SPOT: Spatial Optimal Transport for Analyzing Cellular Microenvironments

Shouvik Mani[1,2,3*] **Doron Haviv**[2*] **Russell Z. Kunes**[2,3] **Dana Pe'er**[2,4]

[1]Department of Computer Science, Columbia University
[2]Computational and Systems Biology, Memorial Sloan Kettering Cancer Center
[3]Irving Institute for Cancer Dynamics, Columbia University
[4]Howard Hughes Medical Institute

Abstract

The spatial context of cells, comprising environmental signals and cell-cell interactions, shapes cell states and contributes to global tissue function. From immune responses to stem cell differentiation [1, 2, 3], to fully understand a biological process, it is imperative we study cells in their spatial context. Although spatial transcriptomic (ST) technologies such as MERFISH or seqFISH [4, 5] measure location and gene expressions at the resolution of individual cells, there is a lack of specialized methods to reason about the cellular microenvironments from these data sets. We present SPOT (Spatial Optimal Transport) which harnesses OT [6] to investigate cellular environments from ST, enabling quantitative analysis of niches from which we can cluster and visualize cells based not on their phenotype, but their neighborhoods. We apply SPOT on mouse primary motor cortex assayed with MERFISH [7] and seqFISH data of organogenesis [8] and find canonical niches which capture cortical layers within the cortex and polarization along the anterior-posterior axis.

1 Introduction

Cells exist and function not in a vacuum, but in complex microenvironments where they interact with other cells and orchestrate tissue function. This intercellular cross-talk can be observed in the lymph nodes, where various immune cell types localize into niches to mount innate and adaptive immune responses against pathogens [9, 10]. In cancer biology, interactions in the tumor immune microenvironment inhibit immune function and promote tumor growth, and are the targets for immunotherapies such as immune-checkpoint blockade [11, 12]. Given their importance, there is growing interest in understanding cellular environments and the role they play in disease and therapeutic responsiveness.

Despite recent advances in ST technologies which provide the opportunity to characterize cellular environments and study cells in their spatial context, there is still a lack of dedicated computational methods for spatial data. Current approaches are usually grounded on cell typing [13], which is itself a non-trivial task for high dimensional spatial data [14]. Cell typing also introduces bias through errors and choice of cell typing resolution, as the model has no knowledge which cell types are related, and buries any continuous signals in the data.

In this paper, we propose SPOT, a method to unravel the environments present in a ST data set, which relies on raw expression profiles, not hampered by cell typing and suited to reveal continuous trends in spatial data. Using Optimal Transport (OT), we derive a principled distance metric which can quantitatively compare niches. We harness SPOT to develop a k-means clustering algorithm which

36th Conference on Neural Information Processing Systems (NeurIPS 2022), Learning Meaningful Representations of Life (LMRL) Workshop.

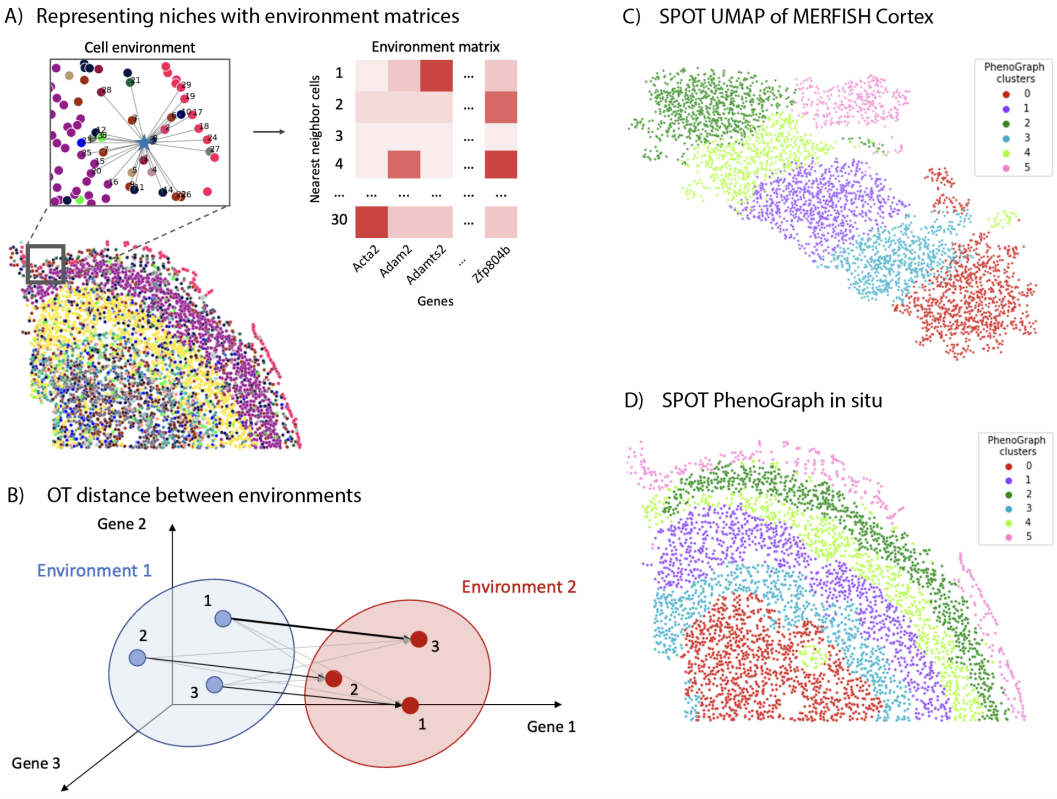

Figure 1: A) Representing cellular environments. A query point (blue star) is placed in coordinate space. A cell x gene environment matrix is assembled containing the gene expression vectors of the nearest cells. B) Optimal Transport to quantify niche differences. Given the set of gene expression vectors in two environments (red and blue), we define the distance between them as the OT divergence between the sets of expression vectors. Intuitively, OT finds a *soft* mapping between points as to best match their sets, shown as dark arrows, and the OT divergence is how good that best matching is. By constructing the SPOT cell by cell distance matrix, we can use algorithms like UMAP (C) and phenograph (D) to understand niches, where two cells are clustered together not if their expressions are similar, but rather if their environments are.

learns a set of canonicals which are representative of distinct environments found in the data set and to visualize environments with an OT UMAP [15], where two cells are close together if their niches are alike, not their expressions.

Applied to MERFISH of primary motor cortex, SPOT learned canonicals representing niches found in distinct cortex layers and revealed heterogeneity within cells of the same cell type in an environment, potentially indicating rare or transient cell states. Further applied on a seqFISH assay of a developing embryo, SPOT was able to reveal trends of gene expression along the anterior-posterior axis.

2 Methods

2.1 Representing cellular environments

To represent cellular environment in ST data, we consider a sample of m cells with spatial coordinates $(x_1, y_1), \ldots, (x_m, y_m)$ and gene expression vectors $g_1, \ldots, g_m$. For each cell, we identify the k cells closest to it in situ and use their expression to construct a $k \times l$ environment matrix $E_i = \left[\{g_j \mid j \in kNN(i)\}\right]$, containing the length-$l$ gene expressions of the k cells nearest to it (Figure 1), resulting in a set of environment matrices $E_1, \ldots, E_n$, representing the immediate niche of each cell. As

environment matrices do not rely on any parameters other than the choice of number of nearest neighbors, and do not use cell typing or artificially averaging the expression across cells in a niche, we see these matrices as the most straightforward and unbiased representation of cellular niches.

Given the environment matrices, the goal of SPOT is to describe a principled way to compare them to one another, defining a distance metric from which we can visualize environments with algorithms like UMAP, tSNE, or Force Directed Layout [15, 16, 17], uncover continuous trends using diffusion components [18], and cluster environments into significant groups, while discovering significant environment as cluster centers.

2.2 Optimal Transport Between Environments

Once we have assembled cellular environments, a natural question to ask is which environments are similar or different from each other. While the method above provides an intuitive approach for representing environments, it introduces challenges when comparing environments. Consider an environment E' that is simply a permutation of the rows of another environment E, and while E' and E are identical, as there is no true order to cells within a niche, their euclidean distance $\|E - E'\|_F$ will be greater than 0. Clearly, the euclidean distance does not meaningfully capture differences between niches.

This question of similarity is at the core of clustering methods such as k-means clustering and graph-based Louvain clustering or visualization algorithms like UMAP, which use a distance metric to map together data points which are similar to each other. In order to faithfully cluster together environments which are phenotypically similar, we need a permutation-invariant distance metric, that is, a metric which assigns small distances between environments with similar gene expressions regardless of how cells are ordered in their underlying representations. We can do this by first matching cells in niches (i.e. rows in environment matrices), and then quantifying how *good* that matching is. This is equivalent to finding the optimal permutation of niche matrices w.r.t to the Frobenius norm, and is known as the linear sum assignment [19] cost:

$$\text{dist}(E_i, E_j) = \min_P \|E_i - PE_j\|_F \text{ s.t. P is a permutation matrix}$$

As permutation are a closed set (i.e. a series of permutations is itself a permutation) any permutation of E_i or E_j will be absorbed into the optimization variable, and the calculated distance will remain the same, thus being a permutation-invariant distance between niches. Desirably, if two environments can be perfectly mapped to one another, their distance will be 0. While linear sum assignment can be solved with the Hungarian Algorithm, due to the hard assignment, this metric is not differentiable w.r.t to the niche matrices E_i or E_j, which prevents optimizing on niches according to it. The Hungarian Algorithm is also to computationally intensive which prohibits its use to calculate the pairwise distance for large spatial data sets.

Instead, we can consider finding a soft matching between cells in niches, where cells are not matched one-to-one, but rather are partially assigned to each other. This translates to changing the constraint in the equation above from P being a permutation matrix to P being doubly stochastic, which has non-negative values and both rows and columns sum to unit. The non-negativity stems from our desire for assignment, and the constraint on rows summing to unit refers that each cell from matrix E_i has to be completely assigned to cells in matrix E_j and vice-versa for the columns.

$$\text{dist}(E_i, E_j) = \min_P \|E_i - PE_j\|_F \text{ s.t. P is doubly stochastic}$$

This optimization problem is identical to finding the Optimal Transport between the set of points in E_i to the set of points in E_j, and we call using it to derive distances between niches Spatial Optimal Transport (SPOT). Unlike the Hungarian problem which is not differentiable and rather slow, OT can be efficiently solved using Sinkhorn Iterations [20] and the solution is fully differentiable w.r.t to E_i and E_j (Figure 1). Using OT, we can efficiently calculate the pairwise distances between the environment matrices of all the cells in a data set, which can be used as a direct input the phenograph, UMAP and diffusion components algorithms (Figure 1), leveraging a plethora of analysis tools to understand cellular environments.

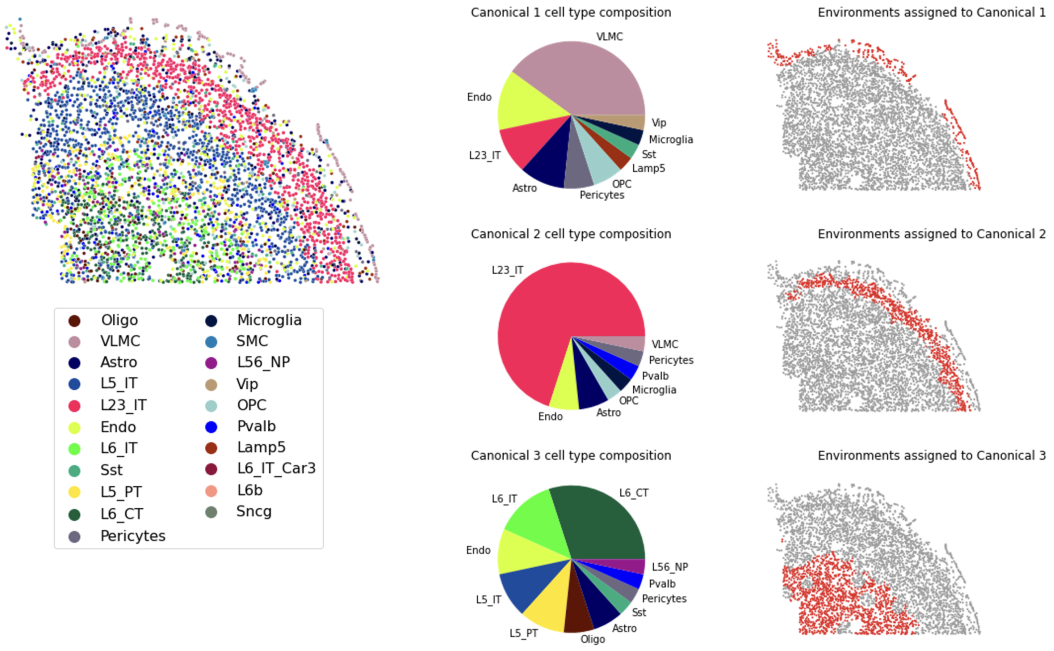

Figure 2: MERFISH data from mouse primary motor cortex (left) shows distinct cellular environments in each layer of the cortex. Environments were clustered using our method to learn canonical environments which have diverse cell type compositions (middle) and are representative of environments in distinct layers of the cortex.

2.3 Finding Canonical Niches in Spatial Data

As SPOT works on direct environment matrices, and is fully differentiable w.r.t to them, we can use it to optimize a k-means algorithm which not only clusters cellular niches based on their similarity, but also provides their "canonicals". Given a number of environment types k' and n environments $E_1, \ldots, E_n$, we want to learn k' canonicals $C_1, \ldots, C_{k'}$ which are representative of the environments. We propose a modified k-means algorithm for this task, designed to minimize the OT distance between environments and canonicals.

Algorithm 1 Clustering environments with an optimal transport distance metric

Input k', $E_1, \ldots, E_n$
Initialize $C_1, \ldots, C_{k'} \sim N(0, 1)$
while $\sum_{i=1}^{k'} \|C_i\|_2$ not converged **do**
  for $i = 1, \ldots, n$ **do**
    $y_i = \underset{j \in \{1,\ldots,k'\}}{\operatorname{argmin}} \operatorname{dist}(E_i, C_j)$ ▷ Assign environments to closest canonical
  end for
  for $j = 1, \ldots, k'$ **do**
    $C_j = C_j - \alpha \nabla_{C_j} \Big[\sum_{i \in \{y_i = j\}} \operatorname{dist}(E_i, C_j) \Big]$
    ▷ Update canonicals to minimize Sinkhorn distances
  end for
end while

Canonicals $C_1, \ldots, C_{k'}$ are randomly initialized. At each iteration, we assign each environment to its closest canonical based on Sinkhorn distances. Then, we perform a gradient descent step to update each canonical to minimize the sum of Sinkhorn distances between itself and the environments

which have been assigned to it. α is the learning rate for the gradient descent and the gradient $\nabla_{C_j}\left[\sum_{i\in\{y_i=j\}} \text{dist}(E_i, C_j)\right]$ can be computed through automatic differentiation [21]. We run these steps until $C_1, \ldots, C_{k'}$ have converged into representative canonical environments.

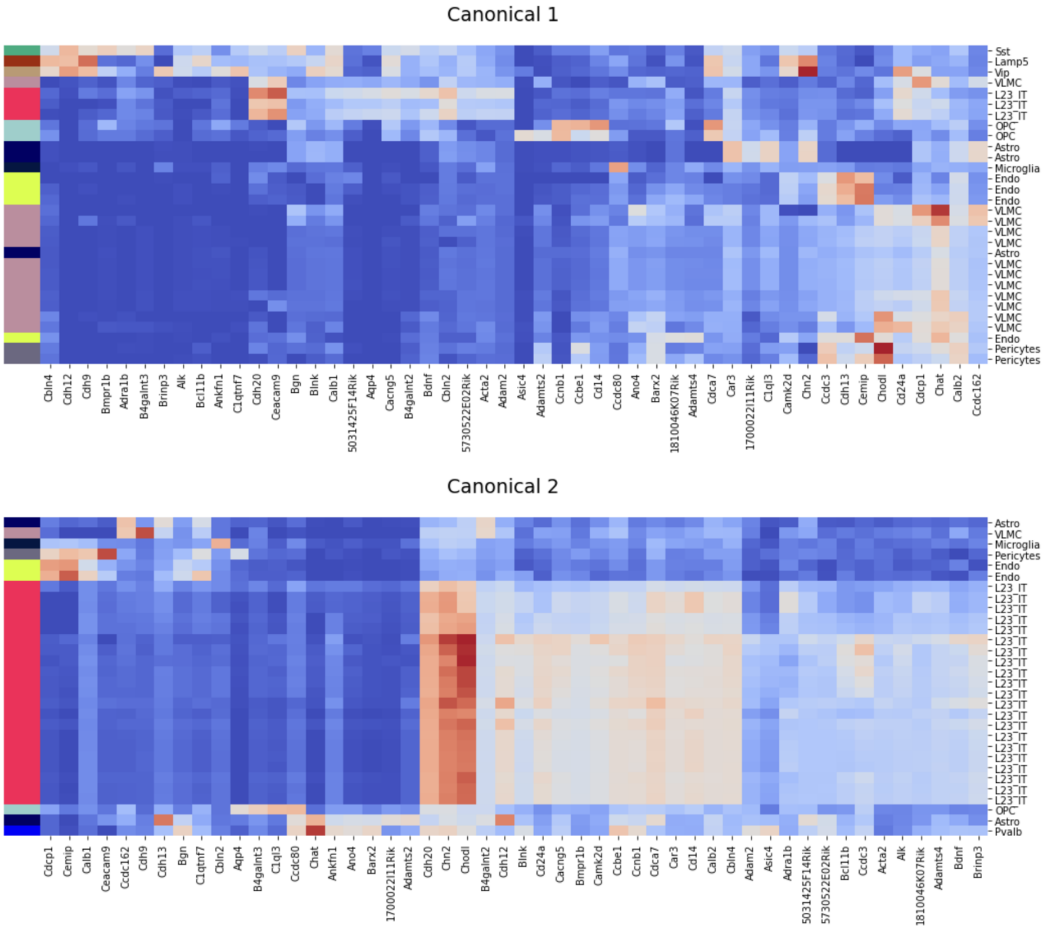

Figure 3: Gene expressions of cells in two of the learned canonicals, showing only the 50 out of 255 most variable genes for brevity. Variations within a common cell type in each canonical suggests that a cell-type only based analysis of microenvironments is insufficient and highlight the granularity of the description of environments achieved with SPOT.

3 Results

We apply our methods to analyze cellular environments in mouse primary motor cortex assayed by MERFISH [22]. The cortex is a highly organized structure with several neuronal cell types that localize into distinct layers. Using the SPOT distance matrix, we calculate a UMAP embedding to the cellular niches, and use our k-means algorithm to determine niche clusters and their canonicals. We find that the resulting canonical environments correspond to distinct layers of the cortex (Figure 2), and have a diverse composition of cell types, even having several variable expression profiles within cell types (Figure 3).

We further applied SPOT on seqISH assay for organogenesis, specifically on the cells which we labeled as "Spinal cord", due to their spatial trajectory along the anterior-posterior axis (Figure 4). With SPOT, we were able to draw the niche FDL, where two cells are embedded close together if their environments are similar, regardless of their own expression. The main axis of the FDL, along

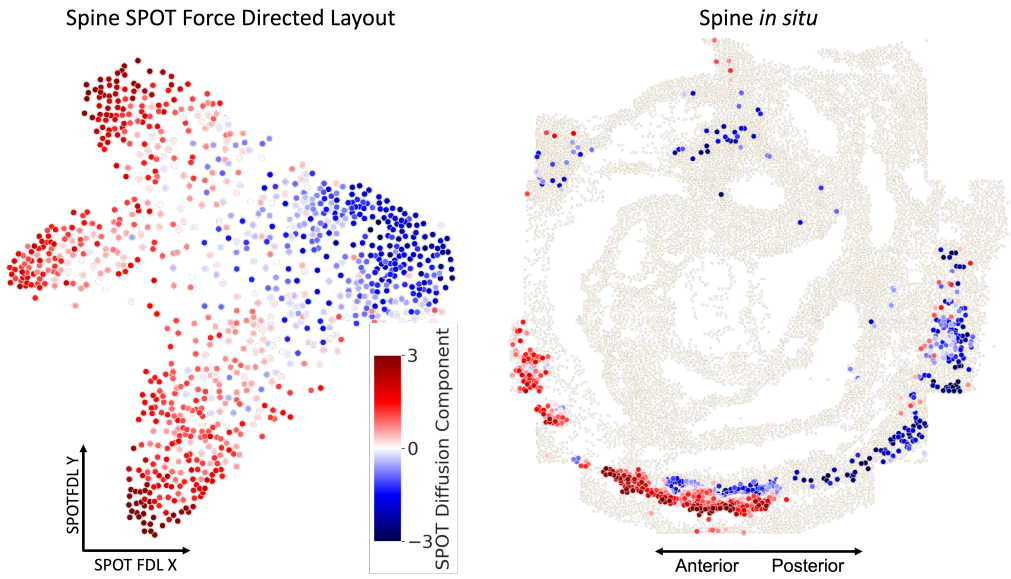

Figure 4: SPOT on Spinal Cord cells from seqFISH of mice organogenesis. From the SPOT calculated niche distance matrix on the spinal cord labeled cells, we were able to both computed a Force Directed Layout, were two cells are close together if their niches are similar, rather than expression profiles, and Diffusion Components, to highlight continuous trends. The first diffusion components along with primary axis along the FDL revealed the well characterized anterior-posterior specification within the embryo, demonstrating the ability of SPOT to group together niches with a similar composition.

with the primary SPOT based diffusion components, mapped to that anterior-posterior axis, validating that SPOT can uncover continuous trends in spatial data.

4 Discussion

We have presented a framework to reason about cellular environments in spatial transcriptomic data. SPOT comes with methods to represent environments, measure their similarity based on Optimal Transport, and visualize environments and cluster them into canonicals representative of environments *"types"* within a data set. SPOT delineates itself from other methods for spatial analyses at it relies on raw expression, and does not need for cell typing, while still being computational efficient and robust. We envision SPOT being used to find complete representations of spatial data, extending models such as scVI [23] to not just reconstruct expression, but also spatial context, utilizing the differentiability and flexibility of SPOT.

References

[1] Jordy Saravia, Nicole M Chapman, and Hongbo Chi. Helper t cell differentiation. *Cellular & molecular immunology*, 16(7):634–643, 2019.

[2] Linheng Li and Ting Xie. Stem cell niche: structure and function. *Annu. Rev. Cell Dev. Biol.*, 21:605–631, 2005.

[3] David T Scadden. The stem-cell niche as an entity of action. *Nature*, 441(7097):1075–1079, 2006.

[4] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. seqfish accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron*, 94(4):752–758, 2017.

[5] Guiping Wang, Jeffrey R Moffitt, and Xiaowei Zhuang. Multiplexed imaging of high-density libraries of rnas with merfish and expansion microscopy. *Scientific reports*, 8(1):1–13, 2018.

[6] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[7] Meng Zhang, Stephen W Eichhorn, Brian Zingg, Zizhen Yao, Kaelan Cotter, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. Spatially resolved cell atlas of the mouse primary motor cortex by merfish. *Nature*, 598(7879):137–143, 2021.

[8] T Lohoff, S Ghazanfar, A Missarova, N Koulena, N Pierson, JA Griffiths, ES Bardot, C-HL Eng, RCV Tyser, R Argelaguet, et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nature biotechnology*, 40(1):74–85, 2022.

[9] Spencer M. Grant, Meng Lou, Li Yao, Ronald N. Germain, Andrea J. Radtke, and Lennon-Duménil Ana-Maria. The lymph node at a glance – how spatial organization optimizes the immune response. *Journal of Cell Science*, 133(5), 03 2020.

[10] Joseph M. Leal, Jessica Y. Huang, Karan Kohli, Caleb Stoltzfus, Miranda R. Lyons-Cohen, Brandy E. Olin, Michael Gale, and Michael Y. Gerner. Innate cell microenvironments in lymph nodes shape the generation of t cell responses during type i inflammation. *Science Immunology*, 6(56), 2021.

[11] Mikhail Binnewies, Edward W. Roberts, Kelly Kersten, Vincent Chan, Douglas F. Fearon, Miriam Merad, Lisa M. Coussens, Dmitry I. Gabrilovich, Suzanne Ostrand-Rosenberg, Catherine C. Hedrick, Robert H. Vonderheide, Mikael J. Pittet, Rakesh K. Jain, Weiping Zou, T. Kevin Howcroft, Elisa C. Woodhouse, Robert A. Weinberg, and Matthew F. Krummel. Understanding the tumor immune microenvironment (time) for effective therapy. *Nature Medicine*, 24(5):541–550, May 2018.

[12] T. L. Whiteside. The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, 27(45):5904–5912, Oct 2008.

[13] Zhenqin Wu, Alexandro E Trevino, Eric Wu, Kyle Swanson, Honesty J Kim, H Blaize D'Angio, Ryan Preska, Gregory W Charville, Piero D Dalerba, Ann Marie Egloff, et al. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering*, pages 1–14, 2022.

[14] Francisco Jose Grisanti Canozo, Zhen Zuo, James F Martin, and Md Abul Hassan Samee. Cell-type modeling in spatial transcriptomics data elucidates spatially variable colocalization and communication between cell-types in mouse brain. *Cell Systems*, 13(1):58–70, 2022.

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

[16] Stephen G Kobourov. Spring embedders and force directed graph drawing algorithms. *arXiv preprint arXiv:1201.3011*, 2012.

[17] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[18] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[19] Roy Jonker and Ton Volgenant. Improving the hungarian assignment algorithm. *Operations Research Letters*, 5(4):171–175, 1986.

[20] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[21] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for

Large-Scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA, November 2016. USENIX Association.

[22] Meng Zhang, Stephen W. Eichhorn, Brian Zingg, Zizhen Yao, Kaelan R Cotter, Hongkui Zeng, Hong wei Dong, and Xiaowei Zhuang. Spatially resolved cell atlas of the mouse primary motor cortex by merfish. *Nature*, 598:137 – 143, 2021.

[23] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.