Scalable embedding model for spatially-resolved transcriptomics data

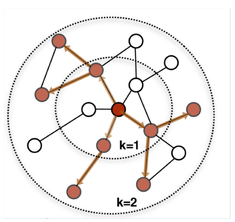
Seo-Yoon Moon*, Ethan Weinberger* , Su-In Lee
(* indicates equal contribution)
symoon@uw.edu, {ewein, suinlee}@cs.washington.edu
University of Washington

MOTIVATION

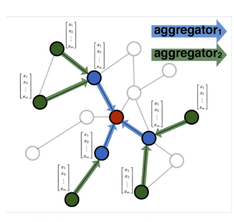
- **Spatially-resolved transcriptomics (SRT)** have enabled gene expression profiling while retaining spatial context in tissues
- As the technology continues to evolve, **higher-resolution datasets from larger tissue regions are becoming available**
- Previous works[1-3] have learned spatially-aware low-dimensional embeddings of SRT data using **graph neural networks** trained on graphs whose edges represent spatial information and histology
- Embeddings can then be used for **automated spatial domain detection** (i.e., finding regions that are spatially coherent in gene expression and/or histology)
- However, previously proposed methods require **full-batch training** and are thus **not scalable to large datasets**
- To facilitate the analysis of larger-scale SRT datasets, we investigated applying GraphSAGE **a scalable graph neural network embedding model**, to SRT data

Method

- GraphSAGE[4] is a **framework for scalable representation learning on graphs**
- For a given node, GraphSAGE **samples a neighborhood around that node** and aggregates information from neighbors to produce an embedding



1. Sample neighborhood

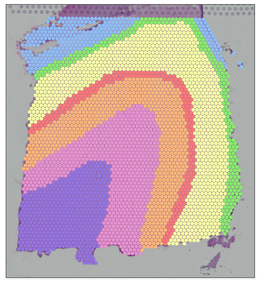


2. Aggregate feature information from heighbors

- By considering only necessary neighborhoods for a set of nodes, GraphSAGE can be trained using mini-batches of data with **constant memory requirements even as dataset size increases**
- We train GraphSAGE by **encouraging neighboring spots to have similar embeddings and non-neighboring spots to have distinct embeddings**

DATA

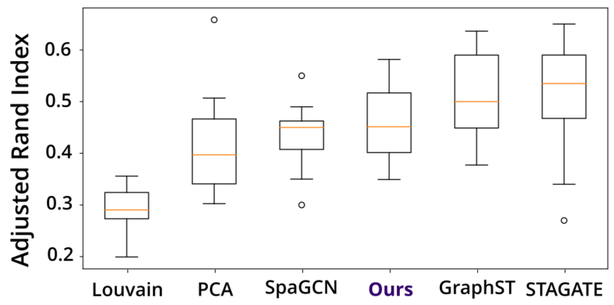
- **Human brain tissue** data from Maynard et al.[5]
- Collected using 10x Genomics Visium platform
- Each spot is adjacent to six other spots
- We construct an unweighted graph that connects each node to its six neighbors



RESULTS

Spatial domain detection

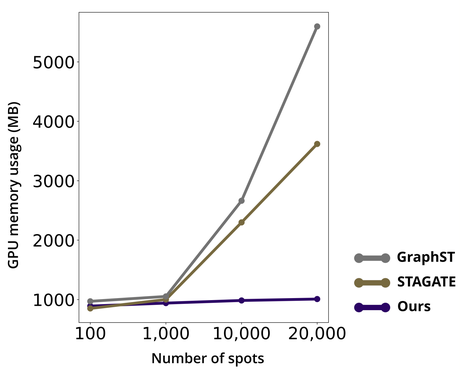
We quantify agreement between the spatial domains obtained by each model and ground truth annotations from human experts using the Adjusted Rand Index



	Louvain	PCA	SpaGCN	Ours*	GraphST	STAGATE
Mean	0.29	0.42	0.43	0.46	0.52	0.51
SD	0.04	0.10	0.07	0.08	0.08	0.11

GPU memory usage

Our model has **near-constant memory usage**, while others' memory grow rapidly as number of spots increase



DISCUSSION

- GraphSAGE showed **comparable performance to the current state-of-the-art** model while using significantly less memory

NEXT STEPS

- **Incorporate histology information in graph construction.** Here we constructed our graph input using only spatial coordinates, but we hypothesize that performance could be further improved by also incorporating histology information.
- **Incorporate decoder with GraphSAGE model.** We can add a decoder network (*e.g.,* that of scVI[8]) to the original model and add a reconstruction term to our loss function. By forcing embeddings to capture information for accurate reconstruction, we may see further performance increases.
- **Additional downstream tasks.** With the decoder, we could also use our model to do several other downstream tasks, such as imputation and differential expression analysis.
- **Incorporate other modalities.** We can extend our model to handle other spatial 'omics datasets such as spatial CITE-seq[6] (RNA + protein measurements), and spatial ATAC-seq[7] (chromatin accessibility)

REFERENCES

[1] Dong, K., & Zhang, S. (2022). Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, *13*(1), 1739. https://doi.org/10.1038/s41467-022-29439-6
[2] Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shinohara, R. T., & Li, M. (2021). 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.
[3] Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., Xu, H., Ong, Z., Sachaphibulkij, K., Chen, A., Zeng, L., Fu, H., Wu, M., Lim, L. H. K., Liu, L., & Chen, J. (2023). Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, *14*(1), 1155
[4] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in neural information processing systems, 30.
[5] Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., ... & Jaffe, A. E. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature neuroscience, 24(3), 425-436.
[6] Liu, Yang, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. *Nature Biotechnology* (2023): 1-5.
[7] Deng, Yanxiang, et al. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* 609.7926 (2022): 375-383.
[8] Lopez, Romain, et al. "Deep generative modeling for single-cell transcriptomics." *Nature methods* 15.12 (2018): 1053-1058.

ACKNOWLEDGEMENTS

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING
NSF
NIH