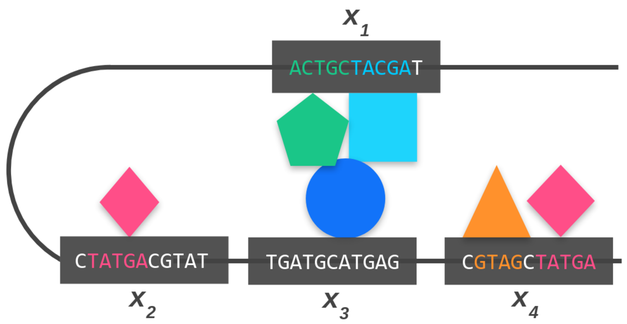
Graph Convolutional Networks for Epigenetic State Prediction Using Both Sequence and 3D Genome Data

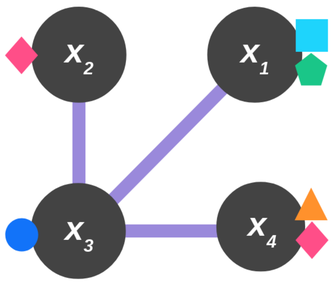
Jack Lanchantin and Yanjun Qi
University of Virginia, Department of Computer Science

Epigenetic State Prediction

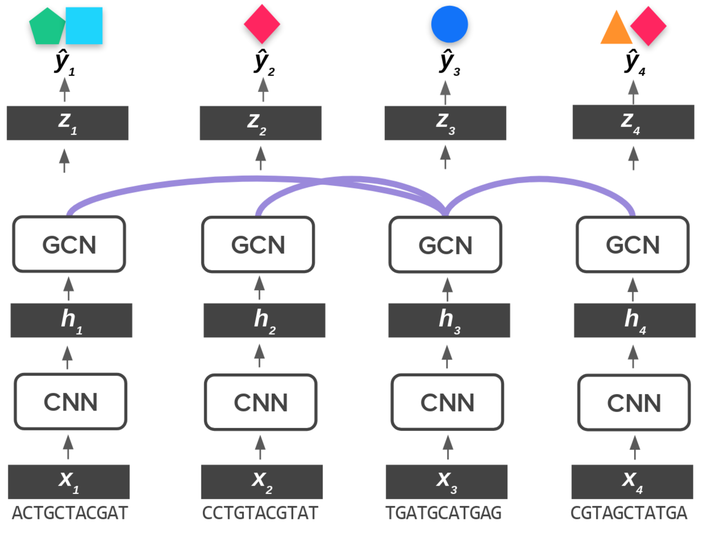


Epigenetic state of DNA window $\mathbf{x}_i$: active chromatin elements at that location (TF binding, histone modifications, and accessibility).

Modeling the 3D Genome Using Hi-C Data



ChromeGCN: Combining Sequence and 3D Genome Data for Epigenetic State Prediction



Method: ChromeGCN

Goal: given DNA window $\mathbf{x}_i$ predict the probability of a set of epigenetic labels $\hat{\mathbf{y}}_i = f(\mathbf{x}_i)$, where $\hat{\mathbf{y}}_i \in \mathbb{R}^L$.

1. Local Sequence Representations Using a CNN

f_{CNN} encodes each local sequence window $\mathbf{x}_i$ into $\mathbf{h}_i \in \mathbb{R}^d$ using a shared convolutional network:

$$\mathbf{h}_i = f_{CNN}(\mathbf{x}_i)$$

2. Long-Range 3D Relationships Using a GCN

f_{GCN} encodes 3D genome relationships between windows $\mathbf{h}_i$ using a graph convolutional network. Relationships are defined by Hi-C edges in the form of adjacency matrix $\mathbf{A}$:

$$\mathbf{z}_i = f_{GCN}(\mathbf{h}_i, \mathbf{A})$$

Each $\mathbf{h}_i^t$ is revised using a function of its neighbors $\mathcal{N}(i)$:

$$\mathbf{z}_i = \sigma\left(\frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \mathbf{h}_j^t \mathbf{W}^t\right)$$

3. Predicting Label Probabilities for Each Window

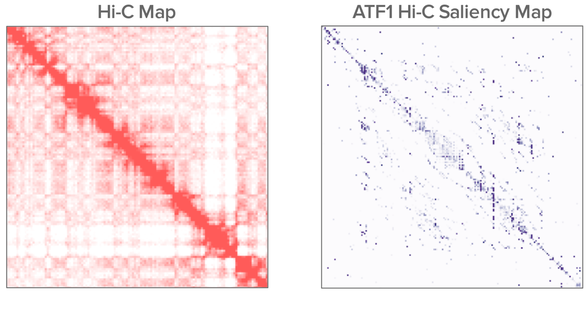
f_{Pred} uses linear classifier layer to classify each $\mathbf{z}_i$ into its output space (a set of epigenetic labels):

$$\hat{\mathbf{y}}_i = f_{Pred}(\mathbf{z}_i)$$

Finding Important Hi-C Edges via Saliency Maps

We propose Hi-C Saliency Maps, a method to identify the important 3D relationships for ChromeGCN's predictions:

$$S_{Hi\text{-}C}^{\ell} = \sum_{i=1}^{N} \mathbf{A} \circ \left|\frac{\partial \hat{\mathbf{y}}_i^{\ell}}{\partial \mathbf{A}}\right|$$

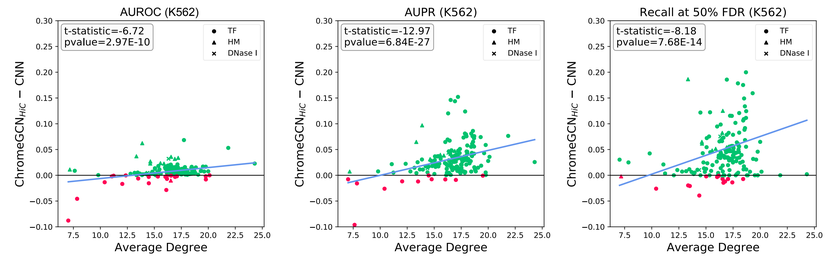


Prediction Performance

	GM12878			K562		
	Mean AUROC	Mean AUPR	Mean Recall at 50% FDR	Mean AUROC	Mean AUPR	Mean Recall at 50% FDR
CNN [1]	0.895	0.350	0.293	0.894	0.325	0.265
DanQ [2]	0.886	0.348	0.290	0.900	0.343	0.290
ChromeRNN	0.906	0.384	0.342	0.910	0.365	0.327
ChromeGCN$_{const}$	0.904	0.377	0.331	0.904	0.358	0.321
ChromeGCN$_{Hi\text{-}C}$	0.904	0.385	0.341	0.903	0.358	0.319
ChromeGCN$_{const+Hi\text{-}C}$	**0.909**	**0.395**	**0.356**	**0.912**	**0.372**	**0.338**

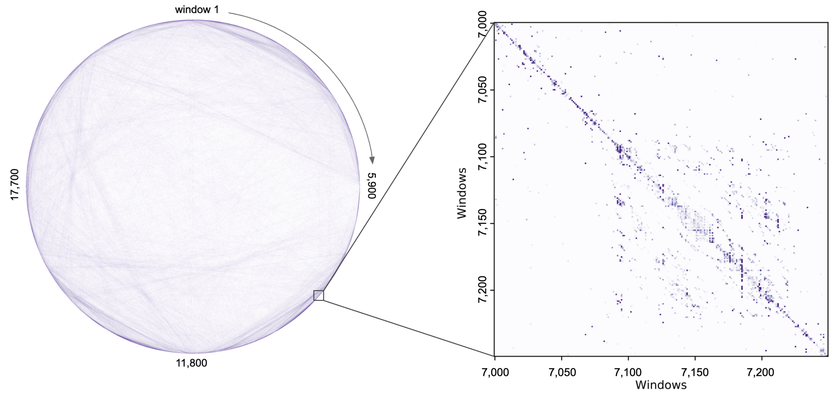
We find that across two cell types, using known long range 3D genome data from Hi-C maps improves prediction accuracy.

Analysis of Using Hi-C Data



Each point represents one epigenetic state label. As the average label degree increases, the improvement of ChromeGCN$_{Hi\text{-}C}$ over the CNN increases.

Hi-C Saliency Map



Saliency Map for all 500k edges in $\mathbf{A}_{Hi\text{-}C}$ for YY1 transcription factor in GM12878 Chromosome 8 (total of 23,600 windows). The darker the line, the more important that edge was for predicting the correct epigenetic state.

References

[1] J. Zhou et al, "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk," *Nature genetics*, vol. 50, no. 8, p. 1171, 2018.

[2] D. Quang and X. Xie, "Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences," *Nucleic acids research*, 2016.