Towards scalable embedding models for spatial transcriptomics data

Seo-Yoon Moon[1], Ethan Weinberger[2] , Su-In Lee[2]

mrn538@snu.ac.kr, {ewein, suinlee}@cs.washington.edu

[1] Seoul National University, [2] University of Washington

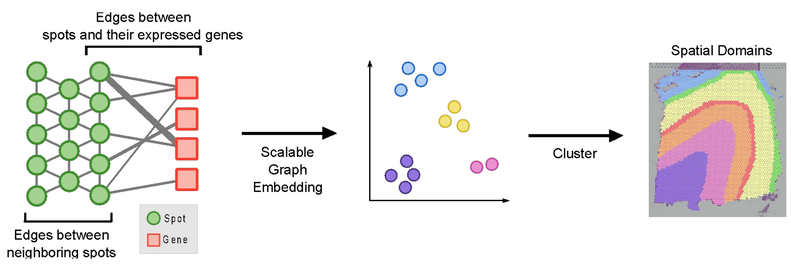
MOTIVATION

- **Spatially resolved transcriptomics (SRT)** technologies enable gene expression profiling while retaining spatial context in tissues.
- An important problem in the analysis of SRT datasets is **spatial domain detection** (i.e., detecting spatial regions with coherent gene expression patterns)
- Moreover, as SRT technology continues to evolve, **higher-resolution datasets from larger tissue regions are becoming available**
- However, previously proposed computational methods[1, 2, 3] for spatial domain detection require **full-batch training** and are thus **not scalable to large datasets**.
- To facilitate the analysis of larger-scale SRT datasets, we investigated **a scalable graph neural network embedding model** for analyzing SRT data.

METHOD

Overview of the model: GLaST

- GLaST (Graph embedding for Large-scale Spatial Transcriptomics data) is an **unsupervised graph neural network** that leverages previous scalable graph embedding method for social network.
- GLaST encourages close nodes to have similar embeddings and distant nodes to have distinct.
- We apply **GLaST to a real-world SRT dataset** and analyze its performance on **spatial domain detection.**

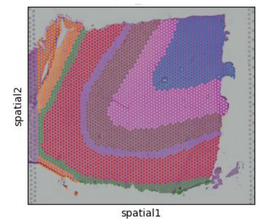


Step-by-step Procedure

- **Step 1. Create a graph** whose nodes consist of both spots and genes. Spots are connected if they are adjacent, and a spot and a gene are connected if the gene is expressed in a given spot
- **Step 2.** Split graph data into **mini-batch** (containing both negative and positive edges) and **train** GLaST.
- **Step 3.** Post-training, feed graph into the encoder and **get spatially aware embeddings of SRT expression data.**
- **Step 4. Cluster** the embeddings to obtain spatial domains.

RESULTS

- Human brain tissue data with ground truth spatial domain labels from human experts.
- Average 3,944 spots for 12 samples.
- Only used 2,000 highly variable genes.
- Each spot is adjacent to 6 other spots, forming hexagonal shape.

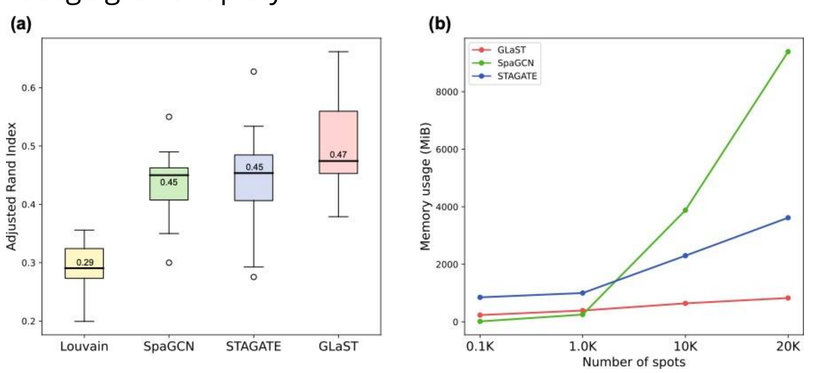


(a) Spatial domain detection

Boxplot of ARI score across three existing models and our method. Our model (GLaST) achieved comparable performance to other models.

(b) GPU memory usage

Our model maintained near-constant memory usage, with increasing numbers of spots, while others' memory usage grows rapidly.



DISCUSSION

- GLaST showed comparable performance to the current state-of-the-art, which implies SRT data can be broken down into minibatches for training just as other forms of data.

NEXT STEPS

- **Attach decoder to GLaST model.** We can add a decoder network to the model and train further with a reconstruction loss term. This might increase the performance, and could enable our model to perform additional downstream tasks, such as imputation and differential expression analysis.
- **Conduct analysis with other spatial omics data.** We can extend our results to other spatial data, including proteomics, epigenomics, etc.

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. https://doi.org/10.1038/s41592-021-01255-8

[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. https://doi.org/10.1038/s41467-023-36796-3

[4] 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.

ACKNOWLEDGEMENTS

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING