Mengyu Li
Email: limengyu516@ruc.edu.cn
Course: Applied Statistics
Instructor: Wei Ma

Final Report
ID: 2020000793
Term: Spring 2022
Due Date: May 10, 2022

Single-cell Developmental Trajectories Inference

1 Introduction

Cellular reprogramming has become a fundamental topic in medical and biological research. Understanding the reprogramming mechanism that guides dedifferentiation during development is a major challenge, which requires researchers to answer questions such as:

- What classes of cells arise at each stage?
- What was their origin at earlier stages? What are their likely fates at later stages?
- What regulatory programs (e.g., transcription factors) control their dynamics?

The first challenge has been largely solved by the advent of the single-cell RNA sequencing (scRNA-seq) technique, while the others remain work-in-progress.

In this project, we aim to address the second question by inferring cell developmental trajectories (in gene-expression space) during reprogramming. Specifically, the question of interest is: given a class of cells at one time point, where will their descendants be at a later time point, and where are their ancestors at an earlier time point? We analyze a scRNA-seq dataset (Schiebinger et al., 2019) collected across 18 days of reprogramming mouse embryonic fibroblasts (MEF) into induced pluripotent stem (IPS) cells. Data and code to reproduce all the experiments are available in https://github.com/Mengyu8042/Cell_trajectory.

2 Data Preparation

Data Information. The scRNA-seq dataset profiles 251,203 cells, collected at 39 time points across 18 days, with samples taken every 12h (every 6h between days 8 and 9). These 18 days are divided into two phases: MEFs were plated in serum, added Dox on day 0 (Phase-1 (Dox)); then

Dox was withdrawn on day 8, and cells were either transferred to serum-free 2i medium (Phase-2 (2i)) or maintained in serum (Phase-2 (serum)). The dataset is publicly available with the accession code GSE122662 in Gene Expression Omnibus (Edgar et al., 2002), from which we have access to two parts of information: a gene expression matrix of size $19,089 \times 25,120$, where rows are genes, columns represent cells, and the (i, j)-th component is the log-normalized expression level of the gene i in cell j; the metadata containing annotation of cells and genes, including cells ID, cells sampling time, cells embedding coordinates, major cell sets, and highly variable genes.

Data Visualization. Data preprocessing was carried out through quality control analysis (i.e., cell-level and gene-level filtering) and gene-level dimensional reduction (i.e., principal component analysis). Then, we visualize the expression profile of cells through force-directed layout embedding (Jacomy et al., 2014) in Fig. 1, annotated according to sampling time (Panel (a)) and culture condition (Panels (b) and (c)).

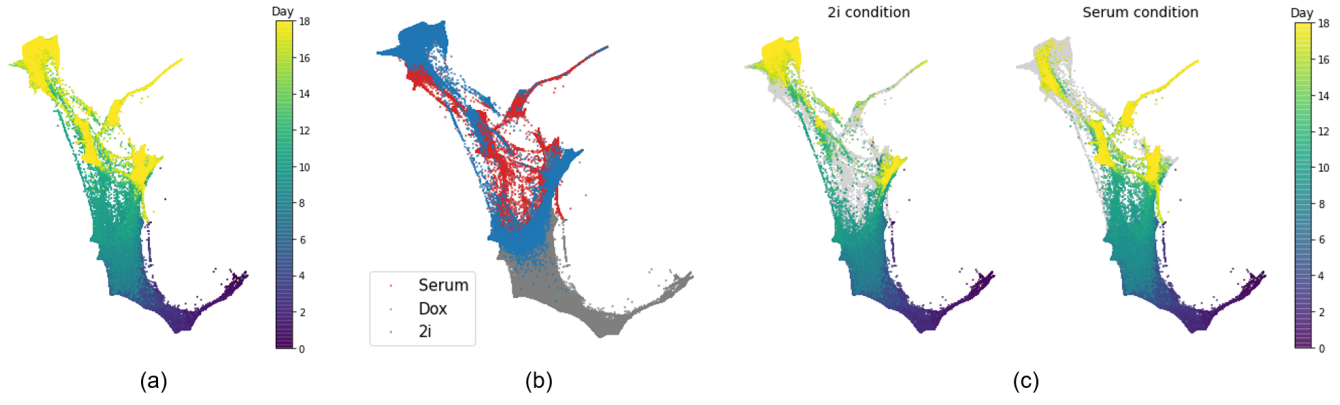

Figure 1: Visualization of scRNA-seq profiles. (a) Cells colored by sampling time. (b) Cells colored by condition. (c) Cells colored by sampling time, with Phase-2 cells from only 2i condition (left) or serum condition (right). Grey points represent Phase-2 cells from the other condition.

As can be seen from Fig. 1, multiple developmental branches were formed during reprogramming. A question of interest is: are these different branches all ending in IPS cells? To answer this question, we display cell set membership in Fig. 2. We can observe that only one branch of cells was dedifferentiated into IPS cells, while the rest went to various types of differentiated cells.

Through exploratory data analysis, we could have a preliminary conjecture about the cells' developmental trajectory. In the next section, we will further explore it through formal modeling.

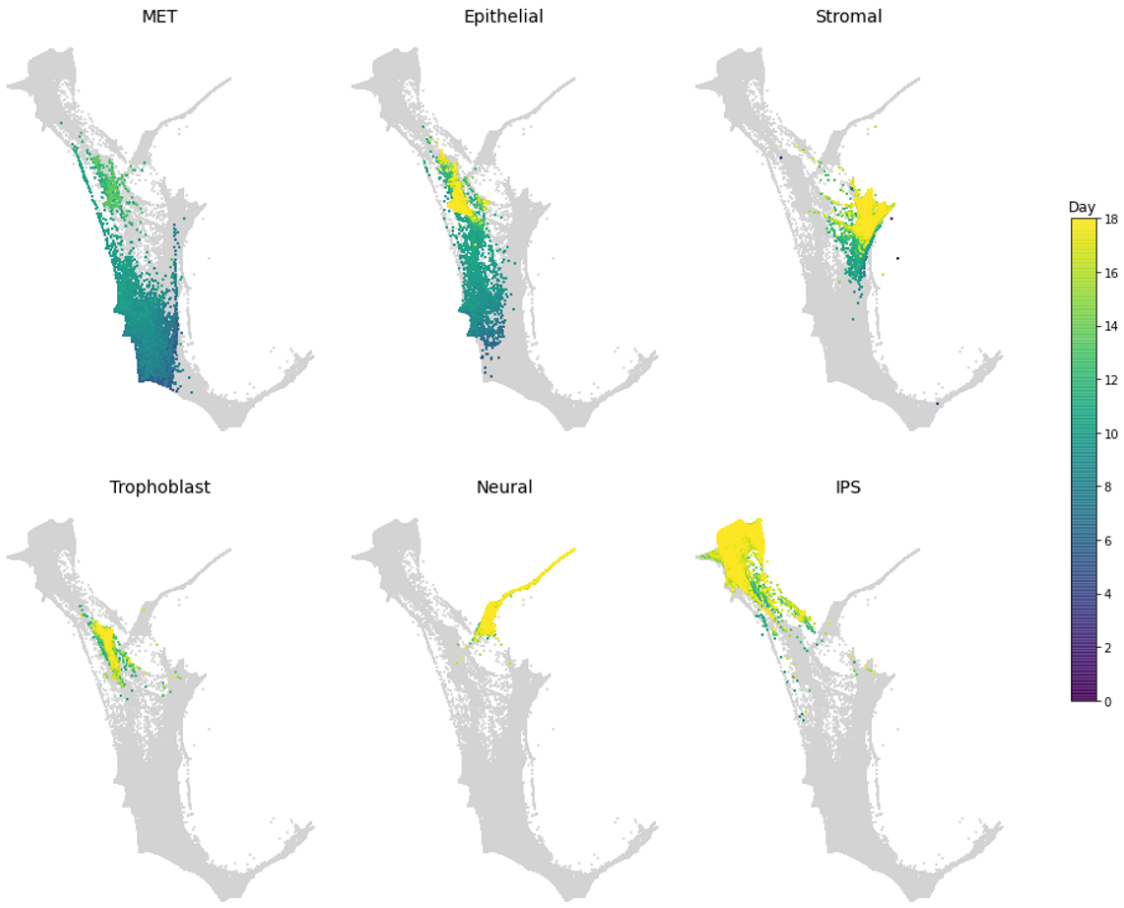

Figure 2: Cell sets membership. Displayed are six major cell sets in Phase-2: MET cells, epithelial cells, stromal cells, trophoblast cells, neural cells, and IPS cells, which are colored by sampling time. Grey points represent cells from other sets.

3 Model

Notations. For a period of time $[0, T]$ and the d-dimensional gene-expression space, suppose that the number of cells at time point t is $n(t)$, then we define the *developmental trajectory* in gene-expression space as a function

$$x : [0, T] \to \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{n(t) \text{ times}}.$$

Thus, the *expression profile* of cells at time t is a matrix of size $d \times n(t)$, defined as

$$x(t) = \big(x_1(t), \ldots, x_{n(t)}(t)\big)\,.$$

Assume that cells at any time are drawn from a probability distribution. As a population of cells grows, the distribution on gene expression space evolves over time. Thus, we define a *developmental process* $\mathbb{P}(t)$ to be a time-varying probability distribution (i.e. stochastic process) over trajectories $x(t)$.

Notice that the unknown developmental process is determined by its temporal dependence structure encoded in the temporal coupling between different time points, which is lost because the cells are killed when performing scRNA-seq. To recover the temporal coupling, we need certain constraints on cellular transitions. A reasonable assumption is that cells don't change expression by large amounts over a short time, i.e., cells move short distances in gene-expression space over a short period. Such assumption motivates us to infer the temporal coupling by using the mathematical tool of optimal transport (Peyré et al., 2019; Zhang et al., 2021).

Method. Given two adjacent time points $t_1, t_2 (t_1 < t_2)$ and their corresponding probability distributions $\mathbb{P}_{t_1}$ (named source distribution), $\mathbb{P}_{t_2}$ (named target distribution), optimal transport finds their coupling π_{t_1,t_2} by minimizing the total transportation cost, i.e., solving

$$\pi_{t_1,t_2} = \underset{\pi}{\operatorname{argmin}} \iint \|x-y\|^2 \pi(x,y)dxdy$$
$$\text{s.t.} \int \pi(\cdot, y)dy = \mathbb{P}_{t_1}, \quad \int \pi(x,\cdot)dx = \mathbb{P}_{t_2}.$$

For biological implications, we make two modifications to the above formula following the work in Schiebinger et al. (2019).

Modification 1: Account for growth. Considering cells may grow at different rates, we rescale the source distribution $\mathbb{P}_{t_1}$ to $\mathbb{Q}_{t_1}$ using the relative growth rate $g(x)$. More precisely,

$$\mathbb{Q}_{t_1}(x) = \mathbb{P}_{t_1}(x) \frac{g^{t_2-t_1}(x)}{\int g^{t_2-t_1}(z) d\mathbb{P}_{t_1}(z)},$$

where the denominator $\int g^{t_2-t_1}(z) d\mathbb{P}_{t_1}(z)$ is for normalizing the total mass to one.

Modification 2: Relax the marginal constraints. Considering the growth rate may be misspecified, we use a robust version of optimal transport, called unbalanced optimal transport (Chizat et al., 2018) as follows,

$$\pi_{t_1,t_2} = \underset{\pi}{\operatorname{argmin}} \iint \|x-y\|^2 \pi(x,y)dxdy + \lambda_1 \mathrm{KL}\left(\int \pi(\cdot,y)dy \| \mathbb{Q}_{t_1}(x)\right) + \lambda_2 \mathrm{KL}\left(\int \pi(x,\cdot)dx \| \mathbb{P}_{t_2}(y)\right).$$

In practice, one can use the output row-sums of $\hat{\pi}_{t_1,t_2}$ as a new estimate for $g(x)$, and iteratively solve the above optimization problem several times.

Descendants and Ancestors. Based on the estimate of temporal coupling between each pair of time points, we calculate ancestor and descendant distributions of major cell sets that will be defined below. Consider a set of cells $C \subset \mathbb{R}^d$ at time point t_j with probability distribution

$$\mathbb{P}_{t_j}(x) = \begin{cases} \frac{1}{|C|} & x \in C, \\ 0 & \text{otherwise.} \end{cases}$$

Regarding the estimated couplings as transition matrices, the *descendants* of C at time t_{j+1} are obtained by pulling C through $\pi_{t_j,t_{j+1}}$, i.e., $\mathbb{P}_{t_{j+1}}^{\top} = \mathbb{P}_{t_j}^{\top}\pi_{t_j,t_{j+1}}$; the *ancestors* of C at time t_{j-1} are obtained by pulling C back through π_{t_{j-1},t_j}, i.e., $\mathbb{P}_{t_{j-1}} = \pi_{t_{j-1},t_j}\mathbb{P}_{t_j}$; and the trajectory of C is the sequence of ancestor distributions at earlier time points and descendant distributions at later time points.

4 Results and Interpretation

Developmental trajectory. Fig. 3 illustrates the major cell sets at day 18 and their ancestors' trajectories, where the lighter color implies that cells were more likely to take the path.

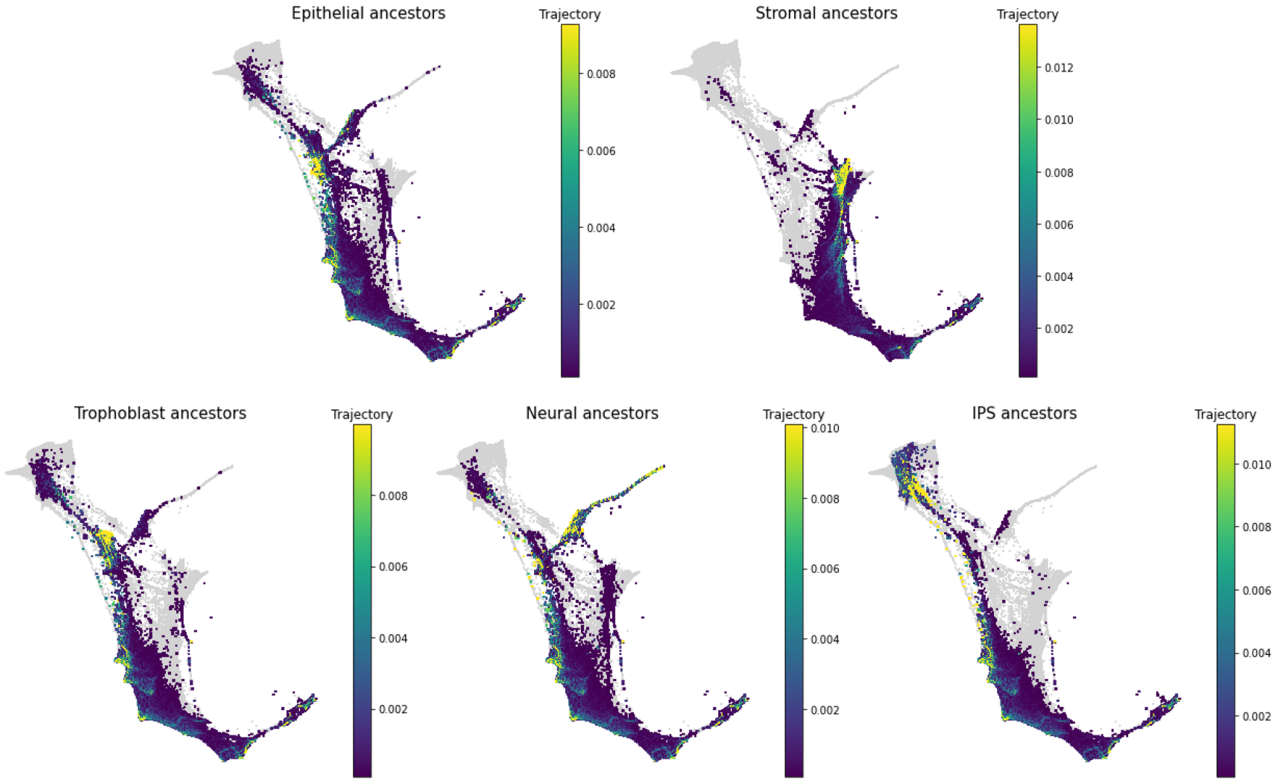

Figure 3: Ancestor trajectory of day 18 major cell sets, colored by (unormalized) probability.

In Fig. 3, we can observe that the vast majority of IPS ancestors went through the leftmost routine instead of the middle path. Moreover, we can speculate that stromal cells stemmed from a different ancestry compared with the rest cell sets. To verify this conjecture, we examine the shared ancestry by calculating the divergence between any pair of cell sets, to study when these two cell sets diverged from a common set of ancestors. The results are presented in Fig. 4.

Shared ancestry. As shown in Fig. 4, IPS cells, epithelial cells, neural cells, and trophoblast cells shared the same ancestry, and they began to diverge at the start of Phase-2; while stromal

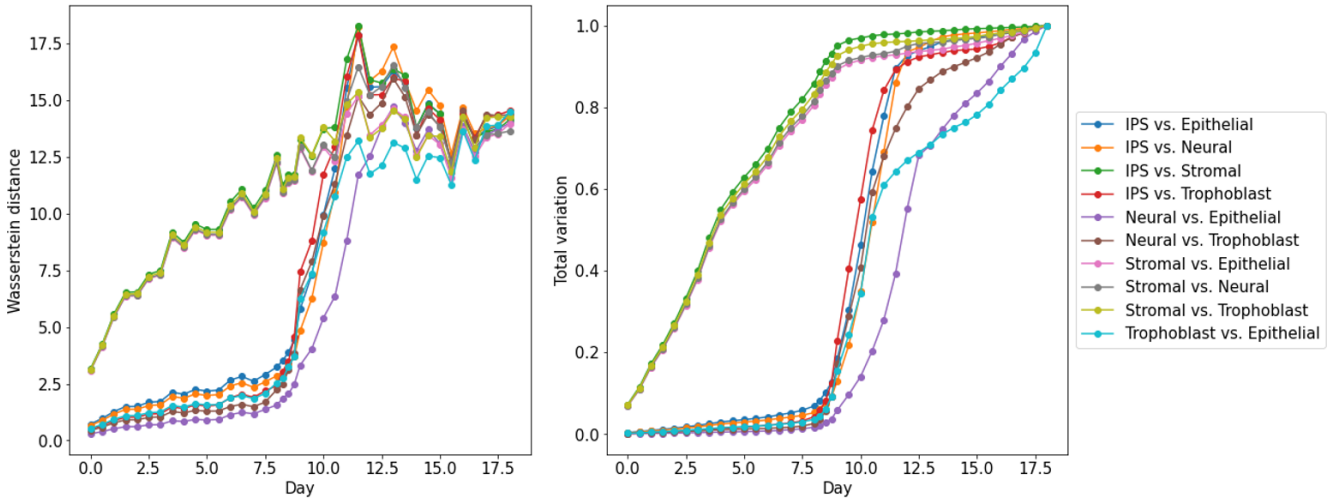

Figure 4: Divergence (Left: Wasserstein distance; Right: Total variation) between any pair of cell sets versus sampling time.

cells appeared to go through a different routine from Day 0.

Transition table. Next, we summarize the temporal coupling via the transition table to explore descendants of major cell sets. In particular, consider cell sets $C_1, \ldots, C_m$ at time t_k and cell sets $D_1, \ldots, D_n$ at time $t_{k+\Delta}$, then the (i, j)-th element of the transition table from t_k to $t_{k+\Delta}$ is

$$\text{mass transported from } C_i \text{ to } D_j = \sum_{x \in C_i} \sum_{y \in D_j} \pi_{t_k, t_{k+\Delta}}(x, y).$$

Fig. 5 shows the transition table from Day 12 to Day 18. We can find that the vast majority of MET cells appearing at the intermediate stage of reprogramming differentiated into neural cells at the end of reprogramming, while others mainly remained in the same class.

Validation via Interpolation. Finally, we validate the model by checking the quality of interpolation. Considering consecutive time points (t_i, t_{i+1}, t_{i+2}) with distributions $(\mathbb{P}_{t_i}, \mathbb{P}_{t_{i+1}}, \mathbb{P}_{t_{i+2}})$, model validation takes several steps:

1. holding out the data from t_{i+1}, estimate the coupling $\hat{\pi}_{t_i, t_{i+2}}$ between t_i and t_{i+2};
2. sample from the estimated $\hat{\pi}_{t_i, t_{i+2}}$, and compute an interpolating distribution $\hat{\mathbb{P}}_{t_{i+1}}$ at time t_{i+1};
3. compare $\hat{\mathbb{P}}_{t_{i+1}}$ to $\mathbb{P}_{t_{i+1}}$ via the Wasserstein distance.

Fig. 6 shows the errors of our interpolation $\hat{\mathbb{P}}_{t_{i+1}}$ compared to several baselines, which include: interpolation using a uniform coupling; distribution at the first time point t_i in the interval (i.e., $\mathbb{P}_{t_i}$); distribution at the last time point t_{i+2} in the interval (i.e., $\mathbb{P}_{t_{i+2}}$). Such results illustrate

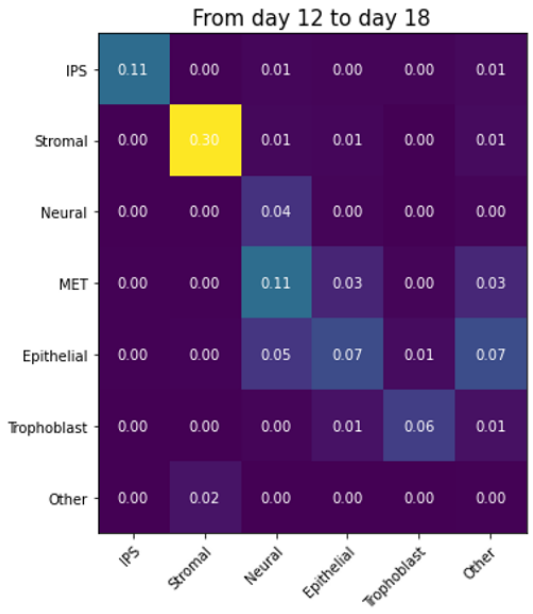

Figure 5: Transition table from Day 12 to Day 18, annotated by transported mass.

that the OT-based model (i.e., the red line) always achieves the smallest error compared to other model-free baselines.

5 Discussion

In this project, we analyze the single-cell developmental trajectories during reprogramming using the mathematical technique of optimal transport. Through exploratory data analysis and model-based data analysis, we obtain the following findings:

- Only a small subset of MEFs became IPS cells after reprogramming; while others mainly went to epithelial cells, neural cells, trophoblast cells, and stromal cells.
- Among these five classes of cells, IPS, epithelial, neural, and trophoblast cells shared the same ancestry, while stromal cells had an alternative routine from the beginning of reprogramming.
- Another major cell set that appeared during reprogramming was MET, which mainly differentiated into neural cells at the end of reprogramming.

Optimal transport produces cell trajectories by connecting samples with straight lines in gene-expression space. However, this is impractical in reality because cells always move smoothly. Thus, a key question remains: how to extend "linear paths" to "non-linear paths"? To fill this gap, Tong et al. (2020) proposed the Trajectory-Net method, which interpolates non-linearly using information from more than two time points by combining continuous normalizing flows and dynamic optimal

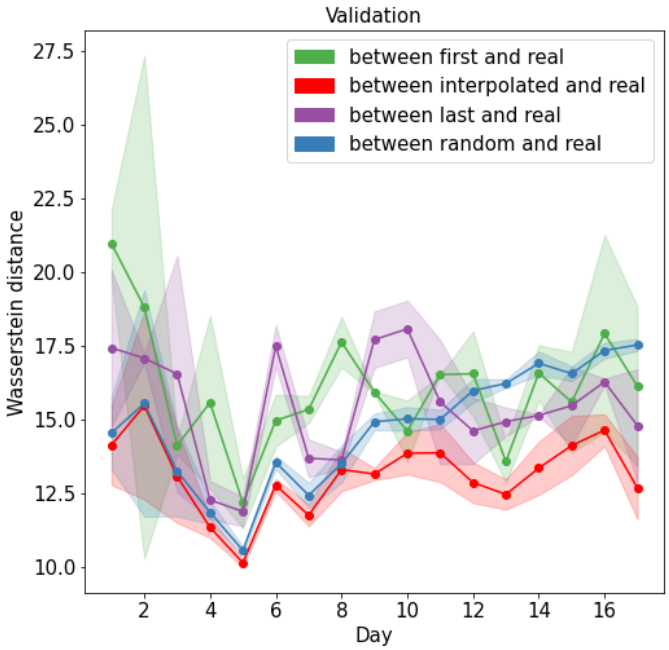

Figure 6: Interpolation errors (with mean and standard deviation) compared to several baselines.

transport. An alternative method may be using nonparametric statistical methods, e.g. smoothing splines and regression splines, to smooth the paths between samples of different time points, which could be the work in the future.

References

Chizat, L., G. Peyré, B. Schmitzer, and F.-X. Vialard (2018). Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation 87*(314), 2563–2609.

Edgar, R., M. Domrachev, and A. E. Lash (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research 30*(1), 207–210.

Jacomy, M., T. Venturini, S. Heymann, and M. Bastian (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one 9*(6), e98679.

Peyré, G., M. Cuturi, et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning 11*(5-6), 355–607.

Schiebinger, G., J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell 176*(4), 928–943.

Tong, A., J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy (2020). Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International Conference on Machine Learning*, pp. 9526–9536. PMLR.

Zhang, J., W. Zhong, and P. Ma (2021). A review on modern computational optimal transport methods with applications in biomedical research. *Modern Statistical Methods for Health Research*, 279–300.