UNIVERSITY of WASHINGTON

Contrastive learning for lineage barcoded scRNA-seq data

Shizhao Joshua Yang, Yixin Wang, Kevin Lin

BIOSTATISTICS
SCHOOL of PUBLIC HEALTH

Motivation

Understanding the history, state, and fate of cells is crucial in biology. **Single-cell lineage tracing (scLT)** technology, using inheritable barcodes, enables the tracking of progenitor and stem cell development, providing insights into gene differential expression (DE) and cell lineage relationships.

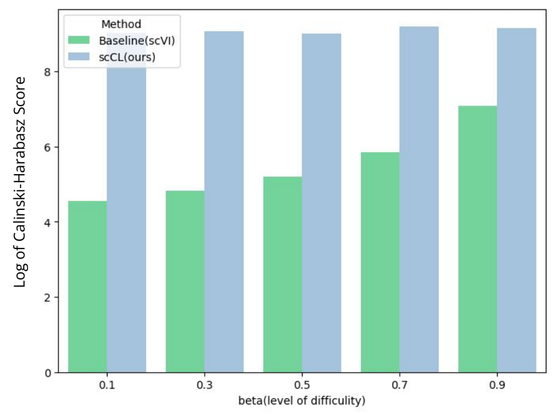
Biddy et al. (2018; Celltag) applied lineage barcodes to learn about cellular reprogramming of fibroblasts. This dataset revealed distinct differences in cell type differentiation behavior across various lineages, including single-fate and multiple-fate (pluripotent) differentiation, as well as the self-renewal of early progenitors. Based on these observations, **a further question arises: Can we learn the gene pathways that differentiate different lineages based on their eventual fates?**

The challenge lies in isolating lineage-specific signals, which are often overshadowed by more dominant signals such as cell type variations. Traditional methods like variational autoencoders (VAEs) excel at reconstructing gene expression but often miss these crucial lineage signals.
To address this, we employ a contrastive learning framework tailored to lineage-barcoded scRNA-seq data. This approach leverages the inherent supervision provided by lineage barcodes, making contrastive learning a powerful tool for capturing and analyzing lineage-specific signals across different time points and states.

Simulation

To test our method's consistency across datasets of varying difficulty, we generated simulation datasets by reassigning lineages to cells a lineage barcoded dataset. The difficulty level is based on how clearly lineage information is represented, measured by the **Calinski-Harabasz score** of scVI's embeddings (higher score is better).

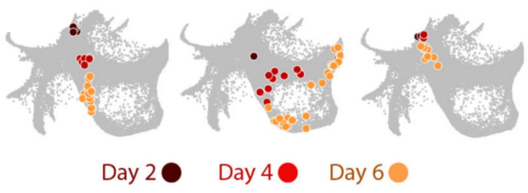
Based on the gene expression matrix, we select a fraction (β) of genes, which are used for a Leiden clustering. We then assign each cell in the same cluster to be part of the same lineage. A higher β is an "easier" dataset since more genes are related to the lineage assignment. We created five datasets, ranging from easy to hard, by varying β values at 0.1, 0.3, 0.5, 0.7, and 0.9.

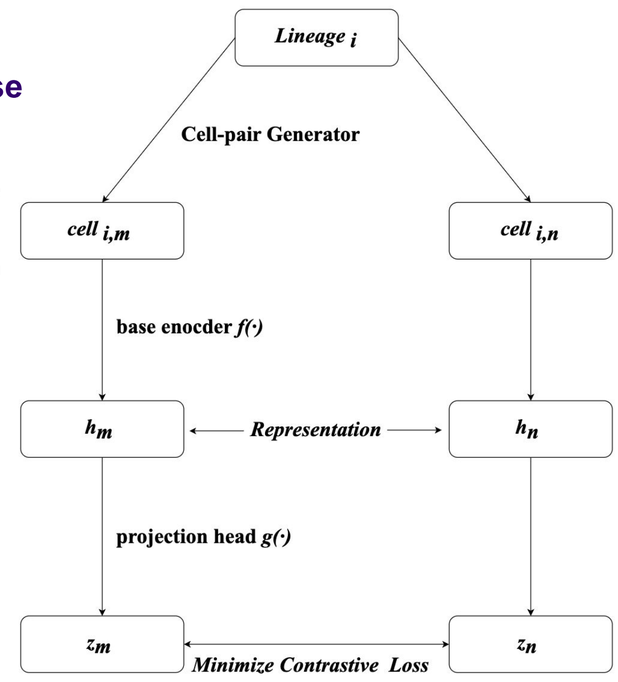


Comparison of scVI and scCL at Different Beta's (Level of Difficulties) (Log Scale)

Framework

The single-cell contrastive learning framework consists of three main components: **cell-pair generator**, **base encoder**, and **projection head**.





Cell-pair Generator: our method defines positive pairs if two cells are from the same lineage. Given a single-cell RNA-seq dataset with *p*=2000 genes with corresponding cell lineage information , each batch select *N* cell pairs from *N* different lineages. Cells from different lineages are considered as negative pairs.

Base Encoder f(): a 3-layer Multilayer Perceptron (MLP) that generates cell representations h from gene expression vectors. Each MLP layer uses ReLU activation and Batch Normalization, which helps stabilize and accelerate training. The input dimension is p (corresponding to the 2000 genes), and the output dimension is 64.

Projection Head g(): consists of a two-layer MLP with ReLU that transforms representations *h* into the space utilized for contrastive loss, denoted by *z*.

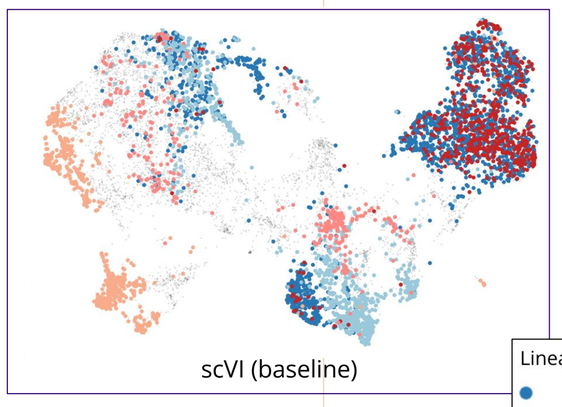
Loss Function:
Suppose for each batch we have *N* cell pairs (*2N* cells in total), we define the loss function for a positive cell pair (*m*, *n*) as:

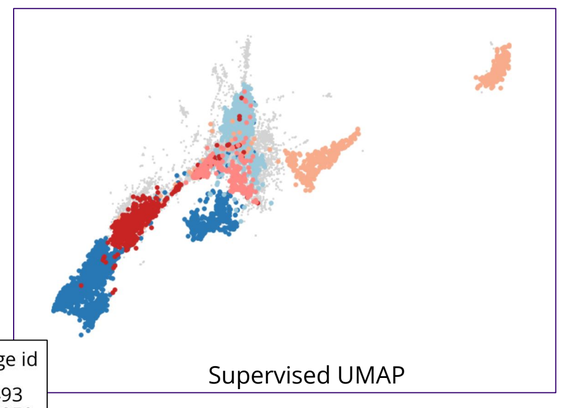
$$loss_{m,n} = -\log\left(\frac{\exp(\mathrm{sim}(z_m, z_n)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq m]} \exp(\mathrm{sim}(z_m, z_k)/\tau)}\right)$$

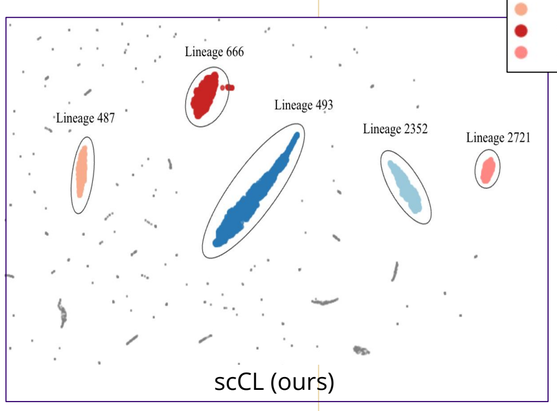
where $\mathrm{sim}(z_m, z_n)$ is the cosine similarity calculated as $(z_m \cdot z_n)/(\|z_m\|_2 \cdot \|z_n\|_2)$, τ is the temperature parameter. Here, $\mathbb{1}_{[k\neq m]}$ in the denominator part is an indicator function that denotes if two cells come from the same lineage. The total loss function then sums $loss_{m,n}$ over all *2N* positive pairs in the batch. **The algorithm therefore aims to maximize the agreement between cells from the same lineage and minimize the agreement between negative pairs.**

Results

UMAP of Top 5 Largest Lineages Highlighted

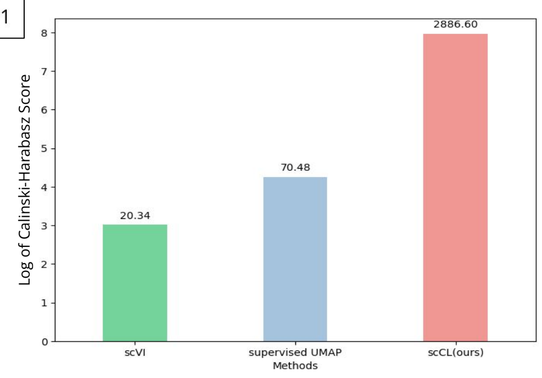






Method	Train Accuracy	Test Accuracy
scVI	44.6%	21.3%
scCL (ours)	93.2%	60.7%

Comparison of Model Performance on Biddy Dataset based on KNN

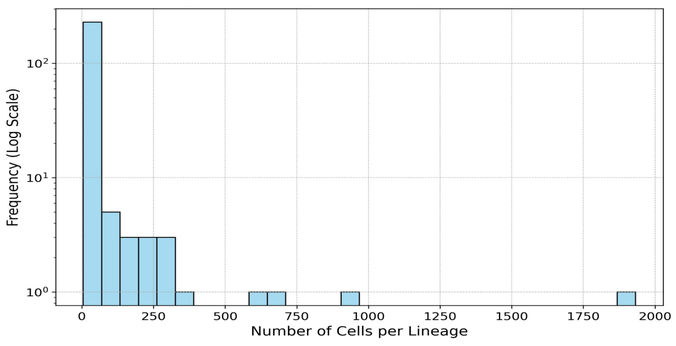


Calinski-Harabasz Score (log scale) of different methods

Our scCL embedding is able to separate the cells by their lineages (i.e., successfully removes information unrelated to lineage)

Dataset

The Biddy dataset, generated from experiments on reprogramming mouse embryonic fibroblasts (MEFs) into induced endoderm progenitors (iEPs) using single-cell lineage tracing (scLT), contains transcriptional profiles and lineage trajectories of over 10k single cells across 240 lineages over a 28-day period.

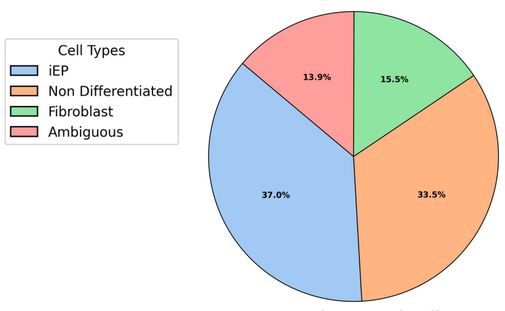


Histogram of Lineage Size of Biddy Data

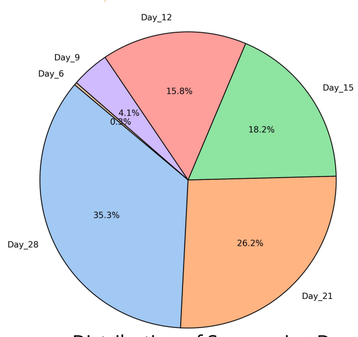
Cell id	Count Matrix
Cell_1	
Cell_2	
Cell_3	
Cell_4	
Cell_5	

Cell Type	Lineage
Type 1	L1
Type 2	L1
Type 1	L2
Type 3	L3
Type 2	L4

an Illustration of input data



Distribution of Cells

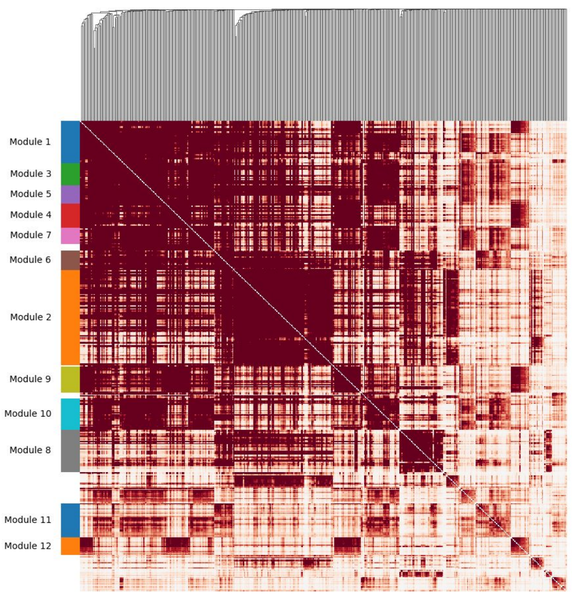


Distribution of Sequencing Days

Biological Insight

We use Hotpost to investigate the genes that are correlated with our scCL embedding. This allows us to find which genes are fate-determining, as different lineages have different compositions of cell types at the later timepoints.

We are looking into what the gene pathways found in our analysis is.



Future work

We plan to do the following for future work:

- Investigate which genes encode lineages and cell fates
- Determining how scCL improves prediction of cell fates over existing embeddings
- Visualizing the landscape of early timepoint cells by their eventual fates using the scCL embedding
- Assessing questions related to cell-of-origin using the scCL framework