Towards scalable embedding models for spatial transcriptomics data

Seo-Yoon Moon*
College of Liberal Studies
Seoul National University
Seoul, Republic of Korea 08826
mrn538@snu.ac.kr

Ethan Weinberger
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98195
ewein@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98195
suinlee@cs.washington.edu

The functions of complex multicellular biological systems depend intimately on the spatial organization of their constituent cells [1]. As a result, significant efforts have been made towards developing spatially resolved transcriptomics (SRT) platforms that can measure gene expression levels while preserving spatial context, and these platforms have provided new insights into the mechanisms of Alzheimer's disease [2], embryonic development [3], and the tumor microenvironment [4].

SRT platforms have traditionally navigated a tradeoff between the number of profiled gene expression features and spatial resolution, thereby limiting the size of individual SRT datasets. For example, platforms based on fluorescent *in situ* hybridization (FISH) [5, 6] profile gene expression at the single-cell level but have traditionally been limited to profiling expression levels of a small subset of tens or hundreds of pre-selected marker genes. On the other hand, sequencing-based methods, such as 10x Genomics Visum [7] and Slide-seq [8]), provide whole-transcriptome measurements yet have traditionally been limited to profiling coarser-grained (e.g. $55\,\mu\mathrm{m}$ for Visium) spots containing multiple cells. However, recent developments have enabled a significant scaling up of these platforms, with recent works presenting FISH platforms that can profile at the near-transcriptome level [9] as well as sequencing-based platforms that operate at near-cellular [10] or subcellular [11] resolution.

An important problem in the analysis of SRT datasets is clustering cells/spots into spatial domains (i.e., spatial regions with coherent gene expression patterns). While a significant line of work exists for clustering analysis of scRNA-seq measurements (see [12] for a review), such methods are not suitable for analyzing SRT data, as they discard spatial information and thereby may produce clusters that are not spatially coherent. Thus, a recent line of work has developed methods that explicitly account for spatial information in the clustering process using e.g. spatial Bayesian priors [13, 14] or graph-neural-network-based approaches [15–17]. Such methods have exhibited notable improvements in performance over non spatially aware methods. However, these methods are not equipped to handle increasingly large-scale SRT datasets generated by newer platforms, as they necessitate loading entire datasets into memory at once and thus exhibit at least linear scaling in terms of memory usage as the number of samples and/or features increases.

As a first step towards resolving this issue, here we present GLaST (Graph embedding for Large-scale Spatial Transcriptomics data), a scalable spatially aware embedding framework designed to facilitate the analysis of large-scale SRT datasets. GLaST leverages previous work in large-scale graph embedding methods originally developed for social network applications [18], which have been adapted previously to analyze dissociated scRNA-seq datasets [19]. We first applied GLaST to a standard spatial domain detection benchmark dataset and found that it achieved comparable performance compared to previous state-of-the-art methods. We then evaluated GLaST's memory consumption for datasets with increasing numbers of samples, and found that it exhibited significantly lower memory usage compared to previously proposed methods for spatial domain detection.

1 Overview of GLaST

As with other graph-based methods for spatial domain detection, we begin by constructing a graph $G = (V, E)$ with vertices V and weighted edges E that captures the similarity between samples both in terms of gene expression and spatial location. Here, our set of vertices consists of one vertex for each sample (i.e., spot or cell) as well as one vertex for each gene expression feature. We then add two types of edges to our graph. First, to capture spatial similarity between samples, we add an edge between two sample vertices if the samples are spatially adjacent. Next, to capture similarity in terms of gene expression, we add edges between sample vertices and gene feature vertices if a given sample expresses a given gene. Sample to gene edge weights was determined using the discretization procedure of [19], which sets the weights in proportion to the normalized gene expression.

After constructing our graph, we then seek to learn an embedding of our dataset $\Theta \in \mathbb{R}^{|V| \times d}$ consisting of embeddings $\theta_v \in \mathbb{R}^d$ for each vertex in our graph. We proceed to learn Θ following the approach proposed in SIMBA [19]. Specifically, for an edge $e = (u, v)$, we let $s_e = (\theta_u)^T \theta_v$ denote the similarity score of e. We then optimize the following multi-class log loss

$$\mathcal{L}_e = -\left(\frac{\exp(s_e)}{\sum_{e' \in \mathcal{N}} \exp(s_{e'})} \cdot w_e\right),$$

*Work performed while a visiting student researcher at the University of Washington.

Preprint. Under review.

where $\mathcal{N}$ is a set of "negative edges" (i.e., connections that do not exist in the true graph G) and w_e is the weight of edge e. By learning Θ to minimize $\mathcal{L}$, we encourage $s_e = (\theta_u)^T\theta_v$ to be large for $(u, v) \in E$ and $s_{e'}$ to be small for $(u, v) \notin E$. Thus, embeddings of sample nodes are encouraged to be close if the samples are both spatially adjacent and express similar sets of genes. We highlight that, unlike previously proposed graph-based methods for SRT data, this procedure is amenable to optimization via minibatch gradient descent and thus does not require loading full datasets into memory during optimization. After optimization, standard visualization methods and clustering algorithms can be applied to embeddings of sample nodes to identify spatial domains. We illustrate this workflow in **Fig. 1**.

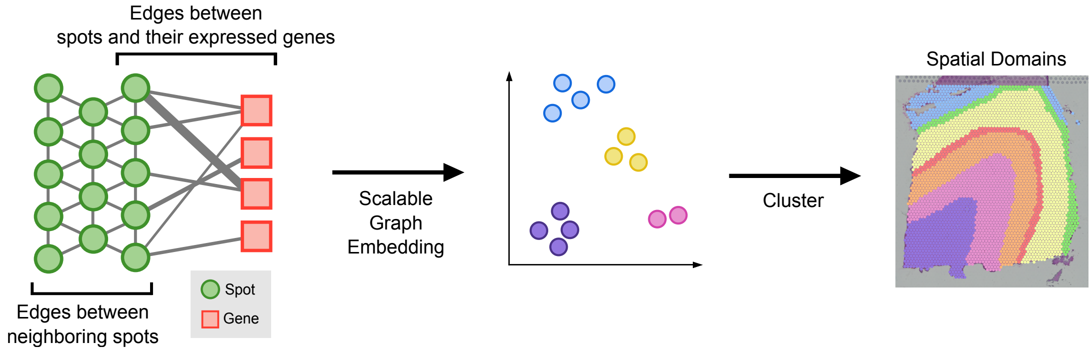

Figure 1: The GLaST framework. A graph is first constructed such that each sample and each gene expression feature is represented as a vertex. Edges are added between sample vertices if the samples are spatially adjacent, and between a sample vertex and a feature vertex if that sample expresses the given gene feature. Low-dimensional representations of the vertices are then learned using scalable graph embedding techniques, and the learned embeddings of sample vertices can subsequently be leveraged for spatial domain detection.

2 Results

Spatial Domain Detection: To evaluate the performance of GLaST, we applied it to a 10x Visium dataset containing SRT measurements of 12 human dorsolateral prefrontal cortex (DLPFC) slices [20], which is commonly used for spatial domain detection benchmarking [13, 15–17]. Each slice in this dataset has up to seven layers of grey and white matter that are manually annotated by human experts, thus providing a set of ground truth spatial domain labels. Following previous work [16], for this dataset, we added edges between samples indicating spatial adjacency in the GLaST input graph if the two samples were connected in a six-nearest-neighbors graph based on the samples' x and y spatial coordinates. In order to quantify GLaST's performance, we clustered the GLaST embedding space for each slice as determined by the optimized Leiden clustering procedure implemented in the single-cell integration benchmarking (scIB) package [21]. Then, we evaluated the agreement between the ground truth spatial domain labels using the adjusted Rand index (ARI). We compared GLaST's performance to that of SpaGCN [15], and STAGATE [16], two graph-neural-network based methods for spatial domain detection that have recently claimed state-of-the-art results, as well as the non-spatially aware Louvain [22] algorithm as implemented in scanpy [23]. We found (**Fig. 2a**) that GLaST achieved competitive performance compared to previously proposed spatially aware methods.

Scalability analysis: We next assessed GLaST and previously proposed spatially aware baseline models' scalability in terms of memory consumption. To do so, we assessed how much memory GLaST, SpaGCN, and STAGATE used when trained on a mouse embryo dataset collected using the Stereo-seq platform [11]. In particular, we examined how each model's memory usage varied when trained with a random subset of spots for varying numbers of spots. We found **Fig. 2b** that GLaST exhibited significantly better scalability (i.e., lower memory usage) compared to baseline models as the number of spots increased.

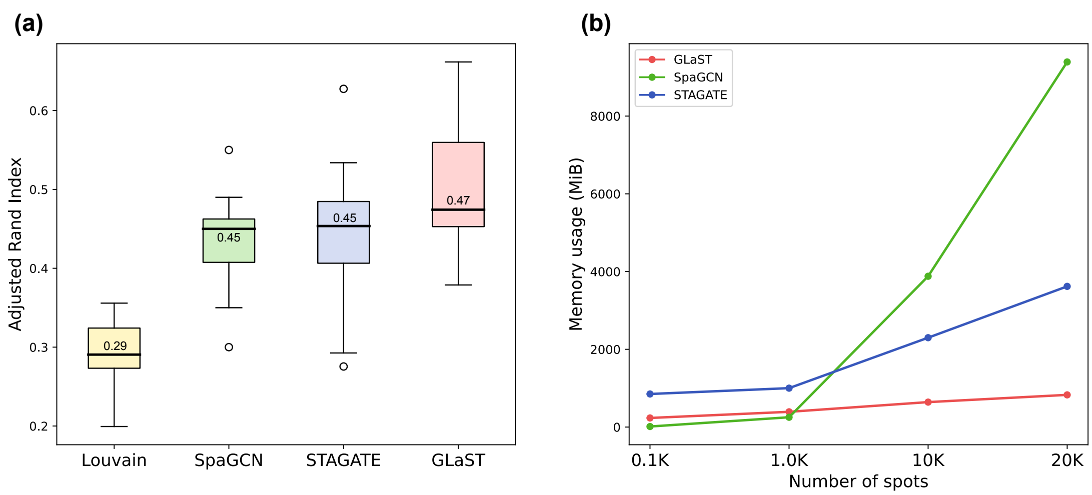

Figure 2: **(a)** Performance of GLaST and baseline methods on the spatial domain detection task as measured by adjusted Rand index (ARI). Median values are shown and are indicated as bold lines. **(b)** Memory usage versus number of spots for the Stereo-seq mouse embryo dataset.

References

[1] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.

[2] Wei-Ting Chen, Ashley Lu, Katleen Craessaerts, Benjamin Pavie, Carlo Sala Frigerio, Nikky Corthout, Xiaoyan Qian, Jana Laláková, Malte Kühnemund, Iryna Voytyuk, et al. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991, 2020.

[3] Abhishek Sampath Kumar, Luyi Tian, Adriano Bolondi, Amèlia Aragonés Hernández, Robert Stickels, Helene Kretzmer, Evan Murray, Lars Wittler, Maria Walther, Gabriel Barakat, et al. Spatiotemporal transcriptomic maps of whole mouse embryos at the onset of organogenesis. *Nature Genetics*, pages 1–10, 2023.

[4] Miranda V Hunter, Reuben Moncada, Joshua M Weiss, Itai Yanai, and Richard M White. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nature communications*, 12(1):6278, 2021.

[5] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[6] Simone Codeluppi, Lars E Borm, Amit Zeisel, Gioele La Manno, Josina A van Lunteren, Camilla I Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmfish. *Nature methods*, 15(11):932–935, 2018.

[7] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[8] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[9] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019.

[10] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

[11] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

[12] Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, pages 1–23, 2023.

[13] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384, 2021.

[14] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22:1–31, 2021.

[15] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.

[16] Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.

[17] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

[18] Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. Pytorch-biggraph: A large scale graph embedding system. *Proceedings of Machine Learning and Systems*, 1:120–131, 2019.

[19] Huidong Chen, Jayoung Ryu, Michael E Vinyard, Adam Lerer, and Luca Pinello. Simba: Single-cell embedding along with features. *Nature Methods*, pages 1–11, 2023.

[20] Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.

[21] Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, Jan 2022.

[22] 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, oct 2008.

[23] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.