2}((1-S) \times D_W + S \times max(0, m - D_W)^2)$

Auxiliary Task II: Siamese Auxiliary with Contrastive Siamese loss



- Bin-level Embedding: Siamese Twin Network
- Map $\mathbf{X}^A$ and $\mathbf{X}^B$ embeddings nearby if not differentially expressed
- further if downregulated or upregulated
- $Loss = \ell_{Diff} + \ell_{Siamese}$

DeepDiff: Variations



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

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

3 **Siamese Auxiliary**: Siamese contrastive loss

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

Two types of auxiliary tasks coupled with the main DeepDiff task of differential gene expression.

Outline

Experiments: DeepDiff Variations

A number of possible combinations of raw features and auxiliary tasks:

- **Raw:d**: Raw Difference Features ($\mathbf{X}^A - \mathbf{X}^B$)
- **Raw:c**: Concatenation of Raw HM features ($[\mathbf{X}^A, \mathbf{X}^B]$)
- **Raw**: Concatenation and difference of raw HM features
- **Aux**: Auxiliary Embeddings as Features
- **Raw+Aux**: Concatenation and Difference of HMs + Embeddings from Auxiliary tasks
- **Aux+Siamese**: Auxiliary Features with Siamese Contrastive Loss
- **Raw+Aux+Siamese**: Raw and Auxiliary Features with Siamese Contrastive Loss

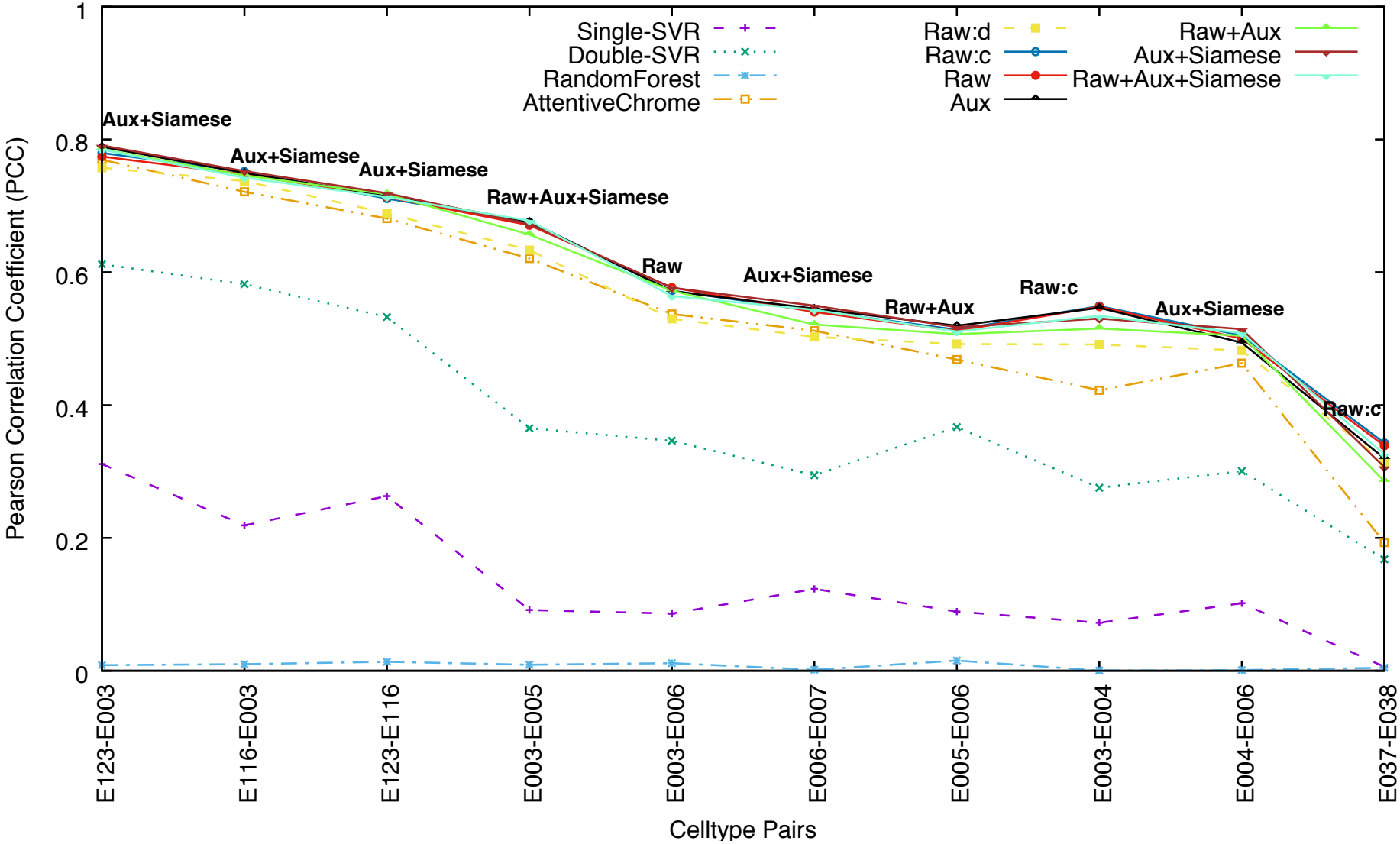
Experiments: Dataset and Setup

- Expression and Histone Modification data from REMC database(Kundaje et al. [2015]).
- HMs: H3K4me1,H3K4me3,H3K9me3,H3K27me3,H3K36me3
- 10 pairs of cell types (10 cases)
- training: 10000 genes
- validation: 2360 genes
- test: 6100 genes

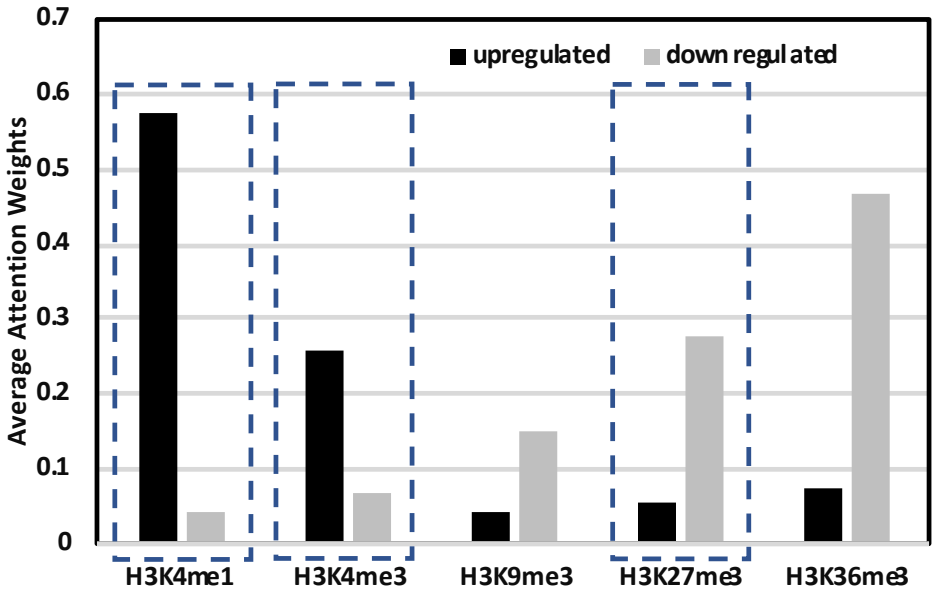
Experiments: Baselines

- Single Layer Support Vector Regression(Cheng and Gerstein [2011])
- Two layer Support Vector Regression(Cheng and Gerstein [2011])
- ReliefF Feature Selection + Random Forest(Li et al. [2015])
- AttentiveChrome(Singh et al. [2017])

Results: Performance Evaluation



Results: Interpreting differential regulation using attention weights



- For the top upregulated genes, H3K4me1 and H3K4me3 get the highest weights.
- For top down-regulated genes H3K27me3 gets the second highest attention weight
- consistent with observations by Grégoire et al. [2016]

Conclusion

- An interpretable, modular and multitasking deep learning framework for predicting differential gene expression from histone modifications
- Outperforms baselines evaluated on 10 different pairs of cell types

Conclusion

- An interpretable, modular and multitasking deep learning framework for predicting differential gene expression from histone modifications
- Outperforms baselines evaluated on 10 different pairs of cell types
Future Work:
- Evaluate Attention more thoroughly for new insights.
- Add TF and other signals.

Acknowledgements

Machine Learning and Bioinformatics Lab

Dr. Yanjun Qi

Dr. Ritambhara Singh

Travel Fund

Thank you!
Code available at– DeepDiffChrome: `deepchrome.org`
Github URL: `https://github.com/QData/DeepDiffChrome`

References

C. Cheng and M. Gerstein. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic acids research*, 40(2):553–568, 2011.

L. Grégoire, A. Haudry, and E. Lerat. The transposable element environment of human genes is associated with histone and expression changes in cancer. *BMC genomics*, 17(1):588, 2016.

A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

J. Li, T. Ching, S. Huang, and L. X. Garmire. Using epigenomics data to predict gene expression in lung cancer. In *BMC bioinformatics*, volume 16, page S10. BioMed Central, 2015.

R. Singh, J. Lanchantin, A. Sekhon, and Y. Qi. Attend and predict: Understanding gene regulation by selective attention on chromatin. In *Advances in Neural Information Processing Systems*, pages 6785–6795, 2017.