Breakout Session 6: Track A

Algorithmic Bias in Single Cell Analysis: A Study of Optimal Transport and Sinkhorn Divergence

Dr. Pilhwa Lee
Lecturer, Morgan State University

NIH Office of Data Science Strategy, NOSI meeting, March 27-28

Project: Characterization of health disparities in African ancestry and reduction of algorithmic bias

Presentation: Algorithmic Bias in Single Cell Analysis: A Study of Optimal Transport and Sinkhorn Divergence

Department of Mathematics
Morgan State University

Pilhwa Lee

Joint work with Jayshawn Cooper and Christina Young

Project summary

This project focusses on characterization of health disparities in African ancestry and reduction of algorithmic bias. We redefine the metrics of artificial intelligence for resolving algorithmic bias decoupled from data bias and explore machine learning techniques on probability spaces and regularization for reduction of “algorithmic bias”. This is expected to enhance the ethics associated with AI and African ancestry, extensible to public portal analytics of “All of Us.”

Dynamic Optimal Transport (OT)

1. The Wasserstein distance - distance between two probability measures and on $\Omega \subseteq \mathbb{R}^n$

- Monge Map $M(\mu, \nu) = \min_{T \in \mathcal{T}(\mu,\nu)} \int \|x - T(x)\|_2^2 d\mu(x)$

  where $\mathcal{T}(\mu, \nu) = \{T : \Omega \to \Omega \,|\, T_\sharp \mu = \mu \circ T^{-1} = \nu\}$

- -Kantrovich $W(\mu, \nu)_p = \left(\inf_{\pi \in \pi(\mu,\nu)} \int d(x,y)^p d\pi(x,y) \right)^{\frac{1}{p}}$

2. Dynamic optimal transport (OT) is the L^2 Wasserstein distance, which includes time, speed, and direction. Represented by the following equation, where and are two distributions at timepoints t_0 and t_1 and $f(x, t)$ represents velocity [Benamou-Brenier, 2000]:

$$W(\mu, \nu)_2^2 = \inf_{(P,f)} (t_1 - t_0) \iint_{t_0}^{t_1} P(x,t)\,|f(x,t)|^2 dt dx$$

where $\partial_t P + \nabla \cdot (Pf) = 0, \quad P(\cdot, t_0) = \mu, \quad P(\cdot, t_1) = \nu$

Waddington OT and Python OT:
pipeline for estimating trajectory probabilities

The Waddington OT developed an algorithm for computing trajectory probabilities at specific times. We modified the pipeline to make the trajectories more accurate and debiased.

For a given set of cells at time t_j. Define the probability vector p_{t_j} as follows:

$$p_{t_j}(x) = \begin{cases} \frac{1}{|C|} & x \in C \\ 0 & \text{otherwise} \end{cases}$$

1. The descendant distribution at time t_{j+1} is calculated by "pushing" the cell set through the transport/cost matrix. Each probability vector is "pushed forward" by multiplying by the transport map on the right

$$p^T_{t_{j+1}} = p^T_{t_j} \gamma_{t_j, t_{j+1}}$$

Therefore, inductively the descendant distribution can be calculated at any later time $t_\ell > t_j$.

RCMI MORGAN
Center for Urban Health Disparities Research and Innovation

CEAMLS MORGAN
CENTER FOR EQUITABLE AI & MACHINE LEARNING SYSTEMS

Waddington OT and Python OT: pipeline for estimating trajectory probabilities

Waddington-OT's approach employed both entropic regularization and unbalanced transport to compute the transport map at time t_i and t_{i+1}. They solve the following optimization problem [Schiebinger, et al, 2019]:

$$\hat{\pi}_{t_i,t_{i+1}} = \arg\min_{\pi} \sum_{x \in S_i} \sum_{y \in S_{i+1}} c(x,y)\pi(x,y) - \epsilon \iint \pi(x,y) \log \pi(x,y)\,dx\,dy$$

$$+ \lambda_1 \mathrm{KL}\left[\sum_{x \in S_i} \pi(x,y) \,\Big\|\, d\hat{P}_{t_{i+1}}(y) \right] + \lambda_2 \mathrm{KL}\left[\sum_{y \in S_{i+1}} \pi(x,y) \,\Big\|\, d\hat{Q}_{t_i}(x) \right]$$

where ϵ, λ_1 and λ_2 are regularization parameters.

Debiasing with Sinkhorn Divergence

As previously mentioned, dynamic OT models tend to lead to algorithmic bias by altering a cell's transport distance. Therefore, when computing transport matrices for the Waddington-OT pipeline, we utilize **Sinkhorn Divergence** (accessible through the Python-OT package) to compute alternative and debiased transport matrices for each pair of timepoints.

2. Sinkhorn divergence is a centering method that can be used to debias the algorithm. This means that $S_\varepsilon(P, Q) = 0 \leftrightarrow P = Q$ [Pooladian, et al. 2022]:

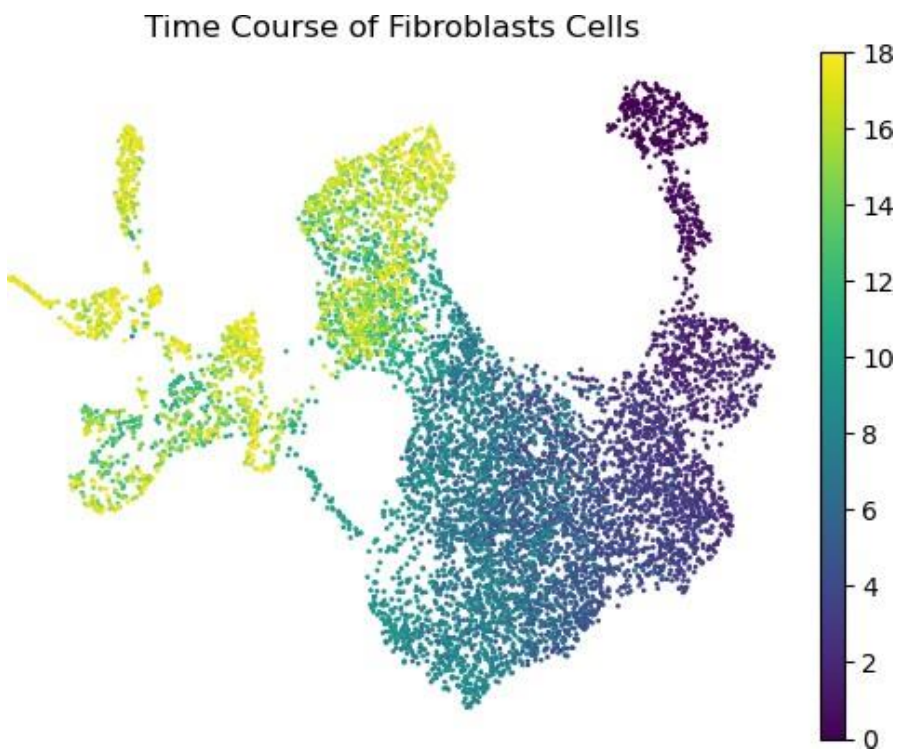
$$S_e = OT_\varepsilon(\mu, v) - \frac{1}{2}OT_\varepsilon(\mu, \mu) - \frac{1}{2}OT_\varepsilon(v, v)$$

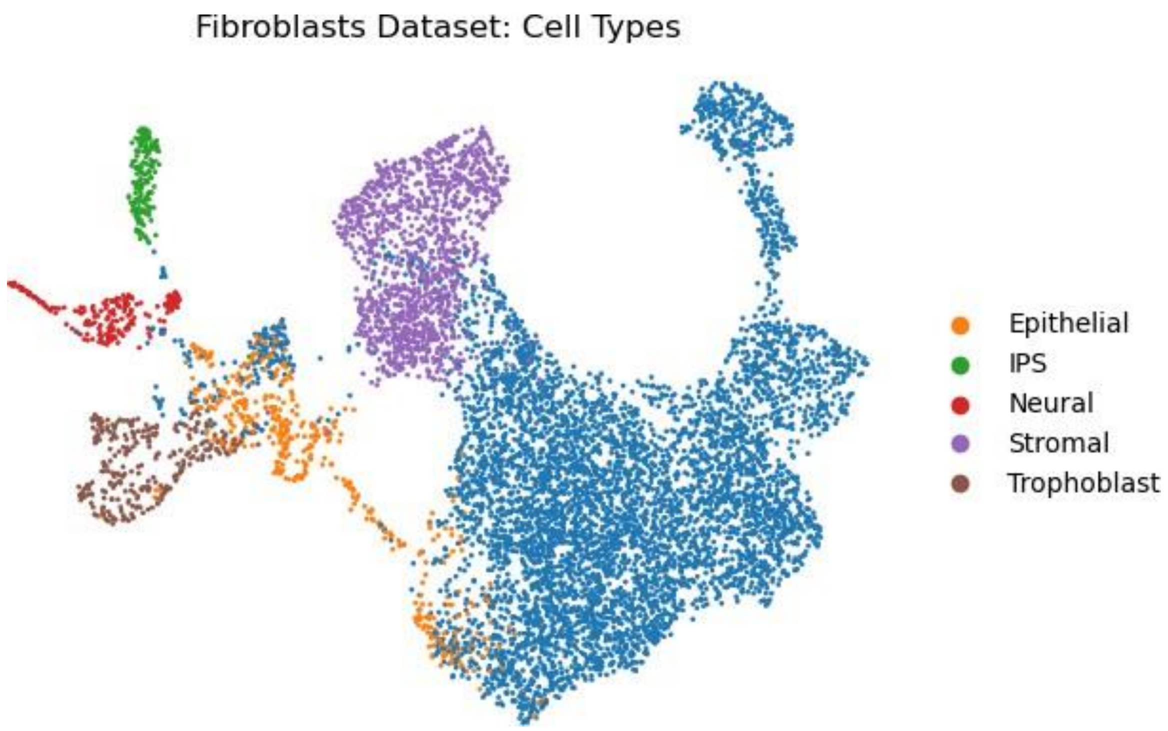
Center for Urban Health Disparities Research and Innovation

CENTER FOR EQUITABLE AI & MACHINE LEARNING SYSTEMS

Fibroblasts iPS reprogramming dataset

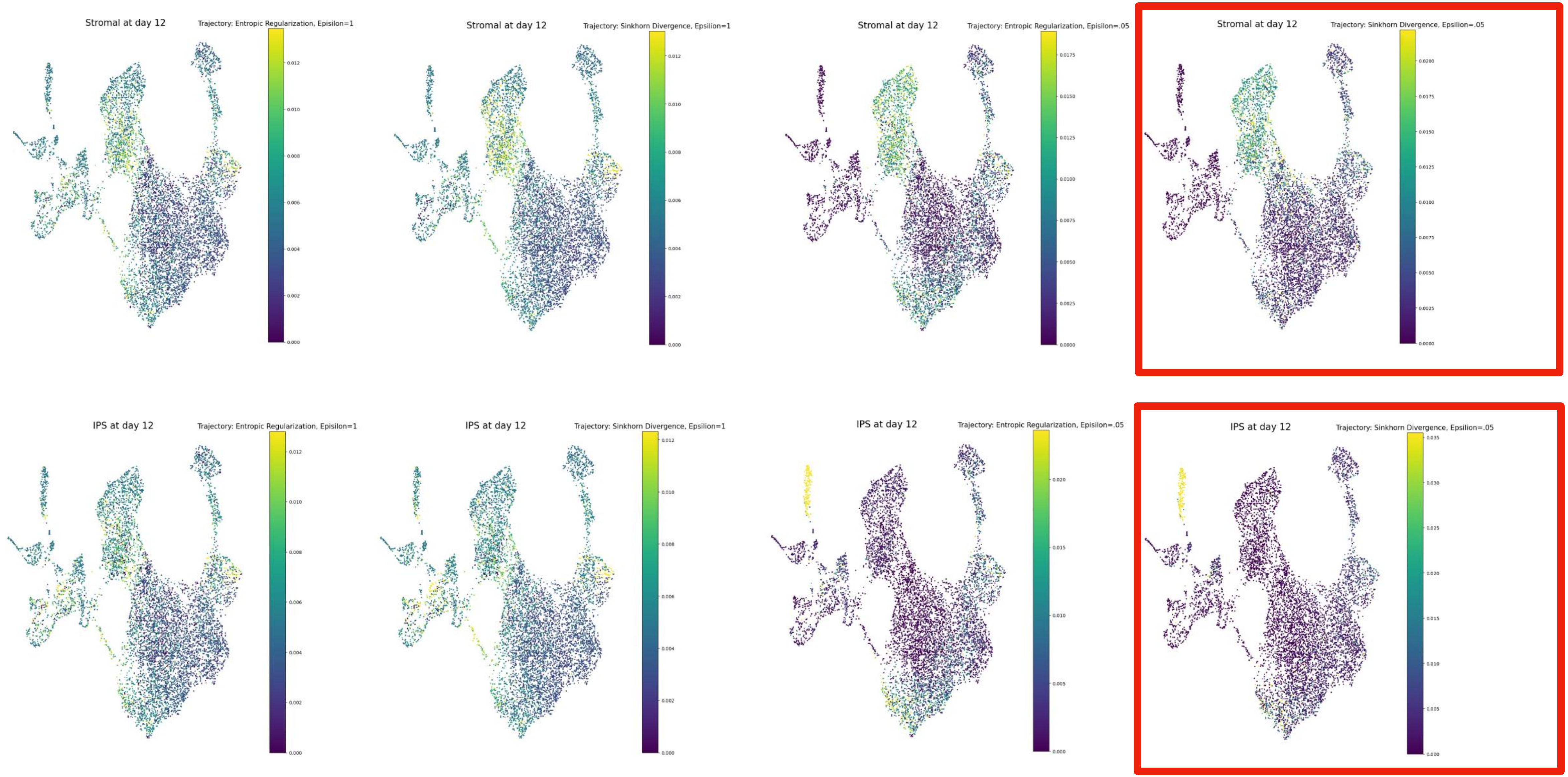
From a dataset of 315,000 cells collected over an 18-day period at half-day intervals, in this experiment the modified Waddington-OT algorithm is applied to a subset of around 8,000 cells [Schiebinger, et al., 2019]:





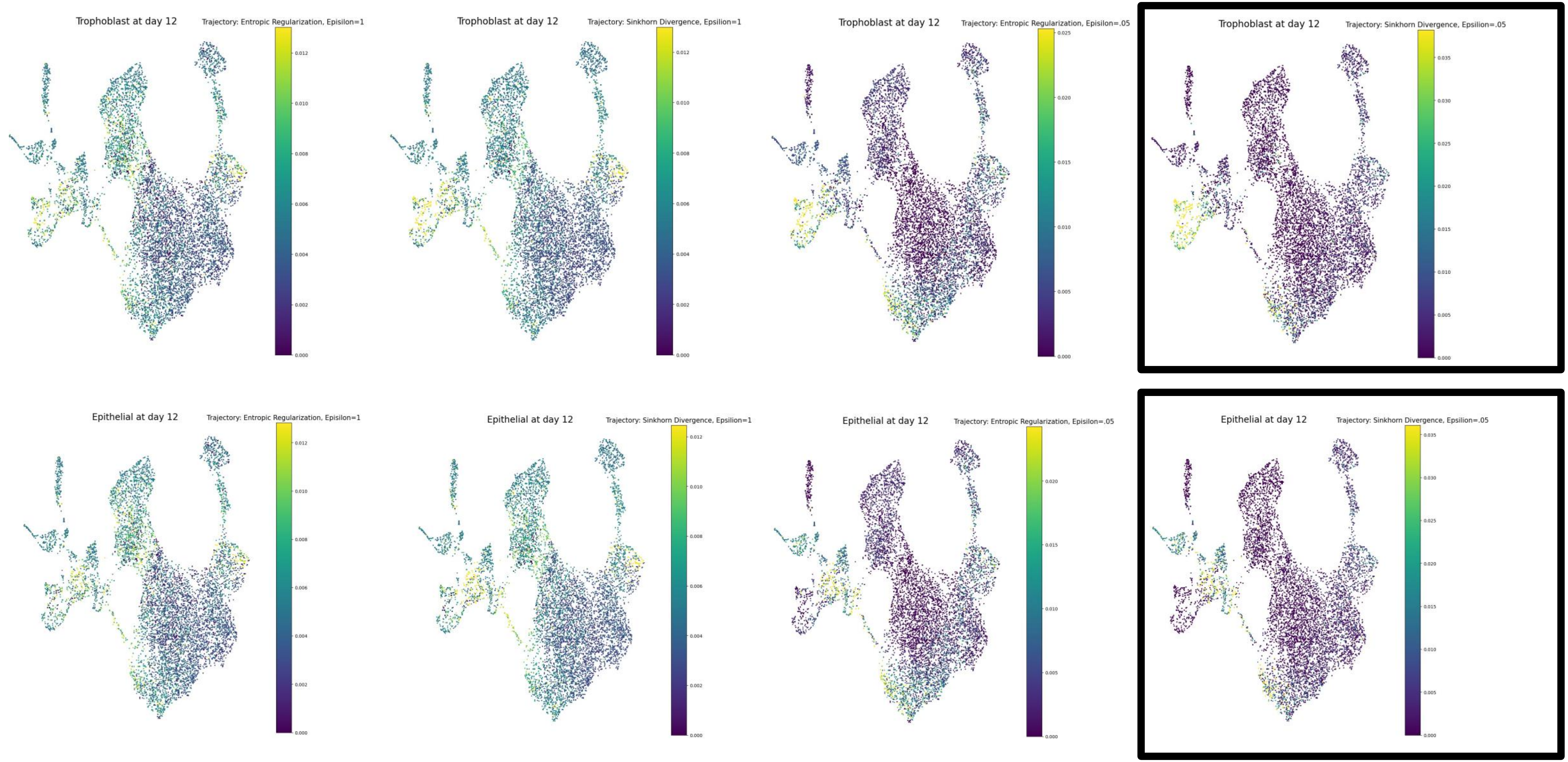
Fibroblasts iPS reprogramming dataset:

Trajectory Probabilities at Day 12. Stomal and iPS cells



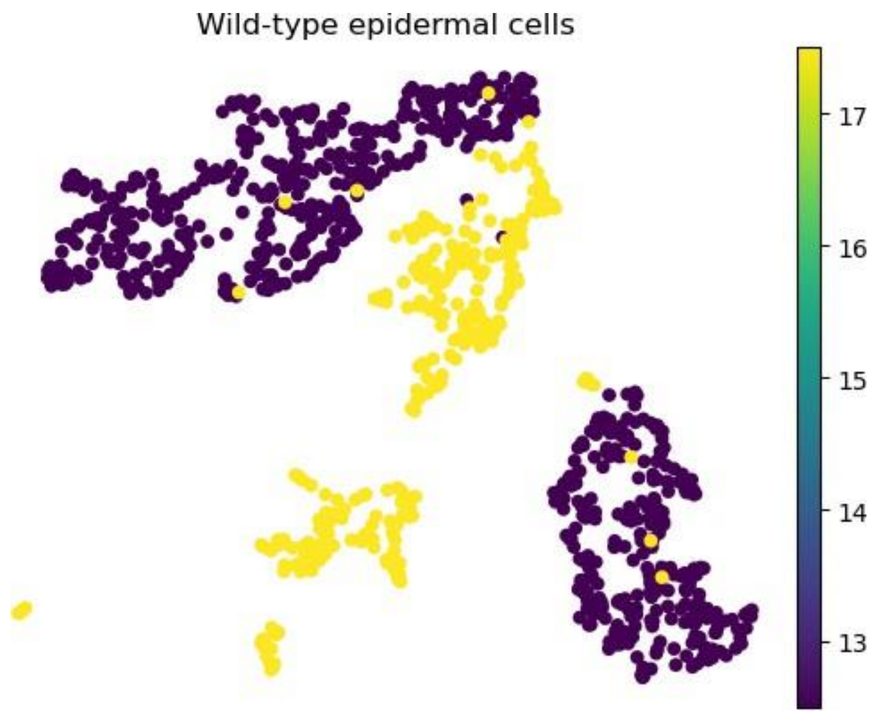
Fibroblasts iPS reprogramming dataset:

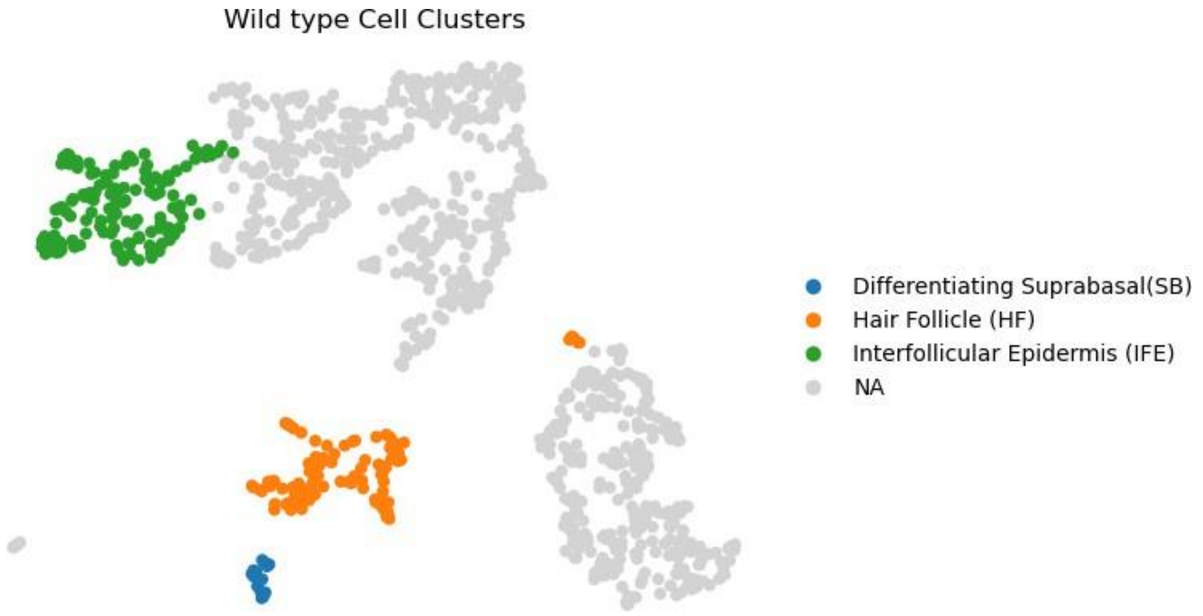
Trajectory Probabilities at day 12. Epithelial and Trophoblast cells



Mouse epidermal cells

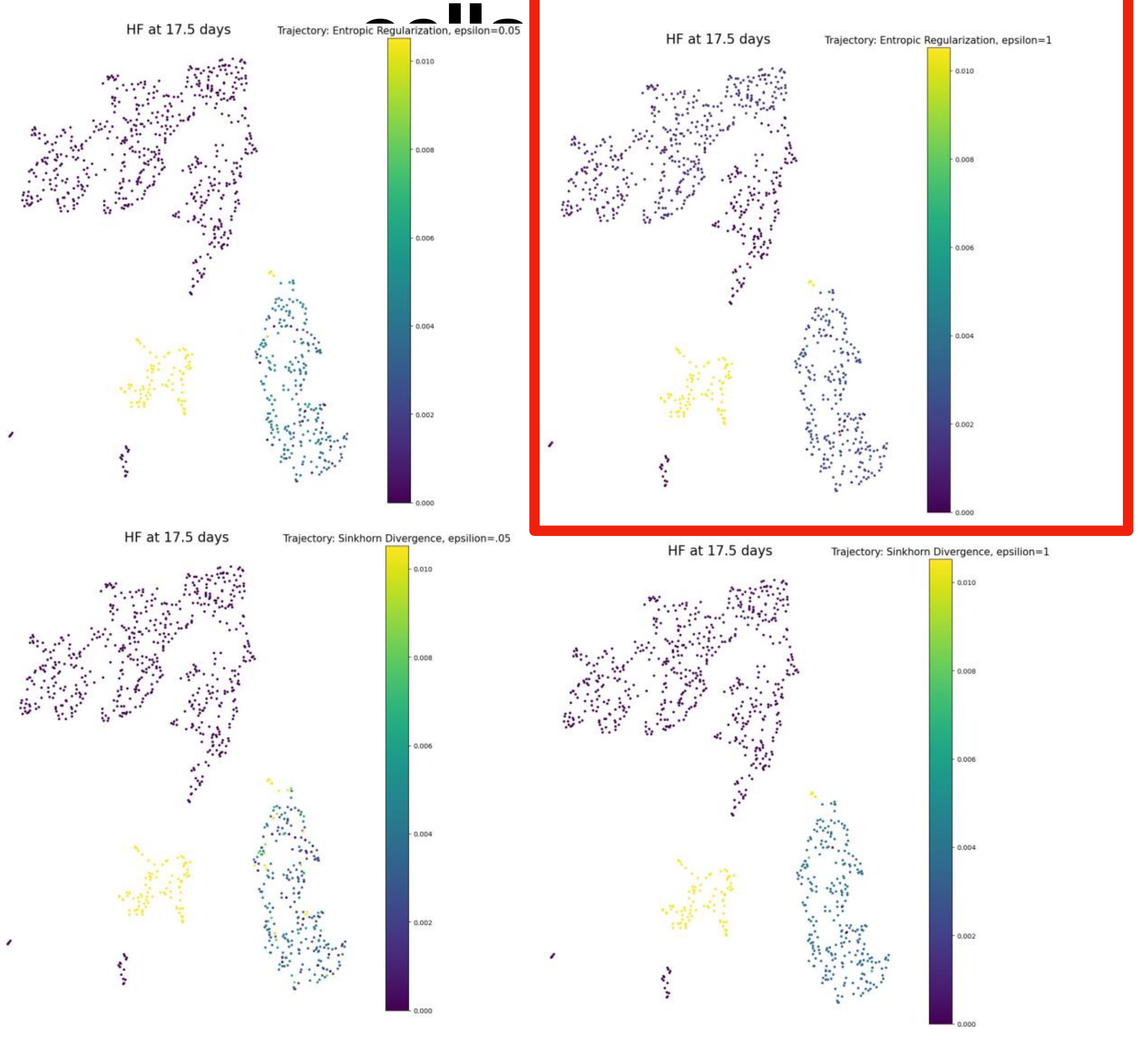
Two embryonic time points (E12.5 and E17.5) were used for single cell RNA sequencing experiments on wildtype epidermal cells [Ellis, et al., 2019]:



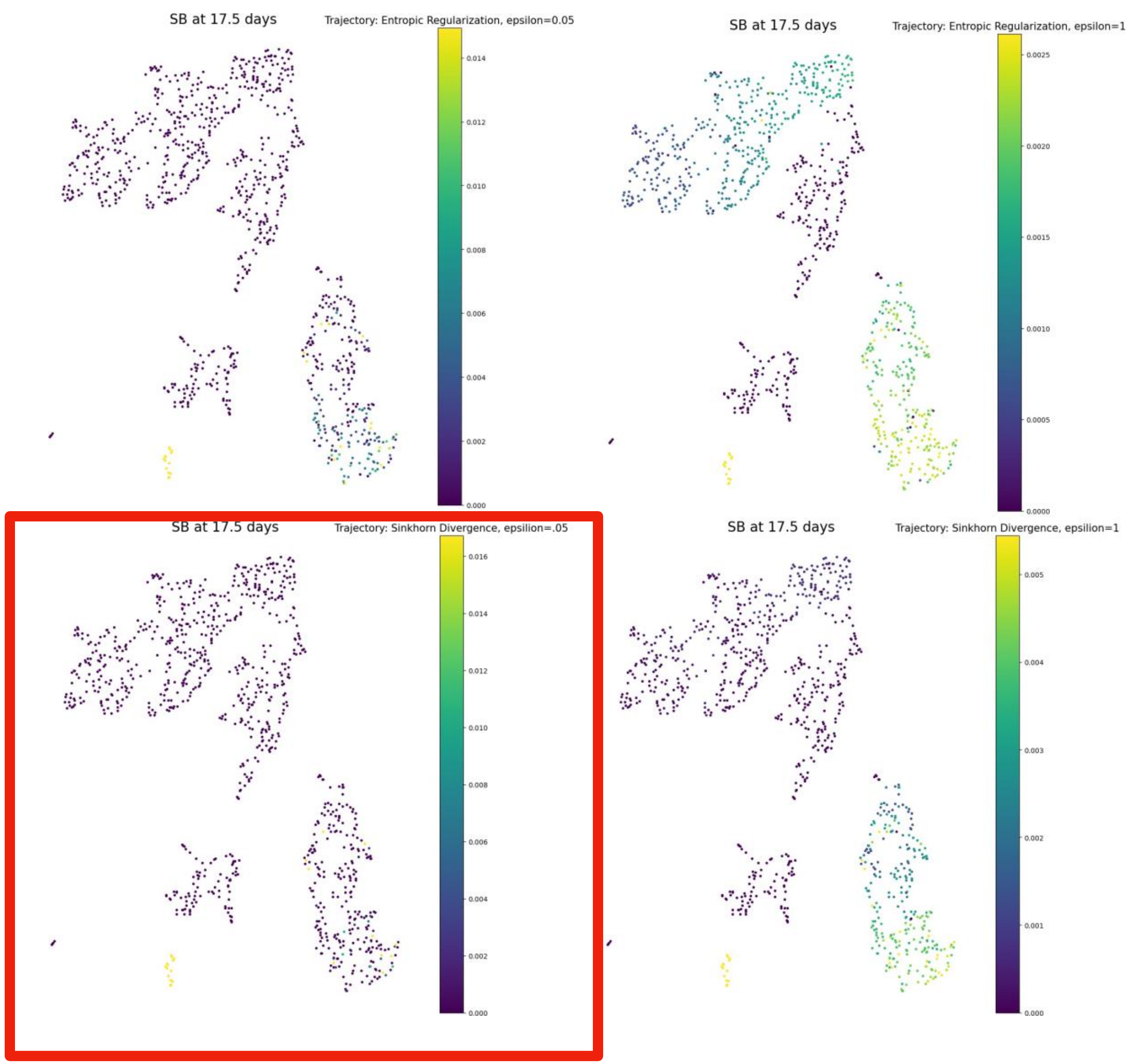


Mouse epidermal cells

Hair Follicle (HF) Trajectories Probabilities at Day 17.5



Suprabasal (SB)Trajectories Probabilities at Day 17.5



Conclusion and Future works

1. In contrast to entropic regularization alone, the sinkhorn trajectory probabilitiesmore often correlate with the original cell set distribution.

2. Next step is to compute the Sliced-Wasserstein distance between the distribution of the original cell set and the distribution of the subset of cells with highest trajectory probabilities, which can be used to define our findings quantitatively. Also, it deserves to optimize the hyperparameters associated with Waddington-OT, i.e. ϵ, λ_1 and λ_2.

Symposium, workshop, seminars

1. One day symposium on AI, Ethics, and Health Disparities at Morgan State

- May 26, 2023
- Participants: around 60
- Keynote speakers:

Dr. David Danks, UCSD, Data Science and Philosophy

The double-edged sword of AI in healthcare

Dr. Melissa McCradden, U of Toronto, Bioethics

An organizational ethics framework for supporting justice, equity, fairness, and anti-bias (JustEFAB) in clinical machine learning systems

2. Health AI Bias and Datathon in Emory School of Medicine

- Aug 19 - 21, 2023
- Detecting and mitigating bias in algorithms on medical imaging datasets: application to Emory pulmonary disease subjects
- 2nd Award in a team competition

3. Interdisciplinary Seminar, Morgan State

- Sep 7, 2023
- Detecting and mitigating bias in algorithms on medical imaging datasets: application to Emory pulmonary disease and breast cancer subjects

Publication

1. **Characterizing cell competition in the developing epidermis using optimal transport algorithms**
   Christina Young, Pilhwa Lee
   Senior Research Presentation, Department of Mathematics, 2023

2. **Algorithmic bias in single cell analysis:**
   Sinkhorn divergence based optimal transport of cellular dynamics
   Jayshawn Cooper, Christina Young, Pilhwa Lee
   JSM 2023 poster presentation

3. **Eulerian and Lagrangian interaction: cell-cell interaction with debiasing by Sinkhorn divergence**
   Samson Alagbe Tetn, Jayshawn Cooper, Christina Young, Pilhwa Lee
   RCMI National Conference, Poster presentation May 2024

Publication in prepartion

1. **Liouville PDE-based sliced Wasserstein: NIPS 2024**

   Fair regression with Wasserstein Barycenter:

   Diabete cytometry date, race and age

2. **Liouville PDE-based sliced Wasserstein: SIAM Journal of Uncertainty Quantification**

   Persistent homology for high dimensional cytometry analysis

3. **Meta-learning of differential gene expression analysis**

4. **Liouville PDE-based meta-reinforcement learning: NIPS 2024**

5. **Koopman operator and Liouville PDE-based meta-reinforcement learning**

6. **Liouville PDE-based Bayesian meta-reinforcement learning**