EVALUATING SPATIAL ENCODING STRATEGIES FOR CELL TYPE ANNOTATION WITH SPATIAL OMICS DATA

Merel Kuijs[1, 2, *], **Alma Andersson**[3], **Ehsan Hajiramezanali**[3], **Tommaso Biancalani**[3], **and Aïcha Bentaieb**[3]

[1]Institute of Computational Biology, Computational Health Center, Helmholtz Munich, Germany
[2]TUM School of Computation, Information and Technology, Technical University of Munich, Germany
[3]Genentech, South San Francisco, CA 94080, United States
[*]Work done while at Genentech
{*andersson.alma,hajiramezanali.ehsan,biancalani.tommaso,bentaieb.aicha*}*@gene.com*

ABSTRACT

Recent spatial omics research leverages the assumption that spatial information enhances model performance on the cell type annotation task. This study investigates and challenges that assumption by conducting benchmark experiments comparing the performance of spatial and non-spatial models. We show that graph-based spatial models do not consistently outperform non-spatial models, provide theories to explain our findings, and make recommendations for future work on spatial encoding strategies.

1 INTRODUCTION

Spatial omics technologies measure the spatial distribution of molecular features (e.g., genes, proteins, metabolites) in tissue sections, placing them in a histological context. The introduction of these technologies has prompted the development of methods for new computational tasks, such as: identification of spatially variable genes (Svensson et al., 2018; Sun et al., 2020; Zhang et al., 2023), characterization of cell-cell interactions (Arnol et al., 2019; Fischer et al., 2021), delineation of spatial domains (Hu et al., 2021; Zhao et al., 2021; Dong & Zhang, 2022; Kim et al., 2022; Varrone et al., 2023), and the discovery of novel cell states (Bäckdahl et al., 2021; Cable et al., 2022). Since their late 20th-century inception, advancements in spatial omics have included readouts at higher spatial resolution, larger fields of view, and the ability to analyze more features. Current technologies offer detailed insights into cellular heterogeneity, tissue architecture, disease pathology, and cellular responses, enhancing our understanding of biological processes (Moses & Pachter, 2022).

Cell type annotation is the task of assigning biologically meaningful labels to each cell, it is also an integral part in any cell-focused analysis (spatial or non-spatial) as these labels are often necessary for biological interpretation of any downstream analysis. Cell types represent groups of cells with similar intrinsic properties (e.g., B-cells or T-cells) that are robust to changes in a cell's environment. Annotation of cell types in spatial omics datasets is typically done using either manual annotation, co-embedding with a reference dataset followed by label transfer (semi-supervised learning tasks) (Xu et al., 2021; Shen et al., 2022), or using classifiers trained on one or more reference datasets (supervised learning tasks) (Alquicira-Hernandez et al., 2019; Abdelaal et al., 2019). Manual annotation is a time-consuming process that tends to be subjective and requires domain expertise but is often considered as the gold standard. Leveraging machine learning methods and reference datasets usually facilitates and scales the annotation process, but relies on accurate models and reference datasets that cover the full cell type label space. With the emergence of foundational models for omics data, the idea of reliable, fully automatic cell type annotation is becoming increasingly tangible (Cui et al., 2023; Heimberg et al., 2023). There is still uncertainty about the best design choices when designing such foundational models (or smaller models) for this purpose. For example, it has been stated that – for spatial data – including information about the spatial relationships among cells in addition to their molecular profiles boosts performance on the cell type annotation task (Brbić

et al., 2022; Shen et al., 2022; Shaban et al., 2024); however, to the best of our knowledge, there hasn't been any systematic evaluation of this hypothesis.

In this paper we seek to address this gap in understanding; we present a benchmarking study designed to empirically test whether *common strategies of explicitly encoding spatial relationships alongside gene expression data significantly improve performance on the cell type annotation task*. Following the state-of-the-art methods (Fischer et al., 2021; Hu et al., 2021; Brbić et al., 2022) for encoding spatial context in spatial omics data, we model the spatial data as a graph: cells are represented as nodes, molecular features as node features, and the cell type of each cell as its node label. In this graph, edges connect spatially proximal cells. We then frame the cell annotation problem as a supervised node classification problem; models are trained to predict node labels given the graph structure and node features. The performance of spatial models (utilizing both node features and graph structure) is compared with that of non-spatial models (only utilizing the node features) across several spatial omics datasets (including transcriptomics and proteomics).

To summarize, our key contributions are as follows:

1. A systematic inquiry as to whether common strategies of explicitly encoding spatial relationships improve model performance.
2. An assessment of model robustness to noise in spatial biology data, including: i) technical noise in features (e.g., amplification bias); and ii) noisy ground truth labels.
3. Recommendations for best practices for the community when working on similar tasks.

2 Related Works

Previous works have explored how spatial and non-spatial models perform in a task related to cell type annotation, namely domain identification (Heumos et al., 2023), which involves assigning spatial domain labels to cells. Domain identification and cell type annotation differ in the types of labels that are assigned. The domain identification task is akin to a region segmentation task in computer vision. Spatial domains, also known as spatial clusters or regions, represent areas in the tissue with unique characteristics, such as different cell type compositions, and can include anatomical regions (e.g., the hippocampus and cortex in the context of the brain). Methods for domain identification vary in their design: some use only molecular features, whereas others incorporate spatial relationships or additional data from histological images (Satija et al., 2015; Pham et al., 2020; Tan et al., 2020; Dries et al., 2021; Hu et al., 2021; Zhao et al., 2021; Xu et al., 2022; Yang et al., 2022; Long et al., 2023). Analogously to the cell type annotation task, it has been suggested that the explicit encoding of spatial relationships between cells could help with the domain identification task.

Cheng et al. (2023) created seven semi-synthetic datasets with ground truth domain labels and compared domain identification models based on accuracy and robustness to perturbations, including image perturbations and sequencing depth variations. The compared methods include both spatial models (Pham et al., 2020; Hu et al., 2021; Zhao et al., 2021) and non-spatial models (Satija et al., 2015; Dries et al., 2021). It was found that *spatial models did not outperform non-spatial models consistently* and that spatial models were more sensitive to perturbations than non-spatial models. In the study, they argue that the investigated models encode spatial structure suboptimally, concluding that the parameterization of spatial context needs improvement.

Liu et al. (2024) benchmarked eight spatial methods based on variants of GNN models (Hu et al., 2021; Dong & Zhang, 2022; Li et al., 2022; Ren et al., 2022; Xu et al., 2022; Zong et al., 2022; Long et al., 2023; Wang et al., 2023), but did not include a non-spatial baseline. The authors concluded that model architecture and *the explicit modeling of spatial context impacted the model's performance*, emphasizing the need for careful consideration and evaluation of model design choices.

3 Experimental Design

Below, we describe the design of our benchmark study and briefly introduce the most common strategies for modeling spatial omics data with graphs.

Study Setup We compare six models: three non-spatial baselines (random assignment, k-nearest neighbors (kNN), and a fully connected neural network (FCNN)) and three spatial models (GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018), and STELLAR (Brbić et al., 2022)). For spatial models, we focus our study on graph-based representations of the spatial data and systematically explore the design space of graph neural networks. Specifically, we contrast variants of graph neural networks (GNNs); neural networks that involve a form of message passing. We compare all models on four public datasets; three transcriptomic datasets (MOSTA, ARTISTA, and STARmap) and one proteomics dataset (TonsilBE). The datasets vary w.r.t. the number of measured features, the number of cell types, the ground truth cell labeling strategy, and tissue type. For a summary of the datasets, see Table 1; more extensive descriptions of each dataset can be found in Appendix 6.1.

Graph Design To represent spatial data, we build a spatial graph in which each observation (i.e., cell) is a node. Each node has a set of associated attributes: the gene or protein expression vector. Edges represent spatial proximity (Fischer et al., 2021; Hu et al., 2021; Brbić et al., 2022). Different criteria have been used to determine connectivity. Popular ones include distance-based criteria, kNN, and Delaunay triangulation. We focus on a kNN approach as it allows us to evaluate the impact of different topologies of the input graph on the cell type annotation task. More specifically, the graph used for the node classification task is created by connecting each cell to its k nearest neighbors (w.r.t. Euclidean distance) in the spatial plane; different graph topologies are obtained by changing the value of k.

Spatial Models We focus our study on three spatial encoding strategies, which correspond to three variants of graph-based spatial models that differ in their neighbor aggregation strategies and expressive power.

- **GCN models**: are, here, models that consist of one or more sequential GCN layers – as presented in Kipf & Welling (2017) – in which each node is updated by aggregating its own and its neighbors' features. In the GCN layer, the contribution of each neighbor is fixed throughout the training but can vary between nodes.
- **GAT models**: are, here, similar to GCN models but use GAT (Graph Attention) layers rather than GCN layers (Veličković et al., 2018). GAT layers are similar to GCN layers, except that they leverage self-attention during the aggregation. In essence, the model learns how to weigh the neighbors' contributions in the aggregation.
- **STELLAR**: is a spatial cell type classification model introduced in (Brbić et al., 2022). It consists of a fully connected layer followed by a ReLU activation and a GraphSAGE (Hamilton et al., 2017) message passing layer. We consider STELLAR as one of the current SOTA methods for this task, as its authors showed how it outperformed existing baselines.

Non-spatial Models The three non-spatial baselines are invariant to edge permutations and added as a baseline to put the spatial models' performances into context.

- **Random**: In the random model, we sample cell type labels from a categorical distribution with the probabilities set to the cell type proportions in the test set.
- **kNN**: In the kNN model, for a given cell to be labeled, we identified the k nearest neighbors in the train set (in feature space) and assigned the most frequent label to the cell. For datasets with high feature dimensions, a PCA step is introduced and tuned per dataset.
- **FCNN**: In the FCNN, we used one or more fully connected (FC) layers. The input is feature expression, meaning there's no notion of spatial relationships.

Hyperparameters All models are optimized with the Adam optimizer and trained with a cross-entropy loss. We run a hyperparameter search on the learning rate, weight decay, number of hidden units, number of layers, dropout rate, model/layer normalization, and type of activation function (see Appendix Table 4). The STELLAR model architecture follows the authors' original implementation (Brbić et al., 2022) but a hyperparameter search was used to identify the best set of hyperparameters to use for each dataset. All models were implemented in PyTorch.

In light of the diverse and high-dimensional nature of spatial omics datasets, spatial models are equipped with an adaptive FC layer, aimed at refining node features into a denser representation.

Table 1: Dataset characteristics.

Dataset	Resolution	Modality	Label Source	#Features	#Cell Types
TonsilBE	Single Cell	Proteomics	Manual Annotations	44	9
STARmap	Single Cell	Transcriptomics	Single Cell Reference	1022	26
MOSTA	Single Cell	Transcriptomics	Manual Annotations	27557	17
ARTISTA	Single Cell	Transcriptomics	Manual Annotations	26540	13

Table 2: Classification performance (average F1 Score and standard deviation across classes) of different non-spatial and spatial models on different cell type labeling datasets.

Model Type	**Models**	**TonsilBE**	**STARmap**	**MOSTA**	**ARTISTA**
Non-spatial	Random	0.11 ± 0.09	0.04 ± 0.04	0.06 ± 0.05	0.07 ± 0.05
	kNN	0.84 ± 0.05	0.39 ± 0.30	0.09 ± 0.18	0.08 ± 0.12
	FCNN	$\mathbf{0.90 \pm 0.03}$	0.63 ± 0.30	$\mathbf{0.60 \pm 0.11}$	$\mathbf{0.68 \pm 0.20}$
Spatial	STELLAR	0.63 ± 0.28	0.63 ± 0.30	0.60 ± 0.28	0.67 ± 0.20
	GCN	0.90 ± 0.04	$\mathbf{0.71 \pm 0.23}$	0.60 ± 0.19	$\mathbf{0.68 \pm 0.20}$
	GAT	0.68 ± 0.12	0.59 ± 0.27	0.60 ± 0.26	0.60 ± 0.23

This layer's inclusion and its dimensionality are optimized as hyperparameters for each model-dataset pairing. For all spatial models, we also optimize the number of neighbors used when constructing the spatial kNN graph. For the purpose of classification, an FC layer is appended to each of the spatial models (`out_features=n_classes`). The only hyperparameter to be explored in the kNN classifier is the value of k (number of feature neighbors). All experiments were repeated ten times to evaluate models' sensitivity to initialization.

Robustness Evaluation In addition to contrasting models based on their classification performance, we examine the robustness of different model architectures to noisy data. Our objective is to shed light on how different architectures, particularly those with spatial encoding capabilities, withstand the adverse effects of noisy data. To systematically explore this, we introduce two specific types of controlled noise during training:

1. **Label Noise**: We simulate real-world scenarios where data annotations are prone to errors by randomly altering the labels for $20, 50, 80, 100\%$ of the instances in a dataset with K distinct categories. This alteration is conducted by reassigning labels from a uniform K-categorical distribution, thereby introducing a measurable level of uncertainty and error into the training process, akin to the noise encountered in large-scale data annotation projects.
2. **Feature Noise**: To simulate the variability and noise inherent in spatial omics data, we add multivariate Gaussian noise to $20, 50, 70, 100\%$ of the features (across all cells). This step is designed to challenge the model's ability to extract robust and discriminative features from noisy inputs, a fundamental capability for any learning system intended for real-world applications.

4 RESULTS

All of our experiments follow the same setup: a model is trained on the training set to predict node labels given the graph; the "best" model w.r.t. hyperparameter choice is taken as the top-performing model on the validation dataset; performance is then reported on the held-out test dataset (see Appendix 6.1 for more information about the datasets and the splits).

Models' performance Table 2 shows the performance results obtained on all four datasets for the best-performing model. Based on our experiments, we see *no strong evidence in favor of the hypothesis that explicitly encoding spatial relationships – using graph representations – alongside molecular expression significantly improves performance on the cell type annotation task*. In fact, the non-spatial baseline FCNN model performs on-par with the spatial models across all datasets

except STARmap, implying, at most, marginal gains by adding spatial information. In addition, spatial models are less robust to minor levels of noise in the input label or feature space, as shown in Figure 1.

Interestingly, the GCN model's superior performance on the STARmap dataset hints at a potentially beneficial alignment of spatial and molecular data under certain conditions. One important distinguishing factor of the STARmap dataset is its unique ground truth labeling strategy. STARmap cell type labels were derived from hierarchical clustering of gene expression and are linked to both cellular differentiation stages and spatial domains (e.g., olfactory versus spinal cord regions) (Wang et al., 2018); this distinguishing aspect may elucidate the observed advantage of spatial models. This suggests that the GCN model's effectiveness is particularly pronounced in scenarios where cell type definitions are inherently tied to spatial domains.

Our conclusion resonates with previous studies examining performance of spatial models on the domain identification task. We can see multiple reasons for this to be true:

1. **Ill-posed task**: Ground truth cell type annotations tend to rely heavily on marker expression and less on spatial cues, and thus induce a strong dependency between feature expression and label. As a consequence, feature expression might be fully sufficient to retrieve the cell type labels, making the spatial information superfluous.
2. **Suboptimal spatial encoding**: In line with recent studies, our approach utilizes graphs and GNNs to represent and encode spatial data. GNNs leverage homophily (also known as local smoothness), the principle that similar nodes tend to connect. However, the prevalent heterophily in cell spatial graphs can dilute distinct features, raising questions about the efficacy of this spatial encoding method (refer to Appendix 6.3). In addition to the homophily assumption, we assume that the graph structure matters, i.e., that it contains valuable information about the relationships between nodes. It's important to note that while spatial adjacency may influence cellular interactions, it might not fully encapsulate the complex environmental factors—like signaling molecules—that are crucial for determining cell function and identity.
3. **Over-smoothing**: The effectiveness of GNNs on certain tasks has been attributed to their smoothing ability. However, GNNs are prone to over-smoothing, a situation in which noise and information are mixed and nodes of different classes become too similar (Hasanzadeh et al., 2020). Over-smoothing even to a low degree could make it difficult to separate cell types that are spatially proximal and similar in feature space.

Impact of graph topology In Table 3, we present the results obtained when varying neighborhood sizes during the design of the input graph. By varying k, the neighborhood size, we can realize modest performance gains. Optimizing k is especially important for the GAT. Attention weights can theoretically help the model to focus on informative neighbors. However, if the attention mechanism struggles to appropriately downweight less relevant neighbors, a node's representation might be diluted with noisy or irrelevant information.

Sensitivity to label and feature noise Our experiments show that the FCNN is consistently more robust to label noise than any of the spatial models (see Figure 1). Introducing label noise (giving similar nodes different labels) likely exacerbates over-smoothing, which affects spatial models only, explaining the performance gap between spatial and non-spatial models.

The impact of feature noise is contingent on the dataset. For example, models trained on the TonsilBE dataset – a dataset characterized by a notably lower feature count than others (see Table 1) – shows heightened sensitivity to noise. The presence of redundant features in other datasets enables models to compensate for perturbed features by leveraging correlated, unperturbed ones, thus preserving noise resilience.

In high-dimensional input settings, the FCNN outperforms spatial models in withstanding feature noise. The inherent design of an FCNN, which processes inputs in isolation without accounting for inter-data relationships, means that any perturbation tends to have a localized effect, minimally influencing the network's overall functionality. In contrast, GNNs, through their message-passing mechanism, propagate the impact of a perturbed node across its neighbors, potentially altering their representations significantly. Notably, GAT models exhibit a pronounced vulnerability to feature noise, underscoring a critical area for future enhancements in graph-based architectures.

Table 3: Impact of different spatial neighborhood sizes on cell type labeling performance. Reported performance values correspond to the average F1 Score and standard deviation across 10 repeated runs for each spatial neighborhood size.

Models	k-neighbors	TonsilBE	STARmap	MOSTA	ARTISTA
STELLAR	3	0.63 ± 0.01	0.62 ± 0.01	0.57 ± 0.01	0.67 ± 0.01
	6	0.62 ± 0.01	0.62 ± 0.01	0.56 ± 0.01	0.66 ± 0.01
	10	0.63 ± 0.02	0.63 ± 0.01	0.57 ± 0.02	0.66 ± 0.01
	20	0.60 ± 0.02	0.62 ± 0.01	0.60 ± 0.01	0.66 ± 0.01
GCN	3	0.89 ± 0.05	0.69 ± 0.01	0.54 ± 0.01	0.68 ± 0.02
	6	0.90 ± 0.04	0.70 ± 0.02	0.58 ± 0.01	0.65 ± 0.01
	10	0.90 ± 0.04	0.71 ± 0.02	0.60 ± 0.01	0.66 ± 0.01
	20	0.90 ± 0.04	0.71 ± 0.02	0.56 ± 0.01	0.65 ± 0.01
GAT	3	0.64 ± 0.12	0.58 ± 0.02	0.60 ± 0.01	0.60 ± 0.01
	6	0.68 ± 0.12	0.58 ± 0.01	0.55 ± 0.01	0.53 ± 0.01
	10	0.68 ± 0.12	0.54 ± 0.01	0.47 ± 0.01	0.47 ± 0.02
	20	0.68 ± 0.12	0.58 ± 0.01	0.47 ± 0.01	0.45 ± 0.02

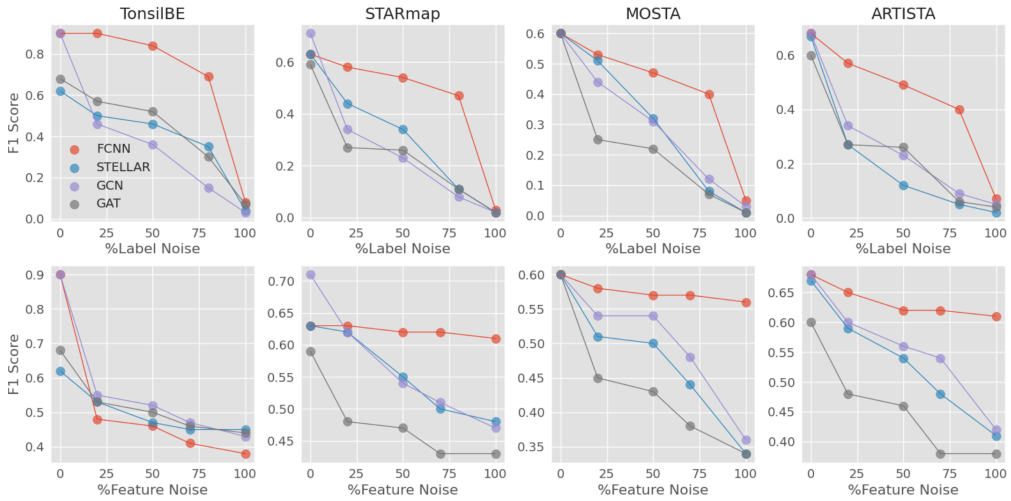

Figure 1: Evaluation of models' robustness to two sources of variation. The top panel compares spatial and non-spatial models under varied label noise, while the bottom panel examines the impact of increasing input noise. The non-spatial model is more robust to both noise types, especially in high-dimensional datasets. Among spatial models, the GAT is notably more sensitive to noise.

5 Discussion

In this study, we have evaluated common strategies for encoding spatial relationships between cells. Our results challenge the prevailing assumption that information about the spatial relationships among cells improves performance on the cell type annotation task. We also show that non-spatial models are more robust to common sources of noise in spatial data. By studying the relationship between spatial information and model performance, our work contributes to a deeper understanding of the limitations of current strategies of encoding spatial information.

Our findings underscore the importance of various factors, including data characteristics, model architecture, and optimization strategies, in determining the optimal approach for node annotation tasks. If annotation uncertainty is high, we recommend authors choose a robust model design such as the FCNN. Furthermore, we encourage authors to perform a permissive feature selection, as we find that feature redundancy improves robustness to feature noise. Lastly, we observe that optimizing only a newly introduced model, but not the baselines, is a common practice that inflates the

performance and perceived benefits of the former at the cost of robustness and reproducibility. To support fair model comparisons in future work, we recommend that authors specify the hyperparameter selection procedure and the range of hyperparameters that was considered.

REFERENCES

Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J. T. Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, 20(1), September 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1795-z. URL http://dx.doi.org/10.1186/s13059-019-1795-z.

Jose Alquicira-Hernandez, Anuja Sathe, Hanlee P Ji, Quan Nguyen, and Joseph E Powell. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, 20(1):1–17, 2019.

Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, and Oliver Stegle. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*, 29(1):202–211.e6, October 2019. ISSN 2211-1247. doi: 10.1016/j.celrep.2019.08.077. URL http://dx.doi.org/10.1016/j.celrep.2019.08.077.

Jesper Bäckdahl, Lovisa Franzén, Lucas Massier, Qian Li, Jutta Jalkanen, Hui Gao, Alma Andersson, Nayanika Bhalla, Anders Thorell, Mikael Rydén, Patrik L. Ståhl, and Niklas Mejhert. Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell Metabolism*, 33(9):1869–1882.e6, September 2021. ISSN 1550-4131. doi: 10.1016/j.cmet.2021.07.018. URL http://dx.doi.org/10.1016/j.cmet.2021.07.018.

Maria Brbić, Kaidi Cao, John W Hickey, Yuqi Tan, Michael P Snyder, Garry P Nolan, and Jure Leskovec. Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods*, 19(11):1411–1418, 2022.

Dylan M. Cable, Evan Murray, Vignesh Shanmugam, Simon Zhang, Luli S. Zou, Michael Diao, Haiqi Chen, Evan Z. Macosko, Rafael A. Irizarry, and Fei Chen. Cell type-specific inference of differential expression in spatial transcriptomics. *Nature Methods*, 19(9):1076–1087, September 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01575-3. URL http://dx.doi.org/10.1038/s41592-022-01575-3.

Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, Xin Wang, Huifang Lu, Xi Chen, Xing Liu, Xin Huang, Zhao Li, Yan Hong, Yujia Jiang, Jian Peng, Shuai Liu, Mengzhe Shen, Chuanyu Liu, Quanshui Li, Yue Yuan, Xiaoyu Wei, Huiwen Zheng, Weimin Feng, Zhifeng Wang, Yang Liu, Zhaohui Wang, Yunzhi Yang, Haitao Xiang, Lei Han, Baoming Qin, Pengcheng Guo, Guangyao Lai, Pura Muñoz-Cánoves, Patrick H. Maxwell, Jean Paul Thiery, Qing-Feng Wu, Fuxiang Zhao, Bichao Chen, Mei Li, Xi Dai, Shuai Wang, Haoyan Kuang, Junhou Hui, Liqun Wang, Ji-Feng Fei, Ou Wang, Xiaofeng Wei, Haorong Lu, Bo Wang, Shiping Liu, Ying Gu, Ming Ni, Wenwei Zhang, Feng Mu, Ye Yin, Huanming Yang, Michael Lisby, Richard J. Cornall, Jan Mulder, Mathias Uhlén, Miguel A. Esteban, Yuxiang Li, Longqi Liu, Xun Xu, and Jian Wang. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, May 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.04.003. URL http://dx.doi.org/10.1016/j.cell.2022.04.003.

Andrew Cheng, Guanyu Hu, and Wei Vivian Li. Benchmarking cell-type clustering methods for spatially resolved transcriptomics data. *Briefings in Bioinformatics*, 24(1):bbac475, 2023.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scGPT: towards building a foundation model for single-cell multi-omics using generative AI. May 2023. doi: 10.1101/2023.04.30.538439. URL http://dx.doi.org/10.1101/2023.04.30.538439.

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.

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.

David S Fischer, Anna C Schaar, and Fabian J Theis. Learning cell communication from spatial graphs of cells. *bioRxiv*, pp. 2021–07, 2021.

Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*, 174(4):968–981.e15, August 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.07.010. URL http://dx.doi.org/10.1016/j.cell.2018.07.010.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1), December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1874-1. URL http://dx.doi.org/10.1186/s13059-019-1874-1.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

Arman Hasanzadeh, Ehsan Hajiramezanali, Shahin Boluki, Mingyuan Zhou, Nick Duffield, Krishna Narayanan, and Xiaoning Qian. Bayesian graph neural networks with adaptive connection sampling. In *International Conference on Machine Learning*, pp. 4094–4104. PMLR, 2020.

Graham Heimberg, Tony Kuo, Daryle DePianto, Tobias Heigl, Nathaniel Diamant, Omar Salem, Gabriele Scalia, Tommaso Biancalani, Shannon Turley, Jason Rock, Héctor Corrada Bravo, Josh Kaminker, Jason A. Vander Heiden, and Aviv Regev. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. July 2023. doi: 10.1101/2023.07.18.549537. URL http://dx.doi.org/10.1101/2023.07.18.549537.

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*, pp. 1–23, 2023.

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.

Junbum Kim, Samir Rustam, Juan Miguel Mosquera, Scott H Randell, Renat Shaykhiev, André F Rendeiro, and Olivier Elemento. Unsupervised discovery of tissue architecture in multiplexed imaging. *Nature Methods*, 19(12):1653–1661, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Jiachen Li, Siheng Chen, Xiaoyong Pan, Ye Yuan, and Hong-Bin Shen. Cell clustering for spatial transcriptomics data with graph neural networks. *Nature Computational Science*, 2(6):399–408, 2022.

Teng Liu, Zhao-Yu Fang, Zongbo Zhang, Yongxiang Yu, Min Li, and Ming-Zhu Yin. A comprehensive overview of graph neural network-based approaches to clustering for spatial transcriptomics. *Computational and Structural Biotechnology Journal*, 23:106–128, 2024.

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.

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

Duy Pham, Xiao Tan, Jun Xu, Laura F Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J Ruitenberg, and Quan Nguyen. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*, pp. 2020–05, 2020.

Honglei Ren, Benjamin L Walker, Zixuan Cang, and Qing Nie. Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nature Communications*, 13(1):4076, 2022.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

Muhammad Shaban, Yunhao Bai, Huaying Qiu, Shulin Mao, Jason Yeung, Yao Yu Yeo, Vignesh Shanmugam, Han Chen, Bokai Zhu, Jason L Weirather, et al. MAPS: pathologist-level cell type annotation from tissue images through machine learning. *Nature Communications*, 15(1):28, 2024.

Rongbo Shen, Lin Liu, Zihan Wu, Ying Zhang, Zhiyuan Yuan, Junfu Guo, Fan Yang, Chao Zhang, Bichao Chen, Wanwan Feng, et al. Spatial-ID: a cell typing method for spatially resolved transcriptomics via transfer learning and spatial embedding. *Nature Communications*, 13(1):7640, 2022.

Hailing Shi, Yichun He, Yiming Zhou, Jiahao Huang, Kamal Maher, Brandon Wang, Zefang Tang, Shuchen Luo, Peng Tan, Morgan Wu, et al. Spatial atlas of the mouse central nervous system at molecular resolution. *Nature*, pp. 1–10, 2023.

Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2):193–200, 2020.

Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. SpatialDE: identification of spatially variable genes. *Nature Methods*, 15(5):343–346, 2018.

Xiao Tan, Andrew Su, Minh Tran, and Quan Nguyen. SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics*, 36(7):2293–2294, 2020.

Marco Varrone, Daniele Tavernari, Albert Santamaria-Martínez, Logan A Walsh, and Giovanni Ciriello. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature Genetics*, pp. 1–11, 2023.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Bo Wang, Jiawei Luo, Ying Liu, Wanwan Shi, Zehao Xiong, Cong Shen, and Yahui Long. Spatial-MGCN: a novel multi-view graph convolutional network for identifying spatial domains with attention mechanism. *Briefings in Bioinformatics*, 24(5):bbad262, 2023.

Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), July 2018. ISSN 1095-9203. doi: 10.1126/science.aat5691. URL `http://dx.doi.org/10.1126/science.aat5691`.

Xiaoyu Wei, Sulei Fu, Hanbo Li, Yang Liu, Shuai Wang, Weimin Feng, Yunzhi Yang, Xiawei Liu, Yan-Yun Zeng, Mengnan Cheng, et al. Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science*, 377(6610):eabp9444, 2022.

Chang Xu, Xiyun Jin, Songren Wei, Pingping Wang, Meng Luo, Zhaochun Xu, Wenyi Yang, Yideng Cai, Lixing Xiao, Xiaoyu Lin, et al. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Research*, 50(22):e131–e131, 2022.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1), January 2021. ISSN 1744-4292. doi: 10.15252/msb.20209620. URL `http://dx.doi.org/10.15252/msb.20209620`.

Yi Yang, Xingjie Shi, Wei Liu, Qiuzhong Zhou, Mai Chan Lau, Jeffrey Chun Tatt Lim, Lei Sun, Cedric Chuan Young Ng, Joe Yeong, and Jin Liu. SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes. *Briefings in Bioinformatics*, 23(1):bbab466, 2022.

Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job Van Der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.

Haoran Zhang, Miranda V Hunter, Jacqueline Chou, Jeffrey F Quinn, Mingyuan Zhou, Richard M White, and Wesley Tansey. BayesTME: an end-to-end method for multiscale spatial transcriptional profiling of the tissue microenvironment. *Cell Systems*, 14(7):605–619, 2023.

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.

Yongshuo Zong, Tingyang Yu, Xuesong Wang, Yixuan Wang, Zhihang Hu, and Yu Li. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. *bioRxiv*, pp. 2022–01, 2022.

6 APPENDIX

6.1 DATA

Below is a more thorough description of the four datasets used in our study:

MOSTA: A single cell resolution spatial transcriptomics dataset generated from mouse embryos – across several developmental timepoints – using the Stereo-seq platform (Chen et al., 2022). The ground truth labels were assigned by clustering of the cells followed by manual annotation using a priori known marker genes. Three sections from three mice in the same developmental stage (E15.5) were split into the train, validation, and test sets (one section each). The data was pre-processed using SCTransform (negative binomial regression) (Hafemeister & Satija, 2019); we applied no further preprocessing. For details, see the original publication (Chen et al., 2022).

ARTISTA: A single cell resolution spatial transcriptomics dataset surveying the expression of 26'540 genes in the axolotl telencephalon, generated using the Stereo-seq platform. Ground truth labels were assigned using the same process as for the MOSTA dataset, i.e., by relying on canonical marker genes. From the comprehensive set of data, which encompasses six developmental stages as well as seven regenerative stages that were activated upon injury, we selected three samples that were collected 20 days post-injury to be split into the train, validation, and test sets (one section each). The data was processed in the same way as the MOSTA dataset. For details, see Wei et al. (2022).

STARmap: A spatial transcriptomics dataset generated from the adult mouse brain and spinal cord using the STARmap technology (Wang et al., 2018). Ground truth cell annotations were assigned using label transfer, where the authors relied on a reference scRNA-seq dataset presented in Zeisel et al. (2018). scRNA-seq annotations were derived from hierarchical clustering of gene expression profiles. We chose three brain sections from the same anatomical plane (sagittal) to be split into our train, validation, and test sets (one section each). This data had also already been pre-processed upon download; pre-processing steps were as follows: normalization such that the sum of feature values for each cell was equal to the median cell-wise total transcript count, log1p-transformation, regressing out the total counts, and finally scaling to unit variance and zero mean for each feature. See Shi et al. (2023) for more details.

TonsilBE: A single cell resolution spatial proteomics dataset generated from healthy human tonsil tissue and diseased tissue from a patient with Barrett's esophagus (BE) using the CODEX technology (Goltsev et al., 2018). Ground truth cell type annotations were assigned manually after clustering the data and assessing the expression of known cell type markers in each cluster. We chose two healthy sections to be split into the train and validation set; a diseased tissue section (BE) constituted our test set. The data had already been pre-processed (with z-score normalization) upon download. For more details, see Brbić et al. (2022).

6.2 DATA AVAILABILITY

Both the MOSTA and the ARTISTA dataset are available in the Spatial Transcript Omics DataBase under https://db.cngb.org/stomics/mosta/ and https://db.cngb.org/stomics/artista/, respectively. The STARmap PLUS data can be downloaded from the Single Cell Portal. Lastly, the TonsilBE data can be downloaded from Dryad.

6.3 SPATIAL DISTRIBUTION OF CELL TYPES

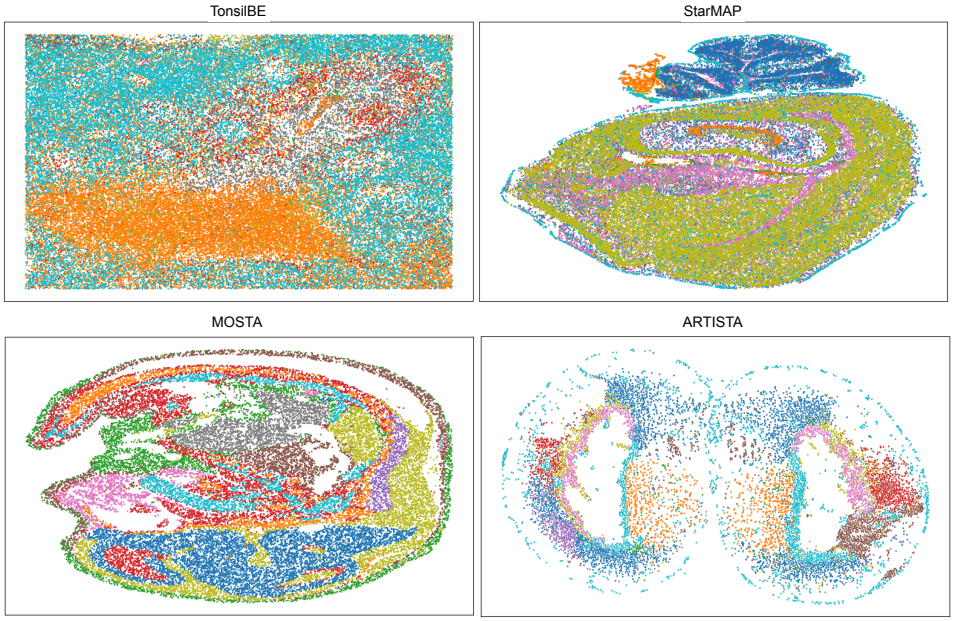

Figure 2: Visualization of the spatial distribution of cell types in the different datasets used in this study. Colors represent different cell types and points correspond to different cells.

6.4 HYPERPARAMETER SPACE

Table 4: Range of hyperparameter values explored during model training.

Hyperparameter	Values Explored
Learning rate	$[1e{-}4, 1e{-}3, 3e{-}3, 2e{-}3, 1e{-}2, 1e{-}1]$
Weight decay	$[0.0, 5e{-}3]$
Dropout percentage	[0.0, 0.5, 0.8]
Number of layers	[1, 3, 5, 8]
Number of hidden units	[64, 128, 256, 1024]
Number of spatial neighbors	[3, 6, 10, 20]