Computational Biology Capstone Project, cse428

Irene Kyeremaa Yeboah

Introduction

Single-cell Hi-C (scHi-C) profiles 3D chromatin architecture and holds information about the cell type specific 3D genome architecture. However, it is sparse to analyze meaningfully, making downstream analyses such as clustering and cell-type identification challenging. In this project, I had explored whether enhancing raw scHi-C matrices using the HiCFoundation model can improve the quality of cell clustering, using the embedding results from Higashi.

Objective

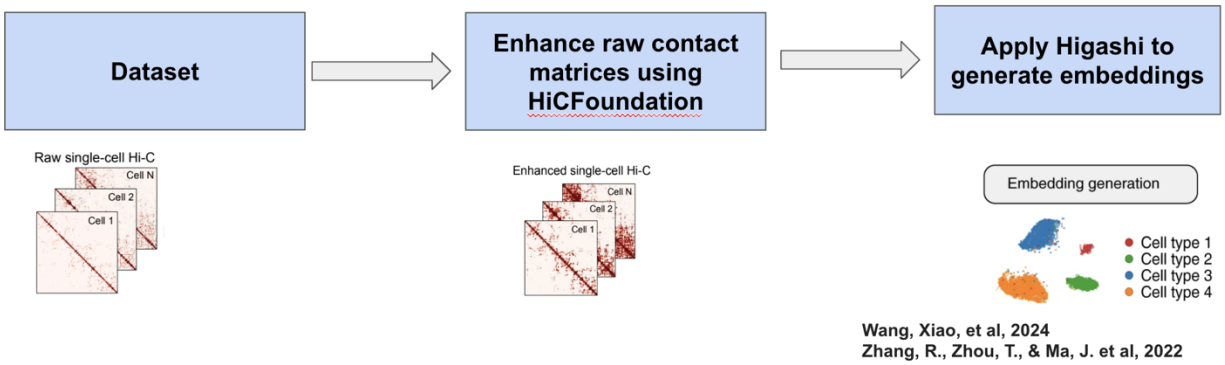
To evaluate the effectiveness of this, we compare two pipelines for cell clustering:

1. Baseline Pipeline: Cluster cells using raw scHi-C matrices.
2. Enhanced Pipeline: Enhance scHi-C matrices using HiCFoundation before clustering.

We hypothesize that enhanced contact maps will better capture biological signals and improve clustering resolution.

Methods

Workflow



Dataset

1. For testing baseline, 4DN Human Cell Line Data from (https://noble.gs.washington.edu/proj/schic-topic-model) with 19,388 cells and 103M chromatin pairs was used
2. For enhanced pipeline,a mouse brain data subsetted from HiRES , a publicly available data from GEO(GSE223917) was used. It contained 400 cells and 21,291,441 chromatin pairs.

Approach

Baseline pipeline

The data that was used here was the same as the mtx files from (https://noble.gs.washington.edu/proj/schic-topic-model/) which had been concatenated and transformed into the Higashi input format, a .txt file with 103,497,337 chromatin interactions. Embeddings were generated and clustering was performed on 16707 cells by 128-dimensional embeddings (16707, 128).

Enhanced pipeline

From the large HiRES data, the mouse brain data which had 400 cells was subsetted. All 400 .pairs where enhanced with HiCFoundation and transformed into the Higashi input format, a .txt. It contained 21,291,441 chromatin interactions across 400 cells. Embeddings were generated using Higashi and clustering was performed on 400cells by 128-dimensional embeddings (400, 128).

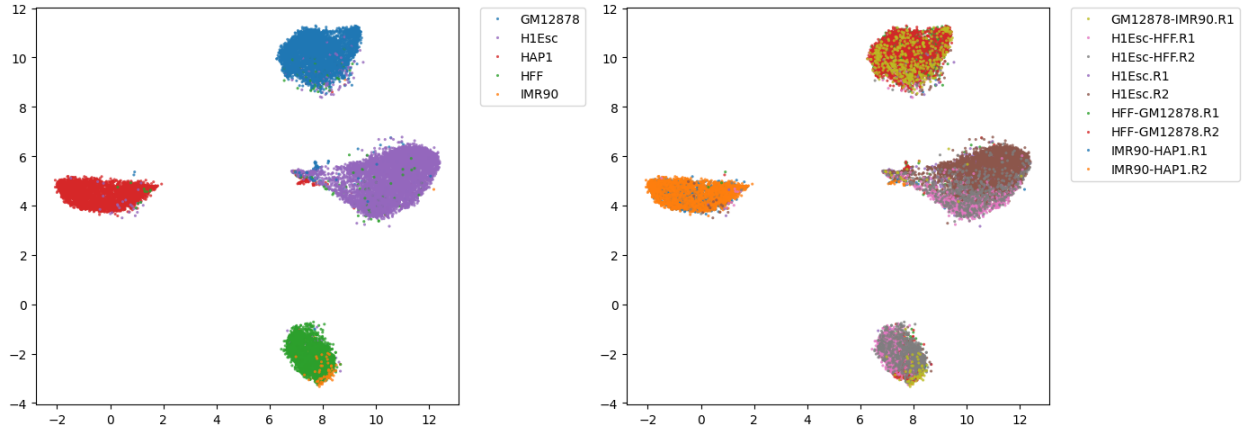
Results

Baseline pipeline

To validate the baseline Higashi pipeline, I reproduced the analysis using the dataset provided by the authors. This dataset includes five distinct human cell types — GM12878, H1Esc, HAP1, HFF, and IMR90 — spanning multiple experimental batches. I used a preformatted .txt file comprising 103,497,337 chromatin interactions across 16,707 cells.

Contact matrices were generated at 5Mb resolution, and the Higashi embedding model was trained for 80 epochs using the default configuration. The resulting 128-dimensional cell embeddings were projected into 2D space using UMAP for visualization.

As shown in Figure below, UMAP embeddings colored by cell type displayed clear and well-separated clusters (left panel), validating Higashi's ability to encode biologically meaningful chromatin structures. Clustering by batch (right panel) showed minimal overlap within cell types.



Enhanced pipeline

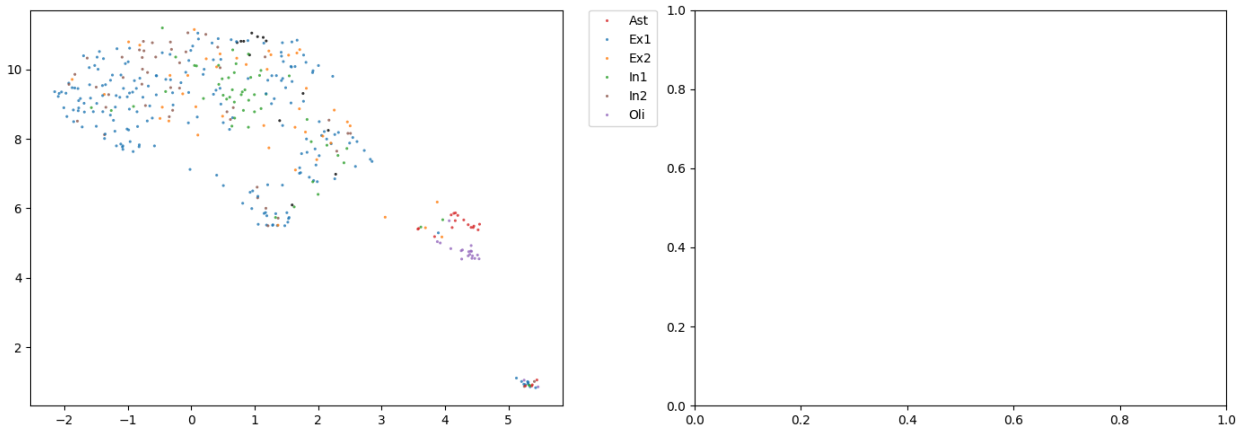
For the enhanced pipeline, I used a subset of the publicly available HiRES mouse brain dataset (GSE223917), selecting 400 .pairs files, each representing a cell. These were enhanced using the HiCFoundation model, which leverages deep learning to impute and densify sparse scHi-C contact maps.

The enhanced contact maps (21,291,441 interactions total) were converted to Higashi format, .txt and used to train a new embedding model. Due to GPU time and storage constraints, the model was trained for only 5 epochs (vs. 80 in the baseline). Contact matrices were built at 1Mb resolution to preserve more structural detail.

UMAP projections of the 128D embeddings are shown in Figure below, with points colored by inferred cell types (e.g., Ex1, Ex2, In1, In2, Ast, Oli). Clusters were not as distinct as in the baseline pipeline. To quantify this:

- Silhouette score: –0.113 (negative → poor cluster separation)
- Adjusted Rand Index (ARI): 0.101 (low agreement with true labels)
- Normalized Mutual Information (NMI): 0.223 (some shared structure, but weak)

These metrics suggest that while some biological signal exists in the enhanced data, the embeddings did not result in well-separated clusters, possibly due to the limited number of cells (n=400) and the shorter training time.



Pipeline	BCE	Accuracy	AUC	Total Sparsity	Embedding training epochs
Baseline	0.3973	84.3%	0.643	0.0128	80
Enhanced	0.7584	52.325%	0.780	0.325	5

Conclusion

This project explored the impact of enhancing sparse scHi-C data using the HiCFoundation model and downstream clustering quality via Higashi embeddings.

The baseline analysis successfully reproduced published results, yielding high classification accuracy and clear UMAP separation between cell types. This confirmed that the Higashi embedding model is effective at capturing biologically meaningful 3D chromatin. In contrast, the enhanced pipeline using HiCFoundation-imputed data from 400 mouse brain cells did not achieve strong clustering performance. Despite improved AUC (0.78 vs. 0.64), the model suffered from low silhouette score (–0.11) and low ARI/NMI, likely due to:

1. Small cell number (n=400) which could be that fewer samples reduce the model's ability to learn distinct chromatin patterns.
2. Short training duration (5 epochs) — due to time and GPU constraints which is insufficient for full convergence.

Better results can be achieved in future experiments by probably increasing the number of cells, training with more samples and increasing the training epochs.