FEATURIZATION OF SINGLE CELL TRAJECTORIES THROUGH KERNEL MEAN EMBEDDING OF OPTIMAL TRANSPORT MAPS

Alec Plotkin[1,2] **Justin Milner**[3,4] **Natalie Stanley**[1,2,5]
University of North Carolina at Chapel Hill
aplotkin@unc.edu, natalies@cs.unc.edu

[1]Curriculum in Bioinformatics and Computational Biology, Department of Genetics
[2]Computational Medicine Program
[3]Department of Microbiology and Immunology
[4]Lineberger Comprehensive Cancer Center
[5]Department of Computer Science

ABSTRACT

Longitudinal single-cell data has spurred the development of computational trajectory models with the power to make time-resolved, testable predictions about cell fates. As "real-time" trajectory inference methods proliferate, there is a growing need for tools that integrate their inherently high-dimensional outputs. In this work, we propose a novel strategy to facilitate downstream analysis of single-cell optimal-transport trajectory models, by constructing feature vectors that contain information about a cell's state across the entirety of its trajectory. This approach leverages kernel mean embedding of distributions to create trajectory features with applications in several domains, including cell clustering and comparison of perturbation response trajectories. We demonstrate how k-means clustering on trajectory features produces interpretable clusters that respect the underlying cell trajectories. Furthermore, we develop a divergence metric between single-cell trajectories based on the maximum mean discrepancy (MMD). We use this trajectory divergence to show that modeling perturbation trajectories may help uncover experimentally interesting perturbations at higher significance levels than by comparing perturbation responses at only a single time point.

1 BACKGROUND

Single-cell sequencing has opened a window into the heterogeneity and dynamics of biology on a cellular scale. One prominent application is to study the dynamics of cell differentiation, development, and response to stimuli. Discoveries in this area have the potential to fuel advances in drug development and cell engineering, with impacts on fields such as regenerative medicine and immuno-oncology.

However, the destructive nature of single-cell sequencing makes it impossible to profile a cell at more than a single time point. This has led to the development of a variety of trajectory inference methods to study dynamic cellular processes. Some of the most frequently used tools attempt to order cells along a pseudotime axis based on the similarity of their gene expression (Haghverdi et al.). While pseudotime methods have proven extremely useful for discovering transient states in differentiation (Laddach et al.), they are limited in their ability to make falsifiable predictions (Weinreb et al., b). Many methods have sought to increase the resolution of inferred trajectories; for example, RNA velocity and related methods estimate the rate of change in cell state with respect to time by solving a system of mass balance equations using spliced versus unspliced transcripts (La Manno et al.). However, there is some debate about the interpretability of inferred velocities (Gorin et al.), and it is difficult to find ground truths to assess their accuracy.

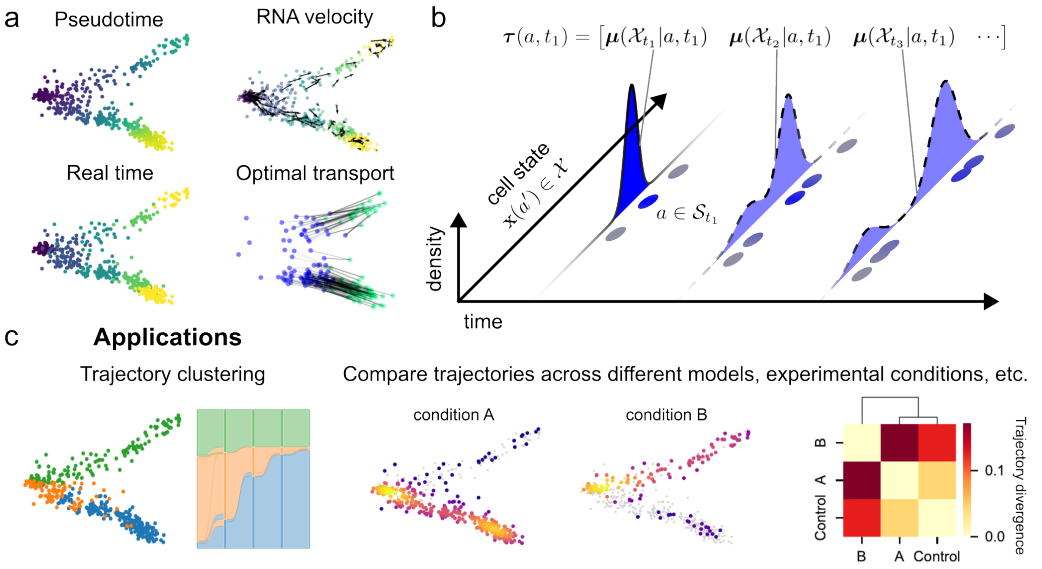

Figure 1: Overview of real-time trajectory featurization method. (a) Visualization of pseudotime, RNA velocity, and optimal transport methods on a simulated dataset of cell differention under diffusion-drift dynamics. (b) Schematic illustrating concept of representing trajectories as a feature vector. (c) Applications of trajectory featurization, including clustering that respects the underlying trajectory model, and comparison of trajectories under distinct experimental conditions.

Another option is to incorporate time point information directly into the trajectory model. Such approaches predict trajectories for each cell by mapping cell distributions from one time point to another (Lavenant et al.). The mapping function is flexible and can be parameterized over discrete cell states, (Schiebinger et al.) or over the continuous underlying gene expression space (Tong et al.). In this work, we focus on discrete maps fit using optimal transport. In addition to predicting cell fates at single-cell resolution, real-time models are also able to model cell proliferation and death as functions of time (Schiebinger et al.). Finally, the accuracy of real-time trajectory models can be evaluated by comparison to ground-truth trajectories, such as experiments incorporating cell-barcoding technologies for lineage tracing (Weinreb et al., a).

Despite the advantages of real-time trajectory inference methods, there are challenges in interpreting their predictions. Whereas trajectories in pseudotime models are defined by a small number of parameters (i.e. pseudotime estimate and branch identity), and velocity vectors can be projected onto a low-dimensional visualization, real-time trajectories are inherently high-dimensional (Fig 1a). The outputs of real-time models may be visualized as fate probabilities for a group of target populations (Klein et al.), but this approach may not work if there is substantial heterogeneity in cell states at the target timepoint. For example, clusters may be highly correlated with time (Kurd et al.), rather than with the biological heterogeneity within each time point. Another challenge of real-time methods is that there is no established way to compare trajectory models fit on disjoint sets of cells. This problem arises when there are distinct experimental conditions that must be modeled separately. Previously, metrics like the Wasserstein distance have been used to compare model predictions to ground truths for each time point (Yeo et al.), however this scales poorly with the size of the dataset. There exists a need to learn representations of real-time single-cell trajectories that facilitate downstream tasks, such clustering to identify distinct pathways of differentiation, and hypothesis testing to identify experimental conditions that significantly affect cell responses in time.

We propose a method that represents the state of a cell across the entirety of its trajectory as a single feature vector. This allows us to compare trajectories of cells that were observed at different time points, as well as to compare trajectories between different models, or disjoint experimental conditions. Our method works by first predicting a cell's ancestor and descendant distributions at each time point, and then embedding these distributions into a vector in a reproducing kernel Hilbert space (RKHS). We show that this information-rich featurization is useful for clustering cell

fates, especially on highly heterogeneous datasets. Furthermore, we demonstrate how an intentional choice of kernel induces sensible distance metrics between trajectories, and explore its utility for quantifying time-dependent experimental perturbations. The proposed method is implemented for trajectory inference models utilizing optimal transport, however it has the potential to be adapted for any trajectory model which utilizes mappings between cell distributions at discrete time points.

2 METHODS

2.1 TRANSPORT MAPS

We consider datasets of cells $\mathcal{X}_s = \{\boldsymbol{x}_s(a) \in \mathbb{R}^g\}_{a=0}^{N_s}$ consisting of N_s cells with g features sampled at time point $s \in \{t_i\}_{i=0}^{T}$. Datasets from multiple timepoints are compiled into a combined dataset $\mathcal{X} = \{\mathcal{X}_{t_i}\}_{i=0}^{T}$. Here, we are working with models that fit mappings of the form $T_{r,s} : \mathcal{X}_r \to \mathcal{X}_s$ between cell sets at time points r and s. Since we are dealing with discrete cell sets, we represent the mapping as a matrix, where the element $\boldsymbol{T}_{r,s}(a, b)$ contains the amount of mass transported from cell state $\boldsymbol{x}_r(a)$ to state $\boldsymbol{x}_s(b)$.

We obtain the stochastic matrix $\boldsymbol{P}_{r \mid s}$ by column-normalizing $\boldsymbol{T}_{r,s}$:

$$\boldsymbol{P}_{r \mid s}(a, b) = \frac{\boldsymbol{T}_{r,s}(a, b)}{\sum_{a'} \boldsymbol{T}_{r,s}(a', b)} \tag{1}$$

Here, $\boldsymbol{P}_{r \mid s}(a, b)$ represents the probability that a cell at state $\boldsymbol{x}_s(b)$ at time s originated from state $\boldsymbol{x}_r(a)$ at time r. [1]

2.2 TRAJECTORY FEATURIZATION

Given a vector-valued function $\boldsymbol{f}_r : \mathcal{X}_r \to \mathbb{R}^d$ defined at time point r, we can now define its pushforward under $T_{r,s}$ as:

$$\left(\boldsymbol{P}_{r \mid s} \sharp \boldsymbol{f}_r\right)\left(\boldsymbol{x}_s(b)\right) \equiv \sum_a \boldsymbol{P}_{r \mid s}(a, b) \boldsymbol{f}_r\big(\boldsymbol{x}_r(a)\big) = \mathbb{E}_{T_{r,s}}\left[\boldsymbol{f}_r \mid \boldsymbol{x}_s(b)\right] \tag{2}$$

This pushforward represents the expected value of $\boldsymbol{f}_r$ for a cell at state $\boldsymbol{x}_s(b)$.

Next, we use the pushforward operator to define features for cell state $\boldsymbol{x}_s(b)$ across all time points in the trajectory model. We concatenate these together to form the trajectory featurization vector $\boldsymbol{\tau} : \mathcal{X}_s \to \mathbb{R}^{Td}$:

$$\boldsymbol{\tau}(\boldsymbol{x}_s(b)) = \left[\left(\boldsymbol{P}_{t_i \mid s} \sharp \boldsymbol{f}_{t_i}\right)\left(\boldsymbol{x}_s(b)\right)\right]_{i=1}^{T} \tag{3}$$

When $t_i = s$, we define $\boldsymbol{P}_{s \mid s} = \boldsymbol{I}_s$, and the pushforward reduces to evaluating $\boldsymbol{f}_s\left(\boldsymbol{x}_s(b)\right)$.

This approach creates a feature vector for each cell that allows its trajectory to be compared to all cells in the dataset, even if they were sampled at different timepoints. Use of an informative featurization function $\boldsymbol{f}$ allows this vector to contain cell state information at each time point of the trajectory, using a relatively low number of features. For example, principal component decomposition is commonly used in single-cell biology, generally with feature dimension $10 \leq d \leq 50$. Compare this to a naive featurization approach utilizing the one-hot encoding $\tilde{\boldsymbol{f}}_r : \mathbb{R}^g \to \mathbb{R}^{N_r}$:

$$\tilde{f}_r\left(\boldsymbol{x}_r(a')\right)(a) = \delta\left[\boldsymbol{x}_r(a') = \boldsymbol{x}_r(a)\right] = \begin{cases} 1 & \boldsymbol{x}_r(a') = \boldsymbol{x}_r(a) \\ 0 & \boldsymbol{x}_r(a') \neq \boldsymbol{x}_r(a) \end{cases}$$

[1] This assumes that $r < s$. We also use the same convention when $r > s$, in which case $p\left(\boldsymbol{x}_r(a) \,\middle|\, \boldsymbol{x}_s(b)\right)$ represents the probability of observing a cell at state $\boldsymbol{x}_r(a)$ at time r by randomly sampling descendants of a cell originating at state $\boldsymbol{x}_s(b)$ at time s.

Applying our approach to $\tilde{\boldsymbol{f}}_r$ will yield a trajectory with $\sum_{i=1}^{T} N_{t_i}$ features. In single-cell datasets, we generally have $N_r = \mathcal{O}(10^3)$ cells per time point, making $\tilde{\boldsymbol{f}}_r$ inefficient compared to a function with fewer features. Another advantage of using our approach with an informative featurization function is that it allows us to compare trajectories defined on different cell sets $\hat{\mathcal{X}}^1$ and $\hat{\mathcal{X}}^2$, as long as they share time points and embedding spaces.

2.3 KERNEL MEAN EMBEDDING

We now discuss a specific choice of featurization function with desirable properties. We showed in (2) that the normalized pushforward induced by $T_{r,s}$ computes the expected value $\mathbb{E}_{T_{r,s}}\left[\boldsymbol{f}_r \mid \boldsymbol{x}_s(b)\right]$. While useful, this also disregards higher-order distributional information about the trajectory. This may not be ideal, for example in differentiation paths with complicated bifurcations.

To address this shortcoming, we turn to kernel-based approaches. Kernel methods have found widespread use in machine learning and have proven useful for applications in single-cell biology (Baskaran et al.). A positive-definite kernel function $K : \mathbb{R}^g \times \mathbb{R}^g \to \mathbb{R}$ can be used to evaluate the similarity between data points, and induces a reproducing kernel Hilbert space (RKHS) with the property $\langle K(\boldsymbol{x}, \cdot), \boldsymbol{f}(\cdot)\rangle = \boldsymbol{f}(\boldsymbol{x})$. If we use a *characteristic kernel*, we can uniquely embed any probability distribution into the RKHS.

A feature map $\boldsymbol{\phi} : \mathbb{R}^g \to \mathbb{R}^d$ is a finite-dimensional approximation of this kernel embedding that retains the reproducing property. In this application, we choose a feature map induced by the radial basis function (RBF) as our featurization function $\boldsymbol{f}_r$. Since the RBF is a characteristic kernel, operations such as addition of feature maps $\sum \phi(x)$ preserve the distributional characteristics of the trajectory. Specifically, the pushforward measure induces the *kernel mean embedding* (KME) $\boldsymbol{\mu}$ of the probability distribution underlying the trajectory:

$$\boldsymbol{\mu}(\mathcal{X}_r \mid \boldsymbol{x}_s(b)) = \mathbb{E}_{T_{r,s}}\left[\boldsymbol{\phi}_r \mid \boldsymbol{x}_s(b)\right] = \left(\boldsymbol{P}_{r \mid s} \sharp\, \boldsymbol{\phi}_r\right)(\boldsymbol{x}_s(a)) \tag{4}$$

The trajectory featurization then becomes:

$$\boldsymbol{\tau}\left(\boldsymbol{x}_s(b)\right) = \left[\boldsymbol{\mu}\left(\mathcal{X}_{t_i} \mid \boldsymbol{x}_s(b)\right)\right]_{i=1}^{T} \tag{5}$$

Using a characteristic kernel has the property that the distance between two kernel mean embeddings $\|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|$ induced by probability distributions P, Q is equal to 0 if and only if $P = Q$. This induces a divergence called the maximum mean discrepancy (MMD), which may be used to compare distributions. For example, it is possible to formulate a statistical test for the difference between two distributions based on their MMD (Gretton et al.).

The trajectory featurization inherits these properties. Given two trajectory feature vectors, the squared $L2$-distance between them is:

$$\left\|\boldsymbol{\tau}\left(\boldsymbol{x}_s(b)\right) - \boldsymbol{\tau}\left(\boldsymbol{x}_{s'}(b')\right)\right\|_2^2 = \sum_{i=1}^{T} \mathrm{MMD}^2\left(\boldsymbol{\mu}\left(\mathcal{X}_{t_i} \mid \boldsymbol{x}_s(b)\right), \boldsymbol{\mu}\left(\mathcal{X}_{t_i} \mid \boldsymbol{x}_{s'}(b')\right)\right) \tag{6}$$

We define this as the *trajectory divergence*. This allows us to compute the distance between two trajectories, for applications such as performing hypothesis tests for experimental conditions versus a control, or comparing a computational trajectory against ground-truth data.

2.4 CLUSTER EVALUATION METRICS

Another application of our method is clustering cells with similar trajectories across time points. Single-cell trajectory feature vectors can be input to any clustering algorithm; in this work, we use k-means. In order to quantify how well different clusterings represent the underlying trajectories, we develop two metrics to measure cluster quality with respect to a trajectory model. The first of these is flow consistency, which we define as the mean probability that a cell is transported to a target within the same cluster as its source:

$$\text{consistency}\left(\boldsymbol{c}_r \mid \boldsymbol{c}_s\right) = \frac{1}{N_s} \sum_{b=1}^{N_s} \sum_{a=1}^{N_r} \delta\left[c_r(a) = c_s(b)\right] \boldsymbol{P}_{r \mid s}(a, b) \tag{7}$$

Here, $c_r(a)$ denotes the cluster assignment for cell a at target time point r, $c_s(b)$ is the cluster assignment for cell b at source time point s, and δ is the Kronecker delta. This metric ranges from 0 to 1, and is higher when cells are most likely to remain in the same cluster over time.

We also define flow entropy, which is agnostic to the source cluster assignment, and is defined as:

$$\text{entropy}\left(\mathcal{X}_r \mid \boldsymbol{c}_s\right) = \frac{-1}{N_s} \sum_{b=1}^{N_s} \sum_{k=1}^{K} p\left(c_r = k \mid \boldsymbol{x}_s(b)\right) \log p\left(c_r = k \mid \boldsymbol{x}_s(b)\right), \tag{8}$$

$$p\left(c_r = k \mid \boldsymbol{x}_s(b)\right) = \sum_{a=1}^{N_r} \delta\left[c_r(a) = k\right] \boldsymbol{P}_{r \mid s}(a, b)$$

The flow entropy is susceptible to variation in the number and size of target clusters, so we normalize it by dividing the entropy under the trajectory model by the overall entropy of clusters at the target time point. The resulting *entropy ratio* varies between 0 and 1, with values closer to 0 representing target clusters that align better with the trajectory model.

3 RESULTS

3.1 TRAJECTORY CLUSTERING DISTINGUISHES DISTINCT LINEAGES IN HSC DIFFERENTIATION

We sought to test our trajectory featurization approach in a differentiation model with clearly defined endpoints, in order to compare clustering on trajectory features to ground-truth cell types. We chose a multi-omic time course dataset of human hematopoietic stem cells (HSCs), which cross sections their development at days 2, 3, 4, and 7 after induction of differentiation (Daniel Burkhardt et al.).

Data were downloaded through the moscot Python package (Klein et al.), and included labeled cell types, as well as uniform manifold approximation and projection (UMAP) dimensionality reductions for both RNA and ATAC modalities. We constructed a shared embedding representation from both data modalities and fit the trajectory model on these shared coordinates using moscot. We then used the fitted model to construct single-cell trajectory feature vectors, using either an RBF feature map (trajectory KME) or PCA decomposition (trajectory PCA) as the featurization function.

We fit k-means clusters on both sets of trajectory features, and compared these against the reference cell types as well as Leiden clusters fit using only gene expression information. Uniform manifold approximation and projection in the gene expression space (UMAP) showed that most cells belong to the HSC reference cell type at day 2, and proceed to differentiate into various lineages from days 3-7 (Fig 2a-b). The Leiden clusters appear to split many of the reference cell types into groups, especially the undifferentiated HSC cluster (Fig 2c). The trajectory KME clusters similarly partition the reference cell types (Fig 2d), however they exhibit less sharp boundaries between clusters than the Leiden clusters. At the termini of the differentiation trajectory, both Leiden and trajectory KME clusters appear to align strongly with the reference cell types. Indeed, Fig 2h shows that most of the trajectory KME clusters at day 7 correspond to a single reference cell type.

We next evaluated the performance of the clustering methods in capturing trajectory information using the flow consistency (7) and flow entropy (8) scores. We see that both trajectory clusters exhibit high consistency and low entropy compared to the Leiden clusters and reference cell types (Fig 2e-f), suggesting that both featurization strategies are successfully encoding information about cellular trajectories. The trajectory KME clusters slightly outperform the trajectory PCA clusters for both metrics, which hints at their greater ability to represent higher-order distributional information. Nonetheless, the trajectory PCA clusters perform quite well, perhaps due to the simplicity of the bifurcations in this dataset. In contrast, the reference clusters show the highest flow entropy of

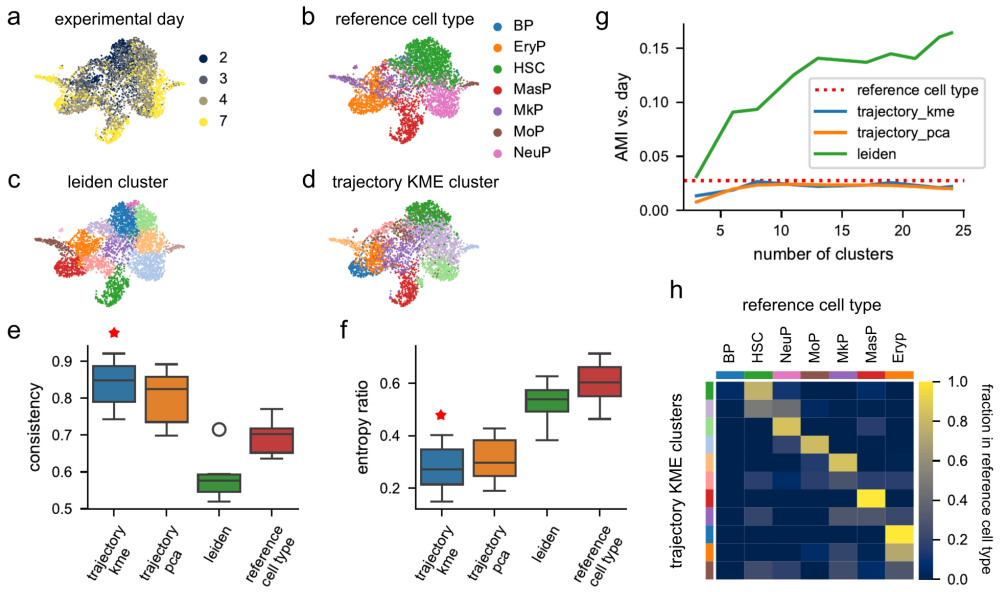

Figure 2: Clustering on trajectory features accurately partitions cell types in HSCs and streamlines cluster flows. (a-d) UMAP dimensionality reductions showing (a) experimental day, (b) reference cell type labels, (c) Leiden clusters, and (d) k-means clusters fit on trajectory KME features. (e-f) Boxplots evaluating how well clusters represent the underlying trajectory model, using (e) flow consistency score and (f) entropy ratio score. The best performing clusters for each metric are annotated using a red star. (g) Line plot showing adjusted mutual information (AMI) between clusters and experimental day, versus the input number of clusters for different clustering methods. The AMI between the reference cell types and experimental day is displayed by a dotted red line. (h) Heatmap colored by the fraction of each trajectory cluster (rows) overlapping with a given reference cell type (columns) at day 7.

any clustering (Fig 2f), likely driven by the ability of the multipotent HSC cell type to give rise to different populations.

Interestingly, the Leiden clusters showed exceptionally low flow consistency (Fig 2e). We hypothesized that this could be because the Leiden clusters are separating by experimental day more than the other clusters. This would cause the trajectories to switch clusters more often between time points, resulting in a lower consistency score. We tested this by measuring the adjusted mutual information (AMI) between each set of clusters and experimental day. We see that the Leiden clusters have higher AMI with experimental day than any of the other clusters (Fig 2g). Furthermore, increasing the number of clusters caused the Leiden clusters to correlate more strongly with time point, but did not affect the trajectory clusters. Trajectory featurization may therefore be a useful strategy for removing confounding effects due to time during single-cell phenotyping.

3.2 Characterizing perturbation responses using trajectory divergence

One of the advantages of using trajectory mean embeddings during data analysis is that they naturally induce the maximum mean discrepancy (MMD) metric on probability distributions. Here, we show how this can be used to analyze combined perturbation and time course single-cell data, to recover perturbations with substantial time-dependent effects on cell state. We performed reanalysis on the dataset from Ishikawa et al., which features differentiation induced pluripotent stem cells (iPSCs) under the influence of CRISPRi guide RNAs targeting 25 transcription factors (TFs). iPSCs were profiled using scRNA-seq on each day from days 2-5 following induction (Fig 3a). Thus, this dataset presents a unique opportunity to show how using trajectory analysis can increase power to detect perturbation responses.

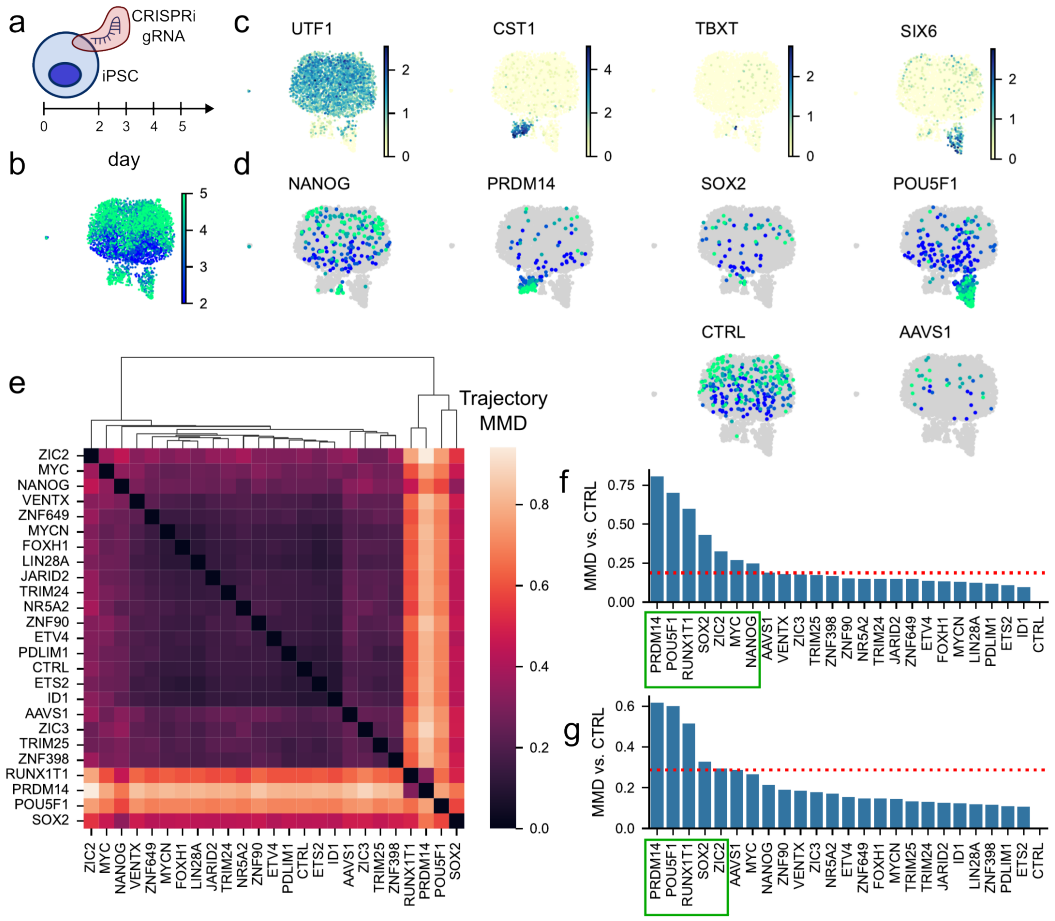

Figure 3: Comparison of iPSC responses to CRISPRi perturbation using trajectory divergence demonstrates enhanced ability to identify significant responses compared to using a single time point. (a) The dataset follows differentiation iPSCs in response to a panel of 50 CRISPRi guide RNAs targeting 25 transcription factors at days 2-5 following induction. (b) UMAP displaying the days at which each cell was collected. (c) UMAPs displaying normalized gene expression of pluripotent (*UTF1*), definitive endoderm (*CST1*), axial mesoderm (*TBXT*), and neural (*SIX6*) lineage markers. (d) UMAPs showing distributions of cells expressing a single gRNA, colored by sampling time. Cells without the gRNA are grayed-out. (e) Heatmap of MMD between trajectory kernel mean embeddings for each pair of perturbations. (f) Trajectory MMD between each perturbation and the control condition. Perturbations with MMDs greater those targeting the non-coding AAVS1 locus (red dotted line) are shown in the green box. (g) MMD between each perturbation and the control condition, computed using cells and features from day 5 only. Perturbations with MMDs greater those targeting the non-coding AAVS1 locus (red dotted line) are shown in the green box.

The gRNA expression from each cell was log1p-transformed, and then binarized using Otsu thresholding. We selected cells expressing exactly one gRNA, and then preprocessed the RNA data using a standard analysis pipeline. UMAP dimensionality reduction showed that the cells start in an intermediate state and branch outwards over time (Fig 3b). Each major branch expressed some lineage marker genes, with the largest branch being defined by the pluripotency marker *UTF1*, and smaller branches defined by the markers *CST1*, *SIX6*, and *TBXT* representing lineage commitment (Fig 3c). A few of the gRNAs are distributed similarly to these lineage markers, suggesting that inhibiting their target TFs promotes differentiation (Fig 3d). In addition, the dataset contained two control gRNAs: the *CTRL* condition, which corresponded to a non-cutting guide, and AAVS1, which corresponds to a non-coding locus in the genome. Both of these show similar trends over time, with a strong preference for the pluripotent state by day 5.

We fit trajectory models using moslin to incorporate the gRNAs as prior information, and computed trajectory featurizations using an RBF kernel approximation. We then calculated the mean embedding of each perturbation's trajectory features, and used this to calculate the trajectory divergence. A few perturbations (RUNX1T1, PRDM14, POU5F1, SOX2) displayed noticeably higher divergence from all the others, suggesting that these are some of the most important regulators of iPSC fate (Fig 3e). However, we were interested in whether it would be possible using the trajectory divergence to detect some gRNAs with less prominent, but still noticeable effects. We took the divergence for each perturbation against the non-cutt CTRL gRNA, and found that seven gRNAs scored higher than the safe-harbor AAVS1 gRNA, which we used as an ad-hoc threshold (Fig 3f). We wanted to know whether we would have been able to detect these gRNAs without using a trajectory model, so we selected features and cells from only day 5 and conducted the same test (Fig 3g). We found that only five gRNAs had higher discrepancy from CTRL than AAVS1. Notably, one of the gRNAs that fell below detection was NANOG, which is visually prominent on the UMAP in Fig 3d, and which is a known regulator of stem cell pluripotency.

4 Discussion

In this work, we introduced a novel strategy for representing real-time single-cell trajectories as information-rich, low-dimensional feature vectors. Our method takes advantage of probabilistic interpretations of optimal transport theory, along with the theory of reproducing kernel Hilbert spaces, to distill both trajectory and phenotypic information into a unified representation. This powerful approach enables us to use optimal transport trajectory modeling as a starting point for various downstream tasks in single-cell analysis, rather than as an endpoint in and of itself.

We paid particular attention to the applications of cell fate clustering and quantification of experimental perturbations, both of which have been subject of recent interest in the field (Lange et al.; Peidli et al.). We demonstrated that clustering on the trajectory feature vectors produces cell subsets that are more consistent over time under the optimal transport maps. Furthermore, we showed that this may be particularly useful in biological systems with heterogeneous cell states, such as CD4 T-cell polarization. We also explored the connection between kernel mean embedding of trajectories and the maximum mean discrepancy, and proposed using the distance between trajectory vectors as a metric for the divergence between experimental responses. We demonstrated how using this approach was able to detect true-positive perturbations with higher efficiency than alternatives that do not take trajectories into account. Given the central role of time across biological processes, the trajectory featurization approach proposed in this paper has the potential to enable the study of single-cell responses at unprecedented temporal and cellular resolution.

Future work may seek to extend this approach from the discrete cell sets in optimal transport, to continuous gene expression space. This would enable integration with generative cell trajectory models such as TrajectoryNet (Tong et al.) or PRESCIENT (Yeo et al.). Another direction for further investigation is the extension of the featurization from discrete to continuous time. This could connect this work to the rich body of work on stochastic process and time-series modeling, including methods such as Gaussian processes.

References

Vishal Athreya Baskaran, Jolene Ranek, Siyuan Shan, Natalie Stanley, and Junier B. Oliva. Distribution-based sketching of single-cell samples. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–10. ACM. ISBN 978-1-4503-9386-7. doi: 10.1145/3535508.3545539. URL https://dl.acm.org/doi/10.1145/3535508.3545539.

Daniel Burkhardt, Malte Luecken, Andrew Benz, Peter Holderrieth, Jonathan Bloom, Christopher Lance, Ashley Chow, and Ryan Holbrook. Open Problems - Multimodal Single-Cell Integration. URL https://kaggle.com/competitions/open-problems-multimodal.

Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. 18(9): e1010492. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010492. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010492.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. 13(25):723–773. ISSN 1533-7928. URL http://jmlr.org/papers/v13/gretton12a.html.

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. 13(10):845–848. ISSN 1548-7105. doi: 10.1038/nmeth.3971. URL https://www.nature.com/articles/nmeth.3971.

Masato Ishikawa, Seiichi Sugino, Yoshie Masuda, Yusuke Tarumoto, Yusuke Seto, Nobuko Taniyama, Fumi Wagai, Yuhei Yamauchi, Yasuhiro Kojima, Hisanori Kiryu, Kosuke Yusa, Mototsugu Eiraku, and Atsushi Mochizuki. RENGE infers gene regulatory networks using time-series single-cell RNA-seq data with CRISPR perturbations. 6(1):1–14. ISSN 2399-3642. doi: 10.1038/s42003-023-05594-4. URL https://www.nature.com/articles/s42003-023-05594-4.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimée Bastidas-Ponce, Marta Tarquis-Medina, Heiko Lickert, Mostafa Bakhti, Mor Nitzan, Marco Cuturi, and Fabian J. Theis. Mapping cells through time and space with moscot. URL https://www.biorxiv.org/content/10.1101/2023.05.11.540374v1.

Nadia S. Kurd, Zhaoren He, Tiani L. Louis, J. Justin Milner, Kyla D. Omilusik, Wenhao Jin, Matthew S. Tsai, Christella E. Widjaja, Jad N. Kanbar, Jocelyn G. Olvera, Tiffani Tysl, Lauren K. Quezada, Brigid S. Boland, Wendy J. Huang, Cornelis Murre, Ananda W. Goldrath, Gene W. Yeo, and John T. Chang. Early precursors and molecular determinants of tissue-resident memory CD8+ T lymphocytes revealed by single-cell RNA sequencing. 5(47):eaaz6894. ISSN 2470-9468. doi: 10.1126/sciimmunol.aaz6894. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7341730/.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, prefix=van useprefix=true family=Bruggen, given=David, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. 560(7719):494–498. ISSN 1476-4687. doi: 10.1038/s41586-018-0414-6. URL https://www.nature.com/articles/s41586-018-0414-6.

Anna Laddach, Song Hui Chng, Reena Lasrado, Fränze Progatzky, Michael Shapiro, Alek Erickson, Marisol Sampedro Castaneda, Artem V. Artemov, Ana Carina Bon-Frauches, Eleni-Maria Amaniti, Jens Kleinjung, Stefan Boeing, Sila Ultanir, Igor Adameyko, and Vassilis Pachnis. A branching model of lineage differentiation underpinning the neurogenic potential of enteric glia. 14(1):5904. ISSN 2041-1723. doi: 10.1038/s41467-023-41492-3. URL https://www.nature.com/articles/s41467-023-41492-3.

Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, and Fabian J. Theis. CellRank for directed single-cell fate mapping. 19(2):159–170. ISSN 1548-7105. doi: 10.1038/s41592-021-01346-6. URL https://www.nature.com/articles/s41592-021-01346-6.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. URL http://arxiv.org/abs/2102.09204.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: Harmonized Single-Cell Perturbation Data. URL https://www.biorxiv.org/content/10.1101/2022.08.20.504663v3.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander.

Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. 176(4):928–943.e22. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.01.006. URL https://www.cell.com/cell/abstract/S0092-8674(19)30039-X.

Alexander Tong, Jessie Huang, Guy Wolf, prefix=van useprefix=true family=Dijk, given=David, and Smita Krishnaswamy. TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics. URL http://arxiv.org/abs/2002.04461.

Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. 367(6479): eaaw3381, a. doi: 10.1126/science.aaw3381. URL https://www.science.org/doi/10.1126/science.aaw3381.

Caleb Weinreb, Samuel Wolock, Betsabeh K. Tusi, Merav Socolovsky, and Allon M. Klein. Fundamental limits on dynamic inference from single-cell snapshots. 115(10):E2467–E2476, b. doi: 10.1073/pnas.1714723115. URL https://www.pnas.org/doi/full/10.1073/pnas.1714723115.

Grace Hui Ting Yeo, Sachit D. Saksena, and David K. Gifford. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. 12(1): 3222. ISSN 2041-1723. doi: 10.1038/s41467-021-23518-w. URL https://www.nature.com/articles/s41467-021-23518-w.