Integrating Spatial Transcriptomics Data for Cross-Species Molecular Region Comparison

by

Bridget Li

B.S. Computer Science and Molecular Biology, MIT, 2023

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND MOLECULAR BIOLOGY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Bridget Li Department of Electrical Engineering and Computer Science May 17, 2024
Certified by:	Xiao Wang Assistant Professor of Chemistry, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Integrating Spatial Transcriptomics Data for Cross-Species Molecular Region Comparison

by

Bridget Li

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND MOLECULAR BIOLOGY

ABSTRACT

Comparative analysis of brain patterns across species can advance understanding of different biological processes and functions. Spatially resolved transcriptomics (SRT) technologies present the ability to measure gene expression of single cells within tissues, enabling the detection of unique spatial molecular patterns in the brain. Several computational methods that rely on cellular neighborhood information have been developed for characterizing molecular tissue regions in SRT data. Here, we show that spatial integration (SPIN) improves the performance of existing methods and enables the clustering of molecular tissue regions. Then, we test SPIN and signal-processing approaches on SRT data from mouse and macaque brains. We integrate the brain atlases of these two species to identify shared and distinct spatial molecular patterns. This work offers new insights into spatial molecular features between mouse and macaque brains and proposes a framework for integrating SRT datasets on a large scale.

Thesis supervisor: Xiao Wang
Title: Assistant Professor of Chemistry

Acknowledgments

I would like to express my gratitude to my direct mentor, Ph.D. candidate Kamal Maher. His teaching and guidance have been invaluable to my research. I am also sincerely grateful to my thesis advisor, Professor Xiao Wang, for her insightful input on my project and support of my career pursuits.

Biographical Sketch

Bridget Li is from Austin, Texas. She received her B.S. in Computer Science and Molecular Biology with a minor in Economics from MIT in 2023. During her undergrad, she conducted both dry-lab and wet-lab biological research in the labs of Professors Forest White, Darrell Irvine, and Ernest Fraenkel. Bridget is passionate about the intersection of computer science and biology and hopes to pursue a career in techbio.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Spatially Resolved Transcriptomics (SRT)

Spatially resolved transcriptomics (SRT) is a burgeoning technique for measuring mRNA expression within tissue sections [1]–[3]. Single-cell RNA sequencing (scRNA-seq) measures averaged gene expression from mRNAs in a cell, but this technique loses positional information. With SRT, one can obtain the original physiological context of those cells in tissue sections [4]. Spatial methods have become essential for analyses in neuroscience, developmental biology, and cancer research [5]. In these fields, spatial information can reveal tissue neighborhoods and local features contributing to biological function.

1.1.1 SRT Technologies

There have been numerous different techniques developed for SRT. Generally, these techniques fall into two classes: imaging-based and sequencing-based [5]. Imaging technologies image in situ hybridization of mRNAs with sequence-specific probes via microscopy. Sequencing technologies synthesize cDNAs from mRNAs and then employ next-generation sequencing to quantify gene-specific sequences. This work focuses on one imaging technology, STARmap PLUS, and one sequencing-based technology, Stereo-seq.

STARmap PLUS

STARmap (spatially resolved transcript amplicon readout mapping) profiles single-cell transcriptional states in three-dimensional brain tissues [6]. STARmap with protein localization and unlimited sequencing (STARmap PLUS) expands upon STARmap by enabling protein detection in the same tissue section [7]. STARmap PLUS was combined with a scRNA-seq atlas to generate a transcriptome-wide spatial atlas of the mouse brain, which is a basis for analysis in this work [8].

Stereo-seq

Stereo-seq (spatial enhanced resolution omics-sequencing) combines DNA nanoball-patterned arrays and in situ RNA hybridization to achieve single-cell resolution and genome-wide coverage [9]. Stereo-seq was used to produce a single-cell transcriptome atlas of the macaque cortex, revealing the organization and evolution of primate cortical regions [10]. This macaque atlas is another key data source drawn upon in this work.

1.2 Molecular Region Identification

Identifying molecularly defined tissue regions, or spatial domains, from SRT data can yield insight into the transcriptional basis of biological organization and function. The structure-function relationship is especially prevalent in the laminar organization of the human cerebral cortex; cells in different layers differ in their expression, morphology, and physiology [11]. The ability to accurately and efficiently identify these molecular tissue regions is important for understanding cytoarchitecture and functions.

Traditional clustering methods, such as k-means and the Leiden algorithm, use only gene expression data to identify clusters of cells [12]. These approaches frequently lead to disjoint regions because they do not utilize spatial data to group nearby cells together [13]. Algorithms that account for similarity among neighboring cells in physical space more

accurately capture the spatial dependency of gene expression. Three commonly used graph-based algorithms for molecular region identification are described here.

1.2.1 STAGATE

STAGATE identifies spatial domains by learning low-dimensional latent embeddings of spatial information and gene expression [14]. STAGATE constructs a spatial neighbor network and uses a graph attention auto-encoder to smooth features over neighboring cells and adaptively learn their similarity. Spatial domains can be identified from the latent embeddings, which preserve local expression patterns.

1.2.2 UTAG

UTAG is an unsupervised deep learning method for the discovery of tissue architecture [15]. UTAG constructs a graph of cellular interactions based on physical proximity. UTAG smooths gene expression features across the graph via message passing, where the physical distances between cells and their neighbors correspond to weights for neighborhood aggregation.

1.2.3 GraphST

GraphST is a graph self-supervised contrastive learning method to cluster cells into spatial domains [13]. GraphST uses a graph neural network encoder to smooth gene expression features based on spatial proximity. Using self-supervised contrastive learning, embedding distance is minimized between spatially adjacent spots and maximized between distant spots.

1.3 Spatial Integration (SPIN)

Spatial integration (SPIN) is an approach developed by Maher et. al to improve the smoothing of gene expression features over tissue [16]. Computational methods for characterizing

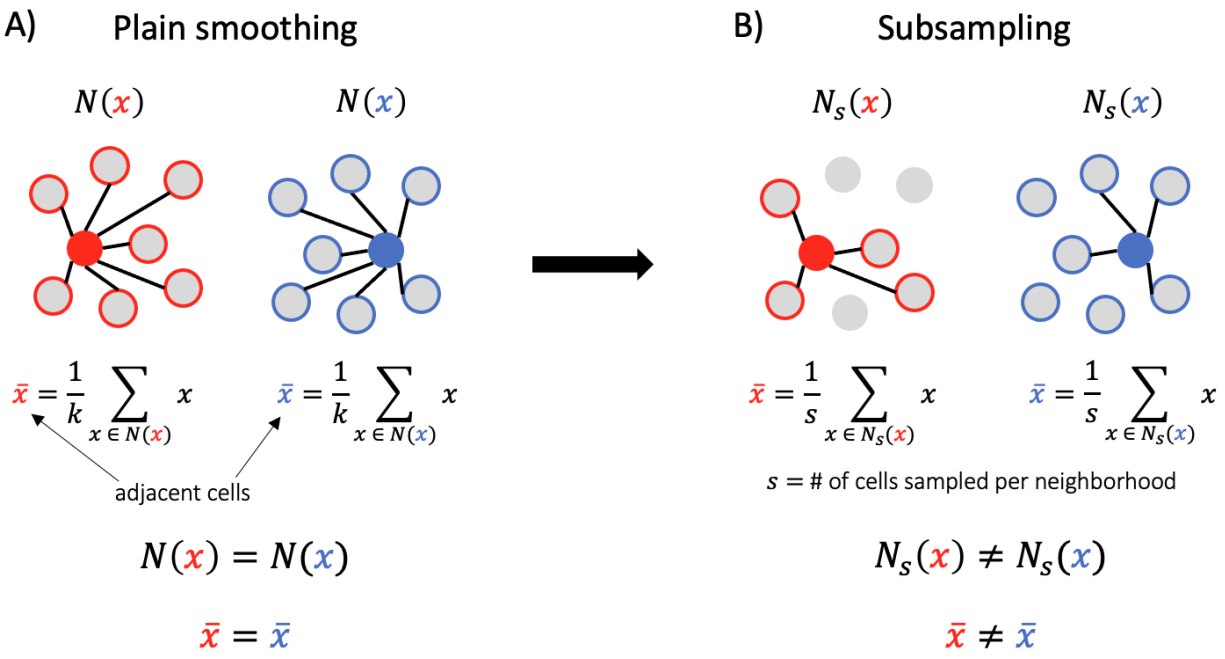

Figure 1.1: Smoothing and the subsampling solution of SPIN. A) Neighboring cells have identical expression features as a result of smoothing. Each dot represents a cell. The red and blue filled dots represent physically adjacent cells. B) With subsampling, adjacent cells can vary their exact neighborhoods during smoothing, resulting in unique expression features.

molecular tissue regions in SRT data rely on smoothing features across neighboring cells. Smoothing was shown to increase autocorrelation between neighboring cells' expression features. Since downstream methods like Leiden and UMAP rely on a nearest neighbors graph, the latent graph resulting from smoothing reflects the spatial closeness of cells rather than spatial transcriptomic features. SPIN involves randomly sub-sampling neighboring cells before smoothing to mitigate autocorrelation between neighbors. This enables more accurate identification of spatial molecular regions defined by gene expression. SPIN was applied to SRT datasets from mouse and marmoset brains to identify region marker genes. A summary of the SPIN approach is shown in Fig. 1.1.

In addition to mitigating autocorrelation for the methods described above, SPIN works simply by randomly subsampling the k-nearest neighbors (kNN) graph and averaging the expression of spatial nearest neighbors. This work includes a quantitative assessment of the

improvements made by SPIN using metrics for physical reconstruction and clustering. SPIN, a simple and efficient alternative to existing methods, can be used for analyzing a variety of transcriptome atlases.

1.4 Cross-Species Comparative Analysis

Cross-species comparative analysis is a powerful method to understand biological process specificity and the evolution of biological systems [17]. Important biological properties are often conserved across species, while dissimilar properties can indicate evolutionary modifications. The brain is of particular interest for cross-species comparisons due to its complexity and function in higher-order information processing [18]. Studying brain tissue can advance understanding of cognitive ability and neurological conditions.

There exists a large body of work focused on cross-species comparison using scRNA-seq data. Comparative studies, often focused on humans, non-human primates, and mice, have revealed adaptations in neuronal types throughout evolution [18]–[21]. Developments in computational methods for this task include clustering cell types, efficiently integrating datasets, and identifying patterns in transcriptome evolution [22]. However, these previous studies do not incorporate spatial information.

Recent work has begun to extend cross-species analysis to SRT data. Fang et al. use multiplexed error-robust fluorescence in situ hybridization (MERFISH) to generate a spatially resolved cell atlas of the human temporal gyrus and drew comparisons to the mouse cortex [23]. Furthermore, Lei et al. use Stereo-seq to map the macaque brain, then compare monkey oligodendrocyte trajectories to those in humans [24]. As more SRT data is collected and made available, further cross-species comparisons can be performed to produce novel biological insights.

Spatial molecular patterns are useful in studying differences in brain organization across species. For example, differences in the locations of inhibitory neurons in mouse and non-

human primate neocortical layers may indicate a relationship between spatial molecular patterns and cognition [25]. It is valuable to perform cross-species comparative analysis on SRT data to understand regional patterns across the brains of related species.

Here, we propose a comparative analysis of mouse and macaque brains using SRT data. We first smooth and integrate the data from the two species, then identify and analyze spatial molecular patterns shared between species or unique to each species. This may lead to new biological insights about the neurological differences between the two species. Furthermore, this research can provide a basis for the computational integration of SRT data across species, which can be applied to an array of different species.

Chapter 2

Methods

2.1 Using SPIN to improve on existing methods

SPIN improves the performance of existing methods for spatial domain identification by mitigating autocorrelation during the smoothing of gene expression features. We quantitatively evaluated the spatial autocorrelation problem and subsampling solution. Spatial autocorrelation is captured by a physical reconstruction metric: the similarity between a kNN graph of cells in physical space and a kNN graph of cells in latent space (Fig. 3.1A). Similarity is given by the Jaccard distance between each row of each graph's adjacency matrices. The physical kNN represents the physical topology of the tissue, while the latent kNN graph represents the molecular relationships between cells as the result of smoothing.

2.1.1 Simulations

We performed simulations that model the laminar organization of the brain (Fig. 3.1B). Each cell was represented by a randomly placed point within a unit circle. Next, we delineated specific tissue regions and assigned each cell to one of these regions. Each cell was also categorized as either a spatial or non-spatial cell type. For each region, half of its cells were designated as a spatial cell type associated with that region (akin to excitatory neurons),

while the other half were randomly assigned a non-spatial cell type (akin to glial cells). For each cell type, we generated four distinct marker genes by incorporating uniform noise into the true cell type labels based on a signal-to-noise ratio of 1. This setup allows the simulation to be easily repeated with varying outcomes by adjusting the random seed, keeping other parameters unchanged.

kNN-based smoothing

In these simulations, we explored the impact of kNN-based smoothing on physical reconstruction. Using k=50 for smoothing, we analyzed 500 simulations and observed that the reconstruction values generally centered around 0.7. We varied the smoothing intensity by changing the subsampling rate from 90% to approximately 1% (single-cell), which is effectively no smoothing, in 10% decrements. We then applied k-means and Leiden clustering on the smoothed features to identify regions, which in this case were assigned spatial cell types. We visualized the tissues in physical and UMAP spaces, with UMAP showing the latent relationships between identified regions. We quantified the performance of the clustering methods using the adjusted Rand index (ARI).

Contemporary smoothing algorithms

We modified our original simulation setup by replacing kNN smoothing with various contemporary smoothing algorithms, including STAGATE, UTAG, and GraphST. Each method calculated a physical nearest neighbor graph, allowing us to apply the same subsampling strategy as with the kNN smoothing. However, instead of using a uniform k=50 for both physical and latent analyses, we adjusted the parameters to k=50 for physical and k=15 for latent, aligning with the default setting in Scanpy's sc.pp.neighbors function. We limited the number of simulations to 10 per method to align with the more resource-intensive experiments conducted on brain and gut tissues. The clustering and quantification follow as described for kNN-based smoothing.

Table 2.1: Mouse and macaque data

species	technology	# specimens	# slices	# genes	size
mouse	STARmap PLUS	1	13	1021	2.4GB
macaque	Stereo-seq	3	163	12934	112GB

2.1.2 STARmap mouse brain

After evaluating the performance of multiple smoothing approaches in simulations, we validated the results using STARmap data of the mouse brain. We employed the same code and methodology in this analysis.

2.2 Cross-species comparison

2.2.1 Smoothing mouse and macaque data

The mouse and macaque datasets are available online. The data is summarized in Table 2.1. We began by downloading and formatting the data into cell-by-gene matrices. The mouse cortex was isolated from the larger brain slices, making the mouse data more comparable to the macaque data, of which only the cortex is available. The Scanpy package was used to perform analyses on matrices [26].

SPIN

SPIN, in combination with Leiden clustering, has already been successfully applied to the mouse dataset to identify spatial molecular regions. We applied SPIN with simple neighborhood averaging to smooth the macaque dataset.

Laplacian smoothing

Another method to identify that does not rely on nearest neighbors is Laplacian smoothing, or "filtering" by signal. We applied Delaunay triangulation to create a spatial graph to

perform filtering on. We then filtered the data using a heat kernel, which smooths graphs based on the heat diffusion equation applied to the cell-by-gene matrix (eq. 2.1). This is a low-pass filter, allowing only large-scale signals below a cutoff frequency to pass.

$$f = \exp(\tau * \frac{X}{G.\text{lmax}}) \tag{2.1}$$

where:

τ = scaling parameter

X = cell-by-gene matrix

G = graph created by Delaunay triangulation

lmax = the largest eigenvalue of the graph G

The mouse dataset was collected using STARmap PLUS, while the macaque dataset was collected using Stereo-seq [8], [10]. These methods both use in situ sequencing but can result in different resolutions of cells. If the measured gene expression is disparate enough in mouse and macaque datasets, we cannot properly identify common spatial molecular regions. To handle this challenge, we experimented with different τ values to optimize the smoothing of mouse and macaque data. The final filtering used $\tau = 60$ for mouse and $\tau = 200$ for macaque, making smoothing of macaque data more aggressive.

2.2.2 Integrating mouse and macaque brain atlases

Following conventional single-cell methods, we normalized, log-transformed, and scaled the data per cell. We then applied principal component analysis (PCA) to the mouse and macaque data to generate a low-dimensional embedding of cells. The principal components (PCs) were integrated using Harmony, an algorithm that projects cells into a joint embedding [27].

2.2.3 Clustering to identify molecular regions

We applied Leiden clustering to integrated PCs to identify molecular regions. Leiden clustering is a popular community detection algorithm based on locally optimal assignment, improving upon Louvain clustering [12]. Leiden clustering works in combination with SPIN but not with Laplacian smoothing as Leiden depends on nearest neighbor-finding. With Laplacian smoothing, we used k-means clustering, which does not depend on cellular neighborhoods. There are six neocortical layers, each defined by different neuron types [28]. From outer to inner, the layers are the molecular layer (L1), external granular layer (L2), external pyramidal layer (L3), internal granular layer (L4), internal pyramidal layer (L5), and the multiform layer (L6). L1, the most superficial layer, is poorly defined by only a few horizontal cells. It is difficult to identify in mouse STARmap data, so we restricted the number of clusters in k-means to 5, representing L2-L6.

2.2.4 Scaling across datasets

The mouse atlas consists of about 1.1 million cells, while the macaque atlas consists of about 48 million cells. This presented difficulty in scaling the processing pipeline to the macaque dataset. To mitigate this issue, we applied filtering separately to each mouse or macaque brain slice at a time. This way, we only filtered at most 500 thousand cells at a time. This filtering step was performed in parallel for all slices. We also cleaned the data to contain only 654 genes shared by all samples.

We then integrated the entire mouse atlas with tractable subsets of the macaque atlas (4-13 slices). We tested using PCA to capture neocortical organization, and then projecting all macaque samples onto the same PCA space. Following this line of reasoning, we could keep the top PCs for integration across a larger number of samples.

2.2.5 Characterizing molecular regions

After obtaining reasonable spatial molecular regions, we analyzed them to form comparisons about brain composition and organization across species. We calculated cell densities for regions. To determine density, we first calculated a convex hull on the points in space to approximate the region's area. We then divided cell count by the region area to estimate cell density.

We performed differentially expressed gene (DEG) analysis to calculate shared and species-specific gene markers for each region. We used an adjusted P-value cutoff of 0.05 to select DEGs. We identified gene ontology (GO) annotations to determine the functions of gene markers. Finally, we performed trajectory inference on an integrated subset of one macaque and one mouse sample to examine the continuous developmental trajectories of cells.

Chapter 3

Results

3.1 SPIN to improve on existing methods

3.1.1 Simulations

kNN-based smoothing

The physical and latent kNN graphs are expected to differ since gene expression does not align exactly with physical space, which we observe for conventional single-cell analysis. However, when the data is smoothed by averaging molecular features across the physical neighborhood, the latent kNN graph starts to mirror the physical kNN graph more closely. The similarity between the two graphs can be quantified using the Jaccard distance between the rows of their adjacency matrices (Fig. 3.1A). This physical reconstruction metric allows us to gauge how well the latent space represents physical adjacency as opposed to spatial transcriptomic patterns.

Comparing fully subsampled (100%) and non-subsampled (single-cell) data revealed that smoothing significantly enhanced reconstruction levels (Fig. 3.1C,D). Moreover, as the extent of subsampling varied, it was observed that the degree of reconstruction decreased proportionally. Therefore, while kNN-based smoothing promotes physical reconstruction in latent

space, this problem can be alleviated by subsampling.

As expected, single-cell clustering roughly delineated regions according to spatial cell types, distinguishing the 12 true cell type clusters (4 spatial and 8 non-spatial) within UMAP space (Fig. 3.1C). Without subsampling, smoothing via k-means accurately formed region clusters, whereas Leiden clustering produced arbitrary spatial patches within the tissue. UMAP visualization subtly transformed the physical cell arrangement into a funnel shape, a visual counterpart to the quantitative physical reconstruction observed in Fig. 3.1D. Moderate subsampling yielded the most precise depiction of spatial molecular features, where the regions displayed a one-dimensional topology extending from the tissue's edge to its center.

Assigning the region identities for each cell in the simulations allowed us to measure the clustering efficacy of k-means and Leiden at various subsampling levels using the ARI. The ARI for Leiden was low with no subsampling, improved with moderate subsampling, and then declined with more extensive subsampling (Fig. 3.1E). Similarly, the k-means ARI decreased under aggressive subsampling, suggesting that this method might eliminate crucial spatial information as it approaches single-cell characteristics. Conversely, the k-means ARI remained high without subsampling, reinforcing the idea that k-means clustering is less affected by physical reconstruction due to its independence from the latent kNN graph.

Contemporary smoothing algorithms

The results of kNN smoothing with k=50 physical neighbors and k=15 latent neighbors show the same patterns as using k=50 for both physical and latent neighbors (Fig. 3.2 top row). Without subsampling, Leiden clustering again produced spatial patches, whereas UMAP reflected the physical connections among cells (Fig. 3.2A). Similar to previous results, subsampling 30% of each neighborhood led to Leiden clusters that mirrored the actual tissue regions, while UMAP accurately depicted the tissue's one-dimensional spatial molecular structure (Fig. 3.2B). Additionally, we noted that physical reconstruction decreased in proportion to the level of subsampling (Fig. 3.2C), and clustering performance enhanced

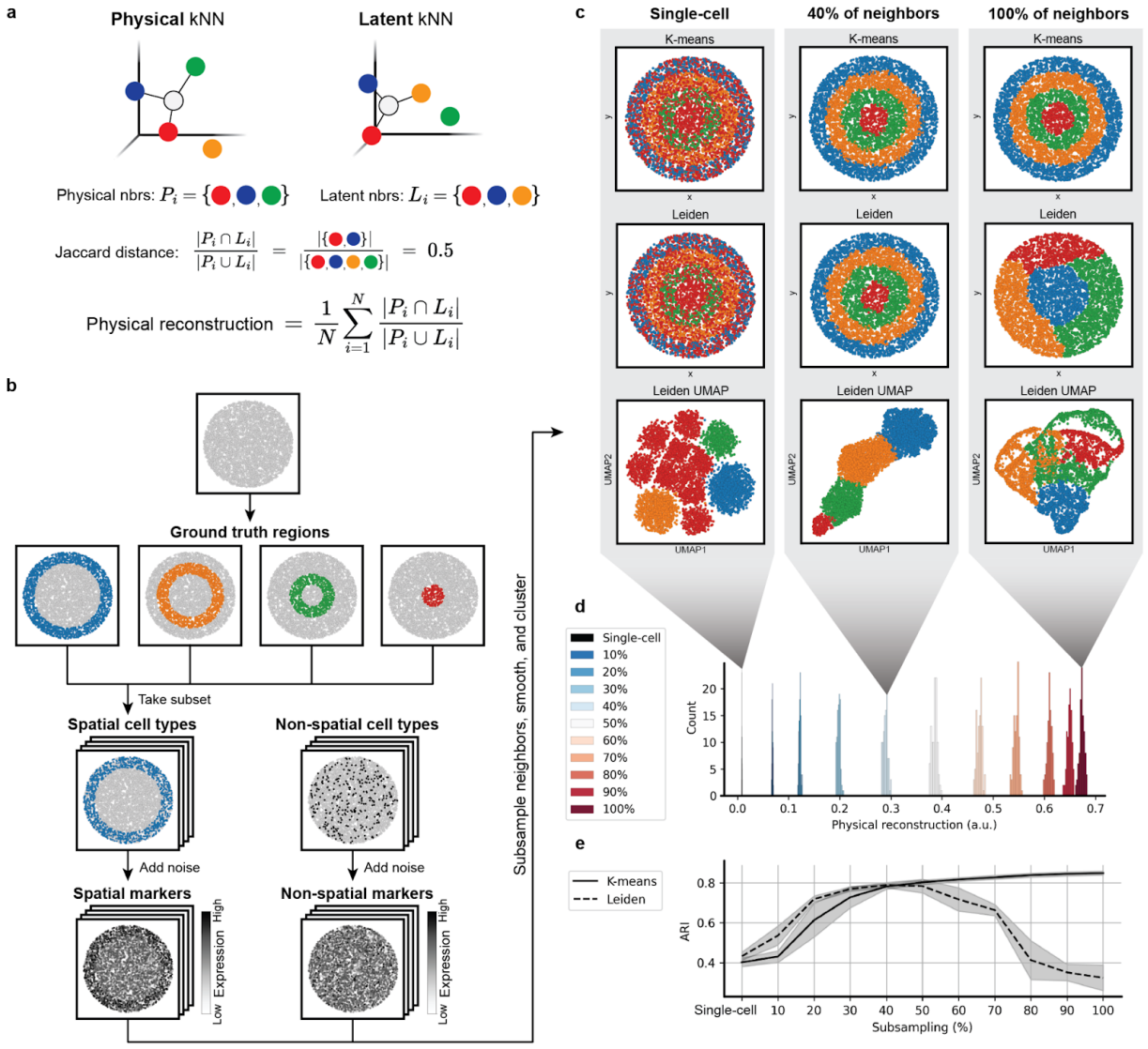

Figure 3.1: Quantification and simulation of physical reconstruction and subsampling. A) The physical reconstruction metric. B) Data simulation protocol. C) Visualization of subsampling and smoothing results using various amounts of subsampling. "Single-cell" corresponds to subsampling 1% of each neighborhood, which amounts to no smoothing at all and thus conventional single-cell analysis. D) Quantification of physical reconstruction given various amounts of subsampling. Histograms represent the distribution of physical reconstruction values for 500 simulations with different random seeds. E) Clustering performance of k-means and Leiden given various amounts of subsampling. Shaded areas above and below curves indicate standard deviation.

with moderate subsampling (Fig. 3.2D).

The subsequent rows of Fig. 3.2 show similar outcomes for various modern smoothing techniques. The results were comparable to those from kNN-based smoothing, although different methods reached optimal performance with varying subsampling levels. Additionally, runtime and memory usage metrics for these specific analyses are detailed in Fig. 3.2E,F. These findings underscore the issues of physical reconstruction and the efficacy of subsampling solutions in simulations.

3.1.2 STARmap mouse brain

We aimed to further validate these findings using mouse brain STARmap data from the original study (Fig. 3.3). The outcomes were consistent with those shown in Fig. 3.1 and 3.2. Despite slight variations in the curve shapes, the quantitative findings resembled those from the simulations (Fig. 3.3C,D). Similar trends were also seen in the runtime and memory usage comparisons (Fig. 3.3E,F).

3.2 Cross-species comparison

3.2.1 Smoothing mouse and macaque data

As the macaque data collected using Stereo-seq has lower spatial resolution, SPIN and Leiden clustering did not function as expected. Compared to marmoset and mouse STARmap PLUS data, where SPIN identifies anatomically accurate cortical layers, the clusters for the macaque slice were formed based on physical adjacency rather than biologically relevant gene expression patterns (Fig. 3.4A). The corresponding UMAP visualization does not show meaningful relationships between regions (Fig. 3.4A). Even after smoothing, region markers specific to certain neocortical layers were not well-defined (Fig. 3.4C). For example, RORB is a marker of L4 neurons, and CCK is most commonly expressed in L2/3 and L5, but the

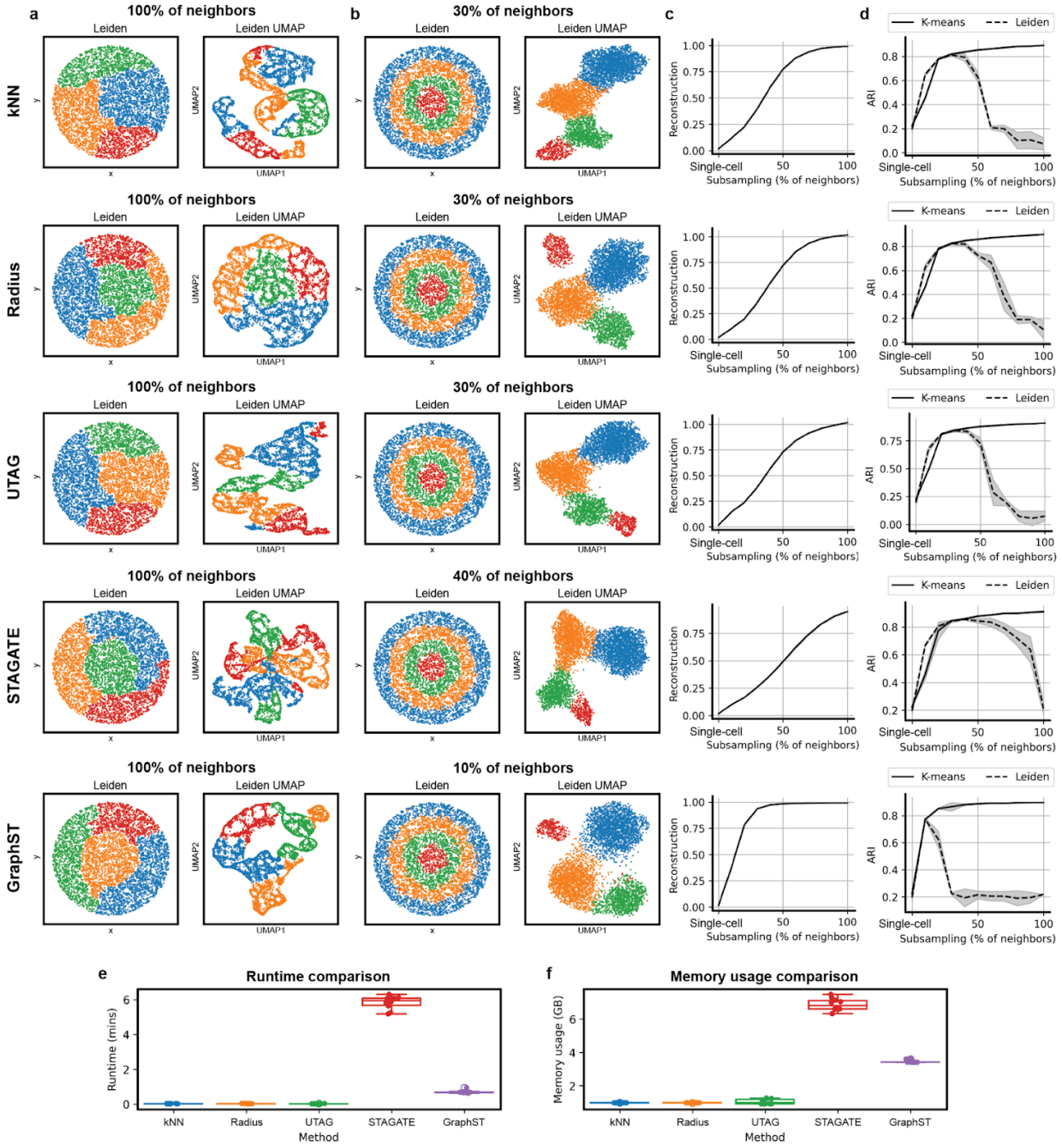

Figure 3.2: Evaluating multiple smoothing approaches with various subsampling levels in simulated data. A) Leiden clustering on non-subsampled, smoothed data. Smoothing was performed by the method indicated on the left-hand side. The Leiden resolution was set to achieve the number of ground truth clusters (4). B) Same as A) but using the amount of subsampling that yielded the highest ARI by Leiden clustering. C) Quantification of physical reconstruction over a range of subsampling levels. Shaded areas above and below curves indicate standard deviation. As the smallest subsample size is one cell, the lowest tick has been labeled "Single-cell". D) Quantification of clustering performance over a range of subsampling levels, formatted as in C). E) Runtime comparison across each method. Datapoints represent the runs shown in A-D). F) Same as E) but comparing memory usage.

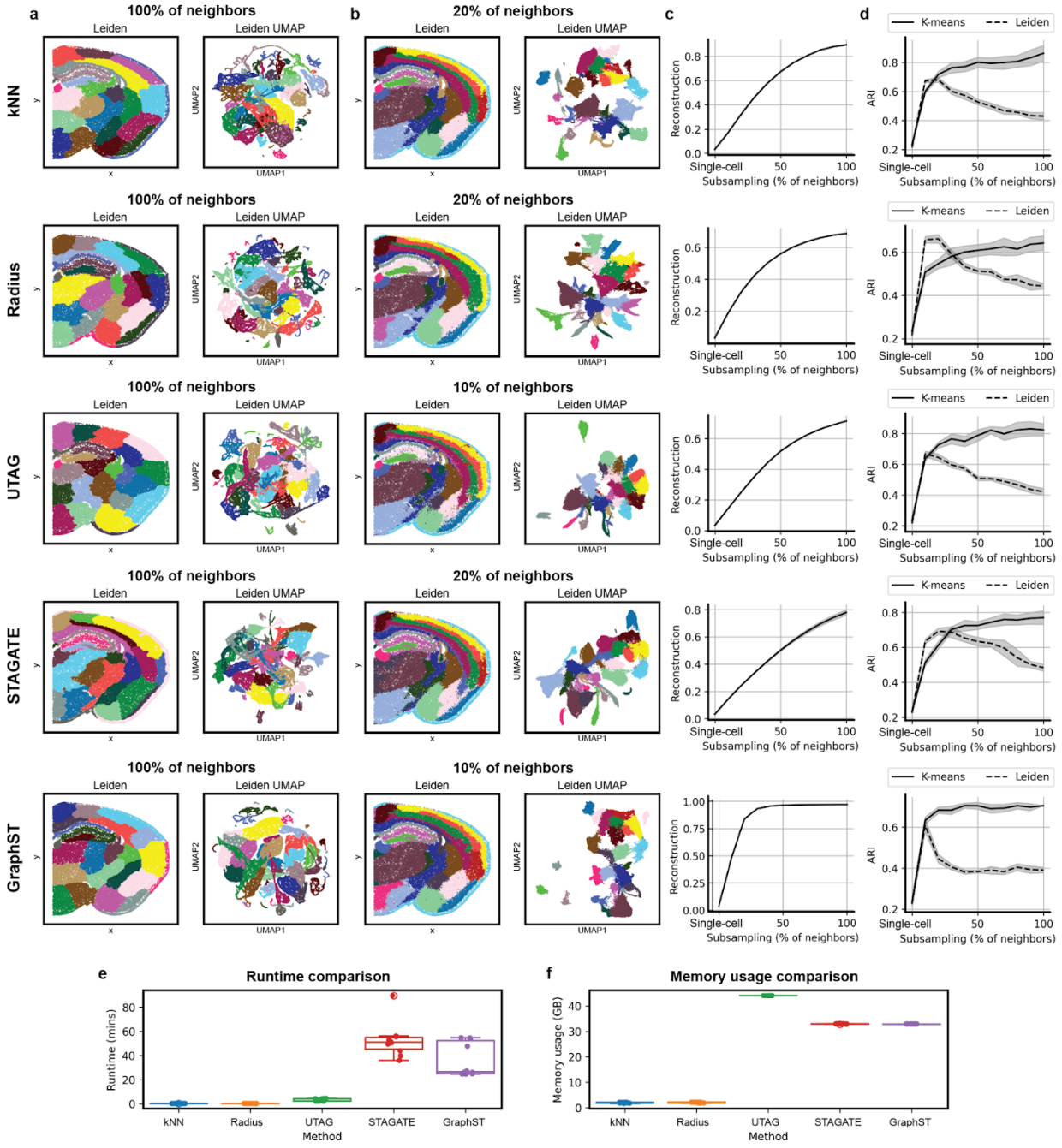

Figure 3.3: Evaluating multiple smoothing approaches with various subsampling levels in mouse brain STARmap data. A) Leiden clustering on non-subsampled, smoothed data. Smoothing was performed by the method indicated on the left-hand side. The Leiden resolution was set to achieve the number of ground truth clusters (23). B) Same as A) but using the amount of subsampling that yielded the highest ARI by Leiden clustering. C) Quantification of physical reconstruction over a range of subsampling levels. Shaded areas above and below curves indicate standard deviation. As the smallest subsample size is one cell, the lowest tick has been labeled "single-cell". D) Quantification of clustering performance over a range of subsampling levels, formatted as in C). E) Runtime comparison across each method. Datapoints represent the exact runs shown in A-D). F) Same as E) but comparing memory usage.

appearance of these markers in their respective layers is unclear after smoothing. Changing the neighborhood sizes and amount of subsampling did not improve the problem.

As such, we resorted to Laplacian smoothing and k-means clustering, which obtained regions that align better with known cortical layers in the brain (Fig. 3.4D). Laplacian smoothing can dramatically amplify low-frequency signals, corresponding to gene expression, with flexible length scales. Since k-means does not construct a nearest neighbor graph, UMAP cannot be performed.

3.2.2 Integrating mouse and macaque brain atlases

Then, we attempted to integrate 10 slices of macaque data with all of the mouse cortex data (Fig. 3.5). This was repeated for different regions of the macaque brain. The integration was somewhat successful, with common regions identified in mouse and macaque brains that match known cortical layers. As the macaque dataset is very large, it is computationally expensive to integrate all samples at once.

Tuning the τ parameter of the heat filter improved region identification in mouse slices, with $\tau = 60$ being selected as optimal (Fig. 3.6). τ was not tuned for macaque data since the originally selected $\tau = 200$ led to accurate clustering.

We tried using PCA to condense the size of the dataset. Using all 13 mouse slices and a representative sample of 10 macaque slices, we calculated 15 PCs to capture neocortical organization. We then projected all mouse and macaque samples onto the same PC space, replacing the cell-by-gene matrix with a smaller cell-by-PC matrix. However, when we attempted to integrate all of the samples with Harmony, the initialization of centroids did not return the expected number of clusters for every sample. Therefore, the shapes of the resulting matrices could not be broadcast together.

Without a method to make the problem more computationally tractable, we used extensive computational resources to integrate all of the original filtered samples, and then apply k-means clustering. An example of one macaque and one mouse sample from the full

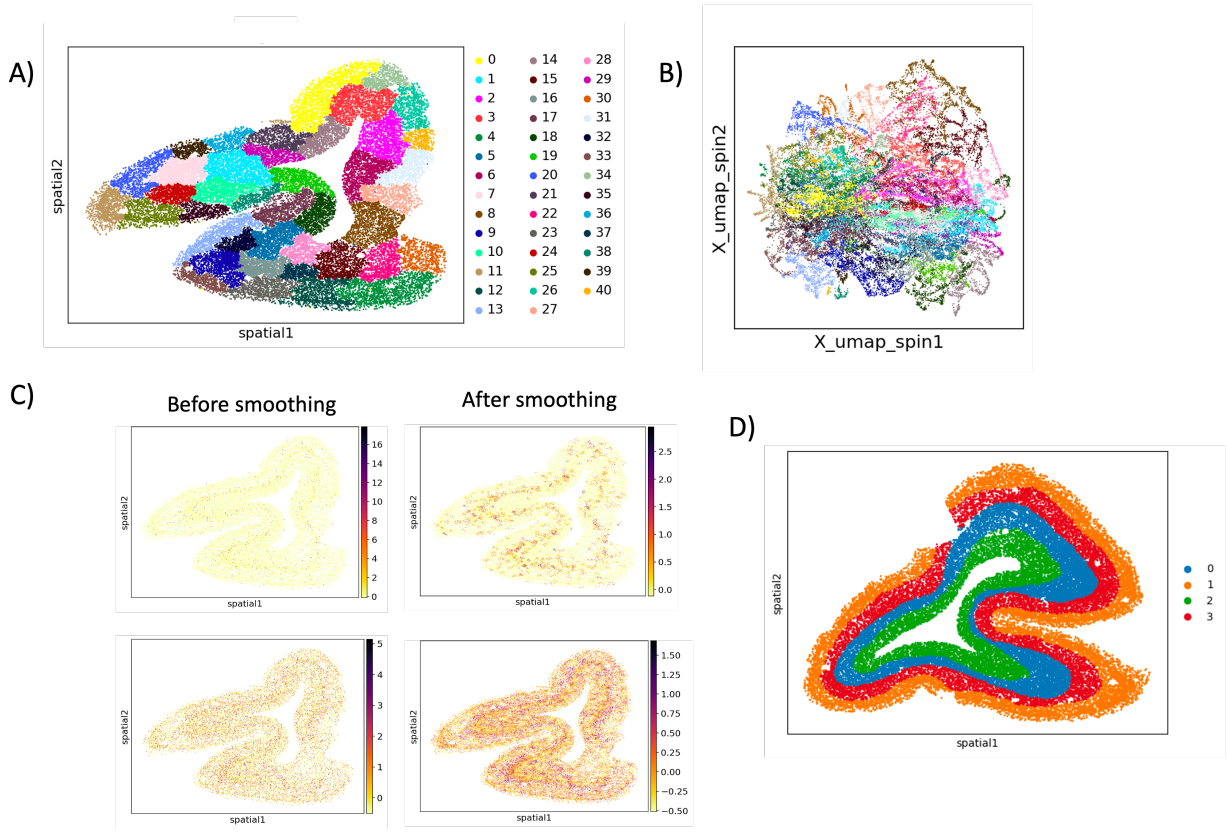

Figure 3.4: Laplacian smoothing performs better than SPIN on macaque data, using slice T147 as an example. A) Clusters identified by SPIN and Leiden clustering in physical space, given by 2 spatial coordinates. B) Clusters identified by SPIN and Leiden clustering in UMAP space, given by 2 latent variables. C) Expression of 2 regional gene markers (RORB and CCK) before and after smoothing using SPIN. D) Clusters identified by Laplacian smoothing and k-means clustering in physical space.

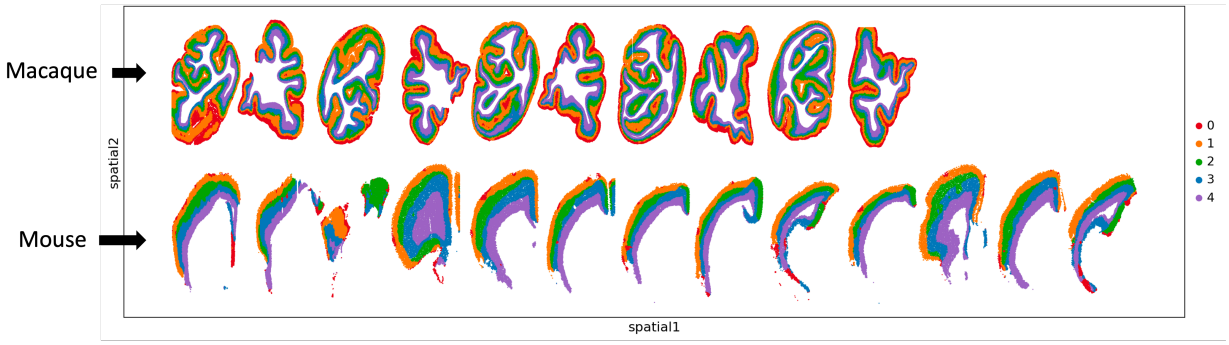

Figure 3.5: Example integration of subset of macaque data and all mouse data. Macaque samples are on the top row, and mouse samples are on the bottom row. Numbers and colors indicate molecular regions identified by k-means clustering. Samples were scaled to the same size for ease of visualization.

integrated dataset is shown in Fig. 3.7. The five regions are numbered 0-4 from outer to inner.

3.2.3 Characterizing molecular regions

We examined DEGs in each region and within regions between species (Fig. 3.8). The region numbering follows as shown in Fig. 3.7. Some regions contain DEGs that are not commonly associated with specific cortical layers, such as IL1RAPL2, NREP, HTR2A, and HPCAL1 (Fig. 3.8A). Meanwhile, a single region may contain markers of multiple layers. For example, the outermost region (0) contains differentially expressed KCNS1 and PVALB, which are commonly found in L5, as well as RORB, which is a marker of L4. The identified regions capture some level of the neocortex's laminar architecture of the neocortex but do not correspond perfectly with the known neocortical layers.

Grouping by species within each region, we find DEGs in the macaque versus mouse data (Fig. 3.8B). Region 0 seems better defined by DEGs in the mouse, whereas region 3 seems better defined by DEGs in the macaque. FAM107A, which is involved in cell cycle regulation and cell proliferation, is differentially expressed in region 0 in the mouse but in region 3 in the macaque. In region 1 of the mouse, the DEGs are all related to neuronal development and structure, while those in the macaque have more general functions not specific to the brain. CCK is expectedly expressed in region 2 of the mouse, as it is a marker of L2/L3. HTR3A, which functions in cognition and mood regulation, and FOXP2, which plays a role in language development and speech production, are differentially expressed in region 4 (the innermost region) of the macaque data. Accordingly, deeper layers of the cerebral cortex are involved in cognitive functions [29]. These genes may contribute to the higher-order processing skills of macaques over mice.

GO term analysis revealed similarities and differences in cellular functions of various regions and between species (Fig. 3.9). Neuron projection, the long-distance communication of neurons in the nervous system, is expectedly found in every region of the cortex. Dendrites

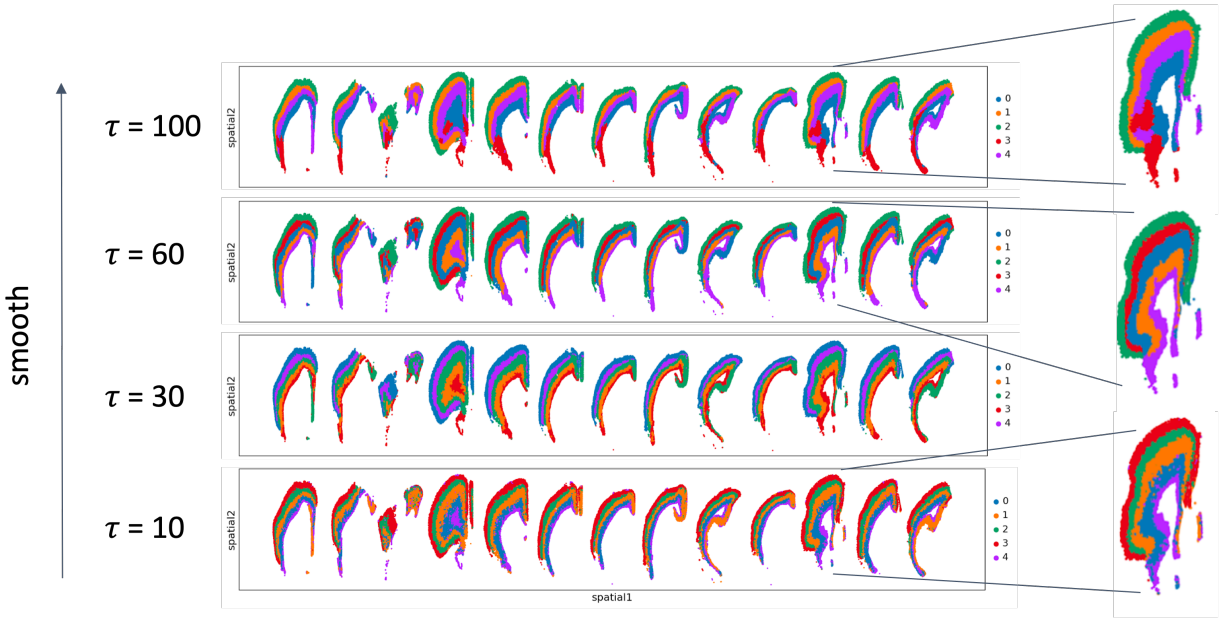

Figure 3.6: Tuning τ in Laplacian smoothing improves region identification in mouse STARmap data. 13 mouse brain slices are shown, with one enlarged as an example. A greater τ indicates a greater level of smoothing.

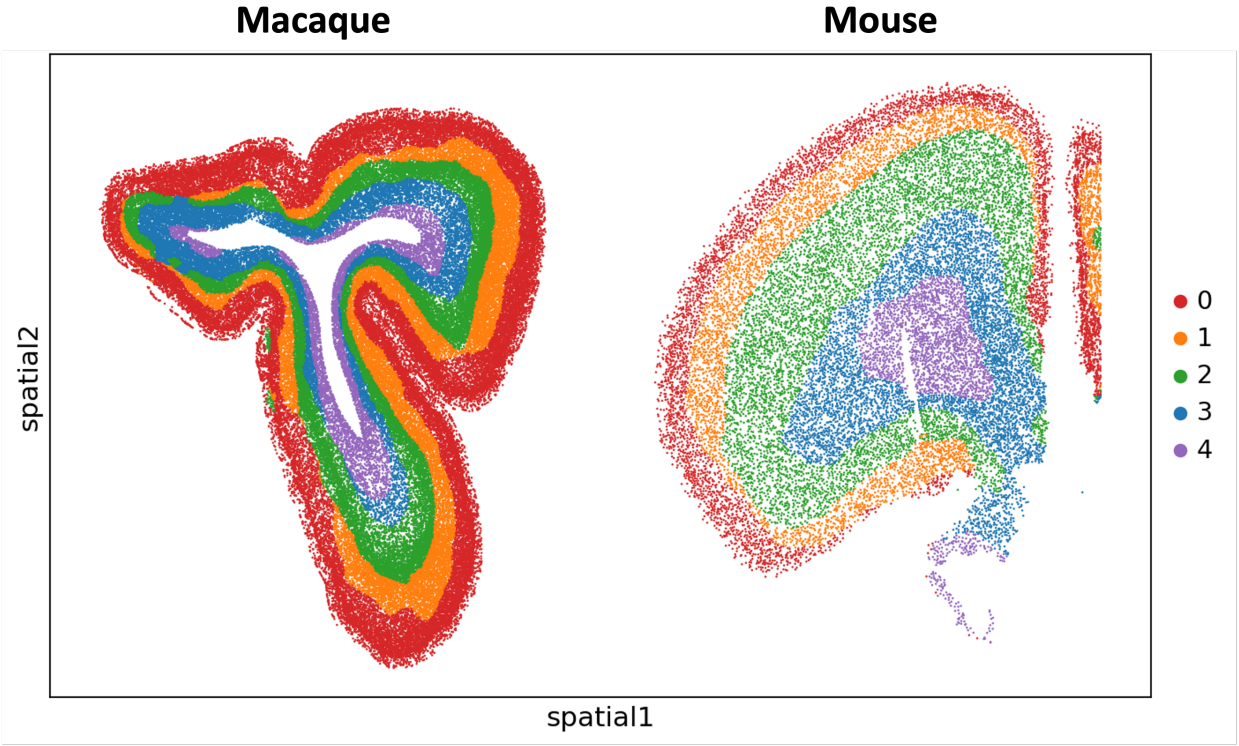

Figure 3.7: One macaque and one mouse sample from the full integrated dataset. Numbers and colors indicate molecular regions identified by k-means clustering. Samples were scaled to the same size for ease of visualization.

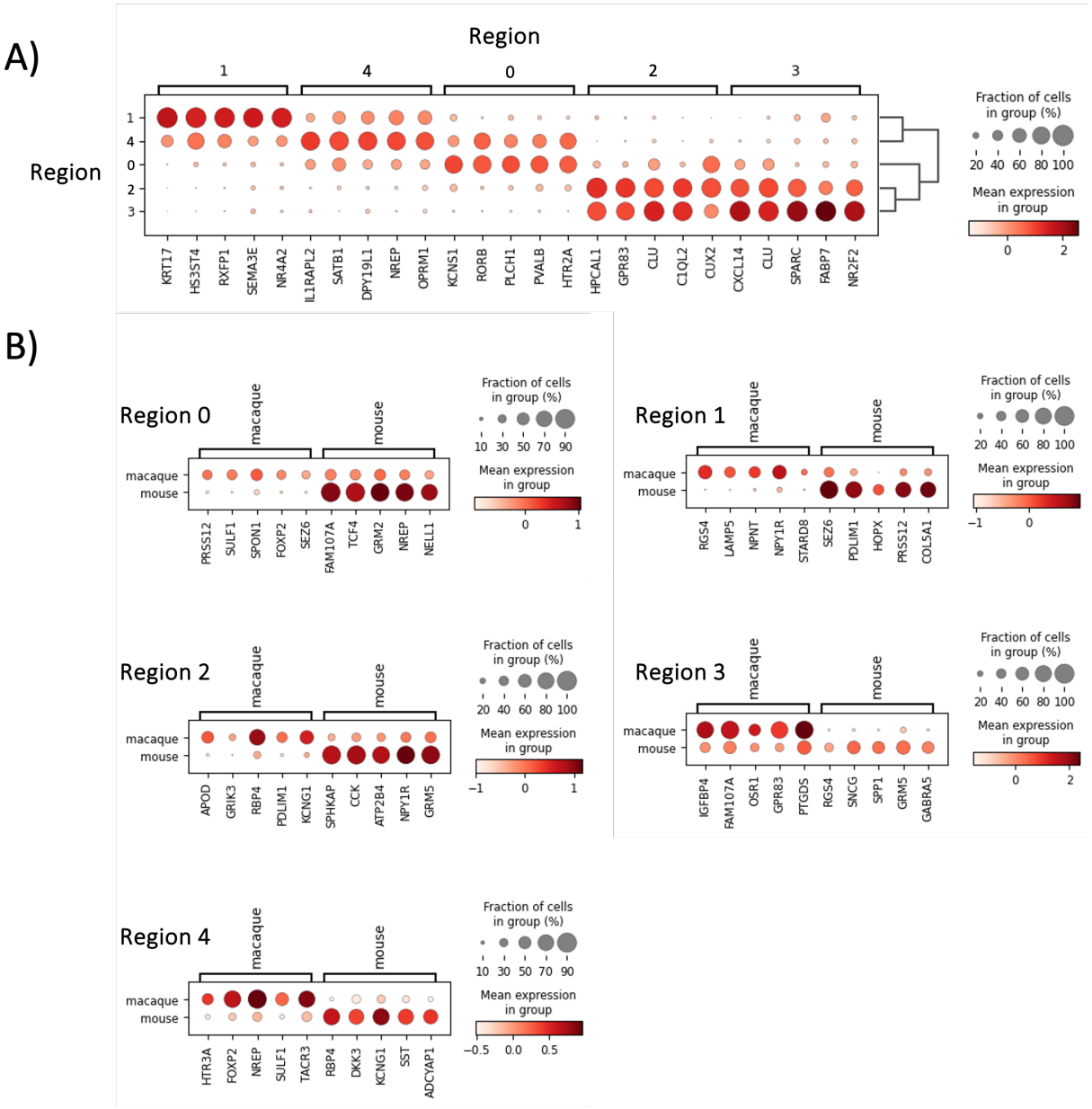

Figure 3.8: Differential gene expression in integrated mouse and macaque samples. DEGs in each group are shown below the dotplots. A) DEGs in each region across both macaque and mouse. Region number is given on the left and top sides of the dotplot. A dendrogram indicates relationships between regions. B) DEGs in macaque versus mouse in each of the five regions. Species are given on the left and top sides of the dotplot.

and axons, which are responsible for receiving signals and transmitting electrical impulses, respectively, are also found broadly across the regions. There are more Functions in transport vesicles and vesicle membranes are identified in the macaque outermost region (0), while these functions appear in region 2 in the mouse.

The density calculation is affected by the spatial resolution. As such, we are only able to compare across regions within a species and not across species. There appears to be more variation in the neuronal density of macaque regions (Fig. 3.10). In both species, region 3, the second-to-most inner region, has a lower average density than the other regions. This is unexpected since deeper layers tend to have higher neuronal density compared to superficial layers [30]. The deep layers contain larger pyramidal neurons, which contribute to their higher density, while the superficial layers contain more sparse populations of smaller pyramidal neurons and various interneurons. However, high error bars in the macaque data mean the difference in density is insignificant.

We generated a diffusion map to visualize the continuous developmental trajectories of cells (Fig. 3.11). However, the diffusion map seems to reconstruct the tissue in diffmap space, which is an artifact of smoothing (Fig. 3.11A). There is minimal change in the first diffusion component (DC1) in the mouse sample, likely because the sizes of each sample are very different (Fig. 3.11B). Since kNN in smoothed expression space is simply kNN in physical space, there is far more variation within the much larger macaque slice than the smaller mouse slice. In the macaque slice, rather than DC1 representing depth along the neocortex, we see a change associated with vertical spatial location.

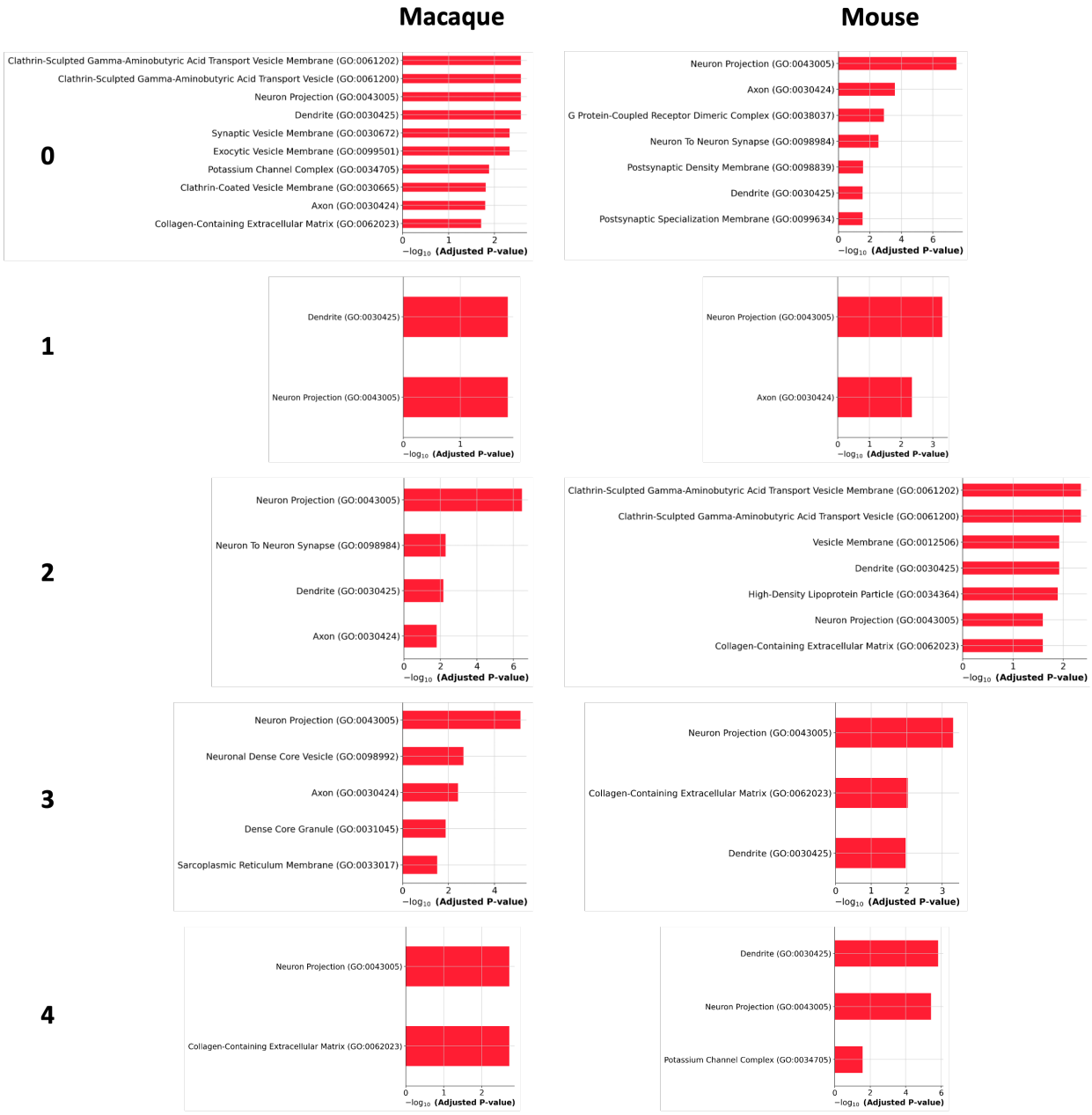

Figure 3.9: Gene ontology analysis of integrated mouse and macaque samples. Region number is shown on the left-hand side, and species is shown at the top.

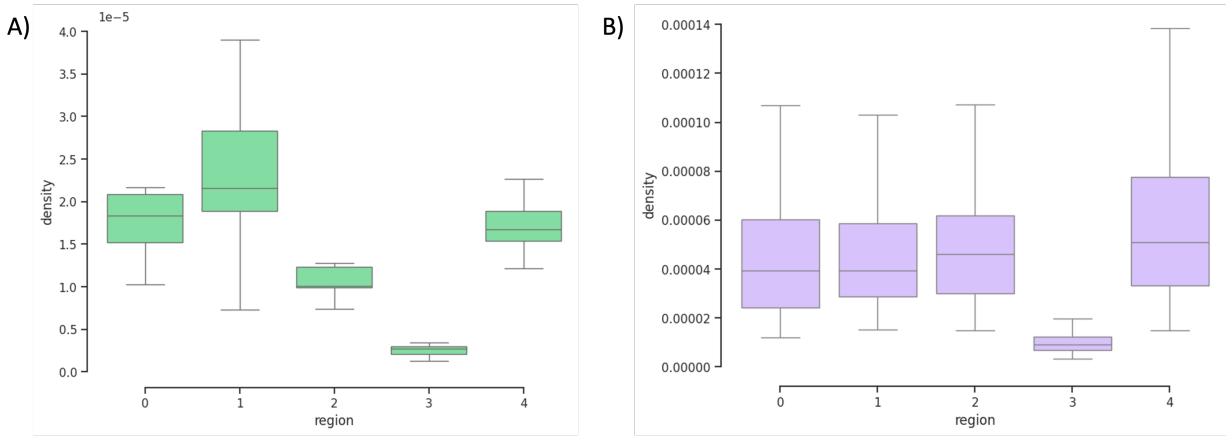

Figure 3.10: Density of molecular regions identified in A) mouse STARmap data and B) macaque Stereo-seq data across all respective samples.

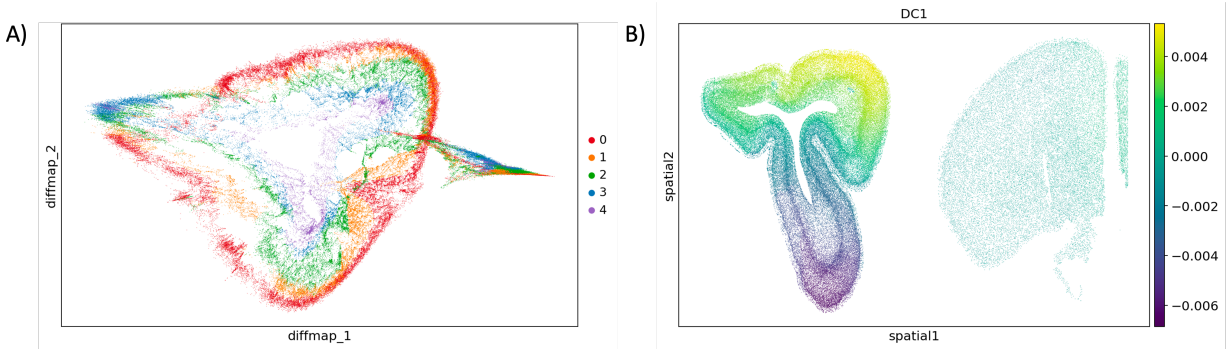

Figure 3.11: Trajectory inference across species. A) Diffusion map built on Harmony-integrated PCs from one macaque and one mouse sample. B) The first diffusion component (DC1) from the diffusion map visualized over tissue.

Chapter 4

Conclusion

Characterizing molecular tissue regions in SRT data can provide insight into biological processes. The subsampling method underlying SPIN enhances the performance of existing methods and serves as the foundation for a new, streamlined approach to spatial characterization. This advancement has enabled us to identify, without supervision, spatial molecular regions across species.

However, we have found limitations in the smoothing and clustering methods at varying length scales. In sparse data, the subsampling technique might be inadequate due to the challenge of capturing relevant regional features with too few cells. In addition, the necessity of selecting a single k value for kNN prevents the simultaneous identification of molecular regions at both small and large length scales. Laplacian smoothing mediates these problems by filtering signals based on their frequency, allowing for aggressive smoothing at different length scales.

Combining and comparing spatial transcriptomic features across emerging large-scale SRT atlases is valuable for gaining insights into regional patterns across the brains of related species. In this study, we conducted a comparative analysis of mouse and macaque brains via Laplacian smoothing of SRT data followed by cellular integration. We found differences in gene expression and cell density that may relate to the improved cognitive abilities of

macaques over mice. This research could establish a framework for computationally integrating SRT data, offering broad applicability to diverse species.

4.1 Future work

We may also apply graph neural network (GNN) approaches, such as STAGATE and GraphST, to smooth gene expression features across space. In our initial approach, we used SPIN with equal weights for edges in our graphs of cells. However, a GNN with an attention mechanism may be useful in incorporating learnable weights of edges in filtering. This approach may yield gene expression signals that are more compatible with SPIN.

We may apply strategies for data enrichment. STARmap PLUS achieves high capture resolution but only captures a small fraction of genes out of the transcriptome. A mouse sample contains around 1,000 genes, whereas a macaque sample contains around 15,000 genes. There are 654 genes that are common to all samples, but only comparing these genes may lead to the loss of valuable biological information. Therefore, we can use gene imputation methods like gimVI or SpaFormer to predict missing genes when performing data preprocessing [31], [32]. However, handling a larger volume of gene data poses significant challenges in terms of computational power and storage requirements.

Another technique we have begun investigating is predicting cell-cell interactions (CCIs) from spatial gene expression data. Information exchange between different cells is a fundamental basis for many biological processes [33]. Since SRT measures the relative position of different cells, there has been interest in using SRT to profile the interaction tendencies of cells [34]. Current CCI tools involve statistical tests and network models [35], [36]. A simple and intuitive alternative may be found in the filtering process. Patterns in how individual cells interact are captured by small, recurrent changes in gene expression, making them "high-frequency." By inverting the heat filter to isolate high-frequency, rather than low-frequency, gene expression signals within a region, we may be able to identify CCIs.

Comparing CCIs between species may facilitate a better understanding of their complex cellular microenvironments.

References

[1] P. L. Ståhl, F. Salmén, S. Vickovic, *et al.*, "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics," *Science*, vol. 353, pp. 78–82, Jun. 2016. DOI: 10.1126/science.aaf2403.

[2] V. Marx, "Method of the year: Spatially resolved transcriptomics," *Nature Methods*, vol. 18, pp. 9–14, Jan. 2021. DOI: 10.1038/s41592-020-01033-y. URL: https://www.nature.com/articles/s41592-020-01033-y.

[3] R. Dries, J. Chen, N. d. Rossi, M. M. Khan, A. Sistig, and G.-C. Yuan, "Advances in spatial transcriptomic data analysis," *Genome Research*, vol. 31, pp. 1706–1718, Oct. 2021. DOI: 10.1101/gr.275224.121. URL: https://genome.cshlp.org/content/31/10/1706.short (visited on 10/18/2022).

[4] J. Du, Y. Yang, Z. An, M. Zhang, X.-H. Fu, Z. Huang, Y. Yuan, and J. Hou, "Advances in spatial transcriptomics and related data analysis strategies," *Journal of Translational Medicine*, vol. 21, May 2023. DOI: 10.1186/s12967-023-04150-2.

[5] C. G. Williams, H. J. Lee, T. Asatsuma, R. Vento-Tormo, and A. Haque, "An introduction to spatial transcriptomics for biomedical research," *Genome Medicine*, vol. 14, Jun. 2022. DOI: 10.1186/s13073-022-01075-1.

[6] X. Wang, W. E. Allen, M. A. Wright, *et al.*, "Three-dimensional intact-tissue sequencing of single-cell transcriptional states," *Science*, vol. 361, eaat5691, Jun. 2018. DOI: 10.1126/science.aat5691.

[7] H. Zeng, J. Huang, H. Zhou, *et al.*, "Integrative in situ mapping of single-cell transcriptional states and tissue histopathology in a mouse model of alzheimer's disease," *Nature Neuroscience*, vol. 26, pp. 430–446, Feb. 2023. DOI: 10.1038/s41593-022-01251-x. URL: https://www.nature.com/articles/s41593-022-01251-x.

[8] H. Shi, Y. He, Y. Zhou, *et al.*, "Spatial atlas of the mouse central nervous system at molecular resolution," *Nature*, vol. 622, pp. 552–561, Oct. 2023. DOI: 10.1038/s41586-023-06569-5. URL: https://www.nature.com/articles/s41586-023-06569-5#ref-CR3.

[9] A. Chen, S. Liao, M. Cheng, *et al.*, "Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays," *Cell*, vol. 185, 1777–1792.e21, May 2022. DOI: 10.1016/j.cell.2022.04.003. URL: https://www.sciencedirect.com/science/article/pii/S0092867422003993 (visited on 09/08/2022).

[10] A. Chen, Y. Sun, Y. Lei, *et al.*, "Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex," *Cell*, vol. 186, pp. 3726–3743, Jul. 2023. DOI: 10.1016/j.cell.2023.06.009. (visited on 07/14/2023).

[11] K. R. Maynard, L. Collado-Torres, L. M. Weber, *et al.*, "Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex," *Nature Neuroscience*, vol. 24, pp. 425–436, Feb. 2021. DOI: 10.1038/s41593-020-00787-0.

[12] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: Guaranteeing well-connected communities," *Scientific Reports*, vol. 9, Mar. 2019. DOI: 10.1038/s41598-019-41695-z.

[13] Y. Long, K. S. Ang, M. Li, *et al.*, "Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst," *Nature Communications*, vol. 14, Mar. 2023. DOI: 10.1038/s41467-023-36796-3.

[14] K. Dong and S. Zhang, "Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder," *Nature Communications*,

vol. 13, p. 1739, Apr. 2022. DOI: 10.1038/s41467-022-29439-6. URL: https://www.nature.com/articles/s41467-022-29439-6.

[15] J. Kim, S. Rustam, J. M. Mosquera, S. H. Randell, R. Shaykhiev, A. F. Rendeiro, and O. Elemento, "Unsupervised discovery of tissue architecture in multiplexed imaging," *Nature Methods*, vol. 19, pp. 1653–1661, Dec. 2022. DOI: 10.1038/s41592-022-01657-2. URL: https://www.nature.com/articles/s41592-022-01657-2.

[16] K. Maher, M. Wu, Y. Zhou, J. Huang, Q. Zhang, and X. Wang, "Mitigating autocorrelation during spatially resolved transcriptomics data analysis," *bioRxiv*, Jul. 2023. DOI: https://doi.org/10.1101/2023.06.30.547258.

[17] X. Zhou and G. Gibson, "Cross-species comparison of genome-wide expression patterns," *Genome Biology*, vol. 5, p. 232, 2004. DOI: 10.1186/gb-2004-5-7-232. (visited on 02/28/2019).

[18] T. E. Bakken, N. L. Jorstad, Q. Hu, *et al.*, "Comparative cellular analysis of motor cortex in human, marmoset and mouse," *Nature*, vol. 598, pp. 111–119, Oct. 2021. DOI: 10.1038/s41586-021-03465-8. URL: https://www.nature.com/articles/s41586-021-03465-8.

[19] Y. Zhu, A. M. M. Sousa, T. Gao, *et al.*, "Spatiotemporal transcriptomic divergence across human and macaque brain development," *Science*, vol. 362, eaat8077, Dec. 2018. DOI: 10.1126/science.aat8077. (visited on 09/21/2021).

[20] R. D. Hodge, T. E. Bakken, J. A. Miller, *et al.*, "Conserved cell types with divergent features in human versus mouse cortex," *Nature*, vol. 573, pp. 61–68, Aug. 2019. DOI: 10.1038/s41586-019-1506-7.

[21] J. H. Lui, N. D. Nguyen, S. M. Grutzner, *et al.*, "Differential encoding in prefrontal cortex projection neuron classes across cognitive tasks," *Cell*, vol. 184, 489–506.e26, Jan. 2021. DOI: 10.1016/j.cell.2020.11.046. URL: https://pubmed.ncbi.nlm.nih.gov/33338423/.

[22] M. E. R. Shafer, "Cross-species analysis of single-cell transcriptomic data," *Frontiers in Cell and Developmental Biology*, vol. 7, Sep. 2019. DOI: 10.3389/fcell.2019.00175. (visited on 02/03/2020).

[23] R. Fang, C. Xia, J. L. Close, *et al.*, "Conservation and divergence of cortical cell organization in human and mouse revealed by merfish," *Science*, vol. 377, pp. 56–62, Jul. 2022. DOI: 10.1126/science.abm1741. (visited on 11/28/2022).

[24] Y. Lei, M. Cheng, Z. Li, *et al.*, "Spatially resolved gene regulatory and disease-related vulnerability map of the adult macaque cortex," *Nature Communications*, vol. 13, p. 6747, Nov. 2022. DOI: 10.1038/s41467-022-34413-3. URL: https://www.nature.com/articles/s41467-022-34413-3 (visited on 12/12/2023).

[25] F. M. Krienen, M. Goldman, Q. Zhang, *et al.*, "Innovations present in the primate interneuron repertoire," *Nature*, vol. 586, pp. 262–269, Sep. 2020. DOI: 10.1038/s41586-020-2781-z.

[26] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, Feb. 2018. DOI: 10.1186/s13059-017-1382-0.

[27] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature Methods*, vol. 16, pp. 1289–1296, Nov. 2019. DOI: 10.1038/s41592-019-0619-0.

[28] N. Palomero-Gallagher and K. Zilles, "Cortical layers: Cyto-, myelo-, receptor- and synaptic architecture in human cortical areas," *NeuroImage*, vol. 197, pp. 716–741, Aug. 2019. DOI: 10.1016/j.neuroimage.2017.08.035.

[29] E. K. Miller, "The prefrontal cortex and cognitive control," *Nature reviews. Neuroscience*, vol. 1, pp. 59–65, 2000. DOI: 10.1038/35036228. URL: https://www.ncbi.nlm.nih.gov/pubmed/11252769?dopt=Abstract.

[30] E. G. Jones, "Synchrony in the interconnected circuitry of the thalamus and cerebral cortex," *Annals of the New York Academy of Sciences*, vol. 1157, pp. 10–23, Mar. 2009. DOI: 10.1111/j.1749-6632.2009.04534.x. (visited on 07/04/2023).

[31] R. Lopez, A. Nazaret, M. Langevin, J. Samaran, J. Regier, M. I. Jordan, and N. Yosef, "A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements," *arXiv*, May 2019. DOI: 10.48550/arXiv.1905.02269. URL: https://arxiv.org/abs/1905.02269 (visited on 12/12/2023).

[32] H. Wen, W. Tang, W. Jin, J. Ding, R. Liu, F. Shi, Y. Xie, and J. Tang, "Single cells are spatial tokens: Transformers for spatial transcriptomic data imputation," *arXiv.org*, Feb. 2023. DOI: 10.48550/arXiv.2302.03038. URL: https://arxiv.org/abs/2302.03038 (visited on 12/12/2023).

[33] Z. Cang, Y. Zhao, A. A. Almet, A. R. Stabell, R. Ramos, M. V. Plikus, S. X. Atwood, and Q. Nie, "Screening cell–cell communication in spatial transcriptomics via collective optimal transport," *Nature Methods*, vol. 20, pp. 218–228, Jan. 2023. DOI: 10.1038/s41592-022-01728-4. (visited on 01/20/2024).

[34] Z. Liu, D. Sun, and C. Wang, "Evaluation of cell-cell interaction methods by integrating single-cell rna sequencing data with spatial information," *Genome Biology*, vol. 23, Oct. 2022. DOI: 10.1186/s13059-022-02783-y. (visited on 12/12/2022).

[35] S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, and Q. Nie, "Inference and analysis of cell-cell communication using cellchat," *Nature Communications*, vol. 12, Feb. 2021. DOI: 10.1038/s41467-021-21246-9.

[36] R. Browaeys, W. Saelens, and Y. Saeys, "Nichenet: Modeling intercellular communication by linking ligands to target genes," *Nature Methods*, vol. 17, pp. 159–162, Dec. 2019. DOI: 10.1038/s41592-019-0667-5.