Diffusion-based Representation Integration for Foundation Models Improves Spatial Transcriptomics Analysis

Atishay Jain[1] Tuan M. Pham[2] David H. Laidlaw [1] Ying Ma[2,3]
Ritambhara Singh[1,2*]
[1]Department of Computer Science, Brown University
[2]Center for Computational Molecular Biology, Brown University
[3]Department of Biostatistics, Brown University

Abstract

In this work, we propose a framework, DRIFT, that integrates spatial information into pre-trained single-cell foundation models by leveraging diffusion on spatial graphs derived from spatial transcriptomics (ST) data. ST captures gene expression profiles while preserving spatial context, enabling downstream analysis tasks such as cell-type annotation, clustering, and cross-sample alignment. However, due to its emerging nature, there are very few foundation models that can utilize ST data to generate embeddings generalizable across multiple tasks. Meanwhile, well-documented foundational models trained on large-scale single-cell gene expression (scRNA-seq) data have demonstrated generalizable performance across scRNA-seq assays, tissues, and tasks; however, they do not incorporate the spatial information in ST data. We use heat kernel diffusion to propagate embeddings across spatial neighborhoods, incorporating local tissue structure into ST data while preserving the transcriptomic representations learned by state-of-the-art single-cell foundation models. We systematically benchmark five foundational models (both scRNA-seq and ST-based) across key ST tasks such as annotation, alignment, and clustering, ensuring a comprehensive evaluation of our proposed framework. Our results show that spatial diffusion significantly improves the performance of existing single-cell foundational models on ST data over specialized state-of-the-art methods. Overall, DRIFT is an effective, accessible, and generalizable framework that bridges the gap toward universal models for modeling spatial transcriptomics.

*Corresponding Author. Email: ritambhara@brown.edu

1 Introduction

Spatial transcriptomics (ST) technologies enable measurement of gene expression while preserving the native tissue architecture, offering a direct view of how cellular identity and spatial organization together shape biological function. By jointly capturing transcriptomic and spatial information, ST datasets reveal the local microenvironment of gene activity and support downstream analyses such as spatially variable gene detection [1, 2, 3, 4, 5], clustering [6, 7, 8, 9, 10], cell–cell interaction inference [11, 12, 13, 14], and cross-section alignment [15, 16]. These capabilities have made ST an essential approach in developmental biology, pathology, and tissue-level systems analysis. However, the rapidly growing complexity and heterogeneity of ST data necessitate generalizing tasks such as automatic cell-type annotation and clustering across a variety of tissues, species, and technologies. Furthermore, ST measurements are typically sparse, noisy, and limited in throughput compared to single-cell RNA sequencing (scRNA-seq). This further highlights the need for robust computational models that can extract biologically meaningful and spatially aware representations that generalize to diverse downstream spatial analyses.

Recently, to address these challenges, ST foundation models have emerged, marking a growing effort to bring foundation modeling into ST. Early examples include Loki [17], Nicheformer [18], SToFM [19], and STFormer [20]. These pioneering studies demonstrate efforts of incorporating ST into model architectures: Loki aligns histology images and spatial gene-expression profiles through cross-modal contrastive learning; Nicheformer applies a transformer trained on ST data, where attention mechanisms capture neighborhood-level dependencies, and SToFM and STFormer introduce spatial attention and positional encoding schemes to jointly model biological and spatial information. However, they rely heavily on large paired multimodal datasets [17] or cell-type priors derived from deconvolution [20], which are often unavailable for many spatial platforms. Moreover, some models do not explicitly model spatial relationships, limiting their ability to capture structural interactions [17, 18]. Additionally, training these architectures from scratch is computationally expensive, often requiring substantial GPU memory and computing resources. Finally, because existing ST foundation models are trained on relatively homogeneous ST datasets, their representations are not broadly generalizable. As a result, these models typically require dataset-specific retraining, limiting their utility as universal foundation models for ST.

On the other hand, scRNA-seq foundation models are trained on millions of scRNA-seq profiles from diverse tissues and species. Models such as Geneformer [21, 22], scGPT [23], and scFoundation [24] learn gene–gene dependencies using transformer-based or masked language-model architectures. The embeddings generated by these models capture pathway-level relationships and regulatory patterns, providing general-purpose representations that can transfer effectively across datasets and biological contexts. Their success demonstrates that large pretrained models can generalize across data sources and experimental conditions with limited fine-tuning, offering a promising foundation for diverse multi-tissue and cross-species analyses [25]. However, since these models are trained solely on scRNA-seq data, they lack spatial context and may overlook the structural organization, which is critical for ST data.

We introduce **DRIFT** (**D**iffusion-based **R**epresentation **I**ntegration for **F**oundation models in spatial **T**ranscriptomics). This scalable diffusion framework denoises expression profiles and integrates the spatial topology of ST data into existing pretrained scRNA-seq and ST foundation models without additional retraining. Foundation models that do not explicitly model spatial information benefit from both denoising and spatial integration, while methods that do so leverage DRIFT's denoised output. DRIFT constructs a spatial adjacency graph among tissue spots and applies a heat-kernel diffusion process that propagates gene-expression signals across local neighborhoods while preserving tissue boundaries. This produces spatially coherent yet biologically meaningful representations that can be directly embedded into pretrained foundation models without retraining, making our approach much more computationally scalable and accessible.

We demonstrate that DRIFT generalizes easily across three critical ST tasks — cell-type annotation, cross-section alignment, and clustering. The singular DRIFT framework outperforms a variety of specialized state-of-the-art methods in all these ST downstream analysis tasks, thus setting a new norm for ST downstream analysis. Furthermore, the DRIFT diffusion leverages existing pretrained scRNA-seq foundation models, achieving improved performance compared to baseline ST foundation models. Therefore, our results establish that the diffusion operator at the core of DRIFT provides a simple, efficient, and general way to couple pretrained scRNA-seq foundation models with ST data by incorporating the spatial topology. For ST foundation models that may embed spatial information, diffusion-based denoising still improves performance, especially for cross-section alignment.

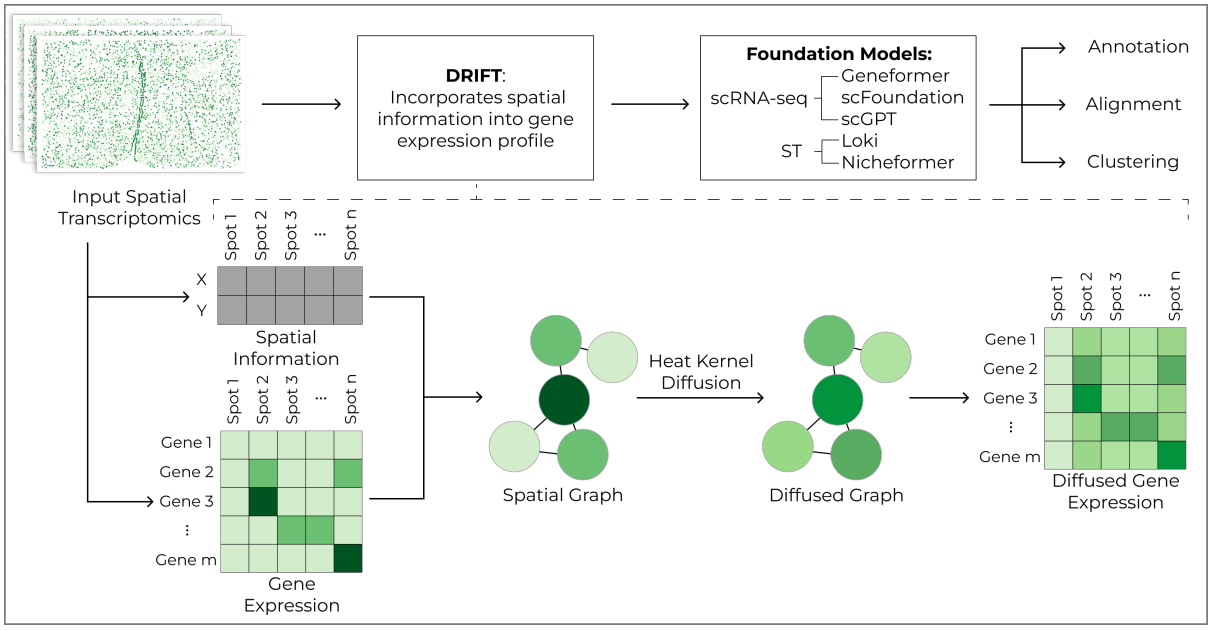

Figure 1: **An illustration of the DRIFT Framework.** Gene expression profiles of individual spots do not include any spatial information. Therefore, we create a spatial graph in which each node represents a spot and each edge represents the spatial proximity between the corresponding spots. We then apply heat kernel graph diffusion on the constructed graph and use the output diffused gene expression profiles for each spot. Through diffusion, each node incorporates information about its neighboring spots. The diffused gene expression values are fed into foundation models instead of the original gene expression for downstream tasks.

2 Method

As illustrated in Figure 1, we introduce a graph diffusion-based framework, DRIFT, to incorporate spatial information into the gene expression profiles. To perform this integration, we construct a spatial graph and apply heat kernel diffusion over it. The diffused gene expression matrices obtained from DRIFT are used as inputs to pretrained foundation models, which are then evaluated for multiple important ST analysis tasks over a variety of ST datasets.

2.1 DRIFT Approach

First, we represent the ST data as a graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a cell or spot, and edges $(v_i, v_j) \in E$ connect spatially proximal spots based on Euclidean distance. Following GraphST [8], we set the three nearest spots in the ST data as each node's neighbors. The edges of this graph are represented by the adjacency matrix $A \in \mathbb{R}^{n \times n}$, where $A_{ij} = 1$ if nodes i and j are spatial neighbors and $A_{ij} = 0$ otherwise.

We then calculate the degree matrix D, which is a diagonal matrix defined as

$$D_{ii} = \sum_j A_{ij}, \tag{1}$$

representing the degree of each node i. Using these, we compute the graph Laplacian as

$$L = D - A. \tag{2}$$

The Laplacian encodes the structure of G and serves as a discrete analogue of the continuous Laplace operator, capturing how information flows between connected nodes.

Heat kernel diffusion models the propagation of information (or "heat") across a graph. It is defined by the heat equation on graphs:

$$\frac{dx_t}{dt} = -Lx_t, \tag{3}$$

Solving this differential equation yields the heat kernel solution:

$$x(t) = e^{-tL} x_0, \tag{4}$$

where, x_0 represents the original (non-diffused) gene expression matrix and x_t denotes the diffused gene expression matrix at time t. t is therefore a hyperparameter that controls the strength of diffusion. Small t values (representing a shorter time for heat diffusion) lead to weaker diffusion and preserve local structure. Meanwhile, large t values result in stronger smoothing across larger spatial distances.

2.1.1 Relation to Spectral Graph Theory

Since G is an undirected graph, its Laplacian L is a symmetric real matrix and is therefore diagonalizable. By the graph spectral theorem, it can be decomposed as

$$L = U \Lambda U^{-1} = U \Lambda U^{\top}, \tag{5}$$

where, U is the orthogonal matrix of eigenvectors and Λ is the diagonal matrix of the corresponding eigenvalues. Substituting equation 5 into the heat kernel (equation 4) gives:

$$e^{-tL} = U e^{-t\Lambda} U^{\top}. \tag{6}$$

This expression reveals that diffusion corresponds to exponentially decaying the eigenvalues of the Laplacian, which is equivalent to a soft low-pass filter on the graph spectrum. Effectively, it smoothens the gene expression based on the graph structure [26]. This idea has been previously explored for scRNA-seq [27] and 3D chromatin organization data [28, 29], further demonstrating that graph diffusion techniques can effectively denoise data, smooth signals across neighboring nodes, and capture spatial or structural relationships. Consequently, the heat kernel can integrate local spatial information while preserving biologically meaningful large-scale spatial structures. The resulting diffused gene expression matrix $x(t)$ thus integrates the spatial context of cells directly with the gene expression matrix, making it compatible with existing scRNA-seq foundation models that only accept gene expression values as inputs. Therefore, our choice of heat kernel diffusion is appropriate for our ST data modeling task. To further justify our choice of the heat kernel for graph diffusion, we show that it outperforms other diffusion methods overall. The other diffusion methods are described, and their results are reported in Supplementary Section S1.

Next, we input the diffused gene expression matrix into the foundation models.

2.2 Foundation Models

We apply DRIFT to a variety of scRNA-seq and ST foundation models to illustrate its ability to leverage the performance of existing models. We first demonstrate that DRIFT incorporates spatial information by pairing it with scRNA-seq foundation models. We further show its denoising property by enhancing the performance of ST foundation models.

Geneformer - Geneformer is a scRNA-seq foundation model that has three models. V1 is trained on 30 million cells [21], while V2 is trained on 104 million cells [22]. Furthermore, V2 has two model architecture configurations: one with 104 million parameters and another with 316 million parameters. In our Supplementary Section S2, we demonstrate that V2, with 316 million parameters, consistently outperforms other Geneformer configurations. Therefore, for any further experiments, we select V2 with 316 million parameters as the Geneformer foundation model.

scFoundation - scFoundation is a large-scale foundation model pretrained over 50 million single-cell transcriptomic profiles with an underlying encoder–decoder architecture.

scGPT - scGPT is a scRNA-seq foundation model trained on over 33 million cells with the aim of simultaneously learning cell and gene embeddings. It utilizes a specially designed attention mask and generative training pipeline to jointly optimize both embeddings.

Loki - Loki is a visual-omics foundation model that learns joint representations between histology images and transcriptomic profiles using contrastive learning based on OpenClip [30]. It is trained on 2.2 million paired image-transcriptomic patches. The framework does not explicitly train spatial information.

Nicheformer - Nicheformer is a transformer-based foundation model trained on 110 million single-cells and ST spots. By training both on scRNA-seq and ST, Nicheformer learns the characteristics of different types of datasets. Similar to Loki, Nicheformer does not explicitly use spatial information for training.

The embeddings from these foundation models are then used in three important downstream ST tasks, as described below.

2.3 Spatial Transcriptomic Tasks and Specialized Methods

To demonstrate the advantage of DRIFT, we measure its performance improvement over baseline foundation models across multiple ST tasks. These tasks are crucial for analyzing ST data and also facilitate cell-atlas construction, understanding disease mechanisms, and biomedical discovery [31, 32]. To further evaluate its performance advantage, we also compare the DRIFT-enhanced foundation models to well-documented state-of-the-art specialized methods for each task.

Cell-type annotation - The aim of cell-type annotation is to identify cell types in a supervised manner, given a reference dataset. This is critical for understanding the heterogeneity of tissues and related biological processes [33]. Annotation is formulated as a few-shot learning task, where we attach a classification head (or probe) to a pretrained foundation model, and only train the weights of the probe to identify the cell types.

The state-of-the-art method for annotation is STELLAR [34] - a deep learning method that generates spatial graphs to represent ST data. It uses a graph neural network (GNN) to learn embeddings of spots in an annotated reference ST dataset. Then it annotates cells in the unannotated ST dataset using the embeddings it generates. We assess the annotation performance through the accuracy and F1-score metrics.

Slice alignment - ST slice alignment is essential for reconstructing coherent anatomical structures from adjacent ST sections, as these slices represent slightly different structures and are subject to technical rotation, stretching, or deformation. Evaluating alignment accuracy, therefore, provides a rigorous test of how well DRIFT-enhanced embeddings preserve spatial geometry and enable meaningful cross-slice correspondence. We treat this as a zero-shot setting, applying Coherent Point Drift (CPD) directly to the original and DRIFT-enhanced embeddings, without any task-specific fine-tuning, across three alignment contexts with varying levels of noise and distortion. Baseline task-specific methods are PASTE [15] and PASTE2 [16]. PASTE uses Fused Gromov–Wasserstein (FGW) alignment, returning probabilistic alignment, while PASTE2 extends this to partial FGW alignment to handle partial overlaps.

We evaluate alignment performance using DRIFT-enhanced embeddings across both simulated and real-world datasets. In the simulation experiment, we simulate each cell/spot's total count using a Negative Binomial fit to the real library-size mean and variance, then generate per-gene counts by drawing from a multinomial whose probabilities are the real gene proportions with a pseudocount δ added.

For simulated data, we evaluate alignment quality using translation error together with PCC and Kendall's τ, since ground-truth spatial coordinates are available. For real-world datasets, where true spatial coordinates are unknown, we instead assess alignment through LTARI, PCC, and Kendall's τ. We average these metrics into a single alignment score reflecting both global and local alignment quality.

Translation Error - The translation error, defined as $\|t_{pred} - t_{true}\|_2$ measures the Euclidean discrepancy between the predicted translation vector and the ground-truth shift required for correct spatial alignment. A lower translation error indicates a more accurate alignment compared to the ground truth. Translation error is computed only when ground-truth spatial coordinates are available (simulated data).

Pearson Correlation - For each aligned pair, we compute the Pearson correlation coefficient (PCC) as follows:

$$PCC = \frac{\sum_{i=1}^{G}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{G}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{G}(y_i - \bar{y})^2}}.$$

PCC captures the global expression pattern and linear correspondence between aligned slices. We report the median PCC across all aligned pairs, with higher values indicating stronger transcriptomic consistency between matched cells.

Kendall's $\boldsymbol{\tau}$ - We quantify rank-based concordance between aligned cells using Kendall's τ: $\tau = \frac{2(C-D)}{G(G-1)}$, where C and D are counts of concordant and discordant gene pairs, and a higher τ indicates

Dataset	ST Technology	Species	Tissue	Number of Slices
10xHPC [41]	10x Visium	Human	Prefrontal Cortex	12
10xHOC [42]	10x Visium	Human	Ovarian Cancer	2
10xHSI[43]	10x Visium	Human	Small Intestine	12
MERHH [44]	MERFISH	Human	Heart	3
MERMPH [45, 46]	MERFISH	Mouse	Preoptic Hypothalamus	5
MERMB [47]	MERFISH	Mouse	Brain	15
MERMBA [48]	MERFISH	Mouse	Aging Brain	12
StereoME [49]	Stereo-Seq	Mouse	Embryo	4

Table 1: **A summary of all the datasets**. The table lists the ST technology, species, tissue, and number of slices in all the datasets.

stronger rank similarity. We report the median τ, with higher values suggesting biologically coherent correspondence between matched cells.

Label Transfer Adjusted Rand Index (LTARI) - We compute LTARI as the adjusted Rand index between true labels $\mathbf{l}$ and alignment-transferred labels $\hat{\mathbf{l}}$ when labels are available. The mathematical formulation is LTARI = ARI($\mathbf{l}$, $\hat{\mathbf{l}}$). A higher LTARI score indicates a more coherent local structure and cell-type organization across alignments.

Alignment Score We compute an overall aggregate alignment score by averaging normalized PCC, Kendall's τ, and LTARI values. By integrating both global and local metrics into a single score, the alignment score provides a quantitative analysis of overall alignment performance.

Clustering - Unsupervised clustering is the task of grouping cells without any prior information, which is an essential step for identifying marker genes [35, 36], and understanding disease pathology [37, 38]. Since it does not explicitly require any labels for training, it is a zero-shot task in which embeddings obtained from the foundation models are directly fed into unsupervised clustering algorithms, such as mclust [39] (a finite normal mixture modeling-based clustering algorithm), to obtain the final clusters. It can also be formulated as a few-shot task in which the foundation model encoder is fine-tuned to reconstruct the input gene expression matrix from the foundation model's embeddings.

To test the clustering performance of DRIFT on this task, we employ the state-of-the-art method - GraphST [8]. GraphST employs a GNN autoencoder architecture and contrastive learning techniques to learn embeddings that are subsequently utilized in unsupervised algorithms. We evaluate the methods using the adjusted Rand index (ARI). ARI is a metric commonly used to assess unsupervised clustering performance [8, 6, 40] because it accounts for the risk of accidental cluster agreement.

2.4 Datasets

To demonstrate DRIFT's generalizability, we benchmark it across eight datasets from various ST technologies and tissues of humans and mice. The ST technology, species, tissue, and number of slices for all the datasets are summarized in Table 1.

2.5 Hyperparameter Tuning

For every task, we tuned the t value (representing the strength of diffusion) in the heat kernel diffusion equation. The values assigned to t during hyperparameter tuning were 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 2, 5, 10, 100. We observed that t values from 1 to 10 performed similarly well overall. For alignment, we recommend selecting the t that yields the highest PCC (forfeiting the need for labels). However, for annotation and clustering, having a ground truth to tune t may be unfeasible. In such case, we suggest a default value of $t = 2$ due to its high performance across tasks, datasets, and foundation models. Results for the hyperparameter tuning of t are summarized in Supplementary Section S3.

Any further hyperparameter tuning necessary for both foundation models and specialized methods during few-shot learning tasks was performed using Optuna [50] with 10 trials. We also tuned PASTE and PASTE2 for the zero-shot alignment task. The hyperparameters and their value ranges are listed in Supplementary Sections S4 and S5.

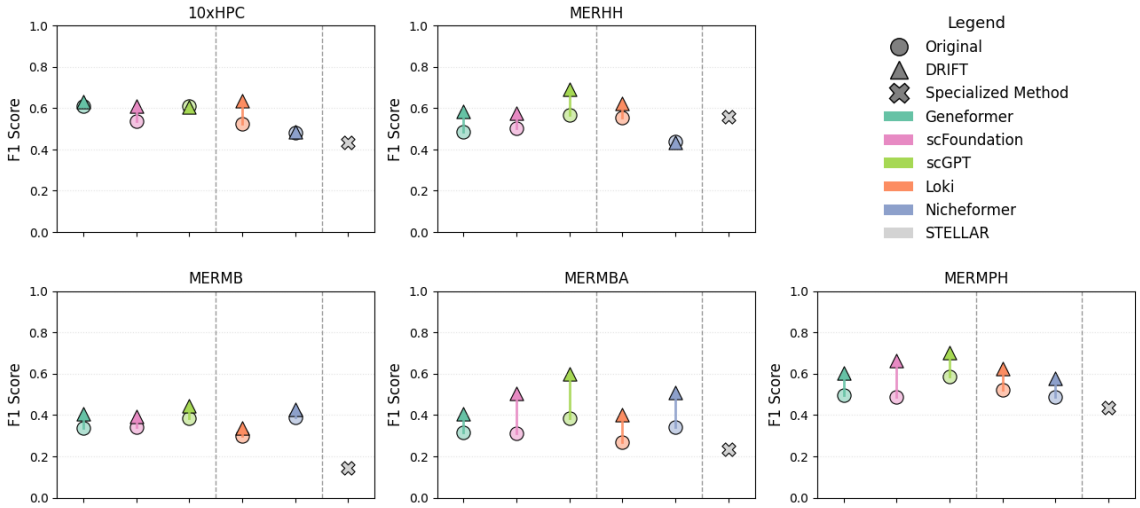

Figure 2: **Cell-type Annotation Results.** A quantitative comparison across multiple datasets. The circles represent the performance of the original foundation models, the triangles represent the DRIFT-enhanced foundation models, and the X represents the specialized method STELLAR. The colors indicate which column corresponds to each foundation model (the order of models in the legend matches column placement). This figure shows that DRIFT improves the F1 score of foundation models on the annotation task. Furthermore, DRIFT-enchanced foundation models, specifically scGPT show state-of-the-art performance, surpassing even the specialized model - STELLAR.

3 Results

3.1 DRIFT demonstrates state-of-the-art cell-type annotation performance

We first tested DRIFT's performance on the cell-type annotation task, formulated as a few-shot learning task. For all foundation models, we attached a three-layer multilayer perceptron (MLP) classification head to predict cell-type labels, while freezing all encoder weights within the foundation model. The MLP was trained using the cross-entropy loss.

In this experiment, we employed datasets 10xHPC (10x human prefrontal cortex), MERHH (MERFISH human heart), MERMPH (MERFISH mouse preoptic hypothalamus), MERMB (MERFISH mouse brain), and MERMBA (MERFISH mouse brain) since they contain expert annotations necessary for evaluating the methods. The five foundation models and one specialized method (STELLAR) were trained and evaluated independently on each dataset. To recreate the realistic scenario of limited annotated data, we used only one annotated slice per dataset to train and hyperparameter-tune the methods, reserving the remaining slices for testing.

Figure 2 displays DRIFT's F1-score performance improvement across all methods and datasets, demonstrating that the heat kernel's denoising and spatial integration properties are beneficial for the annotation task. We first compared the original scRNA-seq foundation models to the corresponding DRIFT-enhanced foundation models. Compared to the original foundation models, DRIFT shows average improvements of 9.32% to 23.96% in accuracy and 18.90% to 28.20% in F1-score. Dataset-specific accuracy results also show the same performance trends and are plotted in Supplementary Section S6.

In particular, scGPT with DRIFT shows the highest performance over all datasets. Furthermore, it consistently outperforms the specialized method, STELLAR. scGPT with DRIFT shows, on average, a 52.52% improvement in accuracy, and a 97.31% improvement in F1-score. scGPT with DRIFT also outperforms the original Loki and Nicheformer models. When compared to Loki, DRIFT-enhanced scGPT shows an average accuracy improvement of 39.43% and an average F1-score improvement of 49.20%. Similarly, when compared to Nicheformer, it shows an average accuracy improvement of 27.45% and an average F1-score improvement of 40.79%. These results demonstrate its state-of-the-art performance for ST annotation.

DRIFT also shows a reasonable performance improvement for both the ST foundation models. When

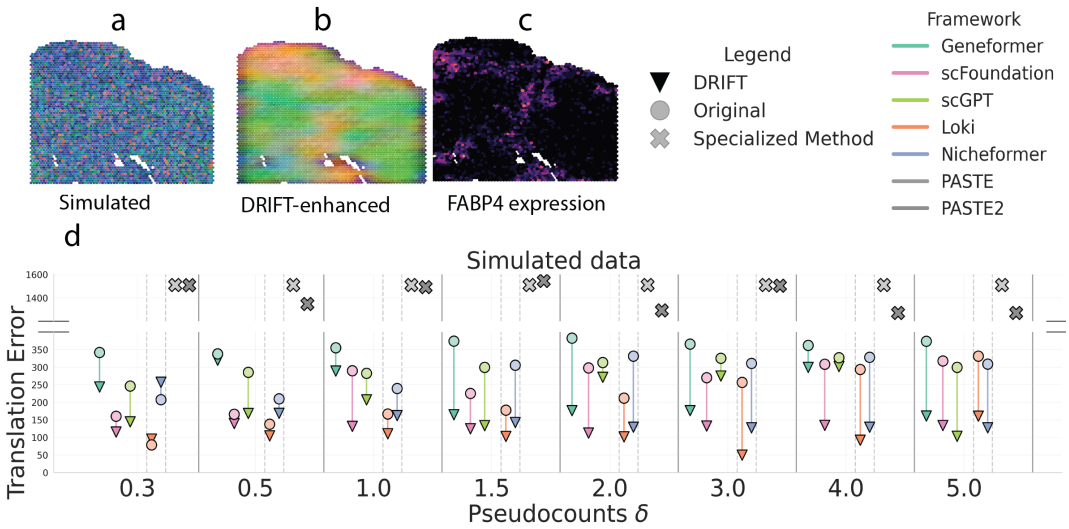

Figure 3: **Simulated Data Alignment Results.** a–c. PCA visualizations of the original and DRIFT-enhanced embeddings for simulated data at pseudocount 4.0, along with the corresponding FABP4 expression patterns. DRIFT produces embeddings with clearer and more coherent spatial organization, recovering biologically meaningful structure associated with ovarian cancer development. d. Translation error across simulated datasets at different values of pseudocounts, demonstrating improved alignment fidelity after applying DRIFT.

compared to the original Loki model, DRIFT-enhanced Loki shows an average accuracy improvement of 19.07% and an average F1-score improvement of 22.89%. Similarly, for Nicheformer, DRIFT shows an average accuracy improvement of 12.03% and an average F1-score improvement of 15.43%. These results illustrate the effectiveness of DRIFT in denoising and enhancing the ST foundation models.

3.2 DRIFT improves ST alignment across different contexts

For alignment tasks, we set up three different experiments and applied the Coherent Point Drift (CPD) algorithm [51] to align embeddings from five DRIFT-enhanced foundation models and their original models. We then evaluated the results using two specialized methods, PASTE and PASTE2, on one simulated dataset and two real datasets, without additional fine-tuning (zero-shot setting). After alignment, we compare paired cells and their respective original transcriptomic profiles on different metrics described in the Methods section.

Based on previous studies [16, 40], we simulated seven paired ST slices from manually aligned 10xHSI (Visium human ovarian cancer) pair of slices with varying pseudocount levels (δ: 0.3 to 5) to introduce different degrees of expression noise. We compared DRIFT and baseline alignments using translation error (lower is better), Pearson Correlation, and Kendall's τ (higher is better).

Across all pseudocount settings, DRIFT consistently reduces translation error across all scRNA-seq foundation models. Geneformer, scFoundation, and scGPT show the most substantial gains, with typical reductions ranging from 17% to 60% across all δ values. ST foundation models also benefit from DRIFT where Nicheformer and Loki show reliable improvements at most pseudocounts, with reductions of 19% to 80% for $\delta \geq 0.5$.

This positive improvement is further supported by systematically higher PCC values (0.5% to 6% gains) and higher Kendall's τ (0.2% to 3.5%). PASTE and PASTE2, on the other hand, have higher translation error and lower PCC and Kendall's τ (Figure 3 and Supplementary Section S7).

Next, since scFoundation saw the greatest percentage reduction in translation error, we performed principal component analysis (PCA) on the original and DRIFT-enhanced embeddings of scFoundation at pseudocount = 2.0 for illustration. While the original embeddings appeared noisy, DRIFT embeddings showed improved spatial organization consistent with important marker gene patterns in ovarian cancer (see Figure 3a-c). Altogether, these results confirm that DRIFT not only reduces spatial reconstruction error but also improves the underlying gene structure.

We then applied DRIFT to reconstruct three-dimensional tissue organization using the StereoME (STEREO-seq) mouse embryo dataset, which provides spatial maps across early organogenesis. We aligned four consecutive E9.5 sections (Slice 1→2, 2→3, and 3→4) and reconstructed the resulting embeddings in 3D. Because ground truth is unavailable in this setting, we evaluated reconstruction quality through visualization and the alignment score. Visually, only the original scFoundation embeddings produced reasonably coherent alignments, whereas the other foundation models frequently yielded inconsistent mappings–such as inversions or rotations–highlighting the difficulty of this alignment task (Figure 4). Incorporating DRIFT noticeably improved performance across nearly all models, producing more stable and biologically consistent reconstructions.

Quantitatively, the scRNA-seq foundation models exhibited strong gains under DRIFT, with Geneformer and scGPT showing high average improvements (2.2-fold and 6-fold, respectively), reflecting DRIFT's ability to correct poor baseline alignments (Supplementary Section 7). scFoundation, which already produces relatively strong alignments, also showed an improvement in performance ($\approx 33\%$), indicating that DRIFT can correct high-quality embeddings. The ST foundation models, especially Nicheformer, also benefited, particularly in Slice 1→2 and Slice 3→4, with an average gain of 7.6-fold by correcting weak baseline performance while Loki's performance increased by over 70% on average. As shown in Supplementary Section 7, some embeddings achieve higher alignment scores than PASTE across slices and remain competitive to or outperform PASTE2 only when combined with DRIFT, which is required for these models to reach their strongest alignment performance.

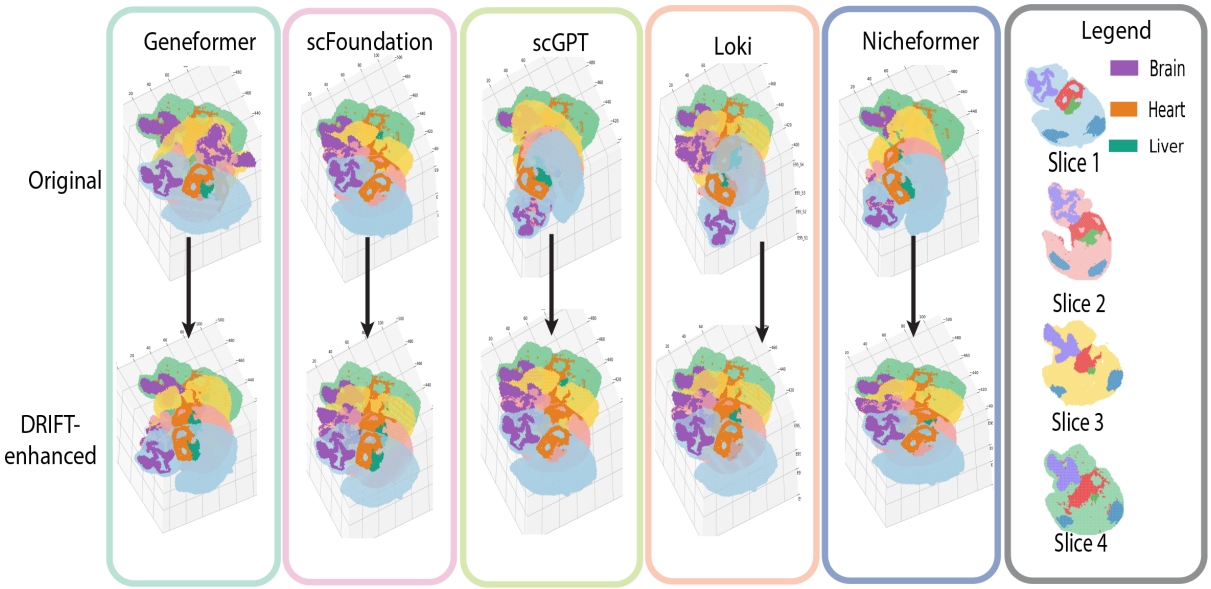

Figure 4: **Mouse Embryo Alignment Results.** First row: alignment using original embeddings. Second row: alignment using DRIFT-enhanced embeddings. DRIFT produces more stable and coherent embryonic alignments, yielding higher-fidelity 3D reconstructions.

Finally, to further validate DRIFT's alignment capability in real-world contexts, we aligned eleven manually annotated adjacent human small intestine sections onto a target section (10xHSI) from the same specimen. Across all intestine alignment experiments, DRIFT produced consistent improvements over the baseline embeddings for every foundation model evaluated. The magnitude of the gain is different for each framework: scFoundation showed the strongest and most stable increase, with several experiments exhibiting significant boosts (up to 33%). Geneformer and scGPT also demonstrated reliable improvements (7% to 17%), suggesting that DRIFT effectively enhances context-aware transcriptomic embeddings (Figure 5). ST foundation models showed similar trends, with Nicheformer benefiting from substantial improvements up to 22% while Loki exhibited smaller but still positive gains across most experiments (between 1 and 5%). DRIFT-enhanced foundation models performed competitively in eight of the eleven experiments – outperforming PASTE and PASTE2 in five cases and matching their performance in an additional three, indicating that diffusion refinement can match alignment tools even under challenging structural variation.

	scRNA-seq foundation models						ST foundation models			
	Geneformer		scFoundation		scGPT		Loki		Nicheformer	
	Original	DRIFT	Original	DRIFT	Original	DRIFT	Original	DRIFT	Original	DRIFT
10xHPC	0.0036	0.0116	0.3079	0.2979	0.3118	0.3525	0.3457	**0.3743**	0.2639	0.0902
MERHH	0.0097	0.0202	0.2234	0.0521	0.3577	0.4561	**0.4611**	0.4055	0.0658	0.0473
MERMPH	0.0234	0.0378	0.1179	0.1374	0.2073	0.2587	0.3059	**0.3276**	0.1039	0.0953
MERMB	0.0090	0.0155	0.3075	0.1444	0.4303	**0.4776**	0.4089	0.4066	0.4447	0.1571
MERMBA	0.3435	**0.4173**	0.3224	0.3540	0.1848	0.2236	0.2529	0.4101	0.3082	0.2217

Table 2: **Clustering Results.** The table lists the mean zero-shot ARI performance of every foundation model and their DRIFT-enhanced versions on each dataset. These results show that the best performing foundation models overall are DRIFT-enhanced scGPT and DRIFT-enhanced Loki.

Altogether, DRIFT denoises spatial embeddings and reconstructs meaningful biological structure, yielding more accurate and coherent alignments.

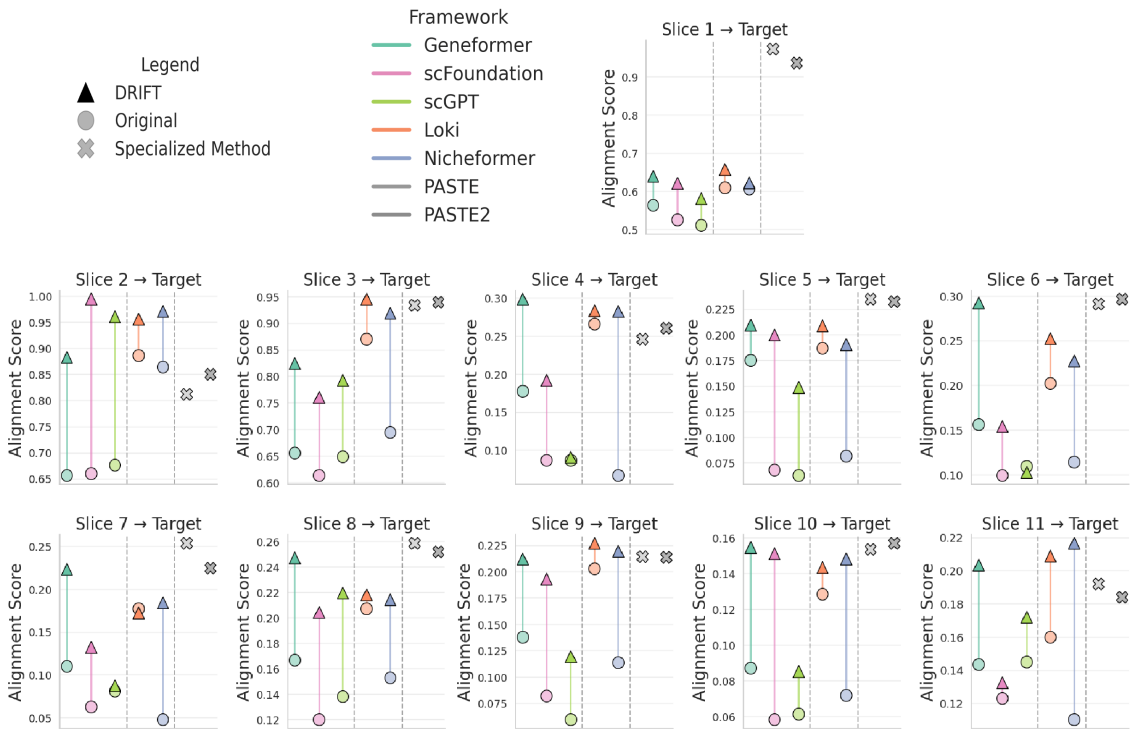

Figure 5: **Human Small Intestine Alignment Results.** Foundation models substantially improve alignment quality across human small intestine sections and become competitive with PASTE/PASTE2 when only paired with DRIFT.

3.3 DRIFT enhances clustering performance

Finally, we evaluated DRIFT's performance on the unsupervised clustering task. Since it is an unsupervised task, we formulated it as a zero-shot problem. We first used the pretrained weights of the foundation models to obtain spot embeddings and subsequently clustered them using the mclust algorithm [39]. We tested the methods on the five datasets used in the annotation task, using expert annotations as ground-truth clusters. The five datasets were 10xHPC (10x human prefrontal cortex), MERHH (MERFISH human heart), MERMPH (MERFISH mouse preoptic hypothalamus), MERMB (MERFISH mouse brain), and MERMBA (MERFISH mouse brain). We clustered each slice independently and quantified the performance of the five foundation models with ARI.

First, we tested the performance improvement of DRIFT for scRNA-seq foundation models. As displayed in Table 2, DRIFT improves the zero-shot performance of the majority of scRNA-seq foundation models. In particular, scGPT with diffused gene expression shows the highest overall performance with a mean ARI of 0.3564 across all datasets. These results further demonstrate that the diffused inputs generated by DRIFT better capture biologically meaningful structure, enabling foundation models to access richer spatial–molecular information even without supervised fine-tuning.

We also tested DRIFT on the ST foundation models, which showed mixed results. For Loki, we observe an overall improvement in performance, whereas for Nicheformer, we observe the opposite. However, Loki with DRIFT shows the highest performance among all ST foundation model configurations, with a mean ARI of 0.3915, further indicating DRIFT's usefulness with ST foundation models.

We then extended this analysis by formulating the clustering task as a few-shot learning problem. This enabled us to compare DRIFT with the specialized method for unsupervised clustering (GraphST) since it does not support zero-shot tasks.

Given that scGPT and Loki performed the best in the zero-shot scenario, we aimed to focus on them for the few-shot evaluation. However, since Loki requires paired images for encoder fine-tuning, we restricted our analysis to scGPT. Following the same experimental design as the cell annotation task, we used one tissue slice per dataset for training, while reserving the remaining slices for testing. We trained GraphST on the single training slice using its standard configuration. For scGPT, we fine-tuned the encoder by minimizing the mean squared error (MSE) loss between the predicted and observed gene expression matrices. For both methods, we employed the default hyperparameters.

From this experiment, we observe that DRIFT not only enhances the performance of scGPT but also improves its performance beyond the state-of-the-art GraphST. Without DRIFT, scGPT's mean ARI is 0.3755, which is lower than GraphST's mean ARI of 0.3981. However, with DRIFT, scGPT's mean ARI performance improves to 0.5365. The dataset-specific ARI scores are listed in Supplementary Section S8. This result demonstrates that DRIFT effectively bridges the performance gap between general-purpose foundation models and specialized spatial methods, allowing previously underperforming foundation models to achieve or exceed domain-specific benchmarks.

4 Discussion and Conclusions

We introduce DRIFT, a graph diffusion–based framework that integrates spatial information into pretrained foundation models. DRIFT leverages embeddings from large single-cell models for spatial-specific tasks — annotation, alignment, and clustering — without requiring architectural changes or retraining. Across simulated and real datasets, DRIFT outperformed base foundational models in various scenarios.

An open research question is testing the effect of approximating the heat kernel used in the diffusion step. While the resources required to run DRIFT are quite low compared to pretraining a new FM, its computational requirements still scale with the increasing size of the slice, which increases heat kernel computation cost. However, approximating the heat kernel using Chebyshev polynomials can resolve this issue.

In conclusion, DRIFT is a simple, effective, and accessible method for coupling pretrained scRNA-seq foundation models with ST data and performing diffusion-based denoising for ST foundational models. By combining pretrained embeddings with diffusion-enhanced information, DRIFT leverages existing models and extends them to a unified framework for spatially informed representations of tissues and developmental organization.

Code and Data availability

The code is publicly available at `https://github.com/rsinghlab/DRIFT`. Dataset MERMB and MERMBA are available at the `https://cellxgene.cziscience.com` [52] website. Dataset MERHH is available at the `https://cells.ucsc.edu` [53] website. Datasets 10xHPC and MERMPH are available on a site hosted by Yuan *et al.* [6] at `http://sdmbench.drai.cn/`. Datasets 10xHSI and 10xHOC are available at Chen *et al.*'s [17] tutorial website: `https://guangyuwanglab2021.github.io/Loki/`. Dataset StereoME is available at `https://db.cngb.org/stomics/mosta/`.

Acknowledgments

This work is supported by NIH award R35HG011939-01 and T32 MH115895.

References

[1] Chen, C., Kim, H. J. & Yang, P. Evaluating spatially variable gene detection methods for spatial transcriptomics data. *Genome Biology* **25**, 18 (2024).

[2] Adhikari, S. D., Yang, J., Wang, J. & Cui, Y. Recent advances in spatially variable gene detection in spatial transcriptomics. *Computational and Structural Biotechnology Journal* (2024).

[3] Yuan, X. *et al.* Heartsvg: a fast and accurate method for identifying spatially variable genes in large-scale spatial transcriptomics. *Nature Communications* **15**, 5700 (2024).

[4] Kats, I., Vento-Tormo, R. & Stegle, O. Spatialde2: fast and localized variance component analysis of spatial transcriptomics. *Biorxiv* 2021–10 (2021).

[5] Zhu, J., Sun, S. & Zhou, X. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome biology* **22**, 184 (2021).

[6] Yuan, Z. *et al.* Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods* **21**, 712–722 (2024).

[7] Zhang, L., Liang, S. & Wan, L. A multi-view graph contrastive learning framework for deciphering spatially resolved transcriptomics data. *Briefings in Bioinformatics* **25**, bbae255 (2024).

[8] Long, Y. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications* **14**, 1155 (2023).

[9] Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications* **13**, 1739 (2022).

[10] Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology* **39**, 1375–1384 (2021).

[11] Yang, W. *et al.* Deciphering cell–cell communication at single-cell resolution for spatial transcriptomics with subgraph-based graph attention network. *Nature Communications* **15**, 7101 (2024).

[12] Cang, Z. *et al.* Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature methods* **20**, 218–228 (2023).

[13] Jin, S., Plikus, M. V. & Nie, Q. Cellchat for systematic analysis of cell–cell communication from single-cell transcriptomics. *Nature Protocols* 1–40 (2024).

[14] Garcia-Alonso, L. *et al.* Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nature genetics* **53**, 1698–1711 (2021).

[15] Zeira, R., Land, M., Strzalkowski, A. & Raphael, B. J. Alignment and integration of spatial transcriptomics data. *Nature Methods* **19**, 567–575 (2022).

[16] Liu, X., Zeira, R. & Raphael, B. J. Paste2: partial alignment of multi-slice spatially resolved transcriptomics data. *bioRxiv* (2023).

[17] Chen, W. *et al.* A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods* **22**, 1568–1582 (2025). URL `https://www.nature.com/articles/s41592-025-02707-1`. Publisher: Nature Publishing Group.

[18] Schaar, A. C. *et al.* Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv* 2024–04 (2024).

[19] Zhao, S. *et al.* Stofm: a multi-scale foundation model for spatial transcriptomics. *arXiv preprint arXiv:2507.11588* (2025).

[20] Li, H. *et al.* Stformer: Spatial-temporal-aware transformer for video instance segmentation. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[21] Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

[22] Chen, H. *et al.* Quantized multi-task learning for context-specific representations of gene network dynamics. *bioRxiv* (2024).

[23] Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods* **21**, 1470–1480 (2024).

[24] Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nature methods* **21**, 1481–1491 (2024).

[25] Wu, J. *et al.* Biology-driven insights into the power of single-cell foundation models. *Genome Biology* **26**, 334 (2025). URL `https://doi.org/10.1186/s13059-025-03781-6`.

[26] Zhang, F. & Hancock, E. R. Graph spectral image smoothing using the heat kernel. *Pattern recognition* **41**, 3328–3342 (2008).

[27] Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).

[28] Ursu, O. *et al.* Genomedisco: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **34**, 2701–2707 (2018).

[29] He, Y. *et al.* Diffusion-enhanced characterization of 3d chromatin structure reveals its linkage to gene regulatory networks and the interactome. *Genome Research* **33**, 1354–1368 (2023).

[30] Ilharco, G. *et al.* Openclip. *Zenodo* (2021).

[31] Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nature Biotechnology* **41**, 773–782 (2023).

[32] Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome medicine* **14**, 68 (2022).

[33] Fan, Z. *et al.* Spascer: spatial transcriptomics annotation at single-cell resolution. *Nucleic Acids Research* **51**, D1138–D1149 (2023).

[34] Brbić, M. *et al.* Annotation of spatially resolved single-cell data with stellar. *Nature Methods* **19**, 1411–1418 (2022).

[35] Hu, J. *et al.* Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods* **18**, 1342–1351 (2021).

[36] Li, J., Chen, S., Pan, X., Yuan, Y. & Shen, H.-B. Cell clustering for spatial transcriptomics data with graph neural networks. *Nature Computational Science* **2**, 399–408 (2022).

[37] Chen, D. *et al.* Single-cell rna-seq with spatial transcriptomics to create an atlas of human diabetic kidney disease. *The FASEB Journal* **37**, e22938 (2023).

[38] Chang, Y. *et al.* Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning. *Computational and structural biotechnology journal* **20**, 4600–4617 (2022).

[39] Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal* **8**, 289 (2016).

[40] Hu, Y. *et al.* Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome Biology* **25**, 212 (2024).

[41] Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience* **24**, 425–436 (2021).

[42] Villacampa, E. G. *et al.* Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genomics* **1** (2021). URL https://www.cell.com/cell-genomics/abstract/S2666-979X(21)00081-1. Publisher: Elsevier.

[43] Mirzazadeh, R. *et al.* Spatially resolved transcriptomic profiling of degraded and challenging fresh frozen samples. *Nature Communications* **14**, 509 (2023). URL https://www.nature.com/articles/s41467-023-36071-5. Publisher: Nature Publishing Group.

[44] Farah, E. N. *et al.* Spatially organized cellular communities form the developing human heart. *Nature* **627**, 854–864 (2024).

[45] Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).

[46] Li, Z. & Zhou, X. Bass: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome biology* **23**, 168 (2022).

[47] Zhang, M. *et al.* Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624**, 343–354 (2023).

[48] Allen, W. E., Blosser, T. R., Sullivan, Z. A., Dulac, C. & Zhuang, X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* **186**, 194–208 (2023).

[49] Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell* **185**, 1777–1792 (2022).

[50] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631 (2019).

[51] Myronenko, A. & Song, X. Point Set Registration: Coherent Point Drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**, 2262–2275 (2010). URL https://ieeexplore.ieee.org/abstract/document/5432191.

[52] Program, C. C. S. *et al.* Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research* **53**, D886–D900 (2025).

[53] Speir, M. L. *et al.* Ucsc cell browser: visualize your single-cell data. *Bioinformatics* **37**, 4578–4580 (2021).