REGULARIZED OPTIMAL TRANSPORT FOR SINGLE-CELL TEMPORAL TRAJECTORY ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

The temporal relationship between different cellular states and lineages is only partially understood and has major significance for cell differentiation and cancer progression. However, two pain points persist and limit learning-based solutions: (a) lack of real datasets and standardized benchmark for early cell developments; (b) the complicated transcriptional data fail classic temporal analyses. We integrate `Mouse-RGC`, a large-scale mouse retinal ganglion cell dataset with annotations for 9 time stages and $30,000$ gene expressions. Existing approaches show a limited generalization of our datasets. To tackle the modeling bottleneck, we then translate this fundamental biology problem into a machine learning formulation, *i.e.*, *temporal trajectory analysis*. An innovative regularized optimal transport algorithm, `TAROT`, is proposed to fill in the research gap, consisting of (1) customized masked autoencoder to extract high-quality cell representations; (2) cost function regularization through biology priors for distribution transports; (3) continuous temporal trajectory optimization based on discrete matched time stages. Extensive empirical investigations demonstrate that our framework produces superior cell lineages and pseudotime, compared to existing approaches on `Mouse-RGC` and another two public benchmarks. Moreover, `TAROT` is capable of identifying biologically meaningful gene sets along with the developmental trajectory, and its simulated gene knockout results echo the findings in physical wet lab validation.

1 INTRODUCTION

Since first introduced in 2009, large-scale single-cell RNA sequencing (scRNA-seq) has presented enormous opportunities for researchers in various research fields (Patel et al., 2014; Satija et al., 2015; Tirosh et al., 2016). It helps reveal detailed information on transcriptional patterns in different cell and tissue types as well as disease models (Elmentaite et al., 2022; Jagadeesh et al., 2022). Equipped with scRNA-seq, we are able to discover significant heterogeneities that would never be found with bulk analysis within the cell population, which contributes to understanding biology questions with higher cellular resolution. The fast-advancing technology and increased recognition of different cell subtypes also naturally lead us to ask: ① *How and when are the cell subtypes established?* ② *Could we predict the developmental trajectory of each cell and predict the cell "fate" based on current status?* ③ *And could we find the key regulator that controls this type of establishment?* Answers to those questions are important for cell differentiation research in developmental biology (Rizvi et al., 2017; Han et al., 2018; Gulati et al., 2020) and can provide promising pipelines to demystify the cellular response during disease progression (Zhang et al., 2021; Jia et al., 2022). In the past decade, although a great amount of effort (Trapnell et al., 2014; Qiu et al., 2017; Ji & Ji, 2016; Street et al., 2018; Cao et al., 2019) has been put into developing trajectory inference methods using single-cell sequencing data, it remains extremely challenging. This is because, with current technologies, we can not trace the same population of cells over developmental time. It only allows us to collect the transcriptional information of cells for a specific time point as a

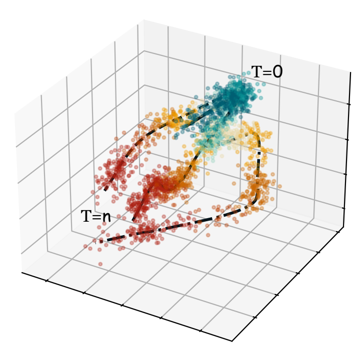

Figure 1: Demo Cell Temporal Trajectories from Time $0 \rightarrow n$. Different colors indicate cells from different time stages.

"snapshot", and then a sophisticated computational modeling (Saelens et al., 2019; Van den Berge et al., 2020) is required to construct cell trajectories over multiple "snapshots", as demonstrated in Figure 1. Existing algorithms reach good performance on simulated datasets (Klein et al., 2023) but are still unsatisfactory on realistic benchmarks.

To enhance the capabilities of learning-based algorithms, we generate `Mouse-RGC`, which is a large-scale integrated mouse retinal ganglion cell dataset. It contains $30,000$ gene expressions from 9 time stages of early cell development. However, naively plugging previous approaches (Street et al., 2018; Klein et al., 2023) fail to generalize well on our benchmark, implying their shortage in handling real cases with much higher data complexity. To develop effective solutions, we recast the biology challenge into a machine learning problem, *i.e.*, *temporal trajectory analysis*, aiming to transport cells across time stages. In detail, our proposed `TAROT` first learns superior cell representations through a tailored masked autoencoder. Then, it performs a regularized optimal transport (OT) to produce mappings between every two-time stages. During the matching, we consider the biological priors of gene expression from both developmental and functional perspectives. Note that directly applying OT will result in inferior results due to neglecting the intrinsic structures in this biology problem. Last, continuous temporal trajectories (*i.e.*, cell pseudotime) are optimized and generated by fitting ordered discrete time stages. Our contributions are summarized below:

- ⋆ We integrate a larger-scale scRNA-seq dataset, *i.e.*, `Mouse-RGC`, with $30,000$ mouse neuron cells annotated cross 9 early developmental time stages. It provides a standardized and challenging benchmark for further research in machine learning (ML) and single-cell transcriptomics.
- ⋆ We recast the analyses of cell developmental differentiation as an ML problem of inferring temporal trajectories. Our proposed `TAROT` consists of an improved design of cell representation extractor and regularized OT with biology priors, delivering substantially enhanced cell lineages.
- ⋆ Based on discrete inferred lineages, we introduce B-Splines optimization to produce continuous cell pseudotime estimations with superior quality.
- ⋆ Extensive experiments validate the effectiveness of our proposals on `Mouse-RGC` and two public datasets. For example, `TAROT` achieves $\{3.10\% \sim 65.03\%, 13.70\% \sim 35.08\%, 6.16\% \sim 27.49\%, 20.82\% \sim 44.28\%\}$ performance improvements on `Mouse-RGC` and `Mouse-MCC` datasets over previous approaches.
- ⋆ Moreover, `TAROT` can locate crucial gene sets that are biologically meaningful for each temporal trajectory. Removing these genes significantly reshapes the simulated cell differentiation, echoed with the wet lab studies on the `Mouse-iPE` dataset.

2 Related Works

Optimal Transport (OT). OT (Villani et al., 2009; Peyré et al., 2019) serves as a powerful tool for comparing two measures in a Lagrangian framework. It has played a beneficial role in widespread applications in statistics (Munk & Czado, 1998; Evans & Matsen, 2012; Sommerfeld & Munk, 2018; Goldfeld et al., 2022) and machine learning (Schmitz et al., 2018; Kolouri et al., 2018) domains. OT can also be used to define metrics such as the Wasserstein distance (Arjovsky et al., 2017; Liu et al., 2019), which has gained tremendous popularity in the training of generative adversary networks (Deshpande et al., 2019; Adler & Lunz, 2018; Petzka et al., 2017; Deshpande et al., 2018; Yang et al., 2018; Baumgartner et al., 2018; Wu et al., 2019), transfer learning (Shen et al., 2018; Lee et al., 2019), and contrastive representation learning (Chen et al., 2021). There also are several preliminary studies that use OT to model the cellular dynamics network (Tong et al., 2020) and cell developmental trajectories (Schiebinger et al., 2019; Klein et al., 2023).

Representation Learning in Single-Cell Genomics. Extracting powerful cell representations is one of the ultra goals for single-cell genomics. It has been investigated for a long history, and various solutions are delivered ranging from classic optimization algorithm (Li et al., 2017; Satija et al., 2015; Zhao et al., 2022; Stuart et al., 2019) to modern deep learning-based approaches (Yang et al., 2022; Hao et al., 2023; Cui et al., 2022; Geuenich et al., 2023; Zhao et al., 2023). For instance, Yang et al. (2022) utilizes the bi-directional transformer to learn robust single-cell representations. To further improve the cell representation quality, more recent studies leverage a variety of advanced pre-training designs, including generative (Shen et al., 2023; Cui et al., 2023), mask language modeling (Hao et al., 2023), multi-task learning (Cui et al., 2022), self-supervised active learning (Geuenich et al., 2023), and contrastive learning (Zhao et al., 2023) objectives.

Lineage and Pseudotime Inference. The increasing availability of scRNA-seq data allows researchers to reconstruct the trajectories of cells during a dynamic process. The relationships between different cellular states and lineages are extremely important for studies on embryonic development (Griffiths et al., 2018; Cang et al., 2021; Mittnenzweig et al., 2021; Kim et al., 2023), cell differentiation (Rizvi et al., 2017; Han et al., 2018; Gulati et al., 2020), cancer progression (Zhang et al., 2021; Jia et al., 2022) and cell fate diversification (Buchholz et al., 2016; Koenig et al., 2022). In the past few years, numerous trajectory inference pipelines have been established, which can be roughly divided into two major categories based on the algorithm they used. The first and perhaps the most commonly used one is minimum spanning tree (MST) based approaches. Monocle and Monocle-2, which are the early used methods, both infer the developmental trajectory of one single cell level and assign the pseudotime of each cell (Trapnell et al., 2014; Qiu et al., 2017). Later, Tools for Single Cell Analysis (TSCAN) (Ji & Ji, 2016) and Slingshot (Street et al., 2018) run the MST algorithm on clusters to construct the cluster-based MST. Then, they orthogonally project each cell onto the paths of the MST to get the pseudotime. Notably, Slingshot utilized a principal curves algorithm to calculate smooth curves from MST, which gives better visualization. The second category is the graph-based trajectory inference method, which employs various algorithms to construct trajectories among cells. One prominent and widely used tool, Monocle3 (Cao et al., 2019), generates trajectories using a principal graph algorithm. Then, it calculates the shortest Euclidean distance of each cell from the root node to assign the pseudotime. However, the self-selected root node required some prior knowledge about the cell identity. Diffusion pseudotime (DPT) (Haghverdi et al., 2016) and URD (Farrell et al., 2018) uses a k-nearest-neighbor algorithm to construct the temporal trajectory of the cells in gene expression space.

Single-Cell Transcriptomics. The heterogeneity analysis is the core reason for performing single-cell sequencing studies. It assesses the transcriptional similarities and differences within the cell populations and helps reveal a higher cellular resolution among cells (Haque et al., 2017; Satija et al., 2015; Tirosh et al., 2016). Using scRNA-seq (Patel et al., 2014), researchers are able to define detailed heterogeneity of immune cells (Shalek et al., 2013; Mahata et al., 2014; Stubbington et al., 2017), cancer cells (Wu et al., 2021; Fan et al., 2020), embryonic stem cells (Jaitin et al., 2014; Klein et al., 2015) *etc*. In the meantime, transcriptional assessments with single-cell sequencing technology also identify rare cell populations that would never been detected using bulk analysis (Miyamoto et al., 2015; Zeisel et al., 2015; Tirosh et al., 2016). In parallel, the gene co-expression patterns that scRNA-seq reveals allow us to define gene modules and point out the underlying mechanism of gene expression regulations (Wagner et al., 2016).

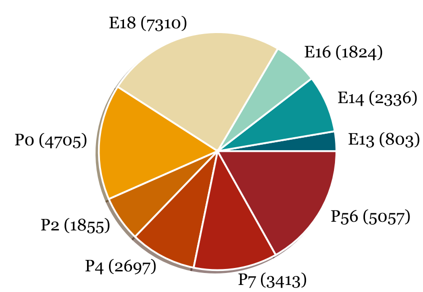

Figure 2: The sample distribution of our `Mouse-RGC` (30K cells) dataset based on developmental time stages. For example, "E18 (7310)" indicates $7,310$ cell samples in time stage E18.

3 Dataset and Machine Learning Formulation

3.1 `Mouse-RGC`: A Large-Scale Dataset of Retinal Ganglion Cells from Mouse

In this section, we introduce all three datasets that are adopted to evaluate `TAROT`'s effectiveness. As for public datasets, we consider a mouse cerebral cortex cell benchmark (Di Bella et al., 2021), *i.e.*, `Mouse-CCC`, and a mouse induced Erythroid Progenitor (iEP)-derived cell benchmark (Capellera-Garcia et al., 2016), *i.e.*, `Mouse-iEP`, which contains $\{66443,\, 1947\}$ cells across $\{11,\, 2\}$ time stages, respectively. The detailed information about our `Mouse-RGC` is presented below.

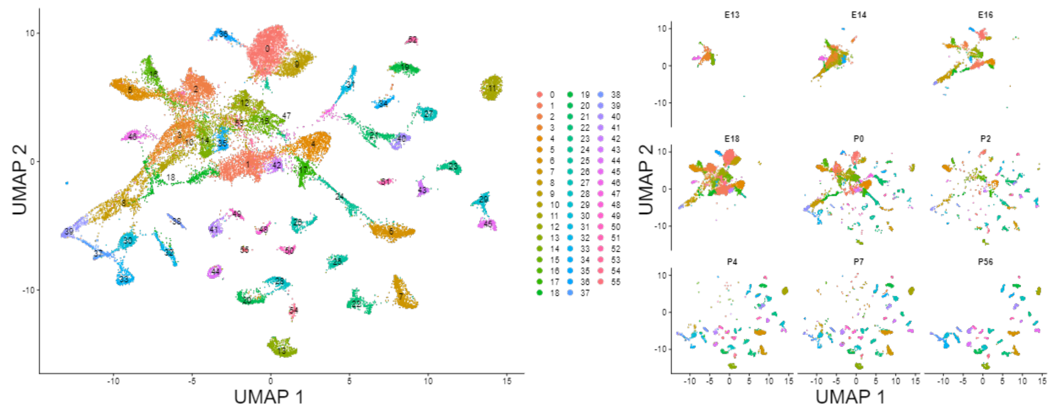

Figure 3: *(Left)*: Clustering `Mouse-RGC` to 56 kinds of cell types and projecting them into a 2D space via UMAP; *(Right)*: Decomposing the clustering results by their time stage labels. Zoom-in for better reliability.

Data Collection. For the `Mouse-RGC` dataset, we extract 30K mouse neuron cells from previously published datasets(Shekhar et al., 2022; Whitney et al., 2023) and newly formed data. The developmental time stages of $\{$E13, E14, E16, E18, P0, P2, P4, P7, P56$\}$ (Figure 4). Then, the corresponding gene expressions are measured by the RNA sequencing technique as previously defined[1]. Single-cell libraries were prepared using the single-cell gene expression $3'$ kit on the Chromium platform (10X Genomics, Pleasanton, CA) following the manufacturer's protocol. To be specific, single cells were partitioned into Gel beads in EMulsion (GEMs) in the 10X Chromium instrument followed by cell lysis and barcoded reverse transcription of RNA, amplification, enzymatic fragmentation, $5'$ adaptor attachment, and sample indexing. On average, around $8,000 \sim 12,000$ single cells were loaded on each channel, and around $3,000 \sim 7,000$ cells were recovered. Libraries were sequenced on the Illumina HiSeq $2,500$ platforms.

3.2 Single Cell Data Process

Preprocess and Properties. After we collected the raw signals, the following single-cell sequencing data processing was done using the Seurat package (Hao et al., 2021). Sample quality control was performed on each sample individually. For each sample, doublets were removed using DoubletFinder (McGinnis et al., 2019). We retained cells that expressed at least $1,500$ genes and less than $11,000$ genes. Meanwhile, we removed cells that have more than 5% mitochondrial genes and genes expressed in fewer than 10 cells. The resulting n cells $\times$ g genes matrix of UMI counts were subject to downstream analysis. The UMI-based gene expression matrix was normalized using sctransform (Hafemeister & Satija, 2019). After that, the batch correction was done with canonical correlation analysis (Hotelling, 1992; Anderson et al., 1958), using the top $4,000$ anchor genes.

Clustering. In this research, we are interested in the evolution of different cell types of mouse neurons. Therefore, we built a nearest-neighbor graph to cluster cells based on their transcriptional similarity. Specifically, the number of nearest neighbors was chosen to be 50, according to the rich experiences of biology scientists. The edges were weighted based on the Jaccard overlap metric, and graph clustering was performed using the Louvain algorithm (Blondel et al., 2008). In the end, as demonstrated in Figure 3 (*Left*), the cell clusters were then projected onto a nonlinear 2D space using the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2018). For temporal trajectory analysis, we further decomposed the cell clusters by their time stage labels like Figure 3 (*Right*). Our goal is to demystify the neuron evolution path across these 9 time stages. Referring Appendix B.2 for more details.

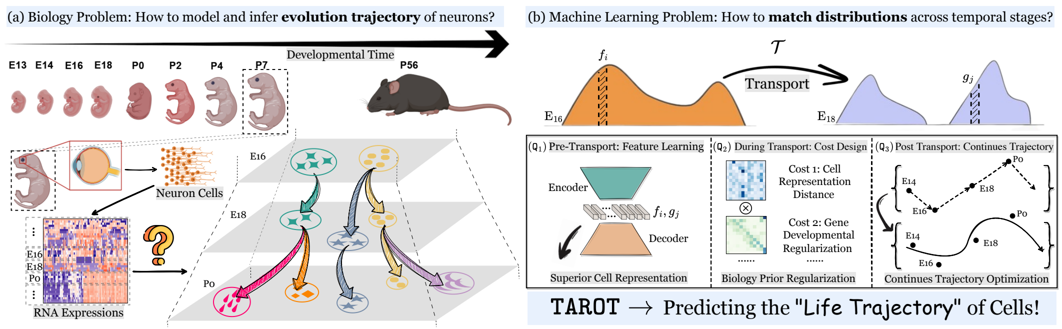

Figure 4: (a) *Biology Problem.* We aim to model and infer the evolution trajectory of neurons. Specifically, the neuron cells are extracted from the developmental time stages {E13, E14, E16, P0, P2, P4, P7, P56} of mouses. Then, RNA sequencing is performed to collect its expression data. (b) *Machine Learning Problem.* We translate the biology problem to an ML problem – matching sample distributions across multiple temporal stages. This challenging transport problem can be further decomposed into three sub-questions, *i.e.*, $\mathtt{Q}_1$, $\mathtt{Q}_2$, and $\mathtt{Q}_3$. To tackle these research questions, our proposed `TAROT` introduces superior cell representations, regularized optimal transport via biology priors, and continuous trajectory optimization, respectively.

3.3 ML Formulation - Matching Sample Distribution across Temporal Stages

Understanding the development of stem cells into fully differentiated cells requires accurate cell lineage and pseudotime. Thus, the fundamental biology problem here is *how to model and infer evaluation trajectory of neurons?* This paper recasts it as a machine learning (ML) problem, aiming to *match cell distributions across temporal stages*.

Notations. Let $\{\mathbf{r}_i\}_{i=1}^n$ denote the raw cell expressions and $\{\mathbf{c}_i\}_{i=1}^n$ are extracted cell representations, where n is the total number of cells and $\mathbf{r}_i \in \mathbb{R}^{1\times g}$. For each cell representation $\mathbf{c}_i \in \{\mathbf{c}_1, \cdots, \mathbf{c}_n\}$, it has the labels of time stage and cluster, obtained from the data pre-processing. Therefore, the total n cells can be divided into k groups, *i.e.*, $\{\mathcal{C}_1, \cdots, \mathcal{C}_k\}$ where $\sum_{i=1}^k |\mathcal{C}_i| = \sum_{i=1}^k n_i = n, \mathcal{C}_i = \{\mathbf{c_1^{(i)}}, \cdots, \mathbf{c_{n_i}^{(i)}}\}$, $|\mathcal{C}_i| = n_i$ is the number of cells in cluster $\mathcal{C}_i$. Considering temporal information like time stages $\{t\}_{t=1}^s$, we use $\mathcal{G}^{(t)} = \{\mathcal{C}_1^{(t)}, \cdots, \mathcal{C}_{k_t}^{(t)}\}$ to represent all cells in the time stage t, where s is the total number of time stages and k_t denotes the number of clusters at time step t. Our goal is to establish a mapping from $\mathcal{G}^t \to \mathcal{G}^{(t+1)}$, which shares certain similarity to the trajectory analysis problem (Helland-Hansen & Hampson, 2009).

Problem Definition. *Given the cluster set* $\mathcal{G}^{(t)} = \{\mathcal{C}_1^{(t)}, \cdots, \mathcal{C}_{k_t}^{(t)}\}$ *as each time stage* $t \in \{1, \cdots, s\}$, *we aim to (1) infer temporal trajectories like* $\mathcal{G}^{(1)} \to \mathcal{G}^{(2)} \to \cdots \to \mathcal{G}^{(s)}$, *based on their gene expression; (2) estimate continuous pseudotime for each sample.*

An Ideal Solution. To infer the temporal trajectory, it requires answering three key questions ($\mathtt{Q}_1$, $\mathtt{Q}_2$, and $\mathtt{Q}_3$) as summarized in Figure 4 (b):

① *Before the Transportation.* It needs to extract high-quality cell representations $\{\mathbf{c}_i\}_{i=1}^n$ from the gene expressions $\{\mathbf{r}_i\}_{i=1}^n$. Both low-dimensional projection methods like PCA (Bro & Smilde, 2014) and UMAP (McInnes et al., 2018), and deep neural networks (Yang et al., 2022; Shen et al., 2023; Cui et al., 2023) can serve as feature extractor.

② *During the Transportation.* It focuses on computing the mapping function $\mathcal{T}_{t,t+1} : \mathcal{G}^{(t)} \to \mathcal{G}^{(t+1)}$, $t \in \{1, \cdots, s-1\}$, given all cell information from the current and history time stages. Each mapping between two-time stages is a bipartite graph and can be derived from distribution matching problems (Gretton et al., 2012) through the Hungarian algorithm or Optimal Transport, *etc*. The crucial challenge here is the design of cost functions for transportation. Naively plugging in distance measurements based on ML intuitions leads

[1]https://rna.cd-genomics.com/resource-rnc-rna-sequencing-introduction-workflow-and-analysis-pipelines.html

to inferior results (Klein et al., 2023), which demands appropriate cost designs to integrate *biology priors* of the neuron developments.

③ *After the Transportation.* Global lineages are deduced according to the pair-wised mapping $\{\mathcal{T}_{t,t+1}\}_{t=1}^{s-1}$. However, it only contains a discrete order of different time stages, which is an irregularly sampled time series due to the constraints of cell data collection. Since cell differentiation occurs continuously, we need to calculate a continuous cell pseudotime based on its inferred lineage. It can be addressed by interpolation approaches (Shukla & Marlin, 2020) like Splines.

4 METHODOLOGY

Overview of `TAROT`. The overall procedures of `TAROT` are described in Figure 4. Our paper tackles the aforementioned ML problem by answering the three key questions. Before transport, we introduce a customized Masked Autoencoder (MAE) transformer to learn adequate cell representation. During transport, we integrate important biology priors into the design of cost functions and leverage them to enable a regularized optimal transport. After transport, continuous trajectories will be produced by performing B-Splines fitting optimization to inferred cell lineages.

4.1 `TAROT`: CELL LINEAGE INFERENCE VIA DISTRIBUTION MATCHING

Regularized Optimal Transport for Matching. Optimal transport (OT) distance is a popular option for comparing two distributions. We consider the discrete situation in our case. For two time steps t_1 and t_2, there are two sets of features $\{\boldsymbol{f}_i\}_{i=1}^{\mathrm{M}}$ and $\{\boldsymbol{g}_j\}_{j=1}^{\mathrm{N}}$. Since we focus on a cluster-level mapping, then $\mathrm{M} = k_{t_1}$ and $\mathrm{N} = k_{t_2}$ are the number of clusters in stage t_1 and t_2 respectively. $\boldsymbol{f}_i$ and $\boldsymbol{g}_j$ are averaged cell representations for each cluster. Note that it is straightforward to extend to cell-level mapping by adopting cell-specific representations. Our discrete distributions can be formulated as $\boldsymbol{u} = \sum_{i=1}^{\mathrm{M}} u_i \delta_{\boldsymbol{f}_i}$ and $\boldsymbol{v} = \sum_{j=1}^{\mathrm{N}} v_j \delta_{\boldsymbol{g}_j}$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the discrete probability vectors that sum to 1, and $\delta_{\boldsymbol{f}}$ (or $\delta_{\boldsymbol{g}}$) is a Dirac δ function placed at support point $\boldsymbol{f}$ (or $\boldsymbol{g}$) in the embedding space. Then, the total cost of transportation is depicted as $< \mathcal{T}, \mathcal{D} >= \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{N}} \mathcal{T}_{i,j} \mathcal{D}_{i,j}$.

The matrix $\mathcal{D}$ is a cost matrix, where each element denotes the cost between feature $\boldsymbol{f}_i$ and $\boldsymbol{g}_j$, like $\mathcal{D}_{i,j} = 1 - \mathtt{sim}(\boldsymbol{f}_i, \boldsymbol{g}_j)$ and $\mathtt{sim}(\cdot,\cdot)$ is a similarity measuring function. The $\mathcal{T}$ is the transport matrix that describes the mapping from $\{\boldsymbol{f}_i\}_{i=1}^{\mathrm{M}}$ to $\{\boldsymbol{g}_j\}_{j=1}^{\mathrm{N}}$. To learn the transport plan $\mathcal{T}$, it will minimize the total cost as follows:

$$\mathcal{F}_{\mathrm{OT}}(\boldsymbol{u}, \boldsymbol{v}|\mathcal{D}) = \mathtt{min}_{\mathcal{T}} \langle \mathcal{T}, \mathcal{D} \rangle \tag{1}$$

$$\text{s.t. } \mathcal{D} \times \mathbf{1}_{\mathrm{N}} = \boldsymbol{u},\ \mathcal{D}^{\mathrm{T}} \times \mathbf{1}_{\mathrm{M}} = \boldsymbol{v},\ \mathcal{D} \in \mathbb{R}_{+}^{\mathrm{M}\times\mathrm{N}}. \tag{2}$$

However, this formulation has a super-cubic complexity in the size of $\boldsymbol{u}$ and $\boldsymbol{v}$, which prevents adapting OT in large-scale scenarios. Sinkhorn algorithm Cuturi (2013) is applied to speed up the computation via an entropy regularization, *i.e.*, $\mathcal{F}_{\mathrm{OT}}(\boldsymbol{u}, \boldsymbol{v}|\mathcal{D}) = \mathtt{min}_{\mathcal{T}} < \mathcal{T}, \mathcal{D} > -\lambda \mathcal{E}(\mathcal{D})$, where $\mathcal{E}(\cdot)$ is the entropy function and $\lambda \geq 0$ is a hyper-parameter. An inadequate choice of λ can either degrade the quality of the OT results if too large or prolong the computation time if too small. To circumvent the significant human effort required to identify an appropriate λ, we integrate a straightforward search algorithm that efficiently identifies a suitable λ for the OT calculation. Please refer to Appendix A for more algorithmic details. The optimization of $\mathcal{F}_{\mathrm{OT}}$ constitutes the base framework of `TAROT`, and more innovative designs are described as follows.

Cell Representations via Masked Autoencoder Transformers (MAE). Previous investigations process gene expression value by biology priors as we used in Section 3.2. However, recent advancements in deep representation learning have shown significant improvements in extracting relevant features from data, which can enhance performance on various downstream tasks. By applying these cutting-edge representation learning techniques to single-cell analysis, we can potentially refine temporal trajectory analysis and gain deeper insights into cellular processes. Moreover, the success of MAE He et al. (2022) in representation learning demonstrates the mask prediction in learning better feature representation in a data-driven way. Therefore, `TAROT` tailors an MAE transformer He et al. (2022) to extract superior cell representations from the gene expressions $\{\mathbf{r}_i\}_{i=1}^{n}$. Figure 5 illustrates the MAE procedure: the raw signals are first masked and fed into the MAE encoder; then,

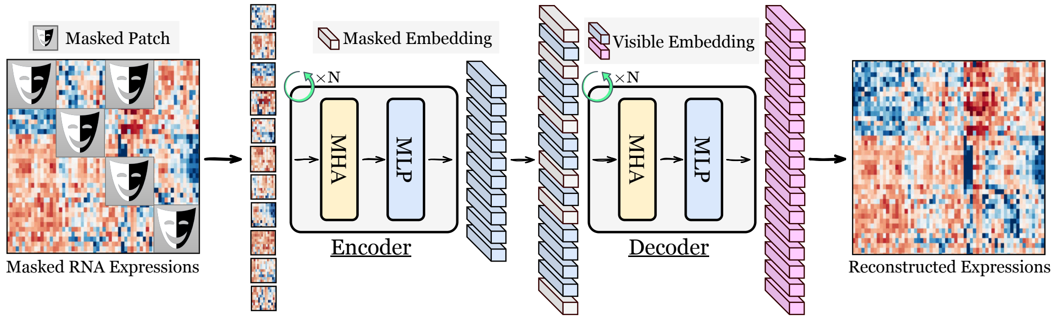

Figure 5: The overall procedure of MAE in TAROT.

masked embeddings are incorporated to align with the full input dimensions; finally, the decoder reconstructs the original input data and computes the MSE training objective. In the inference phase, TAROT adopts the MAE encoder to generate cell representations.

Biology Priors Regularize Cost Function. Another critical component of TAROT is the cost function which include two essential biological aspects: neuron development and gene expression. ① (*Developmental*) *The natural cell differentiation never look back*. In other words, clusters in $\mathcal{G}^{(t)}$ can not be mapped back to ancestor clusters from history trajectories $\{\mathcal{T}_{1,2}, \cdots, \mathcal{T}_{t-1,t}\}$. Specifically, an extra cost penalty $\mathcal{D}^{\texttt{dev}}_{i,j}$ is applied if the cluster j from $\mathcal{G}^{(t+1)}$ at time $t+1$ is an ancestor of the cluster i from $\mathcal{G}^{(t)}$ at time t. ② (*Gene expression*) *The expressions of developmental-related genes satisfy particular patterns.* A specific group of genesoften shows monotonically increasing or decreasing expression during cell differentiation. If a mapping $\mathcal{T}_{t,t+1}$ meet this prior, an additional cost bonus $\mathcal{D}^{\texttt{fuc}}_{i,j}$ will be introduced. Incorporating these biology regulations (①+②), the final cost function is $\tilde{\mathcal{D}} = (\mathcal{D}^{\texttt{dev}} + \mathcal{D}^{\texttt{fuc}}) \odot \mathcal{D}$, with $\odot$ signifying element-wise product and $\mathcal{D} = 1 - \texttt{corr}(\mathcal{G}^{(t)}, \mathcal{G}^{(t+1)})$ representing the cost from cell representatio correlations. Please refer to Appendix B.3 for the definition of $\texttt{corr}(\cdot, \cdot)$.

4.2 TAROT: Pseudotime Calculation via Continuous Trajectory Optimization

Cellular dynamic processes, such as the cell cycle, cell differentiation, and cell activation, can be modeled computationally by pseudotime analysis which orders cells along a trajectory. This method facilitates the reconstruction of the dynamic gene expression profiles that are widely used to study cell differentiation Trapnell (2015); Butler et al. (2018); Crinier et al. (2021), immune responses Yao et al. (2019), disease development Herring et al. (2018), and others. The first stage of TAROT outputs discrete time orders. Then, TAROT executes fitting optimization to get continuous pseudotime to support more fine-grained analysis.

Continuous Trajectory via B-Splines. While previous methods predominantly relied on principal curves to construct continuous cellular trajectories—offering robustness against noisy data—they often overlook critical gene mutations. In the realm of computational analysis, overlooking these mutations and complex genetic variations can significantly impede our understanding of cellular dynamics. Therefore, we explore the possibility of optimization methods in continuous single-cell temporal trajectory construction that can detect mutation signals while maintaining robustness for noisy data. To be specific, we design the trajectory optimization method based on B-Spline in TAROT. The flexible nature of the B-Spline facilitates the integration of trajectory optimization with the connection of discrete temporal orders. A K-degree B-Spline is defined as $\mathcal{C}(u) = \sum_{i=0}^{\mathrm{I}} \mathcal{N}_{i,k}(u) \cdot p_i$, where $\{\mathcal{N}_{i,k}(\cdot)\}_{i=0,k=0}^{i=\mathrm{I}-1,k=\mathrm{K}}$ are bases and the I is the number of control points $\{p_i\}$.

More details about the bases of B-Splines are provided in Appendix A. In TAROT, the previous step outputs the order of each cluster, and we construct the B-Spline from these sequential clusters. For each lineage, we insert J learnable control points $\{p_i^{(j)}\}|_{j=1}^{\mathrm{J}}$ between two fixed control points p_i and p_{i+1}. TAROT treats the averaged cell presentation of each cluster as the fixed control point. And the continuous trajectory optimization (Figure 4 - *Right*) is described as $\texttt{min}_{\{p_i^{(j)}\}_{i=0,j=1}^{i=\mathrm{I}-1,j=\mathrm{J}}} \sum_{k=0}^{n'} \|\mathbf{c}_k -$

Methods	CT ↑	GPT-G ↑	GPT-L ↑	TOC ↑	TTE ↓	CT ↑	GPT-G ↑	GPT-L ↑	TOC ↑	TTE ↓
	Mouse-RGC					Mouse-CCC				
Slingshot	5.28	41.04	42.97	72.22	2.07	10.00	67.90	77.16	68.12	1.13
Monocle-3	7.29	47.04	49.52	48.76	**0.12**	34.17	55.56	61.11	58.68	**0.44**
MOSCOT	67.21	44.32	44.54	55.62	0.78	52.55	67.78	82.44	62.62	1.17
TAROT	**74.53**	**60.73**	**61.17**	**92.10**	0.22	**62.16**	**90.64**	**88.60**	**93.50**	0.58

Table 1: Performance comparisons of TAROT (Ours) vs. diverse representative baselines on Mouse-RGC and Mouse-CCC datasets. Note that Mouse-iEP is mainly used for a real case study of simulated gene knockout.

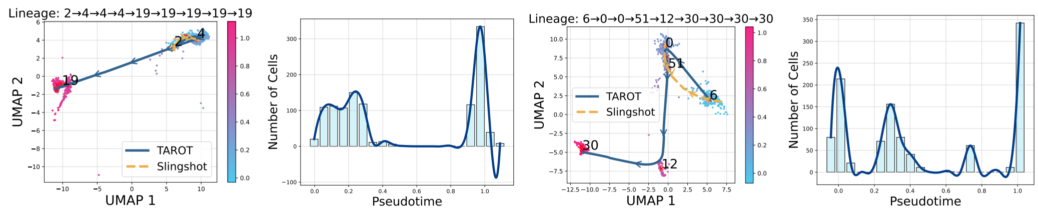

Figure 6: Two inferred lineages and their corresponding pseudotime distributions from TAROT (Ours) and Slingshot (Baseline) on the Mouse-RGC dataset. The color bar indicates the value of cell pseudotime.

$\mathbb{P}(\mathbf{c}_k, \mathcal{C}(u))\|^2$, where $\mathbb{P}(\mathbf{c}_k, \mathcal{C}(u))$ is the projection of cell representation $\mathbf{c}_k$ on $\mathcal{C}(u)$, and n' is the total number of cells on the lineage. Then, the pseudotime $u(\mathbf{c}_k)$ is derived as $\texttt{argmin}_u \|\mathbf{c}_k - \mathcal{C}(u)\|^2, u \in [0, 1]$.

5 Experiments

5.1 Implementation Details

Evaluation Metrics. We introduce five evaluation metrics to measure the quality of temporal trajectories from TAROT and other baselines. Specifically, metrics {❶, ❷, ❸} and {❹, ❺} are created to measure the quality of cell lineage and pseudotime, respectively. ❶ *Correlation Test (CT) for Lineages.* We compute the ratio of lineages that pass the correlation test as $\frac{1}{\mathrm{s}-1}\sum_t^{\mathrm{s}-1}\frac{1}{|\mathcal{T}_{t,t+1}|}\sum_{l_t\in\mathcal{T}_{t,t+1}}\texttt{CT}(l_t)$, where $\mathcal{T}_{t,t+1}$ is set of mappings $\{l_t : \mathcal{C}_i^{(t)} \rightarrow \mathcal{C}_j^{(t+1)}\}$ from time stage t to $t+1$. The $\texttt{CT}(l_t)$ is the indicator function that returns 1 if the spearman correlation between averaged cell representations from $\mathcal{C}_i^{(t)}$ and $\mathcal{C}_i^{(t+1)}$ is the highest one; returns 0, otherwise. ❷ *Gene Pattern Test per Gene (GPT-G)* and ❸ *Gene Pattern Test per Lineage (GPT-L).* Based on the developmental and functional priors, we select an extra group of genes for testing, which are not utilized during the TAROT design. The selection follows the widely adopted standards (Finak et al., 2015). Such genes are experimentally validated to have monotonically increased or decreased expressions along with the cell differentiation (or the cell pseudotime). For each test gene, we first compute the percentage of lineages where the gene exhibits monotonicity. Then, averaging the result across all test genes produces the accuracy of GPT-G. Similarly, we first calculate the percentage of genes that exhibit monotonicity along with a given lineage. Then, averaging the result across all inferred lineage generates the accuracy of GPT-L. ❹ *Time Order Consistency Test (TOC) for Lineage.* It examines whether the optimized cell pseudotime is aligned with the time order in lineages. We focus on the tuning point of lineages where the cell differentiation happens *i.e.*, the cell type changes. If the tuning point cluster is $\mathcal{C}_i^{(t)}$, we compute the accuracy of TOC as $\frac{1}{|\mathcal{C}_i^{(t)}|}\sum_{\mathbf{c}_i\in\mathcal{C}_i^{(t)}}\frac{|\{u(\mathcal{G}_i^{(t-1)})<u(\mathbf{c}_i)\}|+|\{u(\mathcal{G}_i^{(t+1)})>u(\mathbf{c}_i)\}|}{|\mathcal{G}_i^{(t-1)}|+|\mathcal{G}_i^{(t+1)}|}$, where $\{u(\mathcal{G}_i^{(t-1)}) < u(\mathbf{c}_i)\}$ is a set of cells that belong to $\mathcal{G}_i^{(t-1)}$ and has a smaller pseudotime than $\mathbf{c}_i$. The reported accuracy of TOC is averaged across all tuning points and lineages. ❺ *Temporal Trajectory Error (TTE)* is the average distance between cells to their corresponding projection on the temporal trajectory, *i.e.*, $\sum_{i=0}^{n'}\sqrt{\|\mathbf{c}_i - \mathbb{P}(\mathbf{c}_i, \mathcal{C}(u))\|^2}$. Other details like TAROT's training setups are in Appendix B.

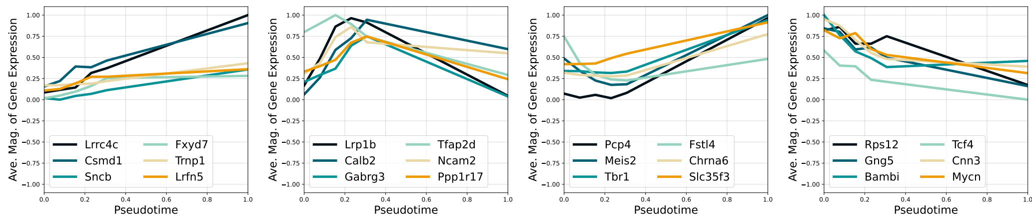

Figure 7: Gene expression dynamics over the cell pseudotime. Four kinds of special gene patterns, from *left* to *right*, are increased, increased then decreased, decreased then increased, and decreased gene waves.

5.2 Superior Performance of TAROT in Lineage and Pseudotime Inference

In this section, we examine the quality of lineage and pseudotime produced by our proposed `TAROT`. Three representative baselines, *i.e.*, Slingshot (Street et al., 2018), Monocle-3 (Cao et al., 2019), and MOSCOT (Klein et al., 2023), are adopted for throughout comparisons. They are distinctive frameworks based on minimum spanning trees, principal graphs, and optimal transport algorithms, respectively. Experimental results on `Mouse-RGC` and `Mouse-CCC` are presented in Figure 6 and Table 1, where several consistent observations can be drawn: ❶ Our `TAROT` demonstrates great advantages with a clear performance margin compared to Slingshot, Monocle-3 and MOSCOT. In detail, for evaluation metrics {CT (↑ %), GPT-G (↑ %), GPT-L (↑ %), TOC (↑ %), TTE (↓)}, `TAROT` obtains $\{65.03\%, 19.70\%, 18.11\%, 20.82\%, 1.75\}$, $\{63.02\%, 13.70\%, 11.56\%, 44.28\%, -0.16\}$ and $\{3.10\%, 16.42\%, 16.54\%, 37.42\%, 0.50\}$ performance improvements on `Mouse-RGC` and $\{52.16\%, 22, 74\%, 11.44\%, 25.38\%, 0.55\}$, $\{27.99\%, 35.08\%, 27.49\%, 34.82\%, -0.14\}$ and $\{9.61\%, 22.86\%, 6.16\%, 30.88\%, 0.59\}$ on `Mouse-MCC`, respectively. Note that a negative TTE gain implies a lower error rate for pseudotime optimization. Such impressive outcomes validate the effectiveness of our proposal. ❷ Although Monocle-3 obtains a lower *Temporal Trajectory Error* (*e.g.*, 0.18 and 0.14 lower), it fails short in terms of *Time Order Consistency Test* (44.28% and 34.82% worse for the accuracy), compared to our `TAROT`. It suggests that Monocle-3 probably sacrifices the correctness of pseudotime to better fit the B-Splines. In contrast, `TAROT` achieves higher time order consistency with a comparable fitting error, making it a superior choice for neuron trajectory analyses. ❸ Figure 6 presents two examples of inferred lineages and their pseudotime distributions, where `TAROT` captures a longer range of neuron developmental trajectories.

5.3 Gene Knockout Simulation - Algorithmic Recourse of TAROT

With the superior cell lineage and pseudotime from `TAROT`, we are curious about (1) whether they capture special gene expression patterns; (2) whether these gene patterns are biologically meaningful; (3) how to manipulate them to influence the cell differentiation.

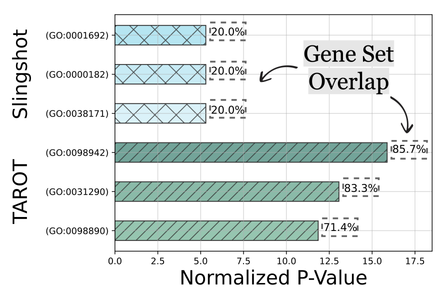

Figure 8: GSEA results of the identified gene sets from Slingshot and `TAROT`. A higher ratio of gene set overlap and a larger normalized p-value ($|\mathtt{log}_{10}(\mathtt{p} - \mathtt{value})|$) suggests a stronger association with biologically meaningful GO terms.

Gene Pattern Identification. It is another important angle to dissect the effectiveness of `TAROT`: examining whether the predicted temporal trajectory can capture clear gene expression patterns. Given one lineage, we record four representative patterns of gene "waves", as presented in Figure 7. We see that in our inferred lineage, the expression values of several gene subgroups consistently increase or decrease, followed by a decrease or vice versa, respectively. For most of `TAROT`'s lineages, a gene set with similar expression patterns can be identified, as shown in Appendix C. The next step is to validate the biological semantics of these located gene groups.

Biologically Meaningful? Do the Pathway Alignment. We use the GSEA (Fang et al., 2023) for the pathway alignment analysis. It is a method to determine whether the input gene set has statistically significant relationships with pathway gene sets of GO terms in biology. We apply GSEA to the selected gene group from `TAROT` and Slingshot, and GSEA considers 22 different mouse gene libraries for the alignment. Figure 8 records the top-3 aligned GO terms with the highest gene set overlap ratio. Meanwhile, their normalized p-values are also reported in the x-axis. TAROT achieves markedly higher values of both metrics indicating its superiority in identifying biologically relevant gene sets.

Algorithmic Recourse for TAROT - Simulating the Gene Knockout. To testify to the importance of found genes for cellular temporal trajectory and differentiation, we perform an algorithmic recourse of TAROT by removing these genes during the trajectory optimization to simulate the gene knockout. TAROT results with and without the gene removal are summarized in Figure 9. We can see the trajectory (or differentiation) is significantly altered after even only removing one gene (e.g., TPT1).

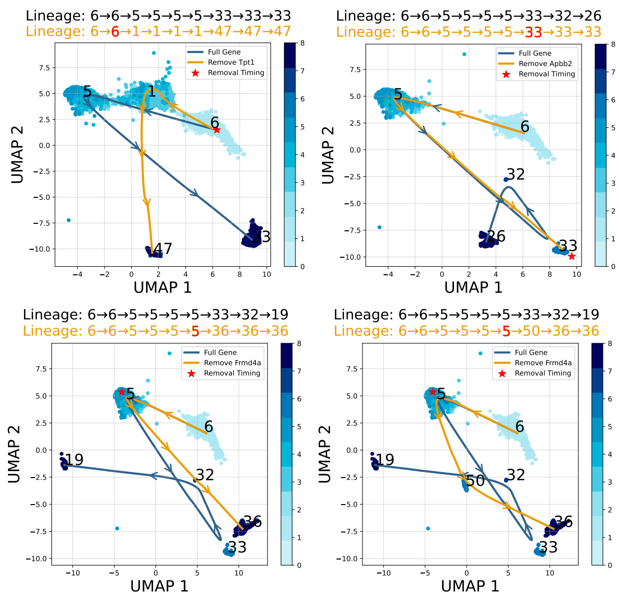

Figure 9: The simulated gene knockout. During the tuning point (red numbers and ⋆) of cell lineage, we knock one of the previously identified genes, leading to totally different cell differentiation.

A Real Case Study of Gene Knockout. The next key question is whether our simulated results echo with the wet lab experiment. Rekhtman et al. (1999) provides a real experimental validation on the Mouse-iPE dataset (Capellera-Garcia et al., 2016) and proves that knockout genes GATA1, SPI1, and LMO2 will discourage the conversion from murine and human fibroblasts to induced erythroid progenitor or precursor cells (iEPs). Impressively, we find that TAROT offers aligned simulation results: *removing these genes impedes the mapping (cell differentiation) to the original iEPs*. In details, the initial mappings from TAROT are {cell: Meg $\rightarrow$ Bas; cell: Neu $\rightarrow$ Mon}. If we remove gene GATA1 and SPI1, the simulated results become {cell: Meg $\rightarrow$ Bas, GMP-like, MEP-like; cell: Neu $\rightarrow$ Mon, Bas, GMP-like}. It implies that TAROT successfully reveals a seesaw-effect regulation between SPI1 and GATA1 in driving the GMP-like and MEP-like lineages.

5.4 Ablation Study

To investigate the contribution of each component in TAROT, comprehensive ablations are conducted on Mouse-RGC. We study the effects of different cell representations, biology prior regularization, continuous trajectory optimization, and the automatic thresholding methods in TAROT, please refer Appendix C.5, Appendix C.6, and Appendix C.4 for more details. We also investigate the relationship between TAROT and cluster quality in Appendix C.3.

6 Conclusions

Modeling and inferring single-cell transcriptional patterns is crucial to understanding cell differentiation in developmental biology. This paper presents a novel angle to formulate this fundamental biology problem into a well-defined machine learning formulation - temporal trajectory analysis. We propose a large-scale single-cell dataset of mouse retinal ganglion (Mouse-RGC) and an innovative algorithmic framework TAROT to: (1) extract superior cell representations; (2) match feature distributions across time stages; (3) optimize and produce continuous temporal trajectories. Extensive investigations validate that our proposals achieve substantial improvements over baseline methods. Lastly, various gene knockout simulations and a real case study are conducted, where the impressive results imply the potential of TAROT in providing meaningful biology landscapes. Future work includes more physical validations of mouse gene knockout and potential applications like gene therapy and cell longevity engineering.

REFERENCES

Jonas Adler and Sebastian Lunz. Banach wasserstein gan. *Advances in neural information processing systems*, 31, 2018.

Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8309–8319, 2018.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008 (10):P10008, 2008.

Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

Veit R Buchholz, Ton NM Schumacher, and Dirk H Busch. T cell fate at the single-cell level. *Annual review of immunology*, 34:65–92, 2016.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

Zixuan Cang, Yangyang Wang, Qixuan Wang, Ken WY Cho, William Holmes, and Qing Nie. A multiscale model via single-cell transcriptomics reveals robust patterning mechanisms during early mammalian embryo development. *PLoS computational biology*, 17(3):e1008571, 2021.

Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

Sandra Capellera-Garcia, Julian Pulecio, Kishori Dhulipala, Kavitha Siva, Violeta Rayon-Estrada, Sofie Singbrant, Mikael NE Sommarin, Carl R Walkley, Shamit Soneji, Göran Karlsson, et al. Defining the minimal factors required for erythropoiesis through direct lineage conversion. *Cell reports*, 15(11):2550–2562, 2016.

Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16296–16305, 2021.

Adeline Crinier, Pierre-Yves Dumas, Bertrand Escalière, Christelle Piperoglou, Laurine Gil, Arnaud Villacreces, Frédéric Vély, Zoran Ivanovic, Pierre Milpied, Émilie Narni-Mancinelli, et al. Single-cell profiling reveals the trajectories of natural killer cell differentiation in bone marrow and a stress signature induced by acute myeloid leukemia. *Cellular & molecular immunology*, 18(5): 1290–1304, 2021.

Haotian Cui, Chloe Wang, Hassaan Maan, Nan Duan, and Bo Wang. scformer: A universal representation learning approach for single-cell data using transformers. *bioRxiv*, pp. 2022–11, 2022.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, pp. 2023–04, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Daniela J Di Bella, Ehsan Habibi, Robert R Stickels, Gabriele Scalia, Juliana Brown, Payman Yadollahpour, Sung Min Yang, Catherine Abbate, Tommaso Biancalani, Evan Z Macosko, et al. Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature*, 595(7868):554–559, 2021.

Rasa Elmentaite, Cecilia Domínguez Conde, Lu Yang, and Sarah A Teichmann. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nature Reviews Genetics*, 23(7):395–410, 2022.

Steven N Evans and Frederick A Matsen. The phylogenetic kantorovich–rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(3):569–592, 2012.

Jean Fan, Kamil Slowikowski, and Fan Zhang. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine*, 52(9):1452–1465, 2020.

Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.

Jeffrey A Farrell, Yiqun Wang, Samantha J Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015.

Michael J Geuenich, Dae-won Gong, and Kieran R Campbell. The impacts of active and self-supervised learning on efficient annotation of single-cell expression data. *bioRxiv*, pp. 2023–06, 2023.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Statistical inference with regularized optimal transport. *arXiv preprint arXiv:2205.04283*, 2022.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Jonathan A Griffiths, Antonio Scialdone, and John C Marioni. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology*, 14(4): e8046, 2018.

Gunsagar S Gulati, Shaheen S Sikandar, Daniel J Wesche, Anoop Manjunath, Anjan Bharadwaj, Mark J Berger, Francisco Ilagan, Angera H Kuo, Robert W Hsieh, Shang Cai, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405–411, 2020.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296, 2019.

Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848, 2016.

Xiaoping Han, Haide Chen, Daosheng Huang, Huidong Chen, Lijiang Fei, Chen Cheng, He Huang, Guo-Cheng Yuan, and Guoji Guo. Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell rna-sequencing. *Genome biology*, 19(1):1–19, 2018.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Le Song, and Xuegong Zhang. Large scale foundation model on single-cell transcriptomics. *bioRxiv*, pp. 2023–05, 2023.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021. doi: 10.1016/j.cell.2021.04.048. URL https://doi.org/10.1016/j.cell.2021.04.048.

Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, 9 (1):1–12, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

W Helland-Hansen and GJ Hampson. Trajectory analysis: concepts and applications. *Basin Research*, 21(5):454–483, 2009.

Charles A Herring, Amrita Banerjee, Eliot T McKinley, Alan J Simmons, Jie Ping, Joseph T Roland, Jeffrey L Franklin, Qi Liu, Michael J Gerdes, Robert J Coffey, et al. Unsupervised trajectory analysis of single-cell rna-seq and imaging data reveals alternative tuft cell origins in the gut. *Cell systems*, 6(1):37–51, 2018.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

Karthik A Jagadeesh, Kushal K Dey, Daniel T Montoro, Rahul Mohan, Steven Gazal, Jesse M Engreitz, Ramnik J Xavier, Alkes L Price, and Aviv Regev. Identifying disease-critical cell types and cellular processes by integrating single-cell rna-sequencing and human genetics. *Nature genetics*, 54(10):1479–1492, 2022.

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343 (6172):776–779, 2014.

Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

Qingzhu Jia, Han Chu, Zheng Jin, Haixia Long, and Bo Zhu. High-throughput single-cell sequencing in cancer research. *Signal Transduction and Targeted Therapy*, 7(1):145, 2022.

Junil Kim, Michaela Mrugala Rothová, Esha Madan, Siyeon Rhee, Guangzheng Weng, António M Palma, Linbu Liao, Eyal David, Ido Amit, Morteza Chalabi Hajkarim, et al. Neighbor-specific gene expression revealed from physically interacting cells during mouse embryonic development. *Proceedings of the National Academy of Sciences*, 120(2):e2205371120, 2023.

Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimee Bastidas-Ponce, Marta Tarquis-Medina, et al. Mapping cells through time and space with moscot. *bioRxiv*, pp. 2023–05, 2023.

Andrew L Koenig, Irina Shchukina, Junedh Amrute, Prabhakar S Andhey, Konstantin Zaitsev, Lulu Lai, Geetika Bajpai, Andrea Bredemeyer, Gabriella Smith, Cameran Jones, et al. Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nature cardiovascular research*, 1(3):263–280, 2022.

Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. Sliced wasserstein distance for learning gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436, 2018.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10285–10295, 2019.

Xiangyu Li, Weizheng Chen, Yang Chen, Xuegong Zhang, Jin Gu, and Michael Q Zhang. Network embedding-based representation learning for single cell rna-seq data. *Nucleic acids research*, 45 (19):e166–e166, 2017.

Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4832–4841, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Bidesh Mahata, Xiuwei Zhang, Aleksandra A Kolodziejczyk, Valentina Proserpio, Liora Haim-Vilmovsky, Angela E Taylor, Daniel Hebenstreit, Felix A Dingler, Victoria Moignard, Berthold Göttgens, et al. Single-cell rna sequencing reveals t helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell reports*, 7(4):1130–1142, 2014.

Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. Doubletfinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell systems*, 8(4):329–337, 2019.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Markus Mittnenzweig, Yoav Mayshar, Saifeng Cheng, Raz Ben-Yair, Ron Hadas, Yoach Rais, Elad Chomsky, Netta Reines, Anna Uzonyi, Lior Lumerman, et al. A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell*, 184(11):2825–2842, 2021.

David T Miyamoto, Yu Zheng, Ben S Wittner, Richard J Lee, Huili Zhu, Katherine T Broderick, Rushil Desai, Douglas B Fox, Brian W Brannigan, Julie Trautwein, et al. Rna-seq of single prostate ctcs implicates noncanonical wnt signaling in antiandrogen resistance. *Science*, 349(6254): 1351–1356, 2015.

Axel Munk and Claudia Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1): 223–241, 1998.

Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

Henning Petzka, Asja Fischer, and Denis Lukovnicov. On the regularization of wasserstein gans. *arXiv preprint arXiv:1709.08894*, 2017.

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

Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10): 979–982, 2017.

Natasha Rekhtman, Farshid Radparvar, Todd Evans, and Arthur I Skoultchi. Direct interaction of hematopoietic transcription factors pu. 1 and gata-1: functional antagonism in erythroid cells. *Genes & development*, 13(11):1398–1411, 1999.

Abbas H Rizvi, Pablo G Camara, Elena K Kandror, Thomas J Roberts, Ira Schieren, Tom Maniatis, and Raul Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature biotechnology*, 35(6):551–560, 2017.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453): 236–240, 2013.

Karthik Shekhar, Irene E Whitney, Salwan Butrus, Yi-Rong Peng, and Joshua R Sanes. Diversification of multipotential postmitotic mouse retinal ganglion cell precursors into discrete types. *Elife*, 11: e73809, 2022.

Hongru Shen, Jilei Liu, Jiani Hu, Xilin Shen, Chao Zhang, Dan Wu, Mengyao Feng, Meng Yang, Yang Li, Yichen Yang, et al. Generative pretraining from large-scale transcriptomes for single-cell deciphering. *Iscience*, 26(5), 2023.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Satya Narayan Shukla and Benjamin M Marlin. A survey on principles, models and methods for learning from irregularly sampled time series. *arXiv preprint arXiv:2012.00168*, 2020.

Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):219–238, 2018.

Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:1–16, 2018.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Michael JT Stubbington, Orit Rozenblatt-Rosen, Aviv Regev, and Sarah A Teichmann. Single-cell transcriptomics to explore the immune system in health and disease. *Science*, 358(6359):58–63, 2017.

Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10): 1491–1498, 2015.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

Koen Van den Berge, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1201, 2020.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160, 2016.

Irene E Whitney, Salwan Butrus, Michael A Dyer, Fred Rieke, Joshua R Sanes, and Karthik Shekhar. Vision-dependent and-independent molecular maturation of mouse retinal ganglion cells. *Neuroscience*, 508:153–173, 2023.

Fengying Wu, Jue Fan, Yayi He, Anwen Xiong, Jia Yu, Yixin Li, Yan Zhang, Wencheng Zhao, Fei Zhou, Wei Li, et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nature communications*, 12(1):2540, 2021.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3713–3722, 2019.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018.

Chen Yao, Hong-Wei Sun, Neal E Lacey, Yun Ji, E Ashley Moseman, Han-Yu Shih, Elisabeth F Heuston, Martha Kirby, Stacie Anderson, Jun Cheng, et al. Single-cell rna-seq reveals tox as a key regulator of cd8+ t cell persistence in chronic infection. *Nature immunology*, 20(7):890–901, 2019.

Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

Yijie Zhang, Dan Wang, Miao Peng, Le Tang, Jiawei Ouyang, Fang Xiong, Can Guo, Yanyan Tang, Yujuan Zhou, Qianjin Liao, et al. Single-cell rna sequencing in cancer research. *Journal of Experimental & Clinical Cancer Research*, 40:1–17, 2021.

Mengyuan Zhao, Wenying He, Jijun Tang, Quan Zou, and Fei Guo. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. *Briefings in bioinformatics*, 23(2):bbab568, 2022.

Suyuan Zhao, Jiahuan Zhang, and Zaiqing Nie. Large-scale cell representation learning via divide-and-conquer contrastive learning. *arXiv preprint arXiv:2306.04371*, 2023.

IMPACT STATEMENTS

Our research primarily advances the scientific understanding of cellular development and its complex temporal dynamics, with potential implications for fields such as regenerative medicine and oncology. While the societal impacts of these advancements may be far-reaching, specific consequences are beyond the scope of this study and require careful consideration by experts in the relevant fields. We encourage interdisciplinary collaboration to explore the practical applications and ethical considerations of our findings in real-world contexts. While our work has various potential societal implications, we do not identify any specific consequences that warrant particular emphasis in this context.

A MORE TECHNIQUE DETAILS

Details about Entropy Weight Search. The entropy weight λ is a critical factor that affects the final Sinkhorn algorithm transport result; an inadequate λ makes the transport prone to random mapping. We design a non-linear entropy weight search algorithm to decide an adequate λ for the Sinkhorn algorithm. The Pytorch-style pseudo code is presented in Algorithm 1.

Algorithm 1 Non-Linear Entropy Weight Search

Require: Initial entropy weight λ.
Require: The best optimal transport cost $\mathcal{F}_{best} \leftarrow \infty$
Require: The current optimal transport cost $\mathcal{F}_{cur}$

$\lambda_i \leftarrow \lambda$
while $\mathcal{F}_{best} \geq \mathcal{F}_{cur}$ **do**
  $\mathcal{F}_{cur}, \mathcal{T} \leftarrow \texttt{min}_{\mathcal{T}} < \mathcal{T}, \mathcal{D} > -\lambda \mathcal{E}(\mathcal{D})$ // Solving the Optimal Transport optimization by the Sinkhorn algorithm.
  if $\texttt{sum}(T) \leq 1$ **then**
    $\lambda_i =\leftarrow \lambda_i * 10$
  else if $\mathcal{F}_{cur} \leq \mathcal{F}_{best}$ **then**
    $\mathcal{F}_{best} \leftarrow \mathcal{F}_{cur}$
    $\lambda \leftarrow \lambda_i$
    $\lambda_i \leftarrow \lambda_i - \lambda_i * 0.1$
  end if
end while
return λ

A.1 DETAILS OF THE BASE FUNCTION IN B-SPLINES

B-Spline is constructed based on the base function, and the base function is defined recursively:

$$\mathcal{N}_{i,0}(u) = \begin{cases} 1, u_i \leq u \leq u_{i+1} \\ 0, \text{otherwise} \end{cases}, \tag{3}$$

$$\mathcal{N}_{i,k} = \frac{u - u_i}{u_{i+k} - u_i} \mathcal{N}_{i,k-1}(u) + \frac{u_{i+k+1} - u}{u_{i+k+1} - u_{i+1}} \mathcal{N}_{i+1,k-1}(u), \tag{4}$$

where the $\{u_i\}_{i=0}^{n}$ are the knots of the B-Spline. For more details about adapting the B-Spline for pseudotime trajectory optimization, please check Appendix B.

B MORE IMPLEMENTATION DETAILS

B.1 TRAINING DETAILS OF TAROT

B.2 CELL CLUSTERING

For each dataset, we perform cell clustering on cells from all time-stages. Initially, principal component analysis (PCA) is applied to reduce the cell feature dimensionality to 55. Subsequently, we utilize the Louvain clustering algorithm with "resolution" set to 1.0 for the `Mouse-RGC` dataset and 1.5 for the `Mouse-RCC` dataset. The hyper-parameter "resolution" is fine-tuned to ensure that the number of clusters matches the number of cell types in each dataset. Notably, the preprocessing for the `Mouse-RCC` dataset mirrors that of the `Mouse-RGC`. To be specific, we retained cells that expressed at least $1,500$ and less than 10500 genes. We remove cells that have more than 5% mitochondrial genes and genes expressed in fewer than 10 cells. The same normalization and batch correction methods used for the `Mouse-RGC` are then applied to the `Mouse-RCC` dataset to identify $5,000$ anchor genes per cell.

B.3 Biology Priors Integration

Developmental Cost. When the cluster j from $\mathcal{G}^t$ at time t, the $\mathcal{D}^{dev}_{i,j}$ will be set to $\max(\mathcal{D})/\mathcal{D}_{i,j} + 1$ if the cluster j from $\mathcal{G}^{t+1}$ is an ancestor of the cluster i from $\mathcal{G}^t$ at time t, otherwise $\mathcal{D}^{dev}_{i,j} = 1$.

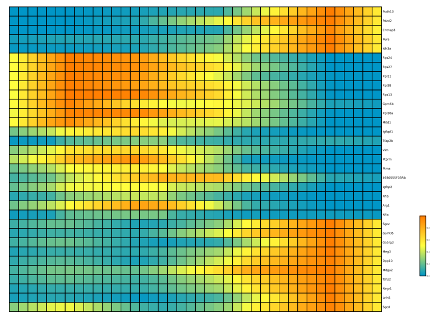

Figure 10: Gene expression dynamics over the cell pseudotime *left* to *right*. From *top* to *bottom* are the heatmap of different genes that increase, decrease, increase followed decrease, and decrease followed increase.

Gene Expression Cost. According to the labeled time stage, we calculate the Pearson correlation between gene expression value and cell time stage. For each gene, we record the Pearson product-moment correlation coefficient and the p-value associated with the gene. The Pearson product-moment correlation coefficient $\texttt{Pcc}_i$ indicates the monotonicity of the gene i, and the p-value $\texttt{Pv}_i$ denotes the degree of the monotonicity. We selected 400 genes with the largest p-value as the *gene group* for $\mathcal{D}^{fuc}$. The calculation of the $\mathcal{D}^{fuc}$ can be found in Algorithm 2, where `G_cur` is $\mathcal{G}^{t+1}$, `G_prev` denotes $\{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^t\}$, `GeneGroup` denotes the *gene group*, `Pcc` denotes the Pearson product-moment correlation coefficient of the *gene group*, `Pv` denotes the p-value of above mentioned *gene group*, and `D_fuc` is the $\mathcal{D}^{fuc}$.

Algorithm 2 The Gene Expression Cost Calculation.

Require: $G_{cur} \leftarrow \mathcal{G}^{t+1}$
Require: $G_{prev} \leftarrow \{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^t\}$
Require: The Pearson product-moment correlation `Pcc`
Require: the *gene group* `GeneGroup`
Require: the p-value of `GeneGroup` `Pv`
Require: $\mathcal{D}^{fuc}$

$\lambda_i \leftarrow \lambda$
$Bonus \leftarrow []$
for *gene* in `GeneGroup` **do**
  if G_{prev}, G_{cur} is monotone monotonic **then**
    $Bonus.append(\texttt{Pv}[gene])$
  end if
  $\mathcal{D}^{fuc}[gene] \leftarrow 1$
  $\mathcal{D}^{fuc}[gene] \leftarrow 1 - mean(Bonus)$
end for
return $\mathcal{D}^{fuc}$

Pre-Transport - MAE training `TAROT` employs a customized transformer network comprising 6 encoder layers and 6 decoder layers. The encoder layers boast a dimension of 256 with 8 attention heads, while the decoder layer has a hidden dimension of 512 and is also equipped with 8 attention heads. Our MAE uses AdamW (Loshchilov & Hutter, 2019) optimizer with the weight decay of $1e^{-5}$, the learning rate of $1e^{-4}$, and the training step of 50K, wherein the initial 2.5K iterations as a warmup. For the single-cell data, we divide each cell's genes into 128 patches, where each patch contains 64 consecutive gene expression values. The final cell representation $\{\mathbf{c}_i\}_{i=1}^n$ is obtained by feeding the encoder output into PCA, which reduces the dimension from 256 to 55.

During Transport - Regularized OT We use Pearson correlation as the vanilla cost function $\texttt{corr}(\cdot,\cdot)$. The final entropy weight λ of each optimal transport is obtained by Algorithm 1.

Post-Transport - B-Splines Trajectory Optimization The curve parameter is predefined before the trajectory optimization. We use 3-degree B-Spline with 300 knots, and the number of learnable control points J is set with 1. The optimization is solved via gradient descent, the learning rate is 1×10^{-2}, and the optimization stop condition is the loss fluctuation is less than $1 \times 10^{-4}\%$ with most $1,000$ optimization steps.

B.4 Details About Metric

For further clarification of differences between GPT-G and GPT-P, we provide the PyTorch-style pseudo codes for both two metrics in Algorithm 3 and 4 respectively.

Algorithm 3 Gene Pattern Test per Gene.

Require: All lineage we have A
Require: All cells C
Require: Specified gene group Sg
1: Gr← []// The gene ratio recorder
2: **for** lineage a in A **do**
3: gr← $\texttt{NGenesInLineage}(a, C, Sg)$// Calcuate the percentage of genes in Sg, which steady increase/decrease over this lineage
4: Gr.$append$(gr)
5: **end for**
6: **return** mean(Gr)

C More Experimental Results

C.1 More Results of Gene Wave Visualization

To illustrate the capability of TAROT in discovering gene sets with specific patterns from lineages, we collect more gene waves with such expression patterns and show them in Figure 11 and 10 in different forms. Results show that TAROT yields more genes with similar expression patterns since the high-quality lineage and pseudotime inference.

Algorithm 4 Gene Pattern Test per Path.

Require: All lineage we have A
Require: All cells C
Require: Specified gene group Sg
1: Pr← [ ]// The path ratio recorder
2: **for** gene a in Sg **do**
3: pr← $\texttt{NPathInLineage}(A, C, gene)$// Calculate the percentage of paths in A, which the specific gene steadily increases/decreases over these paths
4: Pr.$append$(pr)
5: **end for**
6: **return** mean(Pr)

C.2 More Ablation Results

Different Options for Pseudotime Trajectory Optimization The number of learnable points J between two fixed control points greatly affects the B-Spline pseudotime trajectory optimization. We ablate different J to seek a plausible setting for the pseudotime trajectory optimization. Table 4 indicates that more learnable control points deliver improved TTE but sacrifice TOC, which indicates better trajectory fitting does not result in better pseudotime trajectory. Therefore, we use “1” learnable control points per two fixed points. We also compare our method with two other trajectory fitting

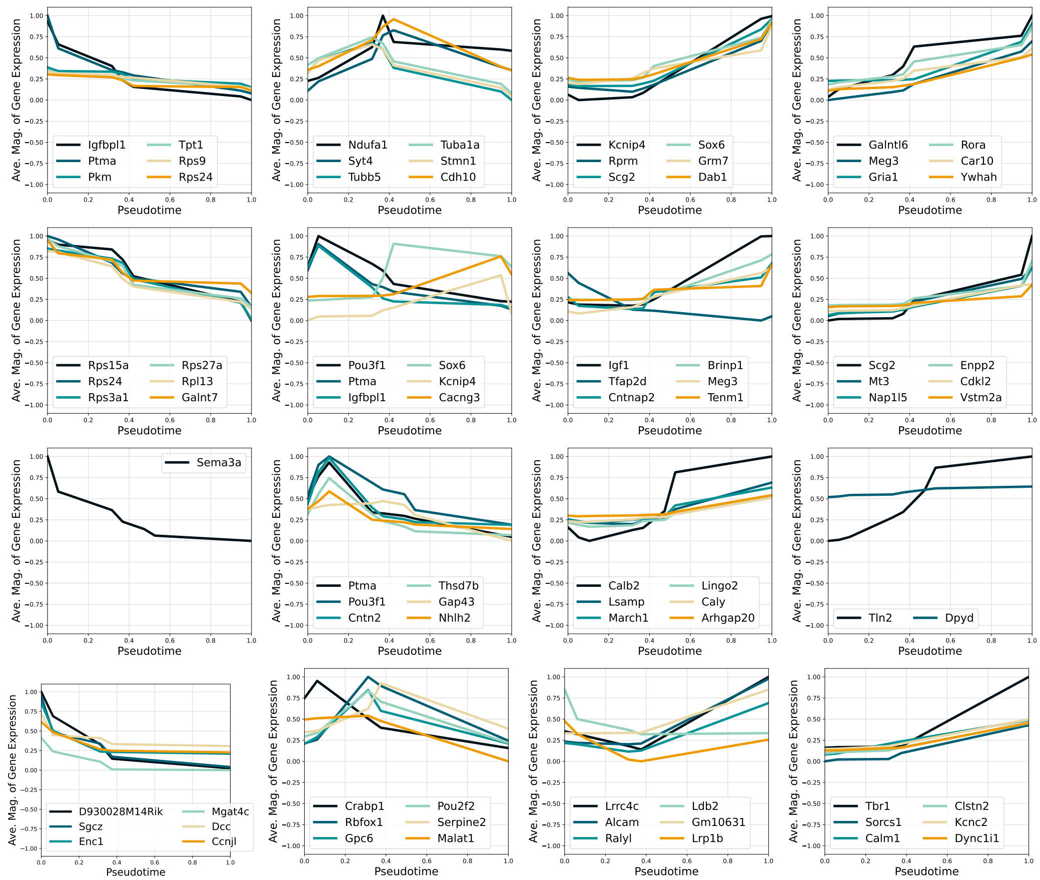

Figure 11: More gene expression dynamics over the cell pseudotime. Four kinds of special gene patterns, from *left* to *right*, are increased, increased then decreased, decreased then increased, and decreased gene waves. Gene waves in different lines are identified from different lineages.

Mouse-RGC	CT ↑	GPT-G ↑	GPT-L ↑	TOC ↑	TTE ↓
P-value	66.58	56.30	56.89	92.86	**0.23**
max. sep.	**74.53**	**60.73**	**61.17**	**92.10**	**0.22**

Table 2: Ablations on automatic thresholding.

methods: the “Poly.” method, which uses the polynomial curve for temporal trajectory fitting (the degree of curve is the number of cell differentiations that happen), and the “Principal” method, which uses the principal curves algorithm, the same trajectory fitting method with Slingshot (Street et al., 2018).

C.3 The Quality of Clustering

At the outset of TAROT, cell clustering is a critical step in data preprocessing. Consequently, we investigate the effects of clustering quality and the application of various cluster methods for TAROT. The results, presented in Table 3, underscore the significance of cluster algorithm for the performance of TAROT. In parallel, variations in the “resolution” parameter of the Louvain method influence TAROT’s performance. Nonetheless, altering the “resolution” modulates the cluster count, complicating the association between clusters and specific cell types, and potentially diminishing the biological insights gleaned.

Table 3: Ablations different cluster hyper-parameter and cluster method. The "`resolution`" parameter influences the cluster count in the Louvain algorithm Blondel et al. (2008). A smaller "`resolution`" yields fewer clusters, whereas a larger "`resolution`" leads to more clusters.

`Mouse-RGC`	CT ↑	GPT-G ↑	GPT-L ↑	TOC ↑	TTE ↓
`TAROT`, `resolution` $= 1.0$	**74.53**	60.73	61.17	92.10	0.22
Different Clustering Config					
`TAROT`, `resolution` $= 0.5$	70.76	**60.97**	**61.20**	92.30	0.25
`TAROT`, `resolution` $= 1.5$	74.44	58.64	58.35	90.63	0.21
`TAROT`, `resolution` $= 2.0$	72.97	60.72	59.99	**92.37**	**0.20**
Different Cluster Algorithm					
`KMean`	72.68	59.22	59.83	90.22	0.23
`Agglomerative clustering`	68.93	46.80	46.87	88.75	0.24

Table 4: Result of different trajectory optimization options.

Methods	TOC ↑	TTE ↓
	`Mouse-RGC`	
Sp-1	**93.41**	0.22
Sp-2	90.73	0.25
Sp-3	92.03	**0.23**
Poly.	63.37	1.24
Principal	72.22	2.07

C.4 AUTOMATIC THRESHOLDING WITHIN `TAROT`

In `TAROT`, the automatic thresholding method is vital to achieving accurate lineage results. We proposed two candidate automatic thresholding techniques: the "P-value" method and the "max. sep." method (*i.e.*, the maximum separation). The "P-value" method utilizes statistical significance to identify mappings with a p-value lower than the threshold of $1e^{-4}$. Conversely, the "max. sep." method selects mappings with close OT costs but notably distinct from other mappings, emulating human intuition. Table 4 reports `TAROT`'s results with two thresholding methods, demonstrating that the "max. sep." selects lineages with higher quality.

C.5 DIFFERENT CELL REPRESENTATION.

A high-quality cell representation is essential for inferring temporal trajectory. To this end, we have implemented the `TAROT` algorithm using various representations derived from a range of sources, including `PCA` (Principal Component Analysis), `UMAP` (Uniform Manifold Approximation and Projection), `VAE` (Variational Autoencoder), and `MAE` (Masked Autoencoder). For the purpose of fair comparison, `PCA` has been applied to the features extracted by both `VAE` and `MAE` to reduce their dimensionality to 55 dimensions. The results presented in Table 6 substantiate the superiority of the `MAE`-based representation.

C.6 DIFFERENT BIOLOGY PRIOR REGULARIZATIONS.

The `TAROT` algorithm's flexibility allow for the integration of various forms of biological prior knowledge during the transport process. We have systematically examined different integration strategies, the details of which are presented in Table 6. Our analysis reveals that both the developmental cost $\mathcal{D}^{\texttt{dev}}$ and the funtional cost $\mathcal{D}^{\texttt{fuc}}$ contribute significantly to the performance of `TAROT`.

Table 5: Ablations on cell representations.

`Mouse-RGC`	CT ↑	GPT-G ↑	GPT-L ↑	TOC ↑	TTE ↓
PCA-55	69.10	53.90	54.04	74.55	0.77
UMAP-2	29.44	16.16	16.14	62.69	0.58
VAE	66.43	53.02	53.07	72.17	0.85
MAE	**74.53**	**60.73**	**61.17**	**92.10**	**0.22**

Table 6: Ablations on biology prior regularizations.

`Mouse-RGC`	CT ↑	GPT-G ↑	GPT-L ↑	TOC ↑	TTE ↓
$\mathcal{D}$	66.28	39.60	39.60	73.41	0.64
$\mathcal{D}^{\text{dev}} \odot \mathcal{D}$	69.89	57.52	57.78	80.48	0.66
$\mathcal{D}^{\text{fuc}} \odot \mathcal{D}$	68.71	56.01	55.76	78.62	0.72
$(\mathcal{D}^{\text{dev}} + \mathcal{D}^{\text{fuc}}) \odot \mathcal{D}$	**74.53**	**60.73**	**61.17**	**92.10**	**0.22**

D ETHICAL STATEMENT ABOUT DATASET COLLECTION

For the data collection of `Mouse-RGC`, mice were maintained in pathogen-free facilities under a 12-hour light-dark schedule with standard housing conditions. Food and water were continuously supplied. Animals used in this study include both males and females.