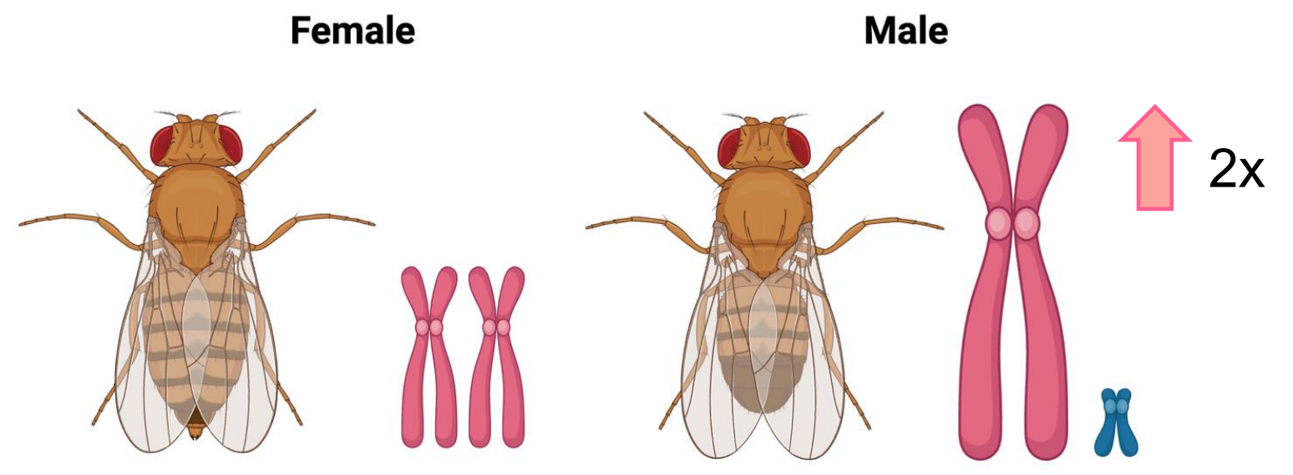
Predicting dosage compensation complex (DCC) binding sites on the X-chromosome vs. autosomes

Sarah Gunasekera[1], **Eric Ewing**[3], **and Erica Larschan**[1,2]
[1] Center for Computational Molecular Biology, Brown University, Providence, RI, 02912
[2] Molecular Biology, Cell Biology, and Biochemistry Department, Brown University, Providence, RI, 02912
[3] Computer Science Department, Brown University, Providence, RI, 02912

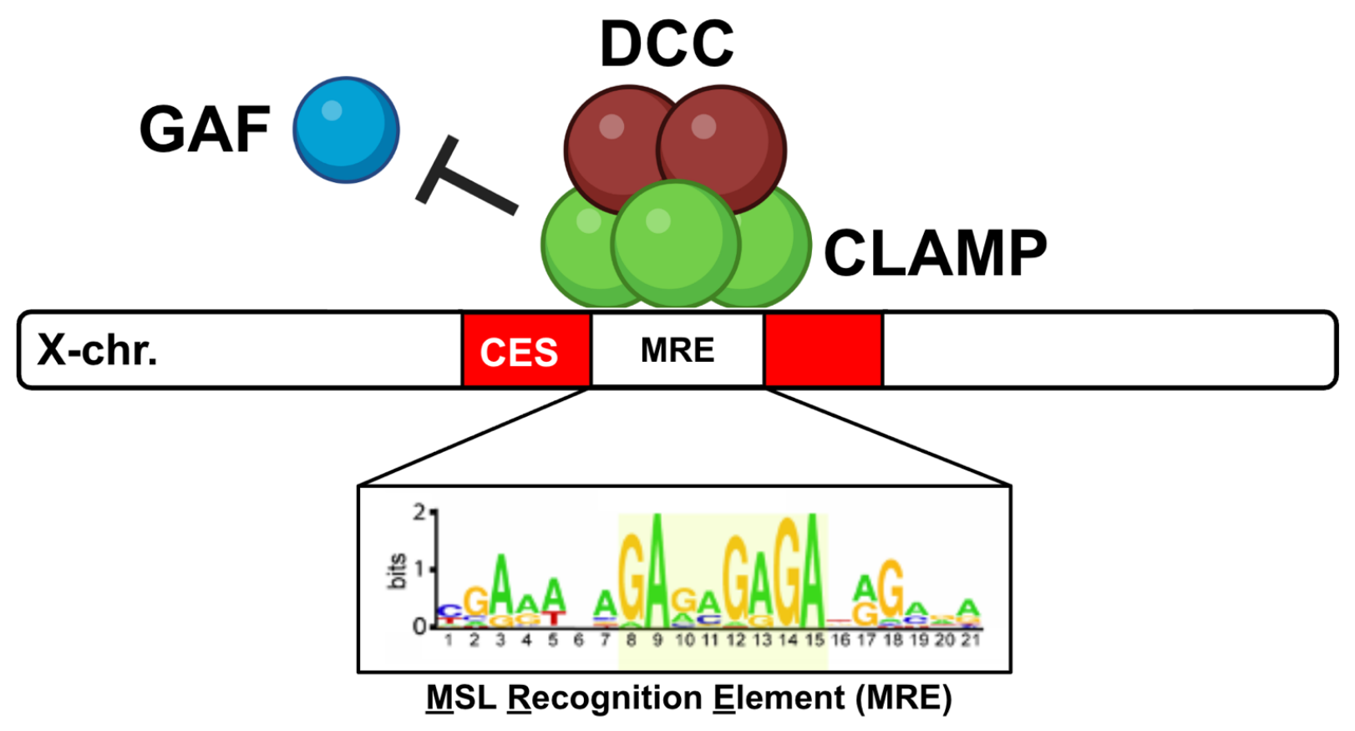
Introduction

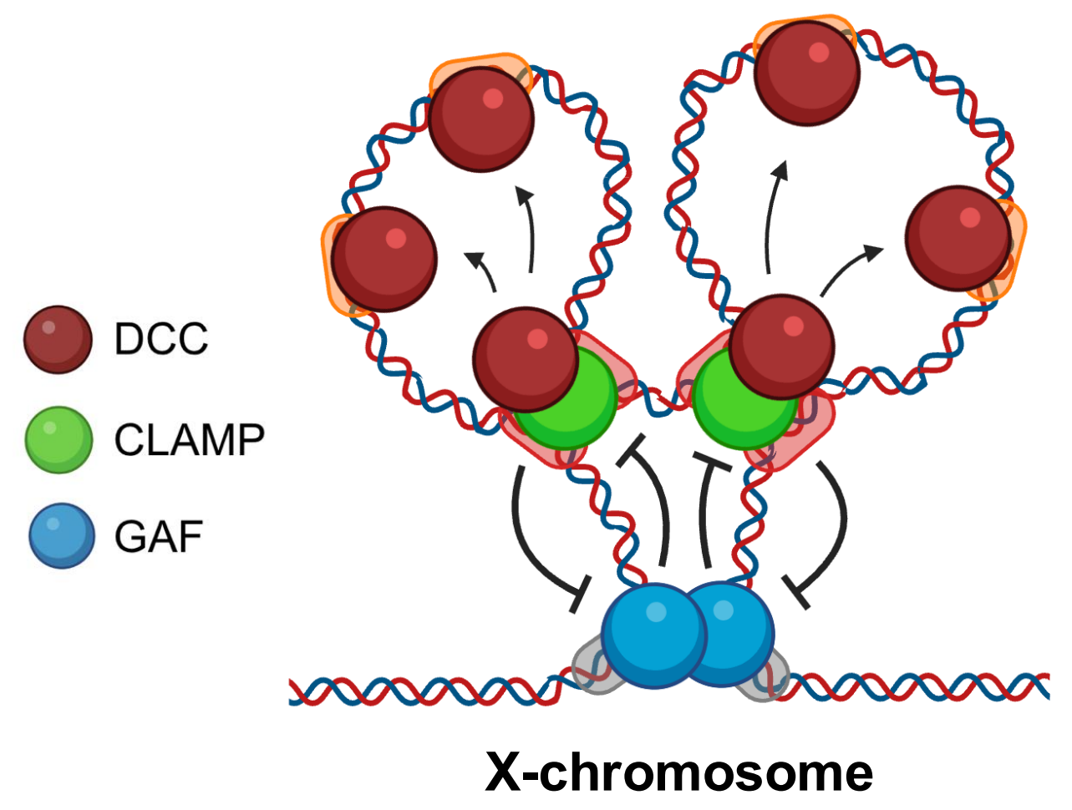
Dosage compensation equalizes X-linked gene expression between males and females in mammals to contribute to sexual dimorphism. In *Drosophila,* there is an observed upregulation of genes 2-fold on the male X-chromosome.

- CLAMP (chromatin-linked adapter for MSL proteins) is a pioneer transcription factor that has a diverse context specific function in transcription and dosage compensation.



- CLAMP binds to GA-rich sequence elements, referred to as MSL recognition elements (MREs), that are distributed across the *Drosophila* genome. However, when the dosage compensation complex (DCC) binds the X-chromosome, CLAMP directs it to Chromatin Entry Sites (CES), which are specific regulatory regions that contain MREs.
- The GAGA-associated transcription factor (GAF) also recognizes MRE sites but differs from CLAMP in which it does not associate with the DCC.
- A proposed binding model from the Larschan Lab has defined GAF binding to moderate larger chromatin loops, while CLAMP mediates smaller chromatin loops that recruit DCC factors on the X-chromosome.





Proposed Questions

Is there a unique epigenomic profile on the X-chromosome vs. Autosomes?

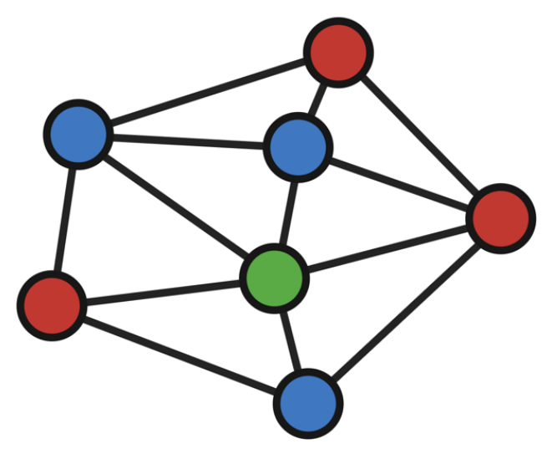
What are the most important 3D interactions on the X-chromosome?

What additional factors are driving dosage compensation to the X-chromosome?

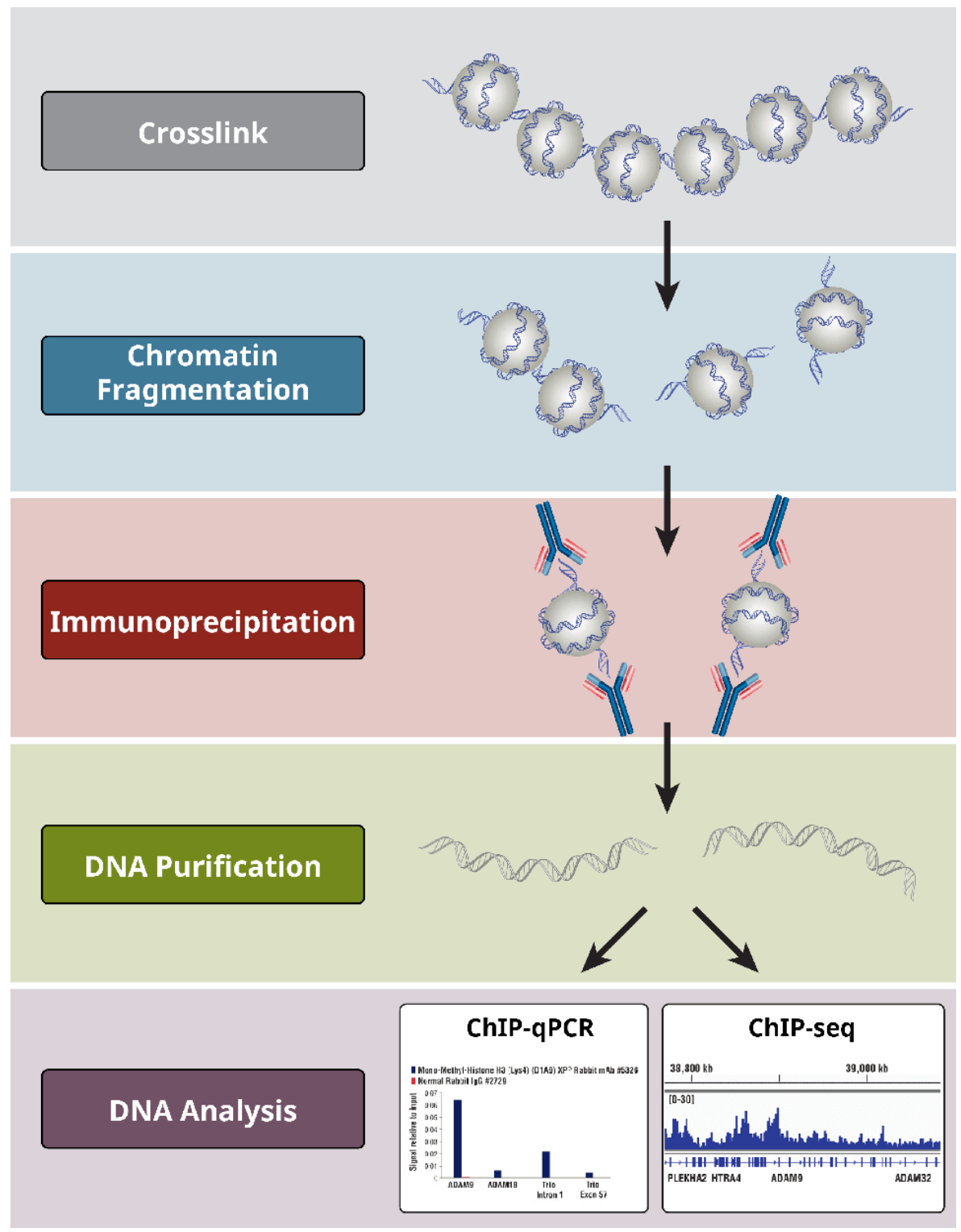
Datasets

Nodes:
- DNA Sequence Data (*Drosophila* genome)
- ChIP-seq Data (Histone Marks)
  - Epigenetic signals

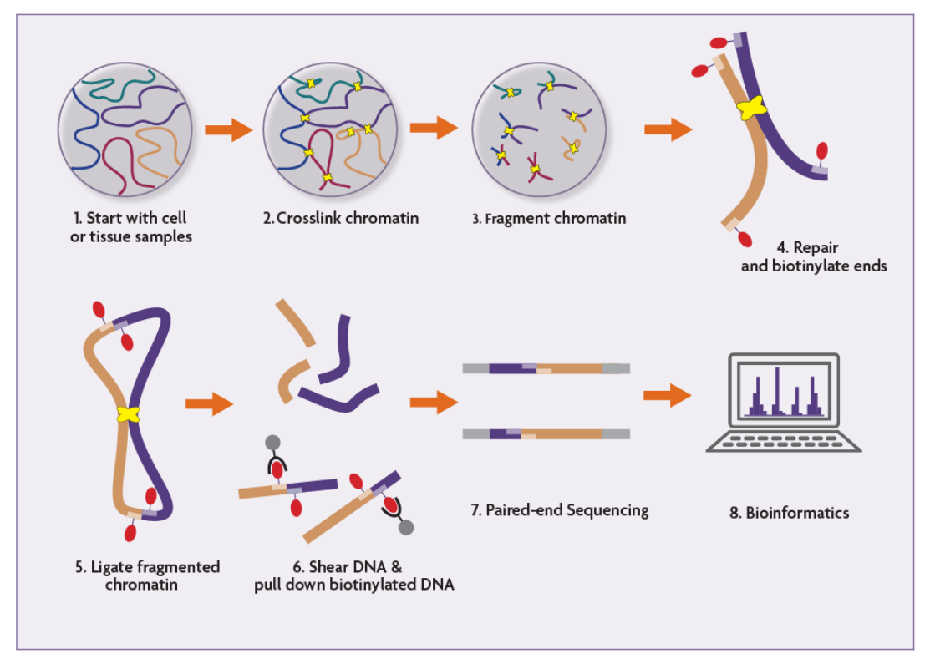
Edges:
- Micro-C Data
  - 3D interactions



ChIP-seq Method



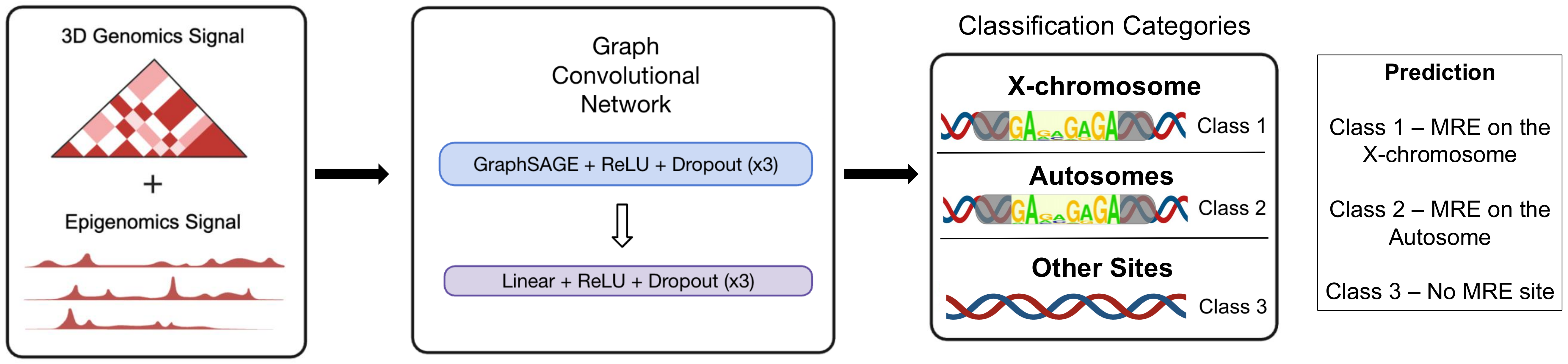
Micro-C Method


1. Start with cell or tissue samples
2. Crosslink chromatin
3. Fragment chromatin
4. Repair and biotinylate ends
5. Ligate fragmented chromatin
6. Shear DNA & pull down biotinylated DNA
7. Paired-end Sequencing
8. Bioinformatics

ChIP-seq Datasets:
h3k27ac, h3k27me3, h3k26me3, h3k4me1, h3k4me2, h3k4me3, h3k9me3m h4k16ac

Graph Neural Network

Graph Convolutional Network Architecture



3D Genomics Signal
+
Epigenomics Signal

Graph Convolutional Network

GraphSAGE + ReLU + Dropout (x3)

Linear + ReLU + Dropout (x3)

Classification Categories

X-chromosome
Class 1

Autosomes
Class 2

Other Sites
Class 3

Prediction

Class 1 – MRE on the X-chromosome

Class 2 – MRE on the Autosome

Class 3 – No MRE site

Hyperparameter Tuning and Model Results

Optimized Parameters

Model Parameters:
- Hidden GNN Size – 128
- Number of GNN Layers – 3
- Hidden Linear Size – 128
- Number of Linear Layers – 3
- Dropout – 0.3
- Normalize – True
- Batch Size – 256
- MRE Negative Samples – 0%*

Optimization:
- Optimizer Type – AdamW
- Learning Rate – 0.0005
- Weight Decay – 0.0005
- Class Weights – True
- Epochs – 35

Neighbor Loader:
- Number of Neighbors – 10

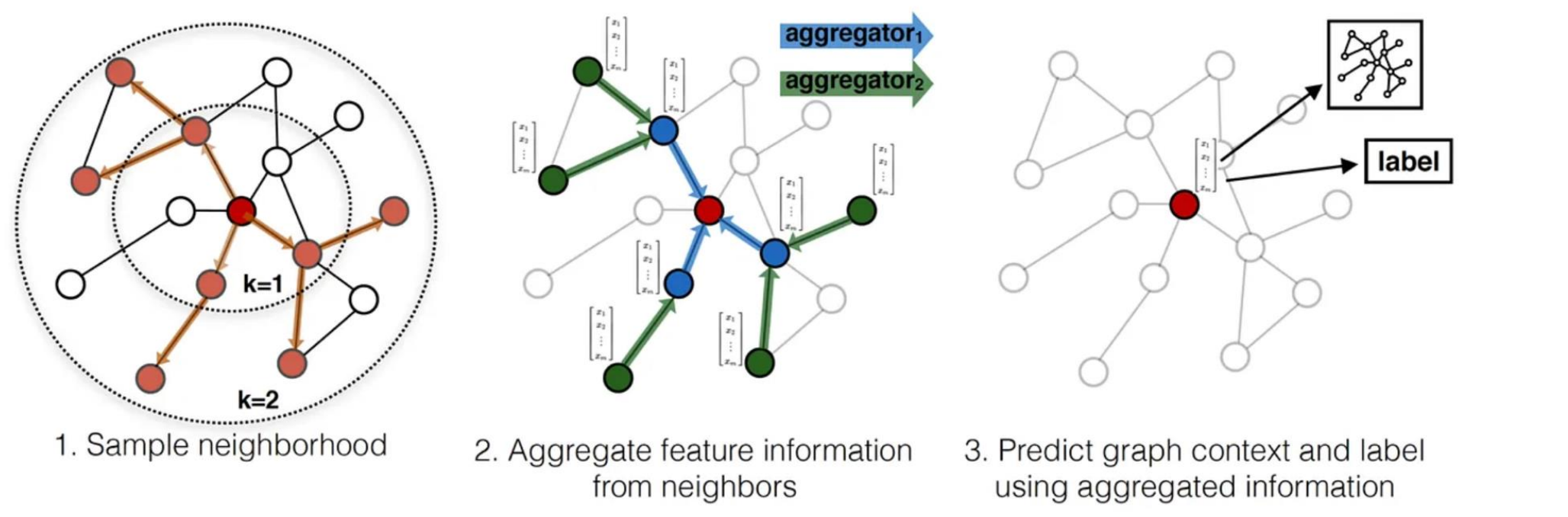

k=1
k=2
aggregator1
aggregator2
label
1. Sample neighborhood
2. Aggregate feature information from neighbors
3. Predict graph context and label using aggregated information

Figure 1: Torch Geometric Neighbor Loader. This method learns a function that generates embeddings by sampling and aggregating features from a node's local neighborhood to reduce the receptive field of each node. The 'Number of Neighbors' parameter was optimized in training and was particularly helpful in reducing the complexity of my graph.

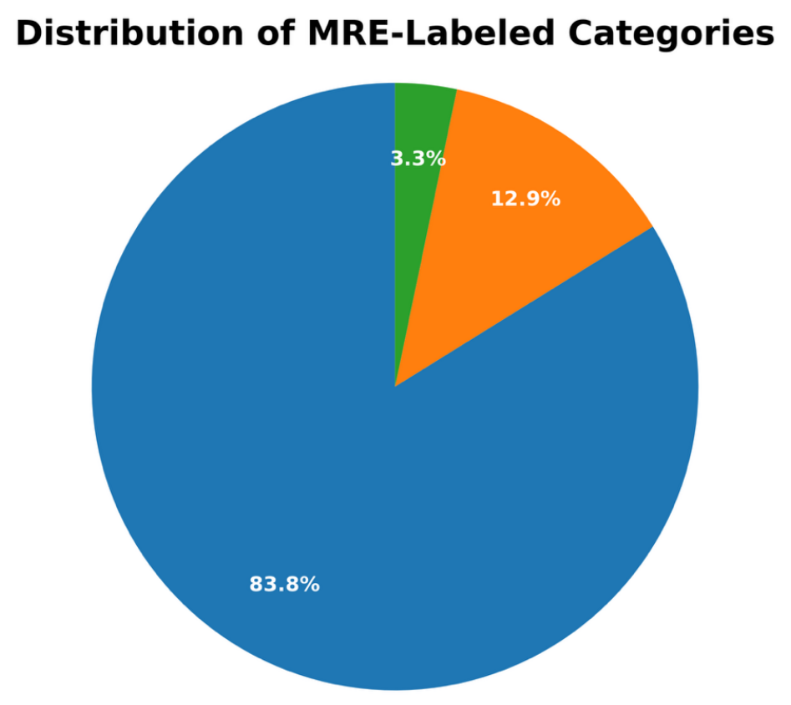


Figure 2: Pie chart to represent the distribution of the three class labels within the dataset.
Blue – non MRE site;
Orange – MRE on the autosome; Green – MRE on the X-chromosome.

The unbalanced distribution highlights the need to introduce 'Class Weights' before training.

WandB Optimal Sweep Results

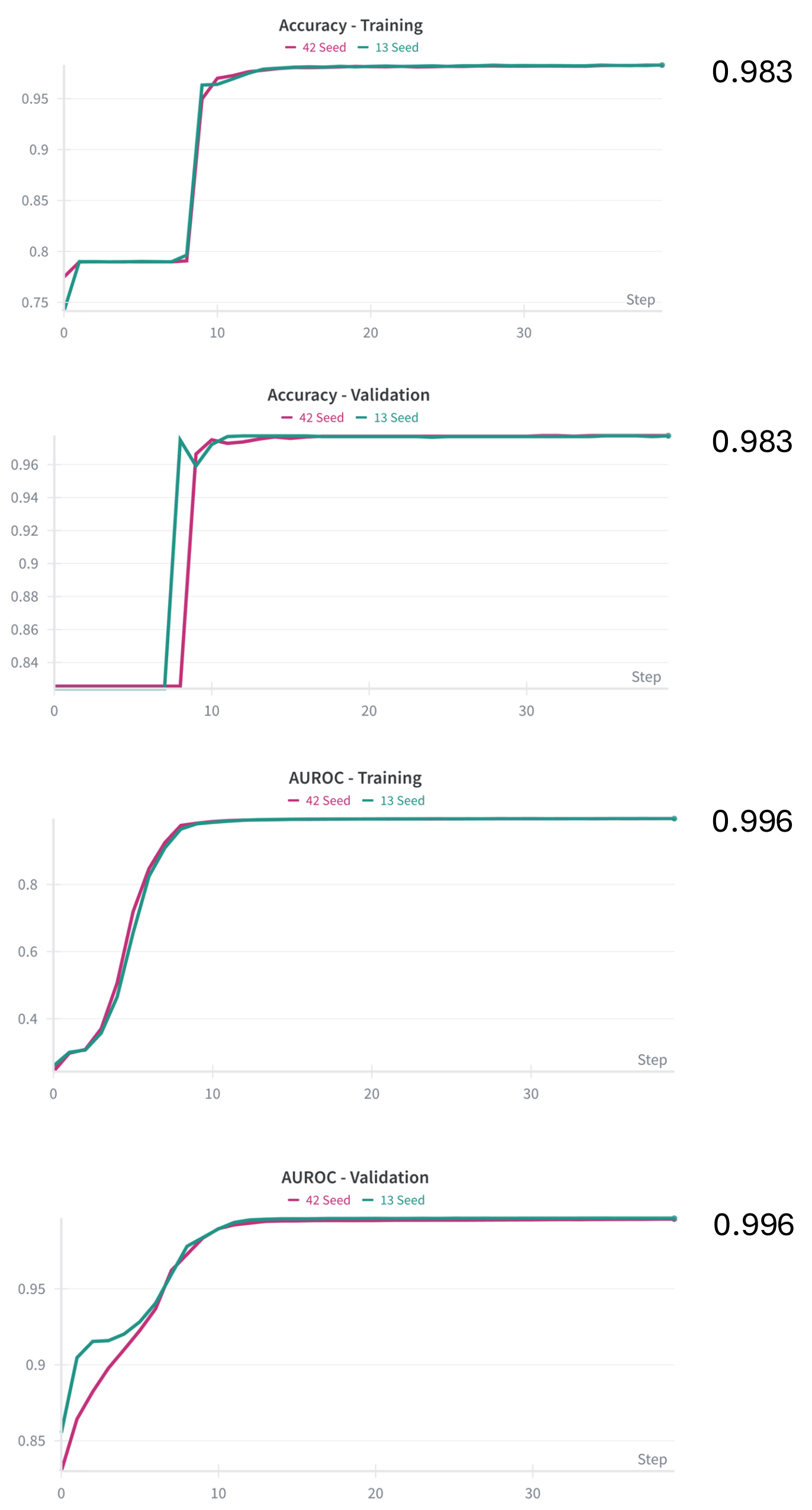


Figure 3: Training and Validation Accuracy and Area under the ROC curve (AUROC) graphs from optimal WandB runs identified through multiple hyperparameter sweeps. After selecting the best performing configurations, two independent seeds were trained to assess reproducibility.

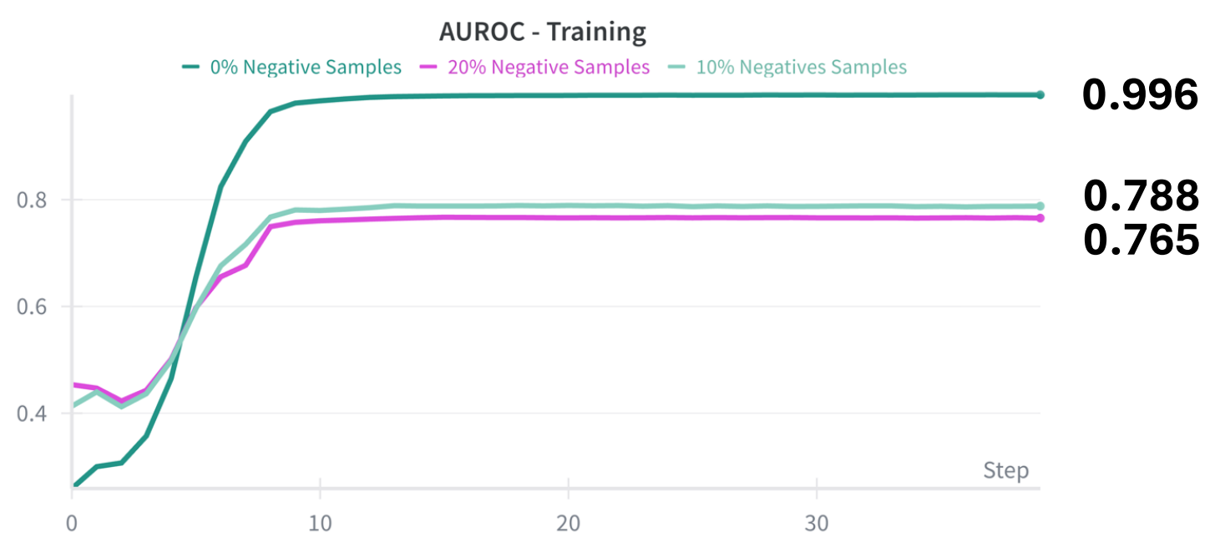


Figure 4: Performance on 'MRE Negative Samples'*, a designated fourth class, was evaluated using the best performing configurations. Increasing the proportion of these samples reduced performance.

Nodes of Importance from GNNExplainer

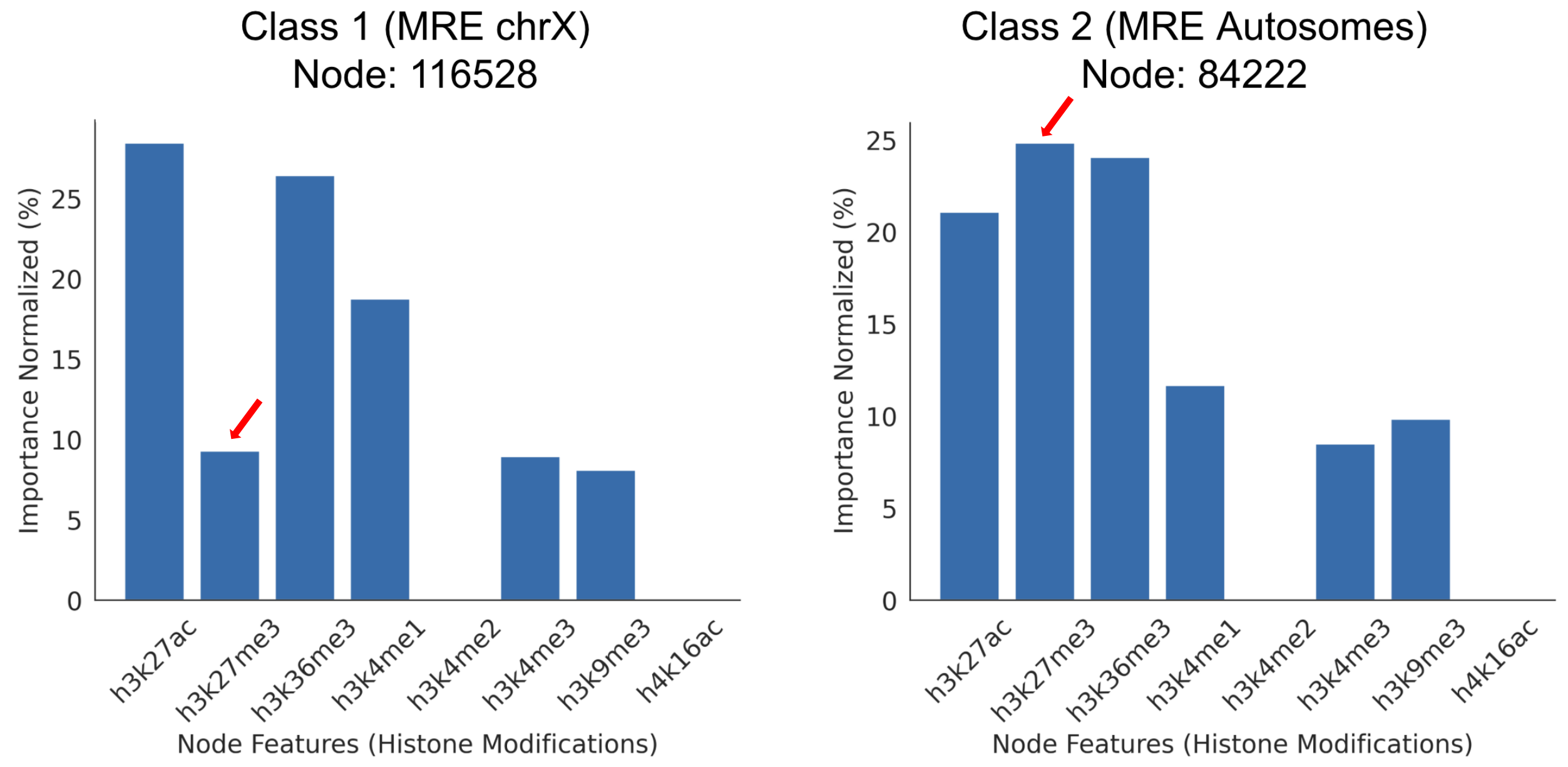


Figure 5: Bar graphs showing the 'important nodes' identified by GNNExplainer for Class 1 and Class 2. The histone mark H3K27me3 exhibits a pronounced difference in importance between the X-chromosome and autosomes in classification.

Discussion

- I learned that graph construction and hyperparameter tuning heavily influence GNN performance and model stability.
- Current limitations include noisy GNNExplainer outputs, a potentially under-constrained classification model suggested by unusually high performance, and performance shifts driven by sensitivity to negative sample proportions.
- Future work includes refining the negative sample strategies, testing graph attention networks, and revisiting the classification scheme to improve biological interpretability of MRE sites (e.g. Class 4: MRE sites within a CES).