

scHi-C

Tensor Decomposition

Rancho Tao
BIMSA Math & AI
[RanchoTao @ gmail.com](mailto:RanchoTao@gmail.com)
RanchoTao.github.io

Overall, we aim to address a problem:

How to Understand Single-Cell Data?

single-cell Hi-C data represents one of the most sparse and complex data types in contemporary bioinformatics,
characterized primarily by **high noise levels**.

What is Single-Cell Data?

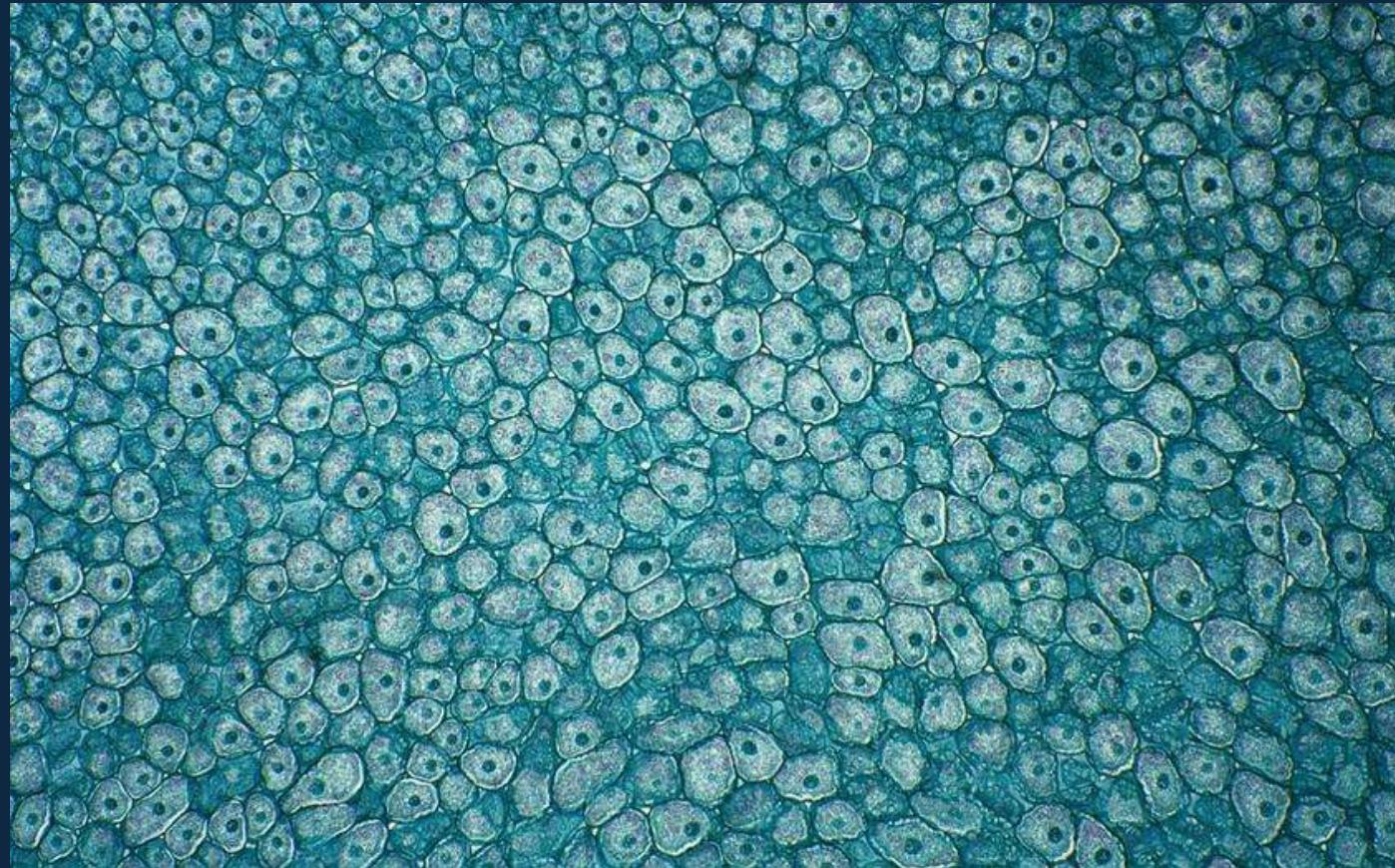
Single-cell data refers to **biomolecular information**
(such as gene expression, chromatin accessibility,
protein content, etc.)
measured at the single-cell resolution.

Bulk Sequencing



Single-cell Sequencing





Why measure
single-cell data?

Within a biological organism,
even within the same tissue !!!
(such as the brain or a tumor),
there exists significant cellular heterogeneity.

The “Appearance” (Data Structure) of Single-Cell Data

From a computational perspective, single-cell data typically manifests as a massive matrix:

Rows: Usually represent cells (potentially ranging from thousands to millions).

Columns: Typically represent features (e.g., 20,000 genes, or millions of bin regions across the genome).

Values: Represent the intensity of that feature in the cell (e.g., expression levels or contact frequency).

It presents two significant algorithmic challenges:

High-dimensional: Extremely numerous features.

Sparse: This is the most challenging aspect of single-cell data, as sampling depth is limited, resulting in the vast majority of values in the matrix being zero (Dropout phenomenon).

How to resolve
this issue?

Tensor Decomposition

Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi
(<https://doi.org/10.1016/j.cels.2022.09.004>)

A comprehensive benchmark of single-cell Hi-C embedding tools
(<https://doi.org/10.1038/s41467-025-64186-4>)

Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome
(<https://doi.org/10.1186/s13059-020-01977-6>)

Multiscale and integrative single-cell Hi-C analysis with Higashi
(<https://doi.org/10.1038/s41587-021-01034-y>)

A fast algorithm based on tensor decomposition (PARAFAC2) introduces the concept of “meta-interactions” directly linking cellular embeddings to 3D genomic features.

Advantages:

9–40 times faster than traditional methods (e.g., Higashi, 3DVI), capable of identifying rare cell types (e.g., cortical neuron subtypes).

Addresses data sparsity through partial random walk (Partial RWR) processing.

Key Applications: Reconstructs neuronal developmental trajectories in mouse developmental brain data and identifies cell type-specific chromatin interactions.

Core Content: Systematically evaluated the performance of 13 embedding tools (e.g., Va3DE, Higashi, scHiCluster) across 10 scHi-C datasets.

Key Findings:

Different tools suit distinct scenarios (e.g., early embryonic development relies on long-range interactions, while complex tissues require short-range interactions).

Deep learning tools (e.g., Va3DE, Higashi) demonstrate robustness on low-coverage data but incur high computational costs.

Random walk and IDF transformation may bias toward long-range interactions, potentially affecting neuronal subtype differentiation.

Avocado: a multi-scale deep tensor factorization method
learns a latent representation of the human epigenome
(<https://doi.org/10.1186/s13059-020-01977-6>)

Core Method: Higashi models scHi-C data as a hypergraph
—connecting cell nodes and genomic bin nodes via hyperedges (representing chromatin interactions)
—and learns embeddings using a hypergraph neural network (Hyper-SAGNN).

Innovative Features:

Dynamic embeddings capture intercellular associations, enabling multimodal data integration
(e.g., joint analysis of Hi-C and methylation).

Contact graph interpolation via hyperedge prediction enhances 3D structural resolution in single cells.

Key Advantages: Identifies cell type-specific TAD-like boundaries and compartment dynamics.

Core Method: Extended hypergraph representation learning supporting multi-scale analysis (compartments, TAD boundaries) and multi-modal integration (e.g., scHi-C + methylation).

Key Findings:

Distinguishes 29 neuronal subtypes in human prefrontal cortex data and identifies ODC cell-specific TAD boundaries enriched for synapse-related genes.

Interpolated contact maps reveal cell cycle-associated chromatin structural changes.

Thank for your listening

Rancho Tao
BIMSA Math & AI
RanchoTao @ gmail.com
RanchoTao.github.io