Computational Methods for Linking Molecular and Anatomical Data with Spatial Transcriptomics

by

April R. Kriebel

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Joshua Welch, Chair
Professor Victoria Booth
Assistant Professor Kenneth Kwan
Assistant Professor Jun Li
Associate Professor Stephen Parker

April R. Kriebel
akriebel@umich.edu
ORCID iD: 0000-0003-0008-9044

© April R. Kriebel 2024

DEDICATION

To my father, for teaching me the dignity of hard work.
To my mother, for showing me the power of unconditional love.
Philippians 4:13

ACKNOWLEDGEMENTS

I have been blessed to have received a tremendous outpouring of love, support, and encouragement throughout my academic career.

My family has always been an incredible source of strength for me. To my parents, who would have supported and loved me no matter what I chose to do: thank you for valuing who I am and thinking that whatever I pursue is worthwhile if I pursue it with kindness. Mom, thank you for being my unfailing rock. Your encouragement and prayers have gotten me through the roughest patches. Dad, thank you for always being so proud of me, even on my worst days. Alexander, for always providing a listening ear and a safe haven. Aaron, may sibling rivalry never die.

Nicolas, my wonderful husband. You have never let me define my worth by what I accomplish and achieve. You encourage me to set boundaries and chase what I love, and you have spent a small fortune in bringing me motivational Starbucks. You made the journey here a happy one.

My friends, who make me laugh and remember joy, even when there is a bug in my code. To my science squad from Spring Arbor, thank you for building a culture where it is okay to be fearless with your ambition. A special thanks to Jez, without whom I would not have survived Real Analysis, Abstract Algebra, or senior year.

To my friends from UMich. Breanna and Kinsey, I truly appreciate the time we spent together and all the camaraderie and fellowship we shared.

Hojae and Hayley, our dissertation writing group brought me so much laughter and support during a stressful time. Thank you for all the encouragement and for sharing the journey with me! I'm very grateful for you both.

Karen Nelson, you taught me that math isn't scary when you understand it and that the things that are difficult to understand are beautiful because they reflect the complexity of their Creator.

To my undergraduate teachers who selflessly gave me their time and energy, especially Garnet Hauger and Timothy Wegner.

Without Aaron Wyman, I would not be here. I would have never dreamed of graduate school or doing research. At a point in my life when I didn't know where I was going, you

gave me the audacious belief that I could do anything. You gave me the courage to pursue my interests and always pushed me to think more deeply about the implications of facts on systems and their perturbations. I am grateful for your influence in my life.

My committee members: Victoria Booth, Kenneth Kwan, Stephen Parker, and Jun Li. Your expertise and intuition gave my work dimensions and substance it had previously lacked. Your insights and critiques have elevated my work and taught me the value of collaboration and creative genius ignited by cross-department interactions. Thank you for taking the time and energy to help me grow as a researcher.

To the members of the Welch lab, thank you for being so kind and welcoming to me. To be surrounded by such a talented and bright group of people is unusual, but when those people are also warm and generous, it is an indescribable gift.

A deep and heartfelt thank you to my advisor and mentor, Joshua Welch. Before coming to your lab, I vaguely understood single-cell data analysis. You supported me as I developed my coding skills, struggled with the biological implications of my findings, worked to derive new algorithms, and submitted my first grant. Under your mentorship, I've grown as a critical thinker and communicator. You gave me the space and freedom to explore different career paths and opportunities for growth, and your standards for rigorous research pushed me every day to show up with my best foot forward. You were always patient and encouraging as I worked to develop my scientific toolbox. It's rare to find a mentor who supports you as much as a person as they do as a scientist. You made it clear that we were people first and graduate students second, always encouraging me to set work hours and maintain a balance of working hard and building a sustainable career. Jeanna, thank you for supporting Josh the way you do. Your daily sacrifices allow him to pour so much into his students, and it has truly impacted my life. To both of you, thank you.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES xi

LIST OF ACRONYMS xii

ABSTRACT xv

CHAPTER

1 Introduction 1

1.0.1 Cellular Diversity Begins at the Molecular Level 1
1.0.2 Defining A Cell Type 2
1.0.3 Assays to Measure Molecular Properties 3
1.0.4 Computational Methods for Integrating Molecular Datasets 6
1.0.5 Spatial Transcriptomics is an Important Aspect of Cellular Identity 9
1.0.6 Assays for Capturing Spatial Transcriptomics Data 9
1.0.7 Computational Methods for the Deconvolution of Spatial Transcriptomics Data 10
1.0.8 Neuron Morphology 12
1.0.9 Cell Types of the Brain 14
1.0.10 Atlas Generation 16
1.0.11 Overview of dissertation chapters 18

2 UINMF Performs Mosaic Integration of Single-cell Multi-omic Datasets Using Nonnegative Matrix Factorization 20

2.1 Preamble 20
2.2 Introduction 20
2.3 Results 23
2.3.1 Integrative nonnegative matrix factorization algorithm for partially overlapping feature sets 23
2.3.2 Including intergenic peak information improves integration of scRNA and snATAC datasets 25

2.3.3 Leveraging additional genes improves integration with targeted spatial transcriptomic technologies . . . 29
2.3.4 UINMF improves integration of multimodal and spatial transcriptomic datasets . . . 36
2.3.5 Incorporating nonorthologous genes improves the integration of cross-species data . . . 40
2.4 Discussion . . . 43
2.5 Methods . . . 44
2.5.1 Increase in Computational Complexity . . . 46
2.5.2 Evaluating Time and Memory Usage . . . 47
2.5.3 Evaluation Metrics . . . 47
2.5.4 Integration of RNA and ATAC profiles from SNARE-seq . . . 48
2.5.5 Integration of scRNA-seq and STARmap . . . 49
2.5.6 Integration of scRNA-seq and osmFISH . . . 50
2.5.7 Cross-Species Integration . . . 51
2.5.8 Integration of SNARE-seq and STARmap . . . 52
2.5.9 Assessing the contribution of unshared features . . . 53
2.5.10 Selecting K and λ . . . 54
2.5.11 Integration of three species . . . 55
2.5.12 Supplementary Tables and Figures . . . 55

3 Constructing a Whole-Brain Cross-Modality Molecular Atlas with Spatial Resolution . . . 67

3.1 Preamble . . . 67
3.2 Introduction . . . 67
3.3 Results . . . 68
3.3.1 Brain-Wide Integration and Spatial Mapping of Molecular Cell Types from Single-Cell Transcriptome and Epigenome Data . . . 68
3.3.2 Neuronal Cell Types Exhibit Highly Characteristic Spatial Distributions . . . 70
3.4 Discussion . . . 75
3.5 Methods . . . 77
3.5.1 Defining Anatomical Regions Across Datasets . . . 77
3.5.2 Online Integration of Single-Cell Datasets . . . 77
3.5.3 Preprocessing for the ABA In Situ Hybridization Data . . . 78
3.5.4 Preprocessing for the Whole-Brain Slide-seq Data . . . 78
3.5.5 SiNMFiD Reference Dataset Downsampling and Gene Selection . . . 79
3.5.6 Leveraging ANLS to Derive a Shared W Among Reference Datasets 79
3.5.7 SiNMFiD Implementation for Deconvolution of Spatial Transcriptomics Data . . . 80
3.5.8 Advantages of supervised iNMF-informed deconvolution . . . 80
3.5.9 Spatial Deconvolution of the Whole-Brain Slide-seq Data Using RCTD 81
3.5.10 Determining Cell Type Radius . . . 82
3.5.11 Performing QC for Filtering Low Quality Fragments from snATAC-seq Data . . . 82

3.5.12 Division of SMART-seq Cells into Primary Cell Classes 83
3.5.13 Supplementary Tables and Figures 83

4 Leveraging Spatial Coordinates to Correlate Anatomical and Physiological Features with Molecular Cell Type Profiles 148

4.1 Preamble 148
4.2 Introduction 148
4.3 Results 149
4.3.1 Joint Registration of Whole Brain Vasculature and Cell Type Location Reveals Differences in Local Vascular Density of Neuron Types 149
4.3.2 Joint Registration with Mesoscale Connectivity Data Reveals Regional Connectivity Differences Among Neuron Types 154
4.3.3 Joint Registration with c-Fos Imaging Implicates Neuronal Subtypes in Stress-Induced Neuron Activity 160
4.4 Discussion 166
4.5 Methods 167
4.5.1 Integration of Molecular Profiles and Vascular Data 167
4.5.2 Integration of Molecular Profiles and ABA Projection Data 168
4.5.3 Integration of Molecular Profiles and c-FOS Data 169
4.5.4 Analyzing the Association Between c-Fos Score and Molecular Cell Types Across Timepoints and Within Specific Regions 170
4.5.5 Assessing the Relationship between Neural Activity and Vascular Density 171
4.5.6 Supplementary Tables and Figures 172

5 Conclusion 175

5.1 Summary 175
5.2 Future Work 177
5.2.1 Future Areas of Exploration for the UINMF Algorithm 177
5.2.2 Future Directions for Research in the Development of the Whole-Brain Cell Atlas 178
5.2.3 Continued Applications of Leveraging Co-Registration to Correlate Molecular and Non-molecular Features of Cellular Identity 179

BIBLIOGRAPHY 182

LIST OF FIGURES

FIGURE

2.1 Overview of the UINMF algorithm for integrating single-cell datasets with partially overlapping features . . . 24
2.2 Addition of intergenic peak information improves integration of RNA and ATAC datasets . . . 27
2.3 Incorporating additional genes improves integration with STARmap data . . . 31
2.4 Incorporating additional genes locates fine cellular subtypes within 3D spatial volume . . . 33
2.5 Incorporating additional genes improves integration with osmFISH data . . . 35
2.6 Incorporating unshared chromatin and gene features to integrate spatial transcriptomic and multimodal data . . . 37
2.7 Incorporating unshared chromatin and gene features improves integration of spatial transcriptomic and multimodal data . . . 39
2.8 The inclusion of non-orthologous genes improves the integration of cross-species data . . . 41
2.9 Addition of unshared features is not time or memory prohibitive . . . 56
2.10 Incorporating unshared features is especially advantegous when there are few shared features . . . 57
2.11 Unshared gene loadings largely contribute to nonneuronal type reconstructions . 58
2.12 KNN imputation for missing genes reflect expected distributions . . . 59
2.13 Using unshared genes has a greater improvement on the integration than use of chromatin features . . . 60
2.14 Cross-Species analysis using only orthologous genes . . . 61
2.15 UINMF integration of mouse, marmoset, and human datasets . . . 62
2.16 BCD Algorithm for UINMF Converges Consistently . . . 63
2.17 Time and memory usage of UINMF is competitive with existing methods when integrating STARmap and Dropviz datasets . . . 64
2.18 Time and memory usage of UINMF is competitive with existing methods when integrating STARmap and SNARE-seq datasets . . . 65
2.19 The choice of K does not severely impact the alignment score . . . 66
2.20 UINMF Alignment Score Is Relatively Robust to Lambda . . . 66
2.21 Summary of shared and unshared features used for integration . . . 66

3.1 Overview of approach for linking molecular and anatomical data types across the whole mouse brain . . . 71
3.2 Neuronal cell types exhibit highly distinctive spatial distributions . . . 73

3.3 UMAP integration of the AI Non-neuronal cells . . . 88
3.4 UMAP integration of the AI Inhibitory and Other Neurons . . . 89
3.5 UMAP integration of the AI Excitatory Neurons . . . 90
3.6 UMAP integration of the AUD Non-neuronal cells . . . 91
3.7 UMAP integration of the AUD Inhibitory and Other Neurons . . . 92
3.8 UMAP integration of the AUD Excitatory Neurons . . . 93
3.9 UMAP integration of the CTXsp Non-neuronal cells . . . 94
3.10 UMAP integration of the CTXsp Inhibitory and Other Neurons . . . 95
3.11 UMAP integration of the CTXsp Excitatory Neurons . . . 96
3.12 UMAP integration of the CB Non-neuronal cells . . . 97
3.13 UMAP integration of the CB Inhibitory and Other Neurons . . . 98
3.14 UMAP integration of the CB Excitatory Neurons . . . 99
3.15 UMAP integration of the HB Non-neuronal cells . . . 100
3.16 UMAP integration of the HB Inhibitory and Other Neurons . . . 101
3.17 UMAP integration of the HB Excitatory Neurons . . . 102
3.18 UMAP integration of the HPFENT Non-neuronal cells . . . 103
3.19 UMAP integration of the HPFENT Inhibitory and Other Neurons . . . 104
3.20 UMAP integration of the HPFENT Excitatory Neurons . . . 105
3.21 UMAP integration of the HY Non-neuronal cells . . . 106
3.22 UMAP integration of the HY Inhibitory and Other Neurons . . . 107
3.23 UMAP integration of the HY Excitatory Neurons . . . 108
3.24 UMAP integration of the MB Non-neuronal cells . . . 109
3.25 UMAP integration of the MB Inhibitory and Other Neurons . . . 110
3.26 UMAP integration of the MB Excitatory Neurons . . . 111
3.27 UMAP integration of the MOp Non-neuronal cells . . . 112
3.28 UMAP integration of the MOp Inhibitory and Other Neurons . . . 113
3.29 UMAP integration of the MOp Excitatory Neurons . . . 114
3.30 UMAP integration of the MOs Non-neuronal cells . . . 115
3.31 UMAP integration of the MOs Inhibitory and Other Neurons . . . 116
3.32 UMAP integration of the MOs Excitatory Neurons . . . 117
3.33 UMAP integration of the OLF Non-neuronal cells . . . 118
3.34 UMAP integration of the OLF Inhibitory and Other Neurons . . . 119
3.35 UMAP integration of the OLF Excitatory Neurons . . . 120
3.36 UMAP integration of the ORB_ACA Non-neuronal cells . . . 121
3.37 UMAP integration of the ORB_ACA Inhibitory and Other Neurons . . . 122
3.38 UMAP integration of the ORB_ACA Excitatory Neurons . . . 123
3.39 UMAP integration of the PAL Non-neuronal cells . . . 124
3.40 UMAP integration of the PAL Inhibitory and Other Neurons . . . 125
3.41 UMAP integration of the PAL Excitatory Neurons . . . 126
3.42 UMAP integration of the RSP Non-neuronal cells . . . 127
3.43 UMAP integration of the RSP Inhibitory and Other Neurons . . . 128
3.44 UMAP integration of the RSP Excitatory Neurons . . . 129
3.45 UMAP integration of the SS Non-neuronal cells . . . 130
3.46 UMAP integration of the SS Inhibitory and Other Neurons . . . 131
3.47 UMAP integration of the SS Excitatory Neurons . . . 132

3.48 UMAP integration of the STR Non-neuronal cells . . . 133
3.49 UMAP integration of the STR Inhibitory and Other Neurons . . . 134
3.50 UMAP integration of the STR Excitatory Neurons . . . 135
3.51 UMAP integration of the TH Non-neuronal cells . . . 136
3.52 UMAP integration of the TH Inhibitory and Other Neurons . . . 137
3.53 UMAP integration of the TH Inhibitory and Other Neurons . . . 138
3.54 UMAP integration of the VIS Non-neuronal cells . . . 139
3.55 UMAP integration of the VIS Inhibitory and Other Neurons . . . 140
3.56 UMAP integration of the VIS Excitatory Neurons . . . 141
3.57 Identified Cortical Layer Distribution in the SS . . . 142
3.58 Identified Cortical Layer Distribution in the MOp . . . 143
3.59 Identified Cortical Layer Distribution in the MOs . . . 144
3.60 Identified Cortical Layer Distribution in the VIS . . . 145
3.61 Interneuron and Neurogenesis Cell Types of the Corpus Callosum in the MERFISH Dataset . . . 146
3.62 Interneuron and Neurogenesis Cell Types of the Corpus Callosum in the ABA ISH Data . . . 147

4.1 Joint Registration of Spatially Resolved Molecular Profiles with Whole-Brain Vasculature Reveals Differences in Local Vascular Density of Neuron Types . . . 152
4.2 Co-Registration of Connectivity and Cell Types Reveals Cell-Type-Specific Connectivity . . . 155
4.3 Neuronal Projections Identify Probable Molecular Cell Type Associations within Distinct Brain Regions . . . 158
4.4 Joint Registration with c-Fos Imaging Data Implicates Neuronal Subtypes in Stress-Induced Neuron Activity . . . 162
4.5 Correlation between Vascular activity score and cell type prevalence . . . 172
4.6 Correlation between neuronal activity and vascular score . . . 173
4.7 Correlation between c-Fos activity score and cell type prevalence . . . 174

LIST OF TABLES

TABLE

3.1 Cell Counts by Dataset and Region Prior to Quality Control 84
3.2 Cell Counts by Dataset and Region After Quality Control 85
3.3 Parameters for Online iNMF Analyses 86
3.4 Parameters for SiNMFiD Deconvolution Analyses 87

LIST OF ACRONYMS

AAV adeno-associated virus

ABA Allen Brain Atlas

ARI Adjusted Rand Index

ATAC Assay for Transposase-Accessible Chromatin

AUD Auditory Areas

AV Anteroventral Nucleus of Thalamus

ANLS Alternating Nonnegative Least Squares

BCD Block Coordinate Descent

BLA Basolateral Amygdalar Nucleus

BRAIN Brain Research Through Advancing Innovative Neurotechnologies

CCA Canonical Correlation Analysis

CCF common coordinate framework

cDNA Complementary DNA

CEA Central Amygdalar Nucleus

CeL Lateral Division of the Central Amygdala

CGE Caudal Ganglioinc Eminence

CT cortico-thalamic

DROPviz Drop-seq Visualization

DNA Deoxyribonucleic Acid

ENCODE Encyclopedia of DNA Elements

ET Extratelencephalic Projection Neurons

FOSCTTM Fraction of Samples Closer than the True Match

FRP Frontal Pole

GABA gamma-aminobutyric acid

HPF Hippocampal Formation

Hi-C High-Throughput Chromosome Conformation Capture

iNMF Integrative Non-negative Matrix Factorization

ISH in situ hybridization

IT Intratelencephalic

KNN K-Nearest Neighbors

LM Lateral Mammillary Nucleus

MBO Mammillary Bodies of the Hypothalamus

MGE Medial Ganglionic Eminence

MM Medial Mammillary Nucleus of the Hypothalamus

MOp Primary Motor Area

MOs Secondary Motor Area

mPFC Medial Prefrontal Cortex

mRNA messenger RNA

NeMO Neuroscience Multi-Omic Analytics

osmFISH cyclic-ouroboros single-molecule fluorescence in situ hybridization method

PCA Principal Component Analysis

PSTN Parasubthalamic Nuclei

PT pyramidal tract

RE Nucleus of Reuniens

RMS Rostral migratory stream

RNA Ribonucleic Acid

rRNA ribosomal RNA

RSP Restrosplenial Cortex

scRNA-seq single-cell RNA sequencing

SiNMFiD Supervised Integrative Nonnegative Matrix Factorization Informed Deconvolution

snATAC single-nucleus ATAC

SS Somatosensory Region

STARmap spatially-resolved transcript amplicon readout mapping

SUM Supramammillary Nucleus

SVZ Subventricular zones

TH Thalamus

TM Tuberomammillary Nucleus

tRNA transfer RNA

UINMF Unshared Integrative Nonnnegative Matrix Factorization

UMAP Uniform Manifold Approximation and Projection

VIS Visual area

VLMC Vascular Leptomeningeal cells

VPM ventral posteromedial nucleus

WNN Weighted Nearest Neighbor

WT Wild Type

Xi Xiphoid Thalamic Nucleus

ABSTRACT

The ideology of, "To understand the whole, we must first understand its parts" lies behind much of recent scientific progress, including the Human Cell Atlas Project. To better understand the body, the Human Cell Atlas Project strives to define what cells compose our bodies. A cell's identity is complex, with some defining characteristics including a cell's transcriptome, epigenome, spatial location within an organ and in relation to other cells, and how that cell functions from a physiological perspective. While most technologies typically assess one or two of these aspects of cellular identity, we cannot yet measure all facets in a single assay. Consequently, existing computational methods must focus on stitching these individual fragments of cellular identity into a single, cohesive cellular profile. The focus of this dissertation is the development of computational tools to facilitate the integration of molecular, spatial, anatomical, and physiological data to define more panoramic cellular profiles.

In Chapter II, we develop UINMF, an integrative non-negative matrix factorization algorithm. Most methods are limited to integrating single-cell datasets using only the features that are shared between all datasets in the integration. Leveraging only shared features within an integration is particularly problematic for spatial transcriptomic datasets with few genes and cross-species integrations that are restricted to the set of one-to-one orthologous genes between species. UINMF accommodates the unshared features into each iteration of the objective function, allowing relevant unshared features to help shape the latent space. We show that the use of UINMF to include unshared features improves cross-modality and cross-species analyses.

In Chapter III, we generate a whole-brain molecular mouse atlas by simultaneously integrating six distinct modalities. We integrate epigenomic and transcriptomic single-cell and single-nucleus datasets in 18 regions across the mouse brain, with each region being analyzed in 3 refined sub-analyses: non-neuronal cells, excitatory neurons, and inhibitory and other neurons. After manually annotating each cluster, we derive probable spatial distributions for the molecular cell type profiles by deconvolving two spatial transcriptomics datasets with two separate deconvolution algorithms: an established algorithm, RCTD, and a novel deconvolution algorithm, SiNMFiD. We created SiNMFiD for the unique challenge of deconvolving

expression data with low resolution. Consequently, we construct a whole-brain mouse atlas with jointly defined molecular profiles and accompanying spatial distributions.

In Chapter IV, we leverage a common coordinate framework to integrate the spatially resolved molecular profiles from our whole-brain atlas with anatomical and physiological datasets. Specifically, we explore the correlation between molecular cell types and vascular density, neuronal projection data, and c-Fos activity after foot shock. We determine which molecular cell type profiles are highly and lowly correlated with vascular density and identify distinct cell types that correspond with specific neuronal projection patterns. Using c-Fos data captured after foot shock, we capture the molecular cell types most associated with increased c-Fos activity over time after a foot shock experiment.

CHAPTER 1

Introduction

1.0.1 Cellular Diversity Begins at the Molecular Level

The Brain Research Through Advancing Innovative Neurotechnologies (BRAIN) Initiative's 2025 scientific vision suggests that the first critical step in understanding the complexity of the brain is assembling a parts list of the brain [2]. To assemble such a list, we begin by looking at the small units that are the most basic, foundational units of life: cells. We compose our parts list of the brain beginning at the cellular level because cells come in neat, organized packages, surrounded by a cell membrane [13]. Most eukaryotic cells contain a nucleus or nuclei, the storage house for genetic information [51]. While most eukaryotic cells share these two characteristics [50], we encounter incredible diversity among cell types, both in the natural world, as well as within the human body. For example, muscle cells are long and slender, specialized to contract [6]. At the same time, neurons are equipped with a cell body, a main presynaptic projection (axon), and several smaller postsynaptic projections (dendrites) [145, 152]. The diversity found within a single living organism is even more remarkable when considering the molecular foundation for these cells. Deoxyribonucleic acid (DNA) is the genetic code captured within a cell's nucleus and replicated across the cells of the same organism. What drives the diversity we observe across the cells of an organism is how that code (DNA) is read and interpreted in each cell.

A primary determinant of how DNA is read is determined by epigenomic modifications. Epigenomic modifications alter gene expression by altering either DNA accessibility or chromatin structure [110]. DNA is stored by being wrapped around histone proteins to form nucleosomes [110], and the occupancy and topological organization of the nucleosomes affect the level of interaction between the Ribonucleic acid (RNA) polymerase and DNA [91]. The RNA polymerase is responsible for initiating transcription, or the process of reading DNA into protein. If the RNA polymerase is unable to interact with the DNA because of nucleosome arrangement, the underlying DNA sequence is not converted into protein because it is inaccessible to the RNA polymerase, a condition termed chromatin accessibility. Between

nucleosomes, DNA accessibility to the RNA polymerase is mediated by transcription factors, architectural proteins, and even RNA polymerases themselves, as the thin DNA strand cannot be shared by two RNA polymerases transcribing the same nucleotide [91]. As such, regions with low chromatin accessibility are unlikely to be transcribed, and regions with high chromatin accessibility are more likely to undergo transcription.

Transcription of DNA results in RNA, of which there are three primary types: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) [192]. mRNA contains the genetic blueprint to make proteins, tRNA translates mRNA into proteins, and rRNA forms ribosomes, which support protein synthesis [192]. In addition to epigenomic influences, alternative splicing also impacts gene expression. Alternative splicing, or post-transcriptional modification of RNA, results in multiple distinct transcripts originating from a single gene [43]. While epigenomic influences impact the level of transcription, alternative splicing controls what form of a gene is being transcribed. The final resulting proteins can be used to produce, process, and secrete other proteins and control cell function by mediating cell metabolism and the exchange of materials between the extracellular environment and the cell itself [16]. Each cell begins with the same genetic code. Yet, the implementation of that genetic code varies between cells due to differing epigenomic modifications, such as chromatin accessibility, as well as post-translational modifications, such as alternative splicing, to cumulatively produce the incredible genetic diversity we observe at the cellular level.

1.0.2 Defining A Cell Type

Molecular variance is not the only variable and measurable aspect of cellular identity. Not only can a cell be defined by its molecular properties, but it can also be classified by its anatomy, function, or connectivity. Murkal et. al. propose that biology is essentially the classification of things into groups, and then extrapolating hypotheses and conclusions about those groups [128]. In regard to cells, the objective is to place the cells into groups, called cell types, that share properties and characteristics at a meaningful level. However, doing so is far from trivial. Given the large number of characteristics that can define a cell group, how do we classify the cells into the right groups, and how do we decide what distinctions are significant enough to be meaningful? Currently, the most popular lens for exploring cell types leverages the molecular profiles of single cells. Using the transcriptomic or epigenomic features captured at a single resolution, dimension reduction algorithms can be used to group the cells into clusters of cells that are typically identifiable as a cell type. However, identifying an appropriate resolution during clustering is another source of debate within the scientific community. Often, increasing the resolution of the clustering algorithm will increase

the number of cell types identified, while decreasing resolutions correspondingly reduce the number of clusters. It can be difficult to ascertain the ideal resolution that identifies all meaningful clusters without simultaneously creating supercilious divisions within the data. A trivial division in the data may create separate clusters for cell states instead of cell types. A cell state is a transient period for a cell, a change that could be induced by disease, pharmacological condition, or various metabolic states [207]. While this might induce a slight shift in the molecular composition of a cell, it does not change the primary identity of the cell. However, failure to use a high enough resolution can result in the inability to discover rare cell types. Mukamel et. al. deems the two warring ideologies in the scientific community as "lumpers" and "splitters", after each group's propensity to either combine or separate clusters they are unsure should be labeled as distinct from another [128].

An ideal solution, proposed by Hongkui Zeng, would be to construct a first edition set of clusterings using molecular data, and then curate the proposed cell types using anatomical, connectivity, and functional data [207]. Such an approach would ensure that the identified molecular profiles are simultaneously representative of true differences within the cell profiles at a level that can be objectively defined as significant. Comparison of what molecular clusters can be consistently identified across many samples and timepoints could also be used to help distinguish between clusters representative of cell state versus cell type [207].

1.0.3 Assays to Measure Molecular Properties

Hongkui Zeng's proposal that cell types be first identified using their molecular properties and then curated using anatomical, connectivity, and functional data [207] necessitates accurate and scalable methods for measuring the molecular properties of cells. To illustrate the current landscape of sequencing methods, we will briefly describe some of the most commonly used available transcriptomic and epigenomic sequencing technologies, as well as exciting innovations in developing multi-modal technologies. To appropriately understand currently available technologies and their advantages and disadvantages, it is necessary to also understand the development of those technologies through the previous decades.

Efforts to measure gene expression began decades ago. The Northern blotting technique, one of the first successful efforts to measure gene transcription levels within a sample, was invented in 1977 [73]. To perform a Northern blot, the RNA is extracted from the sample of interest and then run through a gel-like matrix separating molecules by size [28]. The protocol then demands that the RNA is then transferred out of the gel and exposed to probes that contain complementary sequences for the gene of interest [28]. These probes are fluorescently or radioactively labeled and allow for the visualization of the transcript of interest [28]. The

northern blot is a simple, accurate, and cost-effective method for measuring gene expression [127], but it has an incredibly low throughput, as only one gene can be measured at a time.

The low throughput of the northern blot eventually led to the invention of the microarray, which allowed for the first simultaneous measures of gene expression in 1995 [130]. Microarrays utilize glass slides, or chips, with known sequences placed in precise locations [67]. The unknown transcripts are then reverse transcribed into cDNA, fragmented, and attached to fluorescent markers [67]. The unknown transcripts are allowed to bind to their complementary strands on the microarray chips, and unbounded sequences are washed away [67]. Consequently, observed fluorescence at a specific spot on the chip indicates the presence of a specific gene transcript in the unknown sample. Microarrays were revolutionary because they simultaneously measure gene expression for multiple transcripts within a sample but demand predefined sequences for the construction of the microarray chip and fail to quantify whether the detected transcripts are highly or lowly expressed [118]. Similar transcripts sometimes erroneously hybridize the incorrect complementary sequence, an experimental artifact known as cross-hybridization [118].

The invention of RNA-seq addressed some of the limitations of microarrays since RNA-seq was immune to cross-hybridization and did not necessitate the manual specification of transcripts of interest. RNA-seq protocols use reverse transcription to convert RNA to cDNA, and then fragment the cDNA [196]. Then, sequencing adaptors can be attached, next-generation sequencing technologies can be used to sequence the reads, and bioinformatics tools can be used to process and align the resulting sequence data [196]. RNA-seq robustly captures the full transcriptome profiles of a population of cells, not only those transcripts encoded onto the microarray chip [144]. However, RNA-seq measures the gene expression levels of the sample in aggregate or bulk. For samples composed of heterogeneous cell populations, this results in a substantial loss of information about a unique gene expression pattern relevant to only a single cell type or subgroup of cells.

To understand how disease differentially impacted cell type populations and to estimate what cell types composed a sample, transcript profiling at the single cell level was necessary. In 2009, Tang et al. were the first to capture gene expression information at single-cell resolution by using microscopy to manually select a single cell at a time for sequencing [177]. While successful, this technique was clearly not extensible for large cell populations. Applications of fluorescence-activated cell sorting, used by Smart-seq [143] and Smart-seq2 [137], to separate individual cells increased the number of cells that could be sequenced in a single experiment to the hundreds [82]. However, in 2015, Macosko et al. introduced Drop-seq, an innovative technology that leveraged droplets to capture individual cells, increasing single-cell experimental output thresholds to thousands of cells [116]. Continued

experimentation with single-cell sequencing protocols aimed to increase both the number of cells sequenced and read depth while reducing cost. By 2020, more than 1000 single-cell transcriptomics studies had been published [173], underscoring the scientific communities' interest in understanding the transcriptional landscape of cellular profiles.

The capture of epigenomic features is equally critical in our understanding of cellular function and profiles. The foundational protocol for epigenomic sequencing, transposase-accessible chromatin using sequencing (ATAC-seq), debuted in 2013 [34]. ATAC-seq uses a hyperactive prokaryotic transposase, Tn5, to simultaneously fragment and tag the genome with sequencing adaptors [34]. The ability of Tn5 to access, and subsequently fragment and tag, the genome largely depends on the chromatin state of the genome. Therefore, it is expected that areas of the genome that are open, i.e., low on chromatin, will be frequently tagged and sequenced. In contrast, regions of high chromatin and low accessibility will produce few, if any, reads. The ATAC-seq protocol was novel in two specific ways. First, the sequencing protocol was logistically less complex than other methods, such as FAIRE-seq [64] and DNase-seq [164]. The ATAC-seq protocol required dramatically less time to complete and could be performed on much smaller groups of cells, eliminating the necessity of growing the samples in vitro [34]. Scientifically, ATAC-seq provided the ability to concurrently measure chromatin accessibility, nucleosome positioning, and transcription factor occupancy [34]. ATAC-seq revolutionized the study of epigenomic features by providing a more comprehensive view of epigenomic state but failed to account for cellular heterogeneity.

ATAC-seq measured bulk samples and could not distinguish individual cellular epigenomic states. As epigenomic markers can vary substantially between cell populations, and even within cells of the same broad cell type, it was critical that an assay be developed that could measure epigenomic markers at a single-cell resolution. In 2015, Buenrostro et al. expanded their original ATAC-seq protocol to leverage Fluidigm, a microfluidics platform, to obtain single-cell ATAC-seq reads [35]. Subsequent protocol developments have utilized the same basic ATAC-seq strategy to expand the capabilities of measuring the epigenome at the single-cell scale. For example, single nucleus ATAC-seq was developed primarily to capture epigenomic profiles from tissues that have been frozen [139]. Single-cell/nucleus ATAC-seq allows the exploration of chromatin accessibility within a cell and can be computationally integrated with scRNA-seq to compare epigenomic regulation with transcriptomic output at the cell cluster level. However, directly examining how epigenomic regulation impacts the transcriptome at the single-cell level was not possible using scATAC-seq and scRNA-seq measurements from different cells.

Single modality sequencing techniques, such as scATAC-seq and scRNA-seq, can only capture one type of information encoded within a single cell. To understand how the cell

phenotypes we observe are created and maintained, it is critical to assess how the different facets of cellular identity (i.e., methylation, chromatin accessibility, proteins, transcription, etc.) interact with one another [96]. The desire to probe these interactions led to the creation of multi-modal technologies capable of measuring more than one modality within a single cell during sequencing.

Molecular multi-omic technologies have focused on pairing transcriptomic data with chromatin accessibility or methylation measures. In 2019, Chen et al. presented single-nucleus chromatin accessibility and mRNA expression sequencing (SNARE-seq), a droplet-based multi-omics method for capturing chromatin accessibility and transcriptomic information from the same cell [44]. A year later, a well-based method, simultaneous high-throughput ATAC and RNA expression with sequencing (SHARE-seq), provided joint expression profiles while allowing for higher throughput [115, 53]. To capture methylation and transcriptomic profiles, scMT-seq cleverly separates the cytosolic and nuclear DNA, performing RNA-seq on one and DNA methylation profiling on the other [76]. As the demand for multi-modal molecular datasets has increased, companies such as 10X Genomics have responded by developing commercially available assays, such as the Single Cell Multiome ATAC + Gene Expressing Sequencing kit, increasing the accessibility of multi-omics analyses [12].

Multi-modal analyses are not constrained to examining only molecular features. Perturb-seq subjects cells to CRISPR-cas-9 mutations before capturing their transcriptomic profiles, linking genomic mutations with transcriptional variability [54]. To characterize cells by both their transcriptomic features as well as their proteome, Stoeckius et. al. developed CITE-seq [168], a method that has now been made publicly available. The rapid commercialization of several multi-modal technologies, including CITE-seq and the 10X Genomics Single Cell Multiome ATAC + Gene Expressing Sequencing kit [12], underscores the scientific community's interest in such assays. Because of their recent development, the widespread use of multi-modal data in atlas generation efforts has been limited, but Baysoy et al. conjecture that cell atlases will soon expand to accommodate these incredible new datasets [27].

1.0.4 Computational Methods for Integrating Molecular Datasets

Integration of single-cell datasets is a powerful computational tool and can be effectively applied in various frameworks to facilitate the discovery and validation of different hypotheses. When integrating datasets derived from the same modality, where the data sources are expected to exhibit similar properties, integration can help eliminate batch effects caused by different sequencing protocols or platforms and confirm that the identified clusters and marker genes are consistent across samples [149]. Integrating datasets from the same modal-

ity can also help identify differences, as well as similarities, found between datasets. Using integration, it is possible to distinguish how cluster composition changes between diseased and healthy samples or between cases and controls in the event of an experimental condition. Combining datasets from different species can point to areas of evolutionary conservation or divergence. Integrating molecular profiles derived from the same modality validates cell types, confirms marker genes, and facilitates the examination of disease states.

Dataset integrations are not limited to data originating from the same modality, and data from different modalities, such as scRNA-seq and snATAC-seq, can also be represented in a jointly defined latent space. In such cases, where a cluster of cells contains both transcriptomic and epigenomic molecular measurements, cell type profiles can also be jointly described, adding a dimension to cell type characterization. These joint cell profiles are more robust as they are identifiable across modalities. Relationships between epigenomic regulation and transcriptomic output for each cell type is a further area for exploration when performing cross-modality integrations.

Given how much can be achieved with high-quality data integration, it comes as little surprise that there have been substantial and continued efforts to construct robust computational algorithms that address all necessary concerns when performing data integration. Ryu et al. neatly categorize these integration algorithms into three primary types: linear decomposition methods, generative models, and similarity-based batch correction with dimensionality reduction [149].

Methods that utilize linear decomposition, such as ComBat [212] and limma [163], assume that all cells within a dataset have been equally impacted by a condition or artifact when performing integration or batch correction [149]. Because of this, linear reduction methods have been shown to generally perform poorly against other integration strategies when integrating single-cell datasets [114, 186], and therefore do not need to be detailed further.

Generative models, such as scVI [109] and scGEN [111], primarily use variational autoencoders to capture non-linear effects within the data [149]. A recent benchmarking study identified scVI as one of the top-performing algorithms across five scRNA-seq integration tasks [114]. However, when integrating scRNA-seq data with scATAC-seq data, summarized into windows and peaks, several similarity-based batch correction with dimensionality reduction tools outperformed scVI [114], highlighting the need for a variety of tools to exist within the space, as each has its distinct advantages.

The top three algorithms within the similarity-based batch correction with dimensionality reduction space are Harmony, Seurat, and LIGER [186]. Each performs a different dimensionality reduction technique to project the cells into a reduced space, and then similar cells are identified across batches or datasets to assemble clusters [149]. Harmony uses PCA for

dimensionality reduction, and, because of its efficiency, is often recommended as a first pass analysis [186]. Leveraging Canonical Correlation Analysis (CCA), Seurat creates a shared subspace [170]. Pairs of mutual nearest neighbors are then identified and leveraged as anchor points during dataset integration [170]. Neither PCA nor CCA provides biologically interpretable factor loadings and can fail to remove all sources of dataset-specific technical variation. LIGER, an integrative non-negative matrix factorization algorithm, addresses these concerns [197]. The nonnegativity constraint of the factorization produces interpretable factor loadings, and a distinct matrix V_i is created to capture the dataset-specific artifacts for each sample or modality within the integration. With the increasing size and number of single-cell sequencing datasets, a critical evaluation criterion is the scalability of the integration methods.

The size of single-cell datasets is growing exponentially, necessitating the development of algorithms that process such data efficiently, both in terms of computational speed and memory usage. Online learning is one solution. Gao et al. proposed online iNMF, an expansion of the LIGER toolkit, which processes the data in mini-batches, allowing large datasets to be integrated on a standard laptop with minimal memory usage [59]. Additionally, integrated datasets are increasingly being made available online for perusal and further analysis, as is the case with Cell X Gene [3], the Whole-Brain Slide-seq Portal [4], and the whole-brain MERFISH portal through the Allen Brain Atlas [1]. Memory constraints on personal computers are increasingly becoming a consideration in single-cell analysis, and more algorithms will likely move online in the near future.

Another area of development is the analysis of multi-modal datasets. Most existing integration methods were developed with single modality datasets in mind. While current algorithms can integrate multi-modal features, such as scRNA-seq and snATAC-seq, as if they come from independent datasets, they fail to account for the paired aspect of the information present within the multi-modal datasets. The Weighted Nearest Neighbors algorithm, developed by Hao et al., is novel in this regard [71]. The WNN algorithm begins by separately clustering each modality of a multi-modal dataset and then calculates the weight that each modality should be given in the joint graph [71]. To weigh the information from each modality appropriately, an average profile for each cell's K nearest neighbors is constructed, and its similarity to the measured profile is calculated [71]. In this way, the informativeness of each modalities' embedding can be evaluated and adjusted when constructing the final WNN graph. The WNN algorithm heralds a new era for single-cell dataset analysis as the first algorithm capable of jointly considering paired, multi-modal measurements.

1.0.5 Spatial Transcriptomics is an Important Aspect of Cellular Identity

Spatial location is another critical determinant of cellular identity. Tian et al. suggest three primary derivatives of sampling spatial transcriptomic cellular profiles: the construction of cell atlases, the discovery of cellular gradients, and the establishment of cell-to-cell communication [181]. Cell atlases, defined only by their molecular profiles, allow researchers to understand broadly what cell types are present within a sample or organ. The addition of spatial information to these molecular profiles enables the categorization of cells based on spatial distribution and allows for the identification of cell types that may be molecularly similar but spatially distinct. Similarly, gene expression and, consequently, cell type location, are rarely defined by strict boundaries. Even within a complex organ like the brain, the transition between neuronal layers is gradual, a gradient rather than a harsh boundary. Spatial transcriptomics allows researchers to examine gradient relationships throughout a tissue, providing intuition into how local cell types might interact and the developmental trajectories of the cells in question. Lastly, spatial transcriptomics identifies neighboring cells and relevant cell neighborhoods, directly implicating what cells can most probably share extracellular products and signals. Declared Method of the Year in 2020 by Nature Methods [119], spatial transcriptomics comprises a powerful suite of tools that is just beginning its contribution to our understanding of cellular composition, tissue organization, and cellular signaling.

1.0.6 Assays for Capturing Spatial Transcriptomics Data

Currently, there are two primary classes of spatial transcriptomic technologies: sequencing-based and microscopy-based [181]. Contiguous with their status as pioneering technologies, spatial transcriptomics assays typically must choose to prioritize one or two key parameters, i.e., scale, resolution, and sensitivity [125], while relaxing expectations for non-prioritized characteristics. The ultimate goal is a spatial transcriptomic technology that balances all desirable features equally, but current technological limitations have necessitated focusing on developing excellence in a subset of assay characteristics.

Sequencing-based spatial transcriptomics technologies barcode the RNA with its spatial location and perform next-generation sequencing [125]. The first commercial method that utilized sequencing-based spatial transcriptomics is the 10X Visium platform. The first-generation Visium platform used a spatial microarray robot [181, 46] to barcode the RNA within 55 μm diameter spots. Each spot captured the transcriptomic profiles for 1 - 10 cells, lacking single-cell resolution [9]. Recent updates to their protocol have resulted in Visium

HD, which can measure gene expression in 2 μm blocks across a tissue sample [10]. Another popular NGS-based spatial transcriptomics method is Slide-seqV2, which uses 10 μm beads to cover the surface of a glass coverslip [117]. The location of each bead on the coverslip is assayed before the addition of the tissue sample [125]. Similarly, Stereo-seq uses rolling circle amplification to create DNA nanoballs that achieve close to single-cell resolution capture of transcripts [181, 40]. Historically, NGS-sequencing-based methods have generally lacked single-cell resolution but have provided complete transcriptomic profiling of a sample (i.e., they have not demanded a manually curated gene-panel). However, the rapid evolution of technology continues to bring improved resolution and transcript capture efficiency.

Microscopy-based methods of spatial transcriptomics sequence RNA transcripts within the tissue, using fluorescence and microscopy [181]. Some significant challenges for microscopy methods include field of view, fluorescence signal specificity and sensitivity, and the need for a predefined gene panel [125]. For example, spatially-resolved transcript amplicon readout mapping (STARmap) cleverly used padlock probes to amplify fluorescence signals and manipulated hydrogel technology to provide the first 3D spatial transcriptomic profile [193]. However, the 3D sample profiled only 160 unique gene transcripts, limiting the definition of high-resolution cellular profiles within 3-dimensions [193]. Although Multiplexed Error-Robust Fluorescence in situ Hybridization (MERFISH) currently is constrained to 2-Dimensional signal collection, its high throughput and resolution have earned the technology a spot on Nature's Technologies to Watch in 2018 list [138]. The innovations of MERFISH include combinatorial labeling and sequential imaging, which enabled the profiling of 10,000 targets [46, 42, 193], and an error robust barcoding system that helps to correct sequencing errors before they compound. The advantage of MERFISH and other microscopy-based methods is the imaging of the RNA molecules within intact tissue.

1.0.7 Computational Methods for the Deconvolution of Spatial Transcriptomics Data

Ultimately, current efforts by various groups to establish molecular atlases with a spatial component, such as the whole-brain mouse atlas, have utilized both classes of technologies with favorable results [94, 209]. Areas for technological improvement across all spatial transcriptomic technologies include expanding the throughput of samples and expanding to 3-dimensional tissue sections. Additionally, as multi-omic methods grow in popularity, there will be increased demand for spatial assays that measure epigenomic modifications and protein expression [181]. Existing methods for these measures, such as DBiT-seq [107] and Epigenomic MERFISH [113], will likely be continuously refined [46]. Monitoring tran-

scriptomic, epigenomic, and protein expression within a cell in vivo is a long-term objective of the scientific community [181], and current spatial transcriptomic methods are a vital component to the eventual development of such methods.

Before single-cell sequencing, samples were processed as an aggregate, or bulk sample. While this provided average gene expression across a tissue, researchers were left to guess about the cellular composition of such samples and the gene expression of specific cell types [23]. After the advent of single-cell sequencing, researchers could define more accurate average expression profiles for cell types, and it was possible to discern the ratios of cellular subpopulations. After the advent of scRNA-seq, bulk sequencing was still cheaper, and many experiments had already been performed using bulk sequencing. To retroactively instill further utility in previously studied bulk-sequenced datasets, algorithms like dampened weighted least squares (DWSL) [187] were created. Using single-cell data, these algorithms defined averaged gene expression profiles for subpopulations of cells and then leveraged these profiles to perform deconvolution, an estimation of the cell type proportions present within a sample. Deconvolution allowed researchers to estimate the cellular composition of a bulk sample.

Although spatial transcriptomic assays are consistently being refined and developed, the two primary methods currently available each have drawbacks. Image-based or microscopy-based methods are often limited in terms of the number of predefined genes they are capable of capturing. Next-generation sequencing methods with spatial barcoding, like Slide-seq [148] and the first-generation 10x Visium [12], can extensively sample the transcriptome but have yet to achieve single-cell resolution. Indeed, depending on their resolution, most of these spot-based methods vary in size from 2 - 100 μm, meaning they can contain anywhere from 1 - 100 cells, respectively [210]. The use of NGS spatial barcoding methods results in spots of capture that resemble mini-bulk samples, average gene expression for one or multiple cells. Consequently, the already developed idea of deconvolution, or estimating cell type proportions within a bulk sample, could also be applied to spatial transcriptomic data.

Zhang et al. define several classes of spatial deconvolution algorithms: machine learning algorithms, regression-based strategies, statistical-modeling-based strategies, and data-mapping strategies, each with its own strengths and weaknesses [210]. Machine learning models, such as DSTG[165] and STRIDE [172], used graph convolutional neural networks and topic models. Regression strategies, such as spatialDWLS [55], primarily rely on NMF to deconvolve voxels. To estimate cell type proportions present within a specific sample, statistical modeling strategies such as RCTD [37] and Cell2Location [92] leveraged the Poisson and Negative Binomial distributions, respectively. Tangram [30], a data-mapping strategy, is capable of handling accompanying histology information. A recent meta-analysis found

spatialDWLS, RCTD, and Cell2Location to be the most powerful of these powerful tools for spot-based estimation of cell type proportions [98].

A regression-based strategy, spatialDWLS [55] modifies the DWSL [187] algorithm by expanding it into a series of two steps. In step one, the cell types expected to be present within a voxel, calculated on the basis of gene expression within the voxel, are identified [55]. Step two then applies DWSL to determine the fraction of each cell subpopulation [55]. SpatialDWLS performed well across the real and simulated data challenges presented by Li et al [98].

RCTD [37] and Cell2Location [92], two statistical-modeling-based strategies, also produced effective deconvolutions [98]. Robust Cell Type Decomposition, RCTD, utilizes the Poisson distribution to describe each pixel as a linear combination of cell types [37]. The ingenious inclusion of a gene-specific platform effect parameter helps to accommodate for differing platform effects between the spatial transcriptomic and scRNA-seq data [37]. Alternatively, Cell2Location models voxel composition using a negative Binomial distribution [92]. A hierarchical decomposition prior adjusts the Cell2Location model for the non-independence of cell type abundance within a location, and the model is also capable of jointly processing multiple batches [92]. Lastly, Cell2Location can perform estimations of absolute cell type abundances instead of simply providing ratios [92].

Evaluating the performance of deconvolution algorithms such as RCTD is not trivial. Because the ground truth is not known for spot-based spatial transcriptomic methods, researchers create pseudo bulk profiles by gridding single-cell resolution spatial transcriptomics datasets, such as those created with the STARmap technology [195], into larger spots, and then combining the gene expression profiles of cells captured within a single grid [98]. Often, it is necessary to downsample the resulting pseudo bulk profiles to a level of read depth typically observed in a spot-based spatial transcriptomics assay, as read-depth can substantially impact the accuracy of deconvolution results [98].

1.0.8 Neuron Morphology

Tracing of neuron morphology, or shape, began as early as the late 19th century when Ramón y Cajal, the father of modern neuroscience, used Golgi stains to examine the morphology of neurons [61]. Morphology describes where the processes of a neuron are and the nature of those processes [95]. Intuitively, the notion that the shape of a neuron effects its function is unsurprising, as the physical overlap and proximity of axon and dendritic processes constrain even which neurons are in direct communication with each other. Notably, overlap is necessary for synapse; overlap does not equate to synapse [188]. Morphology im-

pacts where a neuron is receiving signals from, how it processes those signals, and where it then distributes those signals to.

Morphology dictates where signals are sent and received. In the visual areas of the cortex, visual information is received in Layer 4 of the cortex [95]. It is then passed to Layer 2/3, then to Layer 5, and finally out of the visual areas to other cortical and subcortical regions [95]. Therefore, we expect Layer 4 to have relatively short projections compared to neurons from Layer 5, as Layer 4 neurons primarily project as far as Layer 2/3. In contrast, Layer 5 neurons must span entire cortical areas. Morphology powerfully impacts how information is exchanged within the brain through the organization of neuronal projections and, consequently, neuronal synapses.

Morphology extends beyond orchestrating the physical proximity of cell types, however. Morphology is such a powerful reflection of cellular function that many neuronal cell types share identifiable morphological features. Neurons that have large, pyramidal-shaped soma with apical dendrites replete with spines are overwhelmingly excitatory [162]. In contrast, neurons that lack apical dendrites and have smooth, spine-less dendrites are likely inhibitory neurons [162]. Mapping morphological traits to transcriptionally defined cell types is generally feasible at large-scale resolutions, such as grouping inhibitory and excitatory neurons. Yet, at finer resolutions, transcriptomic groups still exhibit morphological heterogeneity [134]. Consequently, capturing and analyzing neuron morphology patterns is a critical and distinct step in advancing our understanding of the brain.

Capturing morphological features is a 3-step process: staining/labeling, imaging, and data processing. Farhoodi et al. describe four primary ways of labeling morphology, each with its unique advantages and disadvantages: heavy metals, genetic markers, immunolabeling, and direct injection [57]. For example, Golgi body staining, the oldest heavy metal staining technique, injects silver nitrate into the tissue, where it is taken up by a subpopulation of neurons [57]. The Golgi stain has been further developed and modified since its inception in 1873 and remains relevant within neuroscience today [84]. Unlike immunolabeling, it is capable of staining neurons for which we have no selected antibodies, and it stains the entirety of the neuron [84]. However, it does not allow for preselection of the stained neurons. Alternatively, genetic markers use viral tracers and transgenic mice to target specific cell populations [103]. Once injected, viral tracers, such as retroviruses and adeno-associated viruses (AAV), integrate into the target neuron's genome directly, along with a fluorescent protein for visualization [103]. Transgenic mice, created by using prokaryotic injection, bacterial artificial chromosomes, or knock-in techniques [103], transgenically induce a fluorescent protein in the promoter region of a known cell type-specific marker gene [57], but comes with increased breeding costs [103]. Selecting the appropriate labeling method for neuron

morphology experiments is a function of the experimental objective and the desired feature of observation. After staining and imaging, typically accomplished with a confocal laser scanning microscope [77], as in FMOST [66], is complete, the next challenge is tracing the neuron's path in 3D.

While seemingly straight-forward, neurons are actually difficult to trace because they have incredibly complex and chaotic projection patterns [132] and suffer from background noise and inconsistent signal strength [107]. Tracing a neuron involves converting the images to point coordinates while preserving information about the thickness of the processes and the connectivity between points [120]. Only after capturing the 3D representation of a neuron can we perform feature extraction, measuring its size (height, width) and complexity (number of branch points) [95]. To process a neuron's morphology from captured images, the images must undergo segmentation, noise reduction, soma segmentation, and global and local tree segmentation [120]. For neuron segmentation, some methods build neurons from local information, like ViterBrain [21], while other methods prioritize both global and local relationships, like the fast marching spanning tree [203] algorithm [107]. Additionally, Liu et al. details the increasing prevalence of deep learning, with convolutional neural networks being frequently used in neuron segmentation problems [107], including tools such as SRSNet [215] and 3D WaveUNet [100]. Neuron segmentation is the last key piece in capturing neuron morphology, and it is critical to have the neuron tracing be as accurate as possible.

Neuron morphology impacts signal transduction, signal processing, and even where a signal is capable of being transmitted to. It is a crucial piece of neuron identity, as neurons can be characterized by their morphology. Critically, while these morphology-based distinctions can be reflected in transcriptomic profiles, there are morphological differences between shared transcriptomic types [134], indicating that transcriptomic profiles and morphology must be jointly considered when distinguishing distinct cell types.

1.0.9 Cell Types of the Brain

Cell types within the brain are impacted and can be defined by anatomical location, molecular composition, morphological formation, proximity to vascular structures, and their response to specific and varied stimuli. To fully understand a neuron, it is necessary to understand the interaction between these various facets of cell identity and how variability in each will impact a specific cell type. The first, and perhaps most intuitive way, of understanding a cell is by its location within the brain, as each brain region completes specific tasks for maintaining homeostasis and daily functioning.

Cumulatively, the brain can be understood as an intricate mass of gray and white matter.

White matter, composed of heavily myelinated axons [122], makes the fiber tracts that run through the brain. Gray matter, colored by the high concentration of cell bodies within it, performs the actual processing of information [122]. The Allen Brain Atlas has grouped the brain into three primary structures: the cerebrum, brainstem, and cerebellum [194]. The ABA groups the midbrain, hindbrain, hypothalamus, and thalamus as part of the brainstem. S. Ackerman describes the primary function of each brain region in detail [15]. The midbrain controls reflexes and is a relay center for sensory and motor nerve impulses. The medulla and the pons can be found within the hindbrain; these are responsible for coordinating movements like swallowing and complex movements, respectively. Similarly, the thalamus helps to sort sensory information and pass it to the appropriate cortical region. Finally, the hypothalamus stimulates smooth muscles, manages endocrine system response, and controls eating and drinking. On top of the brainstem is the cerebellum, or "little brain" in Latin. This large structure is in charge of maintaining balance and posture, executing deliberate movement, and providing the ability to learn tasks by "rote." The brain's last and largest area is the cerebrum, which contains the olfactory areas, hippocampus, and isocortical areas. The isocortex and hippocampus control processes such as perception, cognition, emotion, and learning [206]. Each brain region has its own delicate architecture and cell-type composition. Therefore, the anatomical position of a cell within the brain profoundly influences its cell type profile.

While the differing anatomical regions of the brain have much variance in their cell type composition, researchers have determined the primary cell types that support brain function across the brain. There are two primary types of cells within the brain: glia and neurons. Neurons receive and transmit chemical or electrical signals, while the glial cells support neuronal function [126].

Glia cells can be predominantly categorized as astrocytes, microglia, oligodendrocytes, and ependymal cells [80, 58]. Astrocytes help regulate blood flow, controlling the homeostasis of water and ions [80], and provide structural support at neuron synapses [58]. The housekeepers of the brain, microglia, remove cellular debris and protect against microorganisms trying to infiltrate the brain [58]. Oligodendrocytes are responsible for forming myelin sheaths around axons, insulating the axons to allow conduction [101]. Lining the brain's ventricles, ependymal cells have cilia that they beat to circulate cerebrospinal fluid [58]. Since glial cells function to support neurons, they were initially not investigated as thoroughly as neurons [80]. Recently, reports implicating glia cells in dysregulation and specific disease pathways have increased interest in understanding the function and variety of glia cells [101], especially at a brain-wide level.

Neurons represent a large amount of the cell type diversity found within the mammalian

brain and can be described by their molecular properties, anatomical location, and morphology. While a specific ontology describing molecular subtypes and their distinctive properties has not yet been created, a common way to broadly categorize neurons is by their neurotransmitter(s). When an electric impulse travels down an axon, it triggers the opening of tiny sacs, or vesicles, at the end of the axon [9]. These vesicles contain a chemical known as a neurotransmitter, which is then released into the synaptic cleft separating two cells [9]. The two primary neurotransmitters within the brain are glutamate and gamma-aminobutyric acid (GABA), where glutamate is the primary excitatory neurotransmitter, and GABA operates as the primary inhibitory signal [157]. Less prevalent neurotransmitters still play critical roles within the brain. Dopamine controls reward and learning systems, and serotonin helps modulate neuropsychological processes [157]. Oxytocin helps shape human behavior, and the establishment of social relationships such as bonding [112]. Acetylcholine, which has been intimately linked to memory formation [151], can also be found within the brain.

While neurotransmitters can provide a preliminary categorization of the neurons present within the brain, differences within individual neurons remain complex and nuanced within neurotransmitter categorizations as well. For example, the glutamatergic neurons of the cortex can also be broadly understood by their morphological projection patterns. Intratelencephalic (IT) neurons project ipsi or bi-laterally within the telencephalon but do not extend into the extra telencephalic areas [159]. Pyramidal tract (PT) neurons reach the brain stem [159], and projections to the thalamus are performed by the cortico-thalamic (CT) neurons [202].

Reproducible identification of cell types across labs and experiments is critical for progressing research, yet establishing the definition of such cell types is hampered by the extensive dimensions in which a cell type can be defined (i.e., morphological, molecular, etc.) [208]. Despite the difficulties of establishing cell profiles that transcend these different facets of cell identity, progressing toward robust definitions of cell types within the brain is driven by the need to understand the brain more deeply. To understand the brain and its disease perturbations, we first need to understand its composition.

1.0.10 Atlas Generation

A large feature of this work is the generation of a whole-organ atlas at molecular scale for the cells within the brain. Single-cell atlas generation has become an increasingly discussed goal, perhaps most ambitiously expressed in the goal of the Human Cell Atlas, "To identify and describe every cell type in the human body" [72]. Generating single-cell atlases is a beneficial endeavor to both immediate and long-term scientific discovery.

The immediate advantage of the production of a single-cell atlas is the compilation of a "parts" list for the organ in question, allowing the identification of what cells are present within an organ. This is critical for defining how the healthy organ functions, and, by assessing the deviations from this norm in the presence of disease, we can gain intuition for how a disease impacts an organ on the cellular level. This discovery facilitates the discovery of disease pathways, as well as the development of more targeted and specific therapies. Returning to homeostasis is a more attainable goal when homeostasis has been defined.

Immediate knowledge is a low-hanging fruit when generating cell atlases. In fact, Dr. Stephen Quake describes single cell atlases as the "phenotypic compliments" to whole genome sequencing. Constructing a cell atlas merely opens the door for each cell type proposed to be interrogated and described, similar to how the function of individual genes were, and continue to be, probed and assessed after the completion of the Human Genome Sequencing Project. An important feature of cell atlases is their ability to serve as a reference for newly sequenced datasets.

The ability to upload cells into the reference space for easy annotation is a significant advantage for three primary reasons. The first of which is the reduction in time and effort it will take to analyze a new dataset. Currently, most single-cell datasets are clustered and then manually annotated, a time-consuming and laborious step that could be avoided or minimized by use of a reference atlas. Another key advantage to using a reference atlas is the immediate standardization of cell type labels. Current nomenclature and labeling strategies differ vastly between data-generating labs, making cross-dataset comparison difficult and finding consensus annotations a non-trivial task. If reference labels are widely used, this problem is mitigated. Last, annotating cells via an atlas reference space helps promote quality cell annotation. Frequently, manual annotation can mean prioritizing specific sets of marker genes, navigating unclear cluster labels, and sometimes even guess work for an untrained annotator. Automating the process of cell type annotation brings a measure of standardization to cell type annotation that eliminates guesswork and vagueness.

While the merits of generating whole-organ atlases are clear, optimal protocols for developing such atlases remain ambiguous. There are an incredible series of steps, from tissue collecting and sequencing, to quality control, normalization, clustering and annotation that must all be refined and strategically coordinated. Other, peripheral challenges include the size of the datasets in question, and how best to organize, process, and disseminate the resulting atlases. Broadly, the Human Cell Atlas defines two primary steps within atlas generation: the single-sequencing of cells that are present in an organ or organism, and then the spatial sequencing of that same organ or organism, such that spatial locations for the derived transcriptomic profiles can be assigned [72]. As we use already existing datasets for

the analyses described herein, our concern is not focused on the collection and sequencing of the data, but rather how to successfully leverage computational tools to effectively integrate the single-cell sequencing information from each dataset into a cohesive picture, and transition those molecular identities into a spatial framework. While the initial construction of databases and the dissemination of results frequently goes through revisions with each successive approach, gaining useability and innovative features with each iteration, we are proud to present a robust, initial whole-brain atlas for use by the scientific community.

1.0.11 Overview of dissertation chapters

This dissertation studies data integration for creating more complete cellular profiles. We first derive and validate a novel integration method that leverages unshared features between datasets for improved dataset integration results. Using an existing algorithm, LIGER, we integrate molecular datasets from the whole mouse brain. We then use a novel deconvolution algorithm to identify how these molecular cell types are spatially distributed. Finally, we register the spatial molecular profiles in the same common coordinate framework (CCF) as several anatomical and physiological datasets and implicate the involvement of specific molecular profiles likely to be associated with higher vasculature, specific morphological projections, or increased neuronal activity as a response to physiological stress.

In Chapter II, we describe UNIMF, an integrative non-negative matrix factorization algorithm capable of incorporating unshared features, those not present in all datasets, into the objective function. Including unshared features is particularly advantageous when integrating datasets with fewer measured genes, as is frequently the case with spatial transcriptomics datasets. We demonstrate the utility of UINMF in integrating spatial transcriptomics datasets with standard scRNA sequencing data and employ UINMF in cross-species integrations to accommodate non-orthologous genes.

In Chapter III, we integrate six distinct datasets, comprising both epigenomic and transcriptomic measures, across 18 non-overlapping regions of the mouse brain. For each region, we perform three refined integrations: non-neuronal, glutamatergic neurons, and GABAeric and other neurons. Cumulatively, we perform 54 individual integrations and manually quality control and annotate the resulting clusters. The results of our dataset integrations have been uploaded to Neuroscience Multi-Omic (NeMO) Analytics portal for dissemination into the scientific community. In Chapter IV, we use deconvolution to determine where our defined molecular profiles can be found within the brain. To ensure that we are getting robust and reproducible results, we perform the deconvolution over two types of spatial transcriptomics data, the in situ hybridization (ISH) data from the Allen Brain Atlas (ABA) [97], as

well as a Slide-seq whole brain dataset [94]. We debut our novel deconvolution algorithm, siNMFiD, and show its performance to be comparable to that of existing algorithms.

In Chapter V, we leverage the common coordinate framework from the ABA to co-register the spatially defined molecular profiles in the same coordinate space as several anatomical and physiological datasets. Using a vascular density dataset [182], we identify molecular cell types with increased and decreased association with vascular density. We use the neuron projection dataset [133] produced by the Allen Brain Atlas to investigate molecular cell types associated with specific mesoscale projection patterns between regions. Lastly, the whole-brain c-Fos dataset, which captures neuronal activity across time in response to foot-shock, allows us to identify specific molecular cell types most likely associated with the foot-shock paradigm response.

In Chapter VI, we summarize the utility of our work for use within the scientific community and discuss avenues for future investigations.

CHAPTER 2

UINMF Performs Mosaic Integration of Single-cell Multi-omic Datasets Using Nonnegative Matrix Factorization

2.1 Preamble

The following work has been published as:

Kriebel, A.R., Welch, J.D. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature Communications* 13, 780 (2022). For the work described in the following chapter, Joshua D. Welch conceived the idea of UINMF. Joshua D. Welch and April R. Kriebel developed and implemented the UINMF algorithm. April R. Kriebel carried out data analyses and created the figures. Joshua D. Welch and April R. Kriebel wrote the paper.

2.2 Introduction

Each cell type or state within an organism is distinguished by its gene expression, epigenetic regulation, and spatial location within a tissue. Single-cell sequencing technologies measure each of these features in individual cells, allowing researchers to classify cells in a data-driven manner. Determining what features are common, or different, between each cell type provides researchers insight into the function of the cell. Comparing the profiles of diseased cells with those of healthy cells also reveals disease-related aberrant features. An ideal characterization of a cell type goes beyond analyzing features such as epigenetic and gene expression individually, instead examining their relationships. Jointly examining cellular features holds promise for understanding gene regulatory mechanisms that control cell fates.

Current single-cell sequencing technologies cannot simultaneously measure all relevant aspects of cell state. Recently developed multi-omics technologies are limited to capturing only a couple of the modality measures that encompass cell identity [179]. In particular, emerging techniques such as the Multiome assay from 10X Genomics can measure gene expression and chromatin accessibility from the same cell [44, 105, 115, 63, 99, 48], but do not generally capture methylation or spatial features. Spatial transcriptomics, named Method of the Year 2020 by Nature Methods [11], encompasses a rapidly growing suite of techniques [124, 195, 69, 49] that interrogate gene expression patterns within intact tissue. However, protocols for spatial measurements of epigenomic state are not widely available.

The different types of features measured by different single-cell technologies create unique computational data integration challenges. Some more recent methods, such as Seurat v4's Weighted Nearest Neighbor (WNN) algorithm [204], are designed for datasets containing multiple modalities measured within the same cells, while other approaches focus on integrating modalities from different single cells into a shared latent space. Most existing computational approaches for multi-omic data integration are designed for either vertical or horizontal integration scenarios [146, 20]. Vertical integration approaches are useful for datasets measured across a common set of samples or cells. Well-established methods for multi-omic integration of bulk data, such as similarity network fusion and iCluster [158, 191], fall into this category, as well as recent methods for single-cell datasets with multiple modalities per cell such as MOFA+, totalVI, and the Seurat v4 weighted nearest neighbors algorithm [36, 19, 62]. Conversely, horizontal integration uses a set of common variables or features to integrate over multiple experiments, typically using shared genes as the anchors for integration. Batch effect correction approaches originally designed for bulk sequencing data (e.g., RUV [147, 78] and ComBat [81]) solve a horizontal integration problem. Similarly, dataset alignment algorithms developed for single-cell data, such as Seurat v3, Harmony, and our previous method LIGER [36, 93, 197, 170] also rely on shared features and can thus be considered horizontal integration techniques.

A recent review of single-cell computational integration approaches identified a third category of approaches as those for diagonal integration, in which neither features nor cells are assumed to be shared across datasets [20]. In this review, Argelauget et al. also identified a type of problem that they call mosaic integration–in which only some features or cells are shared across datasets–and for which they did not identify any existing approaches. Our work here aims to address some of the challenges of mosaic integration.

LIGER, Seurat v3, and Harmony are effective tools for a number of single-cell integration tasks, and were chosen as the best methods in a recent systematic comparison of 14 computational methods [186]. However, these three approaches are all constrained to integrate

across features shared between datasets, and require that the input matrices all contain a common set of genes or features that are measured in all datasets. Thus, these methods cannot incorporate features unique to one or more datasets, such as intergenic epigenomic information.

Restricting single-cell integration analyses to features shared across all datasets is problematic because it often necessitates discarding pertinent information. For instance, scRNA-seq measures transcriptome-wide gene expression within individual cells, but spatial transcriptomic protocols often measure only a chosen subset of all genes. Yet for many applications, we want to integrate scRNA-seq and spatial transcriptomic datasets, which have neither the same number of features (genes) nor the same number of observations (cells). By integrating the datasets using only shared features, we fail to capitalize on the higher resolution provided by the scRNA-seq modality. When integrating cross-species datasets, the integration is restricted to orthologous genes, disregarding all genes without unambiguous one-to-one relationships between species. Likewise, when integrating single-cell epigenomic data with single-cell transcriptomic data, horizontal integration approaches do not take into account the important epigenomic features from intergenic regulatory elements. As a final example, existing methods do not provide a way to leverage paired epigenomic information when integrating data types such as SNARE-seq [44, 63] or 10X Multiome with single-cell or spatial transcriptomic datasets. Such integration analyses do not fit neatly into either the horizontal or vertical integration paradigm, requiring the development of different methods.

The critical need to include unshared features in single-cell integration analyses motivated us to extend our previous approach. We developed UINMF, a nonnegative matrix factorization algorithm that allows the inclusion of both shared and unshared features. UINMF can integrate data matrices with neither the same number of features (e.g., genes, peaks, or bins) nor the same number of observations (cells). Furthermore, UINMF does not require any information about the correspondence between shared and unshared features, such as links between genes and intergenic peaks. By incorporating unshared features, UINMF fully utilizes the available data when estimating metagenes and matrix factors, significantly improving sensitivity for resolving cellular distinctions. Note that UINMF is not designed to jointly leverage multiple modalities from the same cell as in Seurat v4 or MOFA; rather, the approach solves a problem more like the Seurat v3 anchors approach.

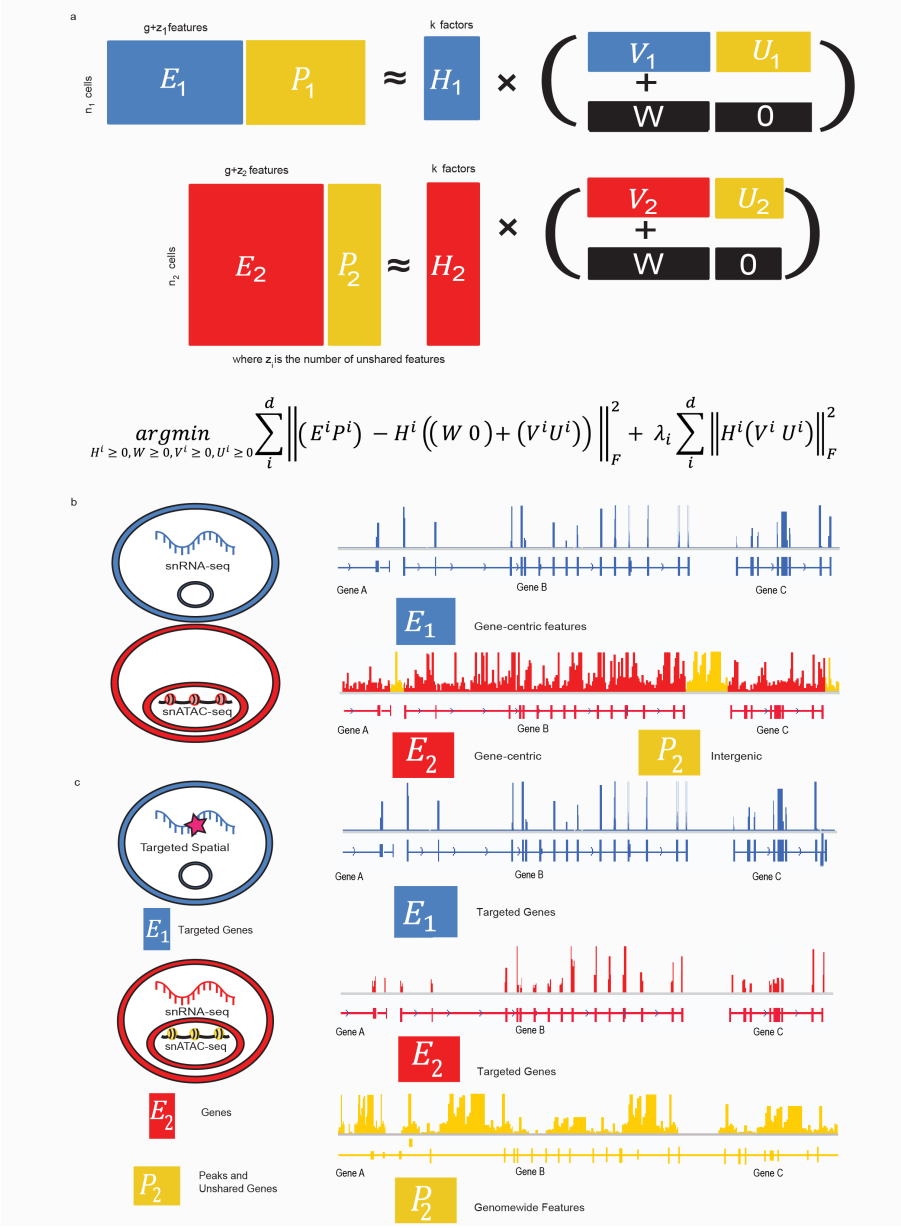
2.3 Results

2.3.1 Integrative nonnegative matrix factorization algorithm for partially overlapping feature sets

The key innovation of UINMF is the introduction of an unshared metagene matrix U to the iNMF objective function, incorporating features that belong to only one, or a subset, of the datasets when estimating metagenes and cell factor loadings. Previously, dataset integration using the iNMF algorithm operated on features common to all datasets [197]. Each dataset (E^i) was decomposed into dataset-specific metagenes (V^i), shared metagenes (W), and cell factor loadings (H^i), and the optimization problem was solved iteratively. By including an unshared metagene matrix (U^i), we provide the capability to include unshared features during each iteration of the optimization algorithm (Fig. 1.1a). Intuitively, the shared features allow for the identification of corresponding cells across datasets, while the unshared features incorporated by U^i simultaneously allow for clearer distinction among the cells within each dataset.

The unshared feature matrix can include extra genes, intergenic features, non-orthologous genes, or any other data type that is measured in one of the datasets. Importantly, UINMF makes no assumptions about the relationship between the unshared features and the shared features; for example, no prior knowledge about linkages between intergenic peaks and genes is required. Instead, such covariance among features is learned during the optimization process, as both shared and unshared features contribute to the reconstruction of the original data through the inferred latent factors. These properties allow for the use of the UINMF algorithm across diverse contexts. For example, U^i can incorporate intergenic information when integrating single-cell transcriptomic and epigenetic datasets (Fig. 1.1b). When analyzing spatial transcriptomic datasets measuring only a few targeted genes, U^i can be used for genes that are not in the targeted panel but are measured in transcriptome-wide scRNA-seq data. With the advent of multi-omic datasets, the flexible nature of U^i is a particular advantage. We can jointly integrate single-modality data with multimodal data, using all of the features present in the multimodal dataset when performing the integration. In this scenario, the U^i matrix can be used for the unshared multimodal data type, such as chromatin accessibility, when integrating SNARE-seq data with spatial transcriptomic or scRNA data (Fig. 1.1c). Another application of the unshared matrix is the ability to include non-orthologous genes into cross-species analyses, leveraging species-specific genes in dataset integration. We demonstrate the functionality of the UINMF algorithm in each of these four possible scenarios, and anticipate that the approach will prove useful for a wide variety of

Figure 2.1: Overview of the UINMF algorithm for integrating single-cell datasets with partially overlapping features



(**a**) Schematic representation of the matrix factorization strategy (top) and optimization problem formulation (bottom). The addition of a factor matrix U_i allows for unshared features to be utilized in joint matrix factorization. Each dataset (E_i) is decomposed into shared metagenes (W), dataset-specific metagenes constructed from shared features (V_i), unshared metagenes (U_i), and cell factor loadings (H_i). The incorporation of the U matrix allows unshared features that occur in only one dataset to inform the resulting integration.(**b**) UINMF can integrate data types such as scRNA-seq and snATAC-seq using both gene-centric features and intergenic information.(**c**) UINMF can integrate targeted spatial transcriptomic data with simultaneous single-cell RNA and chromatin accessibility measurements using both unshared epigenomic information and unshared genes.

future applications.

The UINMF optimization algorithm has a reduced computational complexity per iteration compared to our previous iNMF algorithm [197] on a dataset of the same size (see Methods). Furthermore, the addition of the unshared features in each iteration of the matrix decomposition does not impose a prohibitive increase in time or memory usage in practice. Rather, UINMF demonstrates a moderate increase in time and memory usage due to the unshared features while providing significant benefits compared to algorithms incapable of incorporating unshared features (Supplemental Fig. 1.9).

UINMF, as well as iNMF, requires random initializations, and is nondeterministic in nature. Therefore, unless otherwise noted, we perform UINMF and iNMF with multiple initializations, and use the model of best fit for each respective algorithm. In practice, the use of a single initialization, set with a random seed, is typically sufficient.

As an initial test of UINMF, we investigated how differing numbers of shared and unshared features affect integration quality. To do this, we integrated two single-cell transcriptomic datasets taken from the mouse Primary Motor Area (MOp) [204], sequenced with two different protocols: SMART-seq and 10X Genomics. Using a constant number of total genes (2500), we progressively decreased the number of shared genes from 500 to 100. We then performed two UINMF integrations, one using the SMART-seq data as the source for the unshared features, and the other using the 10X dataset as the source for the unshared features. We calculated the ARI and purity scores for each resulting integration (Supplemental Fig. 1.10).

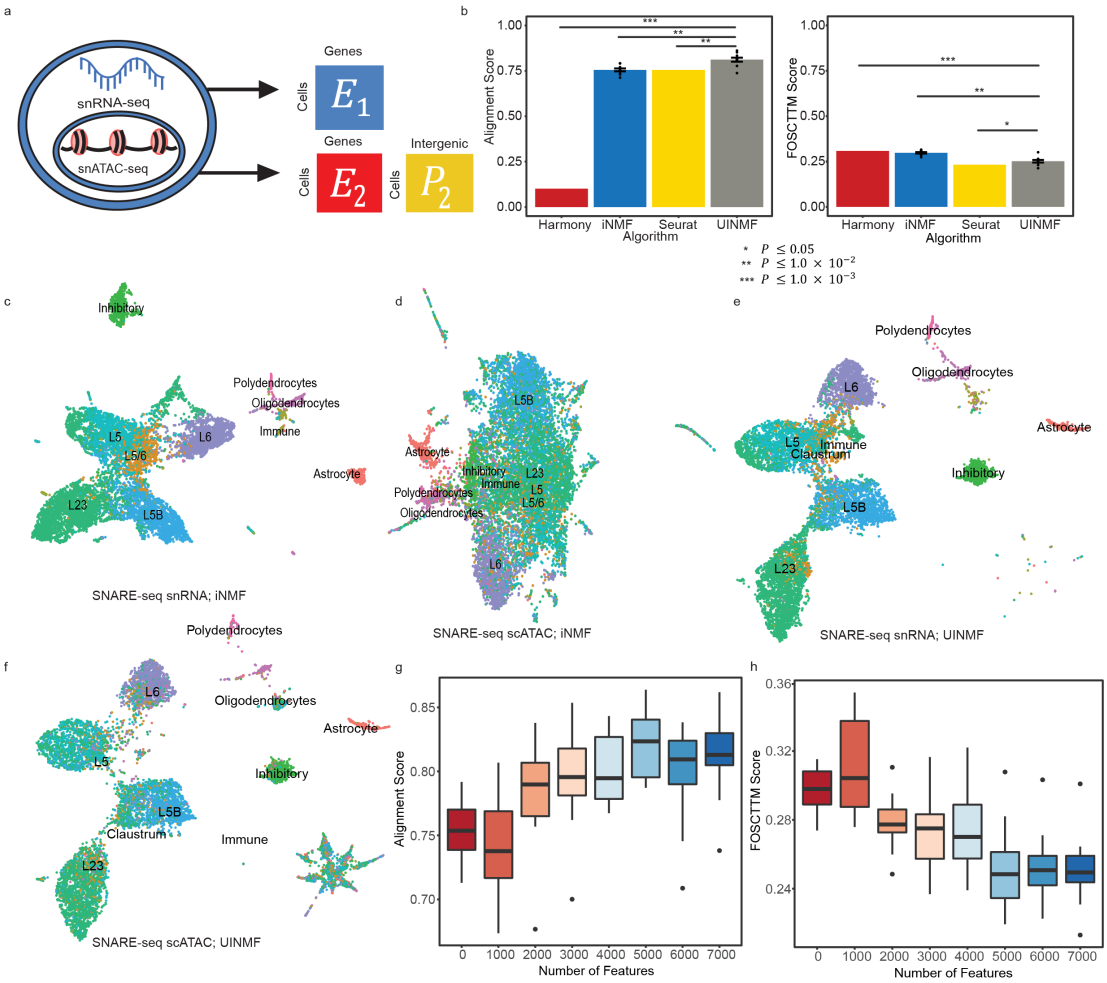
Our results indicate that using unshared genes from the SMART-seq data, known to be higher quality than the 10X data, consistently increases the quality of dataset integration, even with a continued reduction in the number of shared genes. The 10X data, while initially improving dataset integration, is more dramatically impacted by the reduction of shared features. Nevertheless, the benefit of using UINMF to include additional features in dataset integrations is clear, with the amount of benefit derived correlated with the quality of the dataset from which the unshared features are being chosen.

2.3.2 Including intergenic peak information improves integration of scRNA and snATAC datasets

We first investigated how the inclusion of additional features might impact the integration of scRNA and snATAC datasets. In our previous work, we summed ATAC reads that fell within a gene to provide gene-centric ATAC profiles, then used these shared features for integration, neglecting intergenic information [204]. In contrast, UINMF uses the unshared

feature matrix to incorporate the ATAC reads present between genes–intergenic peaks–when estimating the metagenes (Fig. 1.2a). Single-nucleus ATAC data is extremely sparse, with only 1–10% of peaks detected per cell, compared to the 10–45% of genes captured per cell in scRNA sequencing methods [41]. Including the intergenic snATAC data allows more of the detected regions to be used from each cell, a distinct advantage in such sparse datasets. Additionally, the intergenic chromatin peaks provide information about the chromatin state of important *cis*-regulatory elements, such as promoters and enhancers. We hypothesized that the inclusion of this additional information would better resolve molecular distinctions among cells when integrating single-cell transcriptomic and epigenomic datasets.

Figure 2.2: Addition of intergenic peak information improves integration of RNA and ATAC datasets



(**a**) Schematic illustrating how the UINMF algorithm incorporates intergenic peaks when separately integrating the RNA and ATAC measurements from a SNARE-seq dataset. We treat each data type as if it came from an independent source, and perform an integration using regular iNMF and our proposed UINMF method, which incorporates intergenic peaks.(**b**) Average alignment and FOSCTTM (Fraction of Samples Closer Than True Match) scores for iNMF, Seurat v3, Harmony, and UINMF. iNMF and UINMF are both initialized 5 different times over ten random seeds, with UINMF including an additional 7,000 intergenic features into the analysis. For nondeterministic algorithms, data are presented as mean values +/- SEM. To compare algorithm performance, we used a paired, one-sided Wilcoxon test to compare UINMF's alignment and FOSCTTM scores to iNMF ($P = 1.953 \times 10^{-3}, P = 1.953 \times 10^{-3}$), Seurat ($P = 1.953 \times 10^{-3}, P = 0.01855$), and Harmony ($P = 9.766 \times 10^{-4}, P = 9.766 \times 10^{-4}$), with Seurat exhibiting a significantly lower FOSCTTM score. For each algorithm, we compare 10 pairs of data points ($n = 20$). We factorize and cluster the cells using their RNA transcripts (**c**) and chromatin accessibility measures (**d**) separately.

Figure 2.2 *(previous page)*: After integration, we use the known cell correspondences to separately plot the gene expression (**e**) and chromatin accessibility datasets (**f**) from SNARE-seq, colored by the same cell type labels. We assess the contribution of information contained within the intergenic peaks by assessing the alignment (**g**) and FOSCTTM (**h**) scores across a range of included peaks, from 0 unshared features (iNMF) to 7000 unshared intergenic bins, adding 1000 unshared features to each analysis. The bold line indicates the median data value, and the boundaries of each box are defined by the first and third data quartiles (25 and 75%, respectively). The upper (lower) whiskers extend from to highest (lowest) point within 1.5 of the interquartile range. Outliers beyond the whiskers are plotted as points. We calculate FOSCTTM and alignment scores for ten random seeds for each number of unshared features ($n = 10$).

To quantify how leveraging intergenic features improves dataset integration, we analyzed a SNARE-seq dataset [44], which provides gene expression and chromatin accessibility information from the same barcoded cell. Because the RNA and ATAC information is measured within the same single cells, the joint profiles provide ground truth cell correspondence information for assessing integration performance. The RNA and ATAC profiles can be pre-processed and integrated as if they come from separate datasets. Subsequently, the success of the integration can be measured by how closely the ATAC and RNA profiles for the same cell are aligned.

We evaluated the quality of SNARE-seq integrations using the Fraction of Samples Closer Than the True Match (FOSCTTM) [105] metric. The FOSCTTM metric assesses how closely the RNA barcoded cell is placed to its corresponding ATAC barcode in the latent space. Lower FOSCTTM scores are better, indicating that the RNA and ATAC profiles from the same cells have been correctly placed near each other. We also calculated an alignment metric; because the RNA and ATAC datasets come from identical cells, perfect alignment is theoretically achievable and thus the ideal performance is an alignment score of 1.

We assessed the benefit of incorporating intergenic peaks by comparing the UINMF algorithm with our previously published iNMF algorithm [197], which uses only gene-centric features. Over multiple random initializations, iNMF obtained an average FOSCTTM score of 0.2977 and an average alignment score of 0.756 (Fig. 1.2b).

In contrast, UINMF achieved a significantly lower average FOSCTTM score of 0.251 ($P = 1.953 \times 10^{-3}$, paired one-sided Wilcoxon test), as well as a significantly higher average alignment score of 0.812 ($P = 1.953 \times 10^{-3}$, paired one-sided Wilcoxon test). These findings indicate that incorporating unshared features from the chromatin accessibility data improves the integration of scRNA and snATAC datasets.

We also compared UINMF with Seurat v3 and Harmony. UINMF significantly outper-

forms Harmony in both alignment ($P = 9.766 \times 10^{-4}$, paired one-sided Wilcoxon test) and FOSCTTM score ($P = 9.766 \times 10^{-4}$, paired one-sided Wilcoxon test). Compared to Seurat v3, UINMF had a significantly improved alignment score ($P = 1.953 \times 10^{-3}$, paired one-sided Wilcoxon test), however, Seurat v3 did have a significantly better FOSCTTM score than UINMF ($P = 0.01855$, paired one-sided Wilcoxon test).

We also confirmed that the RNA and ATAC profiles were mapped to similar cell types. To do this, we manually annotated the cell type labels using marker genes from the scRNA data only. Before integration, the clusters separated much more clearly from gene expression data alone than from chromatin accessibility data alone (Fig. 1.2c, d). After UINMF integration, the cluster labels aligned well across datasets (Fig. 1.2e, f), indicating that UINMF identified corresponding cell types even though the single-cell correspondence information was not used by the algorithm. In summary, including intergenic chromatin accessibility information in the integration of scRNA and snATAC data significantly improved the integration results.

To examine how the number of intergenic peaks used affected the resulting integration, we assessed the FOSCTTM and alignment scores for increasing numbers of unshared features. We began by performing the analysis with 0 unshared features (iNMF), taking the best of 5 initializations over 10 seeds, and calculating the alignment and FOSCTTM scores. We then added the 1,000 most variable intergenic bins, and repeated the analysis and metric calculations. Iteratively, we added the next 1000 top variable intergenic bins until we reached 7000 unshared features. As the number of variable intergenic bins used increases, the alignment (Fig. 1.2g) and FOSCTTM (Fig. 1.2h) scores show a corresponding stepwise improvement, indicating that each additional set of intergenic bins improves the overall integration. Moreover, the increasingly refined metrics of integration suggest that there is a substantial contribution of information even in the last 2000 variable bins added, as integrating with 7000 intergenic bins has less variance in metric scores than integrating with 5000 intergenic bins.

2.3.3 Leveraging additional genes improves integration with targeted spatial transcriptomic technologies

We expect the UINMF algorithm to be especially effective for integrating targeted spatial transcriptomic datasets, as the number of genes measured in such datasets is often limited. Integrating such datasets with scRNA-seq provides the opportunity to pair transcriptome-wide profiles from dissociated cells with spatial transcriptomic data. This mitigates the loss of sensitivity for distinguishing cell types, while mapping cell types to their spatial positions within a tissue.

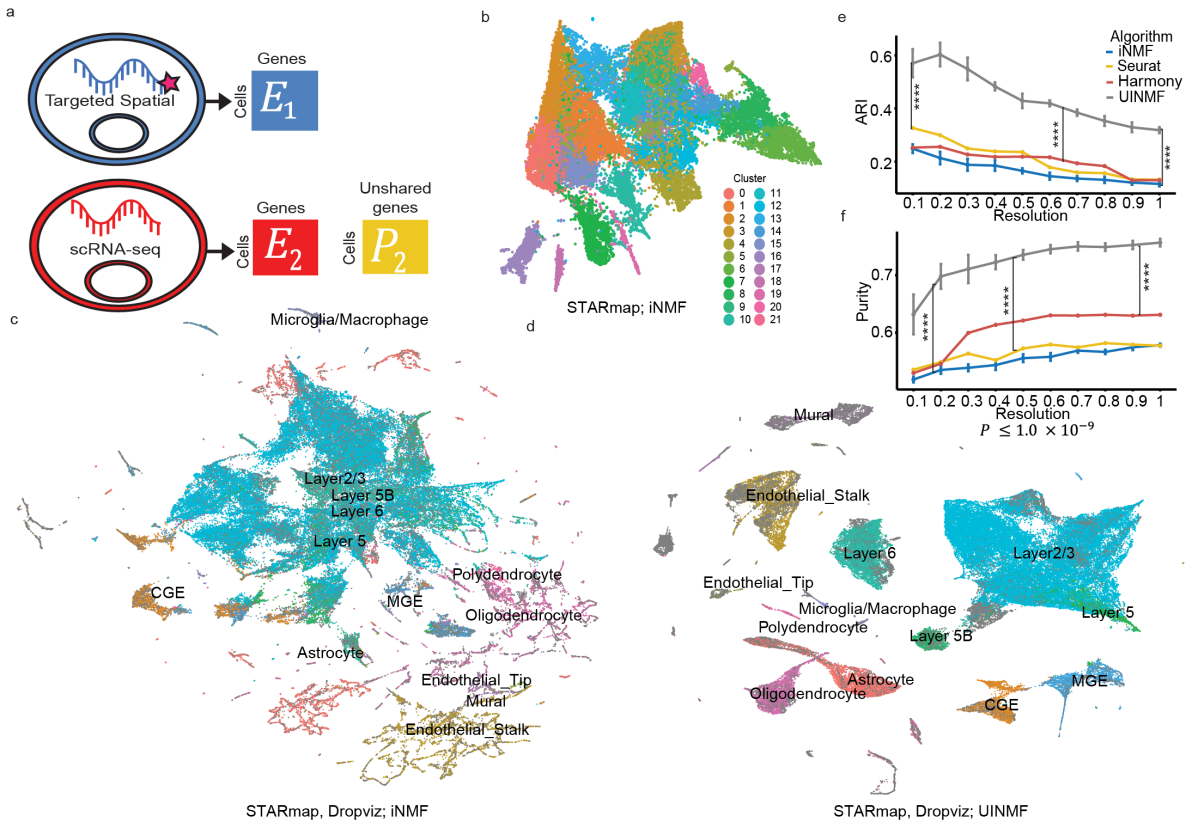
To explore the utility of the UINMF algorithm for integrating targeted spatial transcriptomics and scRNA-seq datasets, we analyzed STARmap data[195] and scRNA-seq data [153] We used a STARmap dataset that contains spatial position and transcription level for 28 genes across 31,294 cells within a 3D block of tissue from the mouse frontal cortex. To our knowledge, this dataset is unique in that it is the only dataset of spatially resolved gene expression for multiple genes within a 3D tissue block.

Even though the 28 genes are selected to distinguish among cell types, the STARmap data fails to clearly separate cortical cell types after matrix decomposition (Fig. 1.3b). Therefore, we integrate the spatial transcriptomics data with scRNA sequencing data from Saunders et al.[153] and perform Louvain community detection. We annotate the scRNA-seq cells using their previously published cellular annotations [153]. We observe improved cluster resolution using iNMF with only shared genes to integrate the STARmap dataset with the scRNA-seq dataset (Fig. 1.3c), but some distinct cell types are mixed together, while others are arbitrarily split. The original cluster labels of the scRNA-seq data are not very well-preserved in the resulting clusters. For example, no distinct boundaries are apparent between the Layer 6, Layer 5, and Layer 5B excitatory neurons. The mural cells are likewise ill-defined. Using UINMF to incorporate 2775 more genes into the integration allows the metagenes to be estimated from a broader array of genes. Consequently, the addition of these unshared features results in dramatically clearer clusters that much better reflect the ground truth labels (Fig. 1.3d). The distinction between the excitatory neuron subtypes become clear, and a defined population of mural cells also becomes distinguishable. Thus, using the unshared features, it is possible to identify cell types that would not be otherwise distinguishable.

To quantify the advantage of the UINMF method, we calculated Adjusted Rand Index (ARI) and purity metrics for multiple initializations of the UINMF and iNMF algorithms across a range of different Louvain resolution parameters (Fig. 1.3e, f). The UINMF algorithm achieves significantly ($P = 3.895 \times 10^{-10}$), paired one-sided Wilcoxon test) higher ARI and cluster purity compared to the iNMF algorithm. UINMF also significantly outperforms both Seurat v3 and Harmony in both ARI and purity metrics ($P = 3.895 \times 10^{-10}$, paired one-sided Wilcoxon test). In short, the addition of the unshared genes significantly improves the integration of STARmap and single-cell RNA-seq data.

We next examined the contribution of each part of the matrix decomposition (W, V^1, V^2, and U) for each factor (methods). Both U and V are regularized in the UINMF objective to encourage the reconstruction to come primarily from the shared factors. Consistent with this regularization, we found that W, the shared metagenes, contributed the most to the matrix reconstruction, and the dataset-specific V^1 and V^2 contributed the least (Supplemen-

Figure 2.3: Incorporating additional genes improves integration with STARmap data



(**a**) Schematic of UINMF integration of the spatial transcriptomic data with the scRNA-seq data, in which U incorporates unshared genes that are captured in scRNA-seq but not in the targeted STARmap data. (**b**) UMAP of STARmap data alone. (**c**) UMAP of STARmap and scRNA integration performed with iNMF using only shared genes. d UMAP of UINMF integration, which incorporates both shared and unshared genes. Both (**c**) and (**d**) are annotated using the original scRNA-seq labels and clusterings derived from either the iNMF (**c**) or UINMF (**d**) algorithm. We compared the adjusted rand index (**e**) and purity (**f**) metrics from UINMF with iNMF ($P = 3.895 \times 10^{-10}, P = 3.895 \times 10^{-10}$), Seurat v3 ($P = 3.895 \times 10^{-10}, P = 3.895 \times 10^{-10}$), and Harmony ($P = 3.895 \times 10^{-10}, P = 3.895 \times 10^{-10}$) for a range of clustering resolutions, using a paired, one-sided Wilcoxon test ($n = 100$ values for each test, with each sample having five values across 10 resolutions). For nondeterministic algorithms, data are presented as mean values +/- SEM.

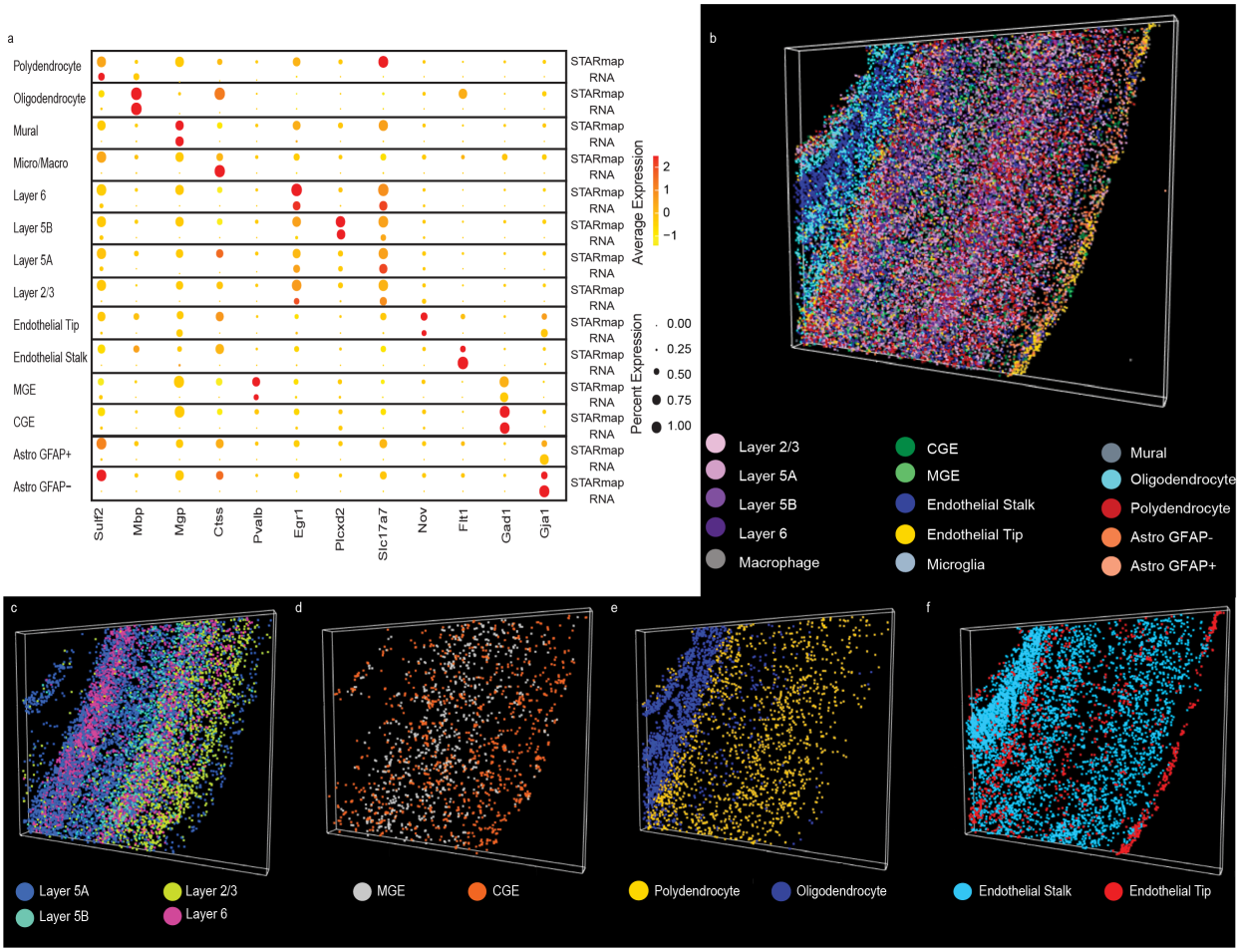
tal Fig. 1.11a). Intuitively, this allows the shared features to compose the largest scaffold of integration, while the dataset-specific metagenes absorb technical artifacts and biological differences (which should have smaller effects than the shared signals). Interestingly, U's contributions to the matrix reconstruction are generally larger than the contributions from V, though still much smaller than those from W.

To investigate whether the unshared features U highlight any cell-type-specific effects, we ranked the contribution of U to each factor, and selected the ten highest factors. We then annotated each factor based on the cell type on which it had the highest loading. The U factors with the 10 highest contributions to the reconstructed matrix were primarily non-neuronal (Supplemental Fig. 1.11b). We suspect that this may be because the selected genes measured by STARmap were primarily marker genes for neuronal cell types. Thus, the inclusion of unshared genes may provide the most benefit for the non-neurons.

To verify that corresponding cell types were clustered together across technologies, we examined the expression of several key marker genes for each labeled cell type by dataset (Fig. 1.4a). Generally, marker genes highly elevated in the cell types of the STARmap dataset are also highly elevated in the corresponding scRNA-seq cell type, indicating that the metagene definitions reflect biological distinctions significant for both datasets. This reinforces that the cell clusterings are not reflective of technical artifacts specific to an individual dataset, nor are they formed overwhelmingly by a single dataset. Rather, the clusters reflect divisions significant to both datasets jointly. Additionally, as we previously noted [197], there is evidence that the STARmap gene capture is somewhat non-specific compared to scRNA-seq for some genes, such as *Sulf2* and *Mgp*.

A key advantage of refining cell types in the STARmap dataset is the ability to use these cell labels within a 3D tissue sample, providing greater insight about how transcriptomic profiles are arranged in vivo. Consequently, we used the derived cell type labels from the UINMF algorithm within the context of 3D space, allowing us to assess their validity on the basis of concordance with known tissue architecture (Fig. 1.4b). This analysis confirms that our cell type annotations accord well with the known structure of the cortex, such as the clear laminar arrangement of excitatory neurons (Fig. 1.4c). It has also been previously established that MGE interneurons originate from the more rostral region of the brain, and the CGE neuron center lies caudal to the MGE center [185]. Likewise, the MGE and CGE determined cell types establish a gradient such that the CGE interneurons increase in more cranial regions of the cortex slice (Fig. 1.4d). The region of white matter that lies beneath the cortex, composed primarily of oligodendrocyte cells, is also identifiable (Fig. 1.4e). Playing a significant role in supporting the brain's vascular systems [102], endothelial cells compose portions of the blood-brain barrier, and the UINMF results indicate that endothelial tip cells

Figure 2.4: Incorporating additional genes locates fine cellular subtypes within 3D spatial volume



(**a**) Dot plot showing marker gene expression in STARmap and scRNA datasets for each joint cluster. The datasets have similar marker expression, indicating that they are well aligned. Plots of three-dimensional spatial locations for different classes of cells colored by cell type: (**b**) all cell types; (**c**) excitatory neurons; (**d**) inhibitory neurons; (**e**) polydendrocytes and oligodendrocytes; and (**f**) endothelial cells.

are located near the outer surface of the brain (Fig. 1.4f).

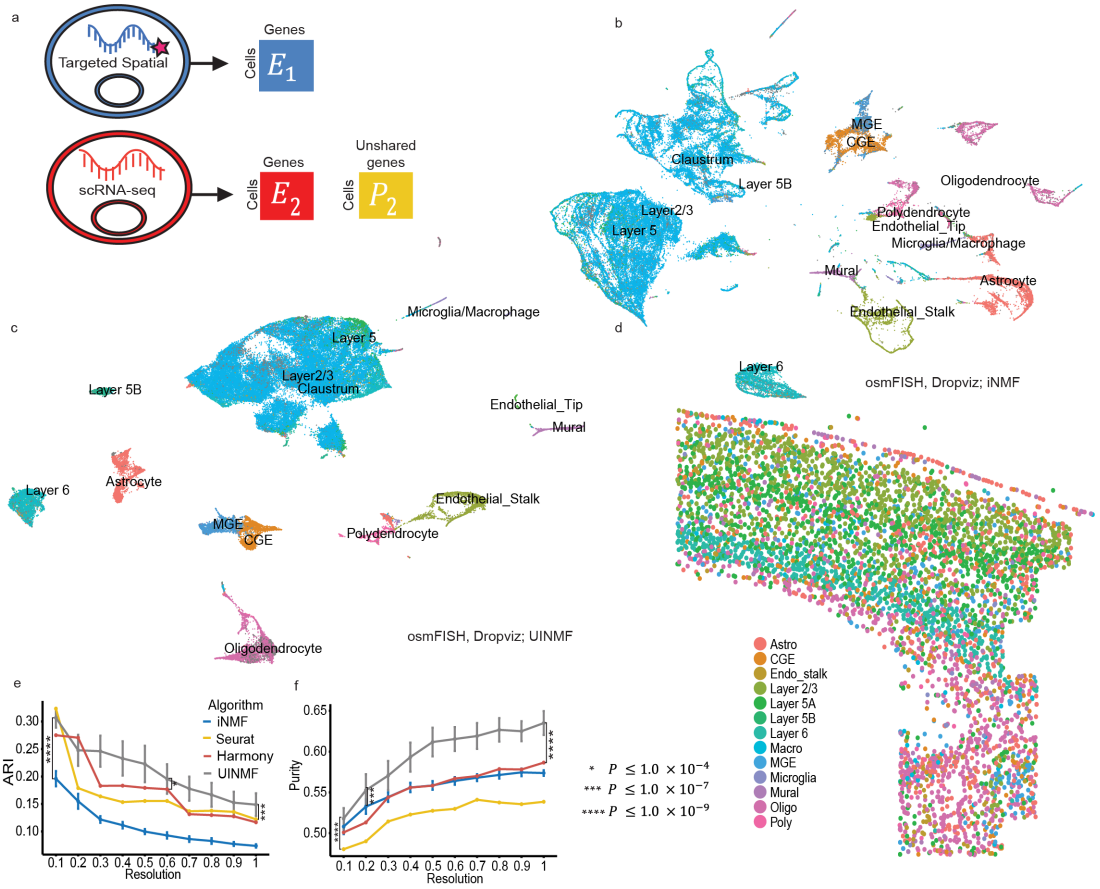
Additionally, we highlight one of the advantages of integrating the STARmap data with scRNA-seq data by imputing values for genes not originally measured in the STARmap assay, e.g., the *Trf* and *Dcn* genes. *Trf* is a recognized marker gene of oligodendrocytes [38], and *Dcn* has previously been found in three types of vascular leptomeningeal cells (VLMCs) [178], which are known to comprise vascular structures. Using KNN imputation, we are able to visualize the imputed gene expression within the 3D volume (Supplemental Fig. 1.12). The expression of *Trf* is primarily confined to the white matter, a region known to be oligodendrocyte-rich, and elevated *Dcn* expression values are visible at the blood-brain barrier.

We further demonstrate the advantages of UINMF for spatial transcriptomic datasets by integrating an osmFISH dataset (5185 cells and 33 genes) [49] from the mouse somatosensory cortex with the DROPviz scRNA-seq dataset from the frontal cortex (70,020 cells). Using iNMF to integrate the osmFISH and DROPviz datasets utilizes the 33 genes measured in the osmFISH protocol as the shared features (Fig. 1.5a). The iNMF integration successfully aligns the two data types, but, similar to the STARmap analysis, leads to both mixing and oversplitting of cell types. The intermixing of cell types, especially the excitatory neurons, fails to resolve a distinct cluster for the Layer 5B Excitatory neurons (Fig. 1.5b). Using UINMF to incorporate an additional 2000 variable genes from the scRNA-seq dataset allows the metagenes to be estimated from many more features, resulting in much more clearly resolved clusters (Fig. 1.5c). A distinct cluster of Layer 5B Excitatory neurons can now be distinguished. Including additional features into the integration not only categorizes broad cell types more effectively, it also allows for the identification of more minute subclasses of cells.

As with the STARmap data, we then plotted the derived UINMF labels within their corresponding spatial context (Fig. 1.5d). The excitatory cells again show a clear laminar arrangement, with layer 6 excitatory neurons forming the innermost layer of the cortex, and layer 5 and layer 2/3 neurons above. Interestingly, layer one contains a number of cells identified as astrocytes (Fig. 1.5d), a finding that has previously been observed experimentally and which has been proposed as evidence for the interaction of glial cells in neuronal signaling [176]. The white matter region, located inferior to layer 6 and known to be composed of oligodendrocytes and polydendrocytes [131, 70], is likewise observable. Lastly, the presence of the caudoputamen and internal capsula region can be identified by the small grouping of inhibitory neurons lateral to the white matter [49].

To quantify the advantage of using UINMF, we measured the ARI and purity scores for both UINMF and iNMF over multiple initializations and multiple clustering resolutions (Fig.

Figure 2.5: Incorporating additional genes improves integration with osmFISH data



(**a**) Schematic of data matrices from osmFISH and scRNA-seq. The osmFISH dataset measures only 33 genes, while the scRNA-seq dataset has many unshared genes that are incorporated during UINMF integration. (**b**) UMAP plot of osmFISH and scRNA integration with iNMF using only shared genes. (**c**) UMAP using UINMF to incorporate an additional 2,000 genes. Both (**b**) and (**c**) are annotated using the original scRNA-seq labels and either the iNMF (**b**) or UINMF (**c**) derived clusterings.(**d**) The spatial arrangement of cell types matches the known tissue structure of the cortex (**e**–**f**). ARI (**e**) and purity (**f**) metrics for iNMF ($P < 2.2 \times 10^{-16}, P = 7.078 \times 10^{-8}$), UINMF, Seurat v3 ($P = 2.505 \times 10^{-9}, P < 2.2 \times 10^{-16}$), and Harmony ($P = 2.982 \times 10^{-5}, P = 1.744 \times 10^{-15}$). Statistical significance was evaluated using a paired, one-sided Wilcoxon test ($n = 200$ values for each test, with each sample having 10 values across 10 resolutions). For nondeterministic algorithms, data are presented as mean values +/- SEM.

1.5e, f). UINMF performed significantly better in terms of both ARI ($P < 2.2 \times 10^{-16}$, paired one-sided Wilcoxon test) and purity ($P = 7.078 \times 10^{-8}$, paired one-sided Wilcoxon test) metrics across the whole range of clustering resolutions. A similar improvement was observed when comparing the UINMF performance to Seurat v3 and Harmony. Note that only a single value is shown for Seurat v3 and Harmony at each resolution because these algorithms are deterministic. UINMF performed significantly better than Seurat v3 and Harmony in both ARI ($P = 2.505 \times 10^{-9}, P = 2.982 \times 10^{-5}$) and purity ($P < 2.2 \times 10^{-16}, P = 1.744 \times 10^{-15}$) metrics.

2.3.4 UINMF improves integration of multimodal and spatial transcriptomic datasets

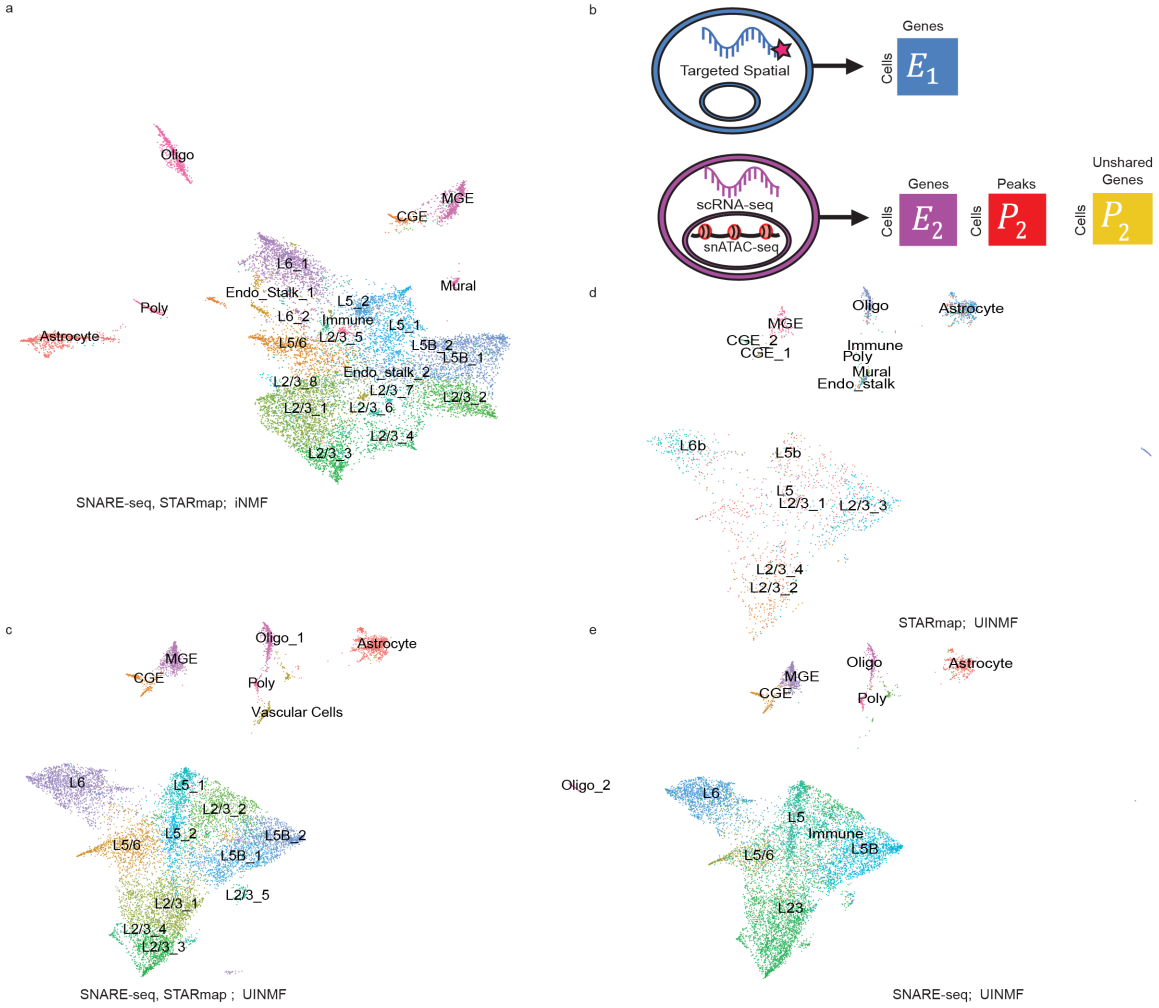
Single-cell multimodal technologies measure epigenomic and transcriptomic profiles from the same cell, providing an exciting opportunity to define cell types from multiple molecular modalities. However, many applications require integrating such multimodal measurements with single-modality datasets. In such applications, the ability of UINMF to incorporate unshared features allows us to capitalize on the multimodal information, rather than using only shared features.

To demonstrate the advantages of such an approach, we used UINMF to integrate STARmap and SNARE-seq data. The STARmap dataset provides gene expression data for 2522 cells across 1020 genes while preserving the 2D spatial coordinates for each cell. The SNARE-seq dataset (10,309 cells) provides simultaneous chromatin accessibility and gene expression levels from the same barcoded cells.

We first integrated the STARmap data and SNARE-seq gene expression measurements only by performing iNMF on the 944 genes shared between the datasets, omitting the unshared genes and completely neglecting the chromatin information. We annotated the cells by using the original annotations from both datasets to jointly define the resulting clusters (Fig. 1.6a). When integrating the datasets with UINMF, we were able to add an additional 2688 highly variable genes present in the SNARE-seq dataset when calculating the metagene loadings. Because the SNARE-seq data is multiomic, we also incorporated the available chromatin accessibility information by including the top 1431 variable chromatin accessibility features within the U matrix (Fig. 1.6b). Thus, the UINMF integration incorporated a total of 4119 features not measured in the STARmap dataset (Fig. 1.6c).

Next, we investigated that the integrity of each individual dataset had been maintained by examining the STARmap and SNARE-seq cells individually by their original labels (Fig. 1.6d, e). There is clear alignment between the original cell labels of each dataset, indicating

Figure 2.6: Incorporating unshared chromatin and gene features to integrate spatial transcriptomic and multimodal data



(**a**) UMAP for iNMF integration of STARmap spatial transcriptomic data and SNARE-seq RNA data only, annotated by jointly examining the SNARE-seq and STARmap labels.(**b**) Schematic of how unshared gene and chromatin accessibility data is incorporated into the integration analysis of STARmap and SNARE-seq using UINMF.(**c**) The integration is improved significantly by the inclusion of ATAC gene-centric features in the U matrix, annotated by jointly examining the SNARE-seq and STARmap labels. (**d**–**e**) The original cell type labels of STARmap cells (**d**) and SNARE-seq cells (**e**) show clear correspondence after UINMF integration.

that the integration defined the metagenes relevant to specific cell populations for both datasets. This suggests that the unshared features can be included into the integration without unduly dominating the metagene calculations.

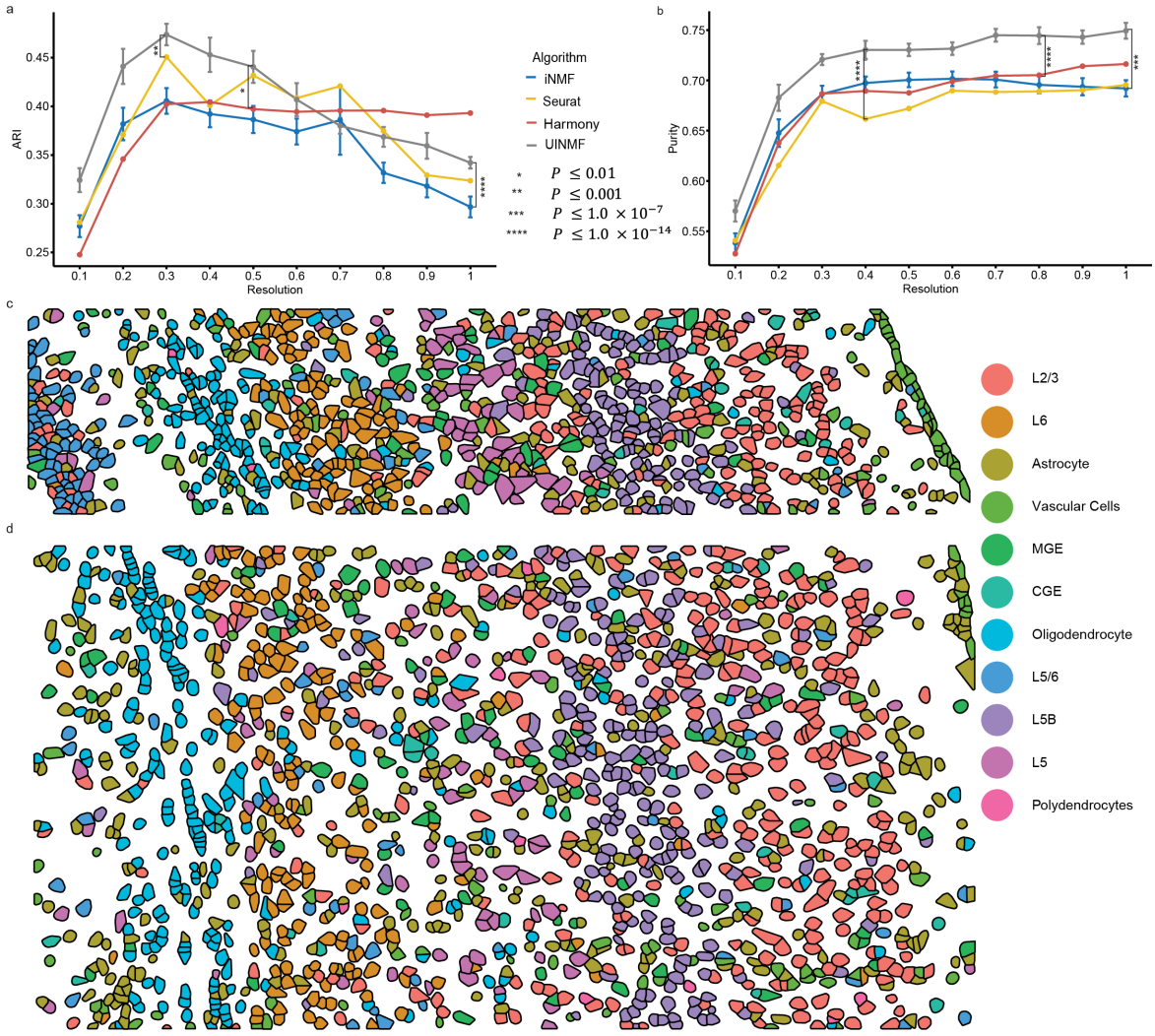
To quantify the derived benefit of including the additional features into the analysis, we then calculated the purity and ARI scores for ten initializations (Fig. 1.7a, b). UINMF significantly outperformed the iNMF algorithm on both ARI ($P = 5.934 \times 10^{-15}$, paired one-sided Wilcoxon test) and purity ($P = 1.046 \times 10^{-11}$, paired one-sided Wilcoxon test). The iNMF algorithm has an ARI competitive with that of UINMF at only a single louvain resolution (0.7). At this resolution, UINMF still has a superior purity score, substantiating the benefit of including additional features using the UINMF algorithm.

Next, we benchmarked the UINMF method against Seurat v3 and Harmony. In comparison to Seurat v3, UINMF has a significantly better ARI score across resolutions ($P = 0.0004184$). Additionally, UINMF also has superior purity over Seurat v3 across all tested resolutions ($P = 2.2 \times 10^{-16}$, paired one-sided Wilcoxon test). Similarly, UINMF has a significantly higher purity score ($P = 2.2 \times 10^{-16}$) at all Louvain resolution in comparison to Harmony. UINMF also has a significantly higher average ARI score than Harmony at all Louvain resolutions ($P = 0.002927$). While Harmony does achieve higher ARI scores at the higher Louvain resolutions, it is critical to note that these are clearly not optimal values for this dataset. Taken together, these results show that the use of UINMF to incorporate added features to the data integration significantly improves the integration of single-cell multimodal data and spatial transcriptomic data.

Because a key motivation for integrating the multimodal and spatial transcriptomic data was bringing enhanced resolution within the context of spatial coordinates, we next plotted the results of the UINMF integration in space. Applying the cell type labels from UINMF to STARmap replicate one (973 cells, Fig. 1.7c) and replicate two (1549 cells, Fig. 1.7d), we found that the UINMF results accord well with the cortical structure. The excitatory neurons are arranged in layers, with L6, L5, and L2/3 clearly visible. Likewise, we also can identify the oligodendrocyte-rich white-matter below the cortex. Additionally, the vascular cell population, which contains endothelial and mural cells, comprise the outermost group of cortical cells.

The SNARE-seq and STARmap integration leverages two distinct types of unshared features, as the U matrix contains both unshared genes as well as snATAC-seq peak information. To determine how the unshared features from each modality contribute to the integration, we repeated the integration using only one type of unshared data. Specifically, we integrated the SNARE-seq and STARmap datasets using UINMF, and only the unshared genes (2688 genes) as unshared features. We also integrated the SNARE-seq and STARmap datasets us-

Figure 2.7: Incorporating unshared chromatin and gene features improves integration of spatial transcriptomic and multimodal data



We compared UINMF results with those from iNMF ($P = 5.934 \times 10^{-15}, P = 1.046 \times 10^{-11}$)), Harmony ($P = 0.002927, P = 2.2 \times 10^{-16}$), and Seurat v3 ($P = 0.0004184, P = 2.2 \times 10^{-16}$) using ARI (**a**) and purity (**b**) metrics. Comparisons were evaluated using a paired, one-sided Wilcoxon test ($n = 200$ values for each test, with each sample having 10 values across 10 resolutions). For nondeterministic algorithms, data are presented as mean values +/- SEM. We also confirmed that the spatial arrangement of predicted cell types in both STARmap replicate one (**c**) and replicate two (**d**) matches the known organization of the cortex.

ing UINMF, and only the unshared peaks (1431 peaks) as unshared features (Supplementary Fig. 1.13). The results indicate that the unshared genes contribute most substantially to the improvement in data integration, but that using the peak data individually and additively enhances the final integration result.

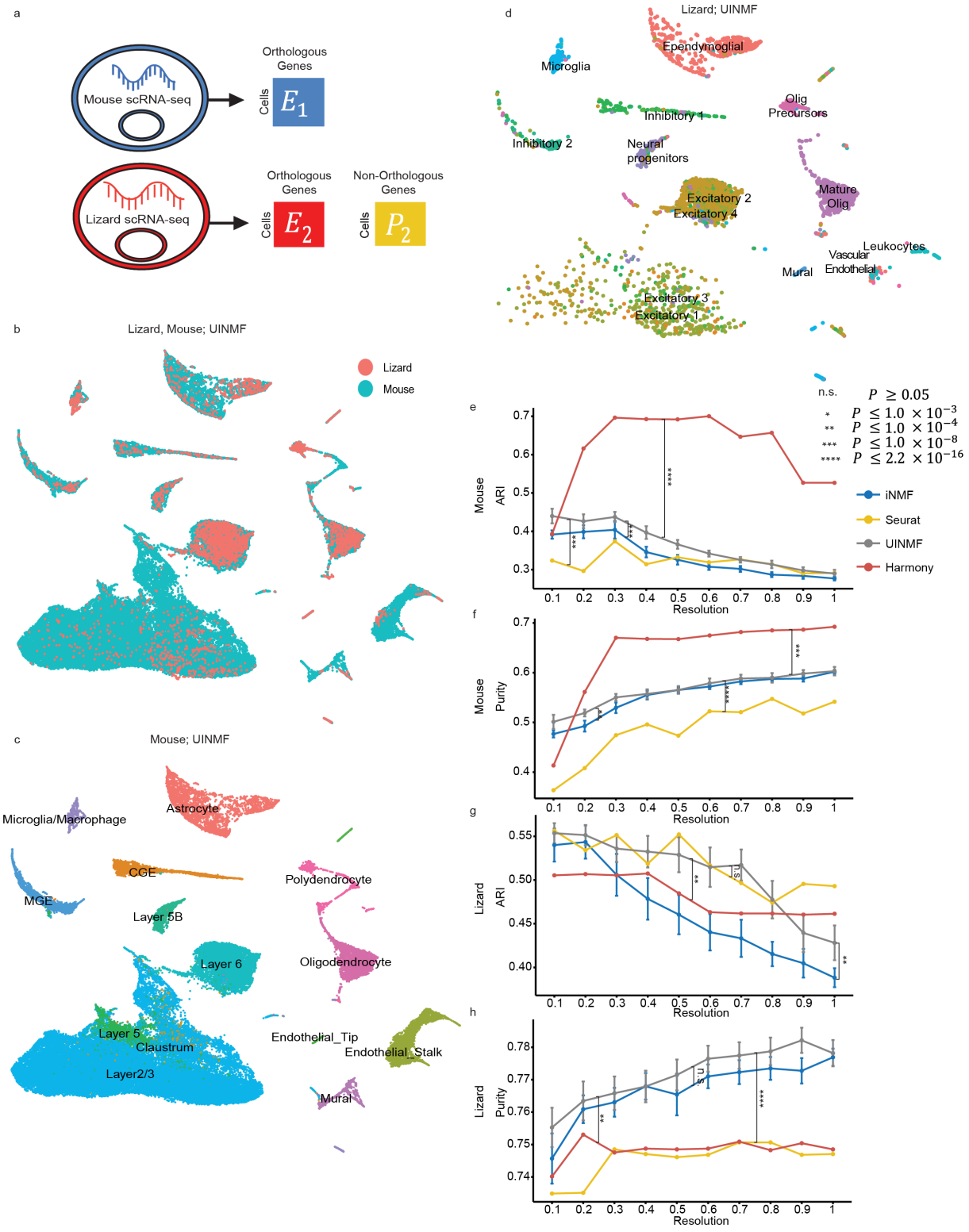
2.3.5 Incorporating nonorthologous genes improves the integration of cross-species data

Previous integrations of cross-species datasets have been limited to genes that are orthologous between species, as non-orthologous genes, by definition, are not shared between datasets [197]. Yet, non-orthologous genes can be key marker genes within a species, providing crucial information for distinguishing cell populations. Using UINMF, we were able explore the potential benefits of including non-orthologous genes in cross-species integration. For this cross-species analysis, we integrated scRNA data (4187 cells) from the pallium of the bearded dragon lizard (*Pogona vitticeps*) [184] with scRNA data from the mouse frontal cortex (71,639 cells) [153]. We first selected 1979 variable genes that were annotated as one-to-one orthologs between the two species. Then we selected an additional 166 non-orthologous variable genes from the lizard dataset. We integrated the datasets using UINMF, with the one-to-one orthologs as shared features, and the non-orthologous genes from the lizard as unshared features (Fig. 1.8a). UINMF successfully aligned the two datasets, as illustrated by the overlapping distributions of the two datasets within the UMAP space (Fig. 1.8b). To confirm the correspondence between the cell types of the two species, we plotted only the mouse cells (Fig. 1.8c) and only the lizard cells (Fig. 1.8d), colored by their originally published labels. Strong correspondence between cell types, including excitatory neurons, inhibitory neurons, and non-neuronal cells, can be observed.

To examine the additional benefit of including non-orthologous genes when performing cross-species integrations,we performed the integration using iNMF to establish a baseline. The baseline iNMF integration was limited to the 1979 orthologous genes, and resulted in a lower quality integration (Supplemental Fig. 1.14). The mural cell populations had decreased alignment between the two species, and many of the astrocytes were misaligned to the excitatory neuron clusters. Furthermore, the lizard's excitatory neuron subtypes were not distinctly separated.

In order to quantify the advantage of UINMF over iNMF in cross-species integrations, we compared the purity and ARI scores of the two algorithms across ten initializations. To ensure that any differences observed were consistent for both species, we used both the published mouse cell labels, as well as the published lizard labels, to benchmark the

Figure 2.8: The inclusion of non-orthologous genes improves the integration of cross-species data



We use UINMF to include both orthologous and non-orthologous genes when integrating the datasets (**a**), and demonstrate the alignment between the two datasets (**b**). We also

Figure 2.8 *(previous page)*: confirmed cell type correspondence by examining only the mouse cells (**c**) and only the lizard cells (**d**), both labeled with their published cell labels. To show the advantage of including the non-orthologous genes, we show the difference in ARI (**e**) and purity (**f**) scores using the originally published mouse labels, comparing UINMF performance to iNMF ($P = 3.626 \times 10^{-9}, P = 6.258 \times 10^{-4}$), Seurat ($P < 2.2 \times 10^{-16}, P = 3.047 \times 10^{-11}$), and Harmony ($P < 2.2 \times 10^{-16}, P = 2.815 \times 10^{-12}$). We compare algorithm performance using a paired, one-sided Wilcoxon test, where $n = 200$ ARI (purity) measures. We also confirm a similar trend in the ARI (g) and purity (h) scores using the original lizard labels to assess performance differences between UINMF and iNMF ($P = 1.145 \times 10^{-6}, P = 0.07157$), Seurat ($P < 2.2 \times 10^{-16}, P = 0.8148$), and Harmony ($P = 1.674 \times 10^{-5}, P < 2.2 \times 10^{-16}$). For nondeterministic algorithms, data are presented as mean values +/- SEM.

algorithms. Including the non-orthologous genes using UINMF resulted in a significant increase in both the ARI ($P = 3.626 \times 10^{-9}$, paired one-sided Wilcoxon test) and the purity ($P = 6.258 \times 10^{-4}$, paired one-sided Wilcoxon test) of the mouse dataset (Fig. 1.8e, f). We also noted a significant increase in the ARI ($P = 1.145 \times 10^{-6}$, paired one-sided Wilcoxon test) of the lizard data set (Fig. 1.8g). Although UINMF does not show a significant increase in the purity ($P = 0.07157$, paired with one-sided Wilcoxon test) of the lizard dataset, UINMF is able to achieve a higher maximum purity score at most resolutions (Fig. 1.8h), and shows significant improvement in all other measures of integration quality.

We also compared the Purity and ARI scores of UINMF to those achieved by Seurat v3 and Harmony. UINMF shows a significant improvement in purity scores over Seurat v3 using both mouse ($P < 2.2 \times 10^{-16}$, paired one-sided Wilcoxon test) and lizard ($P < 2.2 \times 10^{-16}$, paired one-sided Wilcoxon test) labels. When comparing ARI scores, there is not a significant difference between Seurat v3 and UINMF using the lizard labels ($P = 0.8148$, paired one-sided Wilcoxon test); however, the mouse labels show a significant improvement ($P = 3.047 \times 10^{-11}$, paired one-sided Wilcoxon test). Therefore, UINMF significantly outperforms Seurat v3 on three out of the four measures assessed. When examining the difference in performance between UINMF and Harmony, Harmony shows significantly improved ARI ($P < 2.2 \times 10^{-16}$, paired one-sided Wilcoxon test) and purity ($P = 2.815 \times 10^{-12}$, paired one-sided Wilcoxon test) scores using the mouse labels. However, UINMF has superior performance when comparing the ARI ($P = 1.674 \times 10^{-5}$, paired one-sided Wilcoxon test) and purity ($P < 2.2 \times 10^{-16}$, paired one-sided Wilcoxon test) scores using the lizard labels. Harmony's increased performance using the mouse labels, but decreased performance when using the lizard labels, leads us to believe that Harmony has learned principal components driven primarily by the mouse dataset, neglecting to learn factors that likewise distinguish the lizard cells. The large difference in dataset size between the mouse (71,639 cells) and lizard

(4187 cells) supports this hypothesis. UINMF innovatively learns joint metagene factors, resulting in a final integration that respects biological factors in both datasets.

A nice property of UINMF is its flexibility for a range of types of integration analyses, including those with more than two datasets. To demonstrate a case where this is useful, we performed a cross-species integration of the excitatory neurons of the primary motor cortex data of the human (7805 cells), mouse (8242 cells), and the marmoset (8166 cells)[26].

Using 2757 orthologous genes shared between all three species as shared features, we then selected 353 non-orthologous mouse genes, 290 non-orthologous marmoset genes, and 671 non-orthologous human genes as unshared features. UINMF successfully aligned the three datasets (Supplemental Fig. 1.15a), while preserving the originally labeled cell types of each species (Supplemental Fig. 1.15b–d). Interestingly, however, there were two distinct populations of mouse neurons, L5 ET and L5/6 NP, that do not align to the primate data.

We further investigated these clusters to understand why they did not align. The original publication similarly identified primate-specific features of L5 neurons. Specifically, the L5 ET neurons for all species showed cell-type-specific expression of certain ion channel subunits, but the L5 ET neurons of the primates had increased cell-type-specific expression of potassium and calcium channels compared to the mouse L5 neurons. It was hypothesized that these differences drove the observable differences in physiological response between the mouse and primate ET neurons. We thus checked whether our non-aligning clusters reflect these species differences in ion channel expression. We compared the mouse cells that successfully aligned with the primate L5 ET neurons (Cluster 9), with those L5 ET mouse cells that were unaligning (Cluster 13). We observed that the ion channel subunits previously annotated [184] as showing conserved expression across mouse and primates were again consistent in expression, regardless of their alignment status (Supplemental Fig. 1.15e). However, for genes previously identified as showing primate-specific enrichment in L5 ET cells [26], we saw a distinct difference in the expression levels between the unaligning and aligning mouse cells, providing biological support for the algorithmic separation of these clusters (Supplemental Fig. 1.15f).

2.4 Discussion

We have extended our previous integrative nonnegative matrix factorization algorithm [197] by adding an unshared feature matrix. This addition accommodates features that are not present in all datasets and increases the amount of information that is used to define the metagenes. We showed that the inclusion of unshared features provides clear advantages across four different types of integration analyses. First, UINMF can be used to incorporate

intergenic information when integrating transcriptomic and epigenomic datasets. Second, UINMF can incorporate genes not measured in targeted spatial transcriptomic datasets, allowing better resolution of fine cellular subtypes within a spatial coordinate frame. Third, UINMF can utilize all of the information present in single-cell multimodal datasets when integrating with single-modality datasets. Additionally, UINMF can accommodate non-orthologous genes in cross-species integration analyses.

A recent preprint described MultiMAP, another single-cell data integration approach that can incorporate unshared features. Our approach differs from MultiMap in several key ways. First, MultiMAP uses singular value decomposition followed by nonlinear dimensionality reduction [79]. Therefore, the resulting reduced dimensions are not biologically interpretable. In contrast, UINMF learns factor loadings that represent metagenes, groups of genes that often reflect an underlying biological process or technical artifact. Second, MultiMAP makes a key assumption that the data is uniformly distributed on the latent manifold. While this assumption may hold at times when the datasets to be integrated are sampled from the same tissue, several common integration scenarios violate this assumption. For example, nuclear isolation protocols and whole-cell dissociation protocols each have different cell-type-specific biases, leading to different cell type distributions. Similarly, cross-species integrations may present a challenge if the proportions of homologous cell types are different. In contrast, UINMF does not rely on the uniform distribution assumption.

With the rapid development of multimodal and spatial transcriptomic technologies, we anticipate that the UINMF algorithm will prove useful for a wide variety of analyses. As the additional U matrix can incorporate any type of cellular features, with no assumptions about their relationship to gene-centric features, the algorithm is inherently flexible to accommodate a variety of data types and modalities. Future applications could examine the incorporation of data types such as Hi-C [29] measurements, as well as the potential to use UINMF on a diverse collection of in situ hybridization and immunohistochemistry datasets with limited numbers of genes. We expect that, as additional experimental methods for single-cell measurement are developed, our approach will prove increasingly useful for a broad variety of single-cell integration tasks.

2.5 Methods

We extend our previously published ANLS algorithm for solving the iNMF problem [197] so that we can now incorporate unshared features when integrating across datasets. The unshared feature matrix can accommodate any type of unshared feature, whether gene-centric or otherwise. In this paper, we incorporated intergenic peaks from snATAC-seq data

and additional genes not measured in all datasets, although many other applications are possible. For each data set $E^1, E^2, ...E^n$, we normalize the data, and select m variable genes (shared across all datasets), and z^i variable features (not shared across all datasets), such that after scaling $E^i \in \mathbb{R}_+^{n_i \times (m+z_i)} (i = 1, ..., N)$. For a given K, the user-defined number of metagenes, and λ_i, the UINMF optimization problem is:

$$\text{argmin}_{H^i \geq 0, W \geq 0, V^i \geq 0, U^i \geq 0} \sum_{n=i}^{d} \|(E^i P^i) - H^i((W0) + (V^i U^i))\|_F^2 + \lambda_i \sum_{n=i}^{d} \|(H^i(V^i U^i))\|_F^2 \quad (2.1)$$

where P^i is the unshared feature matrix, H^i are the cell factor loadings, W is the shared metagenes, V^i are the dataset-specific, shared metagenes, and U^i are the dataset-specific, unshared metagenes.

We then derive a coordinate block descent (BCD) algorithm[88] for solving the UINMF optimization problem. The BCD approach divides the parameters into blocks, and then finds the optimal parameters by updating each block while holding the others fixed. Because each block-wise optimization sub-problem is convex, iterating these updates is guaranteed to converge to a local minimum[88]. To solve the UINMF optimization problem, we use matrix blocks, one block for each of $H^i \in \mathbb{R}_+^{n_i \times k}$, $W \in \mathbb{R}_+^{k \times m}$, $V^i \in \mathbb{R}_+^{k \times m}$, $U^i \in \mathbb{R}_+^{k \times z_i}$ $(i = 1, ..., N)$.

Each sub-problem is a nonnegative least squares optimization, which we solve numerically using an efficient C++ implementation of the block principal pivoting algorithm[86]. To update H^i, we solve

$$H^i = \text{argmin}_{H \geq 0} \left\| \begin{pmatrix} \begin{pmatrix} W^T \\ 0^{z^i \times k} \end{pmatrix} + \begin{pmatrix} V^{i^T} \\ U^{i^T} \end{pmatrix} \\ \sqrt{\lambda_i} \begin{pmatrix} V^{i^T} \\ U^{i^T} \end{pmatrix} \end{pmatrix} H^{i^T} - \begin{pmatrix} E^i \\ 0^{n_i \times (m+z^i)} \end{pmatrix}^T \right\|_F^2 \quad (2.2)$$

while holding the other parameters fixed. Similarly, to update the other parameters, we solve the following subproblems:

$$W = \text{argmin}_{W \geq 0} \left\| \begin{pmatrix} H^1 \\ H^i \end{pmatrix} W - \begin{pmatrix} E^1 - H^1 V^1 \\ E^i - H^i V^i \end{pmatrix} \right\|_F^2 \tag{2.3}$$

$$V^i = \text{argmin}_{V \geq 0} \left\| \begin{pmatrix} H^1 \\ \sqrt{\lambda_i} H^d \end{pmatrix} V^i - \begin{pmatrix} E^1 - H^1 W \\ E^i - H^i W \end{pmatrix} \right\|_F^2 \tag{2.4}$$

$$U^i = \text{argmin}_{U \geq 0} \left\| \begin{pmatrix} H^1 \\ \sqrt{\lambda_i} H^d \end{pmatrix} U^i - \begin{pmatrix} P^i \\ 0^{n_i \times z^i} \end{pmatrix} \right\|_F^2 \tag{2.5}$$

We iterate these updates until convergence, which we determine by calculating the decrease in the objective function at the conclusion of each iteration (Supplemental Fig. 1.16). We consider the algorithm to have converged when the decrease in the objective value function between the previous and current iteration, weighted by dividing by their mean, is less than the epsilon parameter set by the user. For all UINMF analyses in this paper, we set the convergence threshold to $\epsilon = 1.0 \times 10^{-10}$ and set the maximum number of iterations to 30 for each analysis, unless otherwise noted.

Note that iNMF uses a constant penalty term λ for all datasets. When implementing UINMF, however, we introduced a separate λ_i parameter for each dataset, such that the penalty applied to E is weighted by λ_1, E_2 has a regularization weight of λ_2, etc. The inclusion of λ_i allows for the tuning of dataset penalization at the user's discretion. We found that in some cases, using a different penalty value for different datasets could improve results. However, we simply used $\lambda_i = 5$ (the default value in our iNMF and UINMF implementations) for all analyses except the STARmap and DROPviz integration. The STARmap and DROPviz integration does achieve the best results with different λ values for each dataset.

2.5.1 Increase in Computational Complexity

The difference in computational complexity between UINMF and iNMF increases with the number of unshared features used in UINMF, but the difference between the runtime of the two algorithms is not prohibitive in practice (Supplemental Figs. 1.17, 1.18). To assess the theoretical difference in computational complexity between the algorithms, assume the same total number of features are present in the datasets input into each algorithm. Let iNMF operate on a dataset that has g shared features, and let $g = m + z$, where m is the number of shared features and z is the number of unshared features of the UINMF dataset. Let K be the user-defined number of metagenes. For each iteration, UINMF solves for $U^{z \times K}$

and $V_i^{m \times K}$ separately, but iNMF performs the same number of calculations to solve for $V_i^{g \times K}$, since $g = m + z$. When solving for the shared metagene matrix, W, iNMF solves the optimization problem for a $g \times K$ matrix, whereas UINMF must only solve a $m \times K$ matrix. Because the shared metagene matrix has less features in UINMF ($g > m$), each iteration of the algorithm actually constitutes less computational complexity than iNMF given the same total number of features.

2.5.2 Evaluating Time and Memory Usage

To benchmark the memory usage and run time of UINMF in comparison to current methods, we measured the performance of Seurat v3, Harmony, iNMF, and UINMF when integrating datasets of 20,000, 40,000, 60,000, 80,000, and 100,000 cells. To obtain datasets with these cell populations, we sampled equivalent numbers of cells from the Dropviz scRNA-seq and STARmap datasets. As there are only 32,845 cells in the STARmap dataset, for the 80,000 and 100,000 cells data integrations, we sampled the remaining cells all from the DROPviz dataset. The integrations for the four algorithms were performed on the same sample datasets. We used the 28 shared genes between the datasets as the shared features for all four algorithms. We set $\lambda = 5$ and $K = 27$ as the parameter for both algorithms. Using UINMF, we incorporated an additional 2775 unshared genes when performing the dataset integration. We ran each iNMF and UINMF analysis 5 times to assess run-to-run variation in time and memory usage but found that these numbers were very stable. For the Seurat v3 integration, the number of dimensions was set to 27 for both the *FindIntegrationAnchors* (Seurat 4.0.0) as well as the *IntegrateData* (Seurat 4.0.0) functions. The number of principal components within *RunPCA* (Seurat 4.0.0) was also set to 27 for Harmony as well as Seurat v3 (Supplemental Fig. 1.9).

2.5.3 Evaluation Metrics

The alignment score, based on Butler et al. [36], is a measure that captures how well two samples align uniformly within a latent space. A score closer to zero indicates a poor alignment, or mixing of the two samples, whereas a score closer to one is indicative of datasets that share underlying cell types.

Fraction of Samples Closer Than the True Match (FOSCTTM) scores measures how closely two measurements of the same cell are placed within the latent space [105]. We calculate the FOSCTTM score by finding the distance between the scRNA-seq cell and the snATAC-seq label for each barcode. We then divide by the total number of barcoded cells to derive the average FOSCTTM score.

Cluster purity is calculated based on a reference clustering. Each cluster is assigned a type based on the predominant label for that cluster. The cells that correspond correctly to this label are counted. We calculate purity by summing the correct number of labels across all clusters, and dividing by the total number of labeled cells present. Consequently, a score closer to 0 indicates that the cells are not being accurately grouped into clusters by cell types, and a score of 1 indicates perfect grouping by cell type.

To calculate the Adjusted Rand Index (ARI), we measured the similarity between two clusterings by whether we observe matching or non-matching clustering between pairs of samples. The ARI score can range from 0 (no matches) to 1 (perfect matching).

2.5.4 Integration of RNA and ATAC profiles from SNARE-seq

For a baseline, we integrated the scRNA and snATAC datasets that resulted from the original SNARE-seq publication [44] with iNMF, using the top 2589 variable genes and their associated snATAC peaks. We took the best optimization of 5 random initializations, each performed with $K = 30$ and $\lambda = 5$, across ten random seeds. We performed quantile normalization with parameter $knn_k = 20$. We then calculated the average FOSCTTM and alignment scores for each of the 10 random seeds. To assess the additive properties of including the intergenic peaks into the integration, we used the U-matrix to hold 1000, 2000, 3000, 4000, 5000, 6000, and 7000 of the top variable intergenic bins. To select these bins, we binned the genome into sections of 100,000 bp. We counted the number of peaks within each bin, where a peak was considered to be inside a bin if over 50% of the peak's base pairs overlapped the bin. We then filtered the bins, removing bins that were empty, or overlapping with ENCODE Blacklisted regions [18], genes or promoters. After normalization, we then used the *FindVariableFeatures* function (Seurat 4.0.0) to select the top variable bins. We saw the greatest improvement in performance using 7000 intergenic bins, as this allowed the capture of a large amount of the data variance, without oversaturating the factor loadings with noise. We performed UINMF with $K = 30$, $\lambda = 5$, and $knn_k = 20$. Using a paired, one-sided Wilcoxon test, we compared the FOSCTTM and alignment scores of UINMF and iNMF.

To compare UINMF performance to that of Seurat v3 and Harmony, we integrated the two datasets using the 2589 shared features. For Seurat v3, we used a PCA dimension of 30. To compare UINMF and Seurat v3 performance, we performed a paired, one-sided Wilcoxon test on the alignment and FOSCTTM scores of UINMF and Seurat v3.

Similarly, to run Harmony, we used the 2589 shared features and 30 principal components for RunPCA (Seurat 4.0.0). We also used 30 neighbors for the FindNeighbors function. We

calculated the alignment and FOSCTTM, and then performed a paired, one-sided Wilcoxon test to assess the difference between UINMF and Harmony performance. To cluster just the RNA and ATAC cells, we similarly used $K = 30$, $\lambda = 5$, $knn_k = 100$, and a Louvain resolution of 1. To derive ground truth cell type labels, we clustered and annotated the RNA-seq cell barcodes using marker genes. We use these consistent labels as ground truth labels for each UMAP presented in Fig. 1.2.

2.5.5 Integration of scRNA-seq and STARmap

We integrated the STARmap spatial transcriptomic dataset (31,294 cells, filtered for hippocampus and claustrum cells) and a scRNA-seq dataset (70,514 cells (1125 removed for non-expressing)) using iNMF. For this iNMF analysis, we were limited to the 28 genes measured in the STARmap dataset. Both iNMF and CCA/PCA (used by Seurat v3) are limited to no more components than the number of genes, while UINMF can estimate more dimensions because it also incorporates unshared genes. We thus used $K = 27$ dimensions for both iNMF and Seurat v3. For iNMF we also used $\lambda = 5$, and a quantile normalization with K-nearest neighbors of 20. Using UINMF, we included an additional 2775 of the most variable genes. The parameters for the UINMF integration were $K = 40$ and $\lambda = 10$ for the STARmap data and $\lambda = 1$ for the scRNA-seq data; $knn_k = 20$ for the quantile normalization; and Louvain resolution of 1. For both iNMF and UINMF, we perform 5 initializations with the same random seed and pick the best one. We calculate the cluster purity for both algorithms using the scRNA-seq labels, and use the highest number of cells present to annotate the clusters by cell type. These annotations were then applied within the context of 3D space using the originally provided STARmap coordinates. For five different random seeds, we measured the difference between the cluster purity and ARI of iNMF and UINMF using a paired, one-sided Wilcoxon test.

For Seurat v3, we similarly used the 28 shared genes and a PCA dimension of 27. We then calculated the Purity and ARI scores at each Louvain resolution from 0.1 to 1, in increments of 0.1. We performed a paired, one-sided Wilcoxon test to assess the difference between UINMF and Seurat v3 performance quantified by purity and ARI scores.

Similarly, to run Harmony, we used the 28 shared genes and 27 principal components for *RunPCA* (Seurat 4.0.0). We also used 27 as the number of dimensions used for the *FindNeighbors* function. For each Louvain resolution from 0.1 to 1, in increments of 0.1, we calculated the purity and ARI scores. We then performed a paired, one-sided Wilcoxon test to assess the difference between UINMF and Harmony performance quantified by purity and ARI scores.

To analyze the effect of the additional features on the memory usage and run time of the algorithm, we performed five initializations for 0, 1000, 2000, 3000, and 4000 extra features (Supplemental Fig. 1.17). For consistency, we used $\lambda = 5$ and $K = 27$ for all time and memory analyses.

After performing the dataset integration, it was possible to use *imputeKNN*, with *knn_k* = 20, to impute gene values for genes originally unmeasured by the STARmap technology. By using the gene profiles of the closest scRNA cells to impute probable gene expression values for the STARmap cells, we are able to predict expression for genes not originally assayed with the 3D tissue space. Specifically, we show the results of KNN imputation for the *Trf* and *Dcn* genes (Supplemental Fig. 1.12.)

To examine the individual contribution of each factor (V^i, U, W) to the matrix reconstruction, we develop the *calcNormLoadings* function. Since $W \in \mathbb{R}_+^k \times m$, $V^i \in \mathbb{R}_+^{k \times m}$, and $U^i \in \mathbb{R}_+^{k \times z_i}$ are characterized by different numbers of features, we standardize the factor loadings by dividing by the number of features pertinent to each matrix. For each factor l in K, we measure the contribution to reconstruction as

$$U_l = \|H[,l] \times U^i[l,]\|_F^2 / z_i \quad (6)$$
$$W_l = \|H[,l] \times W^l[l,]\|_F^2 / m \quad (7)$$
$$V_l^1 = \|H[,l] \times V^1[l,]\|_F^2 / m \quad (8)$$
$$V_l^2 = \|H[,l] \times V^2[l,]\|_F^2 / m \quad (9)$$

The resulting ranked norms provide intuition as to which factors each matrix contributes most strongly to (Supplemental Fig. 1.11).

2.5.6 Integration of scRNA-seq and osmFISH

Using the osmFISH dataset (33 genes, 6471 cells) we excluded hippocampal and cells from the top left of the tissue slice originally labeled "excluded" in the original osmFISH publication, as well as cells with zero detected genes. A total of 5185 cells passed these filtering criteria. The scRNA-seq dataset, originally 71,639 cells, was reduced to 70,020 (1,619 cells non-expressing for the 33 genes of interest were removed). We ran the iNMF algorithm using the 33 shared genes, $\lambda = 5$, $K = 32$, and took the best optimization of 5 random initializations for each of the 10 seeds. Note that, as with the STARmap data, iNMF and Seurat v3 are limited to no more components than the number of genes, while UNIMF can estimate more factors due to the use of unshared genes. Using UINMF, we integrated the osmFISH spatial transcriptomic data with the scRNA-seq using the 33 shared genes as well as the 2000 most variable unshared genes. We used 10 different random seeds, with $\lambda = 5$ and $K = 40$, and took the best optimization of 5 random initializations for each seed.

We calculated the Purity and ARI for each algorithm at Louvain resolutions 0.1 through 1.0, in 0.1 increments. We used the published DROPviz labels as our reference clustering, and assessed the significance of our findings using a paired, one-sided Wilcoxon test. We used *knn_k* = 150 for the *quantile_norm* function (rliger 1.0.0), *n_neighbors* = 150 for the *runUMAP* function (rliger 1.0.0), and a resolution of 0.5 for Louvain community detection (rliger 1.0.0).

For the Seurat v3 integration, we also used the 33 shared genes and a PCA dimension of 32. Using louvain resolutions from 0.1 to 1.0 in 0.1 increments, we assessed the performance of Seurat v3 using purity and ARI scores. We compared Seurat v3 and UINMF using a paired, one-sided Wilcoxon test.

Using Harmony, we integrated the two datasets using the 33 shared genes and 32 as the number of principal components and dimensions for the *RunPCA* and *FindNeighbors* functions (Seurat 4.0.0). We quantified Harmony performance by calculating purity and ARI scores for louvain resolutions from 0.1 to 1.0 in 0.1 increments, and compared the results against UINMF using a paired, one-sided Wilcoxon test.

2.5.7 Cross-Species Integration

The Lizard Pallium dataset[184] originally had 4202 cells, but we limited the cells used to the 4187 cells deemed high quality in the original publication. We integrated this dataset with the scRNA-seq dataset from the mouse brain[153] (71,639 cells). We used the original publication's 1-to-1 ortholog labels to select orthologs common to both the mouse and lizard dataset. Using these as our shared features, we normalized and scaled the data. We then selected 1979 shared genes, using a variance threshold of 0.3. For UINMF, we used the same variance threshold to select 166 non-orthologous genes from the lizard. We optimized UINMF and iNMF with $K = 30$, $\lambda = 5$, and took the best of 5 random initializations for 10 random seeds. We performed quantile normalization for each optimized object using the mouse dataset as a reference. To ensure that any differences in purity and ARI score were not driven by a single species, we calculated the ARI and Purity scores using both the lizard and the mouse cell labels separately as ground truth. We performed these calculations for each of the ten seeds at louvain resolutions from 0.1 to 1.0 in increments of 0.1. To examine the difference between UINMF and iNMF performance, we performed a paired, one-sided Wilcoxon test between the ARI and Purity values, for the mouse and lizard labels, respectively. When generating the UMAPs shown, we used the default louvain resolution of 0.25, and the default nearest neighbors of 10.

For Seurat v3 integration, we used the 1979 highly variable, orthologous genes identified

between the datasets, and a PCA dimension of 30. Using louvain resolutions from 0.1 to 1.0 in 0.1 increments, we assessed the performance of Seurat v3 using purity and ARI scores with both the lizard and mouse annotations. To compare the performance of Seurat v3 and UINMF, we used a paired, one-sided Wilcoxon test.

Similarly, when evaluating the performance of Harmony in integrating the two datasets, we use the same 1979 orthogolous genes. We use 30 as the number of principal components and dimensions for the *RunPCA* and *FindNeighbors* functions (Seurat 4.0.0). We quantified Harmony performance by calculating purity and ARI scores for louvain resolutions from 0.1 to 1.0 in 0.1 increments, and compared the results against UINMF using a paired, one-sided Wilcoxon test.

2.5.8 Integration of SNARE-seq and STARmap

To integrate the multi-omic SNARE-seq dataset (10,309 cells) with the spatial transcriptomic STARmap dataset (2522 cells), we used the number of shared genes (944 genes), as well as 4,119 unshared features. To generate the unshared features, we selected genes with a variance threshold higher than that of 0.1, and then removed genes shared between datasets, for a total number of 2688 unshared genes. We selected the chromatin accessibility peaks with a variance greater than 0.01 (1431 peaks). For iNMF benchmarking, we used only the 944 shared genes. We optimized UINMF and iNMF with $K = 30$, $\lambda = 5$, and took the best of 5 random initializations for 10 random seeds. We performed quantile normalization for each optimized object using the SNARE-seq dataset as a reference. Using the SNARE-seq cell labels as ground truth, we calculated the ARI and Purity scores for each of the ten seeds at louvain resolutions from 0.1 to 1.0 in increments of 0.1. We set nearest neighbors to 100 when generating the UMAPs shown, and the louvain resolution to 1. To assess the difference between UINMF and iNMF functioning, we performed a paired, one-sided Wilcoxon test between the ARI and purity values. The UMAPs shown have a louvain resolution of 1.0, and are labeled by jointly examining the original STARmap and SNARE-seq cell labels.

To quantify the contribution of each type of unshared feature to the resulting improvement in dataset integration quality, we also performed the UINMF integration using only the 1431 snATAC-seq peaks, and using only the 2688 unshared genes. For each analysis, we optimized the UINMF optimization equation using $K = 30$, $\lambda = 5$, and took the best of 5 random initializations for 10 random seeds. We performed quantile normalization for each optimized object using the SNARE-seq dataset as a reference. Using the SNARE-seq cell labels as ground truth, we calculated the ARI and purity scores for each of the ten seeds at louvain resolutions from 0.1 to 1.0 in increments of 0.1 (Supplemental Fig. 1.13).

To assess the performance of Seurat v3, we used the 944 shared genes, and set the PCA dimensions equal to 30. We then calculated the purity and ARI scores for each Louvain resolution from 0.1 to 1.0, in increments of 0.1. We conducted a paired, one-sided Wilcoxon test between UINMF and Seurat v3 performance in terms of ARI and purity scores.

Using Harmony to perform the integration, we used the 944 shared genes and the default 30 principal components for *RunPCA* (Seurat 4.0.0). Similarly, we used the default (30) for the number of dimensions in the *FindNeighbors* function (Seurat 4.0.0). Using louvain resolutions from 0.1 to 1.0 in 0.1 increments, we assessed the performance of Harmony using purity and ARI scores. We compared Harmony and UINMF using a paired, one-sided Wilcoxon test on both the purity and ARI scores.

We assessed the impact of the additional features on runtime and memory usage by performing UINMF using 0, 1000, 2000, 3000, and 4000 extra features (Supplemental Fig. 1.18). Note that performing UINMF with 0 unshared features is equivalent to performing iNMF. For consistency, we used $\lambda = 5$ and $K = 30$ for all time and memory analyses.

2.5.9 Assessing the contribution of unshared features

In order to better quantify how the number of shared and unshared features used within an analysis interact to produce the observable differences in the quality of integration, we perform a stepwise analysis. We select two single-cell transcriptomic datasets taken from the mouse Primary Motor Area (MOp)[204], with two different protocols: SMART-seq and TenX. We then randomly downsampled the TenX dataset such that the datasets had an equivalent number of cells (6288 cells). We then defined the 2500 most variable genes between datasets.

From these 2500 most variable genes, we randomly sampled a set of 500 genes. Using these as shared features, we integrated the datasets using iNMF, letting $K = 30$, and $\lambda = 5$, selecting the best optimization score from three random initializations. After performing matrix decomposition, we performed quantile normalization using the SMART-seq dataset as our reference, and performed Louvain community detection with a resolution of 0.1. Using the SMART-seq labels as ground truth, we calculated the resulting purity and ARI scores. We then decreased the number of shared genes by 100, and repeated the integration, and metric calculations. We continued this process for 500, 400, 300, 200, and 100 shared genes. We repeated this process for ten different sets of randomly selected held-out genes.

To assess how including the unshared features benefits the dataset integration, we integrate the datasets using UINMF. The same 500 shared genes are used as shared features, and the remaining 2000 variable genes not selected are used as unshared features for the

SMART-seq dataset. For consistency, we use $K = 30$, $\lambda = 5$ as the parameters for UINMF optimization, selecting the optimal optimization of three initializations. Similarly, we perform quantile-normalization with SMART-seq as the reference dataset, and calculate the Louvain community detection with a resolution of 0.1. After calculating the purity and ARI scores, we repeat the UINMF analysis for the same 400, 300, 200, and 100 shared genes, adding the unshared genes to the unshared features, such that the number of total features for the SMART-seq dataset always totals to the 2500 top variable genes. We repeat this process for the ten sets of randomly selected held-out genes used for the iNMF analysis.

Lastly, we repeat the UINMF analysis, except instead of including the unshared features from the SMART-seq dataset, we include the unshared features as part of the TenX dataset. We also repeat this process for the ten sets of randomly selected held-out genes used for the iNMF analysis (Supplemental Fig. 1.10a, b). Using $n_neighbors = 20$, we plotted the resulting UMAPS from each analysis using 500 shared features and only 100 shared features (Supplemental Fig. 1.10c–f).

2.5.10 Selecting K and λ

The default parameter settings for iNMF are $K = 30$ and $\lambda = 5$. To allow for a fair comparison between iNMF and UINMF, we used these parameters settings for both algorithms when performing the SNARE-seq integration with intergenic peaks, the SNARE-seq integration with the 1020 gene STARmap dataset, and the cross-species analysis. For the spatial transcriptomic dataset integrations using osmFISH (33 genes) and STARmap (28 genes), iNMF required the selected K to be less than the number of genes. Since one of the advantages of our algorithm is that it does not have this constraint, we selected a value of $K = 40$ for both spatial transcriptomics integrations, the largest K we could select without severely impacting the alignment scores (Supplemental Fig. 1.19). To select λ for the spatial transcriptomics datasets, we selected $\lambda = (10, 1)$ for the STARmap integration with the scRNA-seq data. The higher penalty is assigned to the STARmap data, and the use of the vectorized lambda showed improved alignment scores over five initializations (Supplemental Fig. 1.20). We also showed the results of using $\lambda = 5$ for the osmFISH dataset in order to highlight that default choice of lambda still provides a significant improvement compared to iNMF, and that the results were not driven by the use of a vectorized λ (Supplemental Fig. 1.20).

2.5.11 Integration of three species

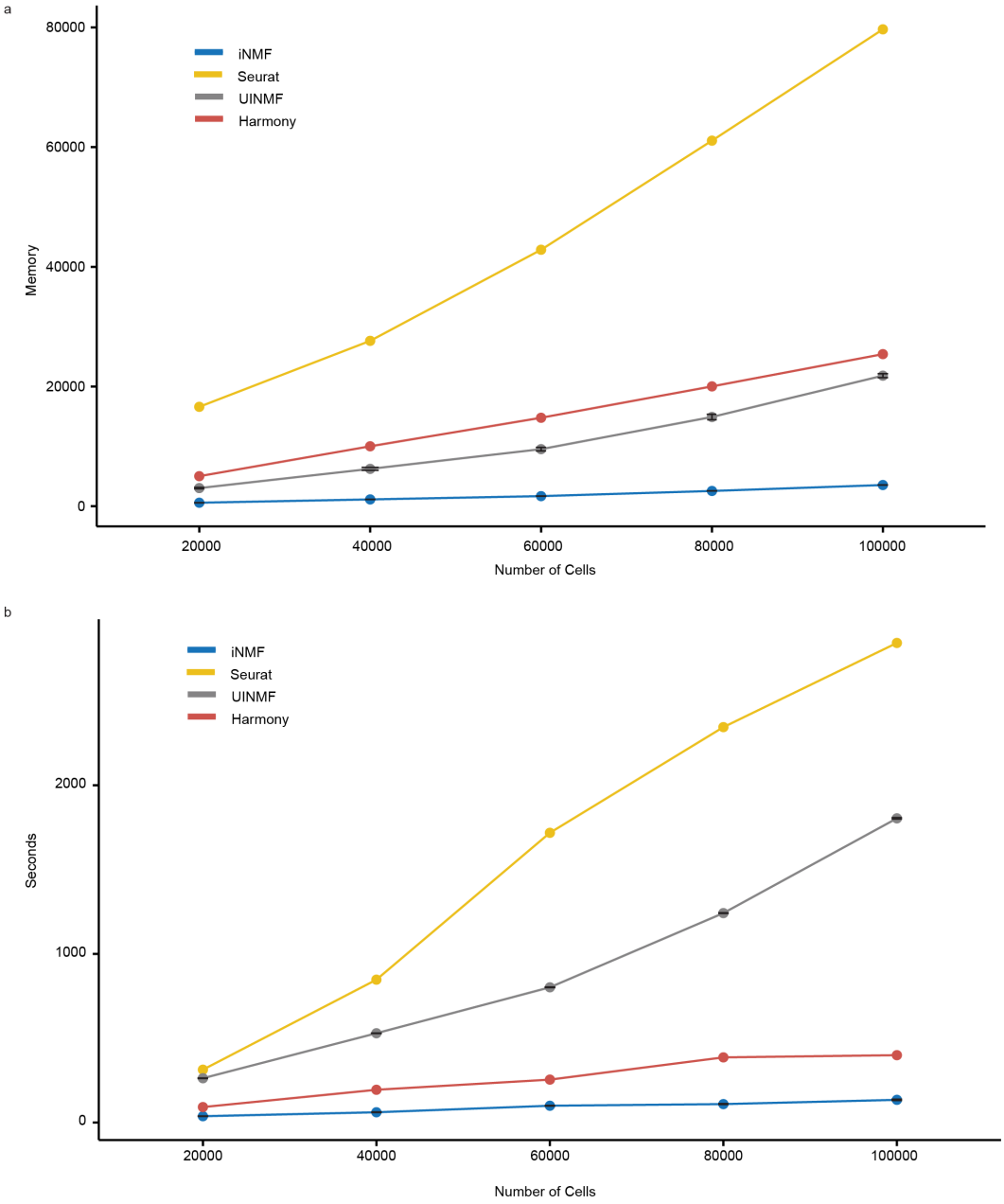
To illustrate the extensibility of the UINMF algorithm to accommodate multiple datasets of unshared features, we performed a cross-species integration of the primary motor cortex data of the human, mouse, and marmoset, sequenced with droplet-based Chromium v3. To elucidate reproducibility and consistency, we downsampled the available excitatory neuron population from each species following the methods of the original publication[26]. This sampling technique resulted in the selection of 8242 mouse excitatory neurons, 8166 marmoset excitatory neurons, and 7805 human excitatory neurons.

To identify non-orthologous and orthologous genes, we downloaded the NCBI table of orthologous genes between species, and identified genes listed as orthologous between all three species, resulting in 14,448 orthologous genes between datasets. Using a variable gene threshold of 0.2, we selected 2757 shared, orthologous genes between the species. To select unshared orthologous genes, we used a threshold of 0.3 for the human data, 0.25 for the mouse, and 0.2 for the marmoset dataset. We used a variable threshold for the non-orthologous genes such that we could retain a relevant population of unshared features for each dataset, despite the differences in sequencing quality and depth. These thresholds resulted in the selection of 353 non-orthologous mouse genes, 290 non-orthologous marmoset genes, and 671 non-orthologous human genes.

We then performed the integration with 3 distinct initializations, keeping the best optimization value, using $K = 20$, $\lambda = 10$. For quantile normalization, we used the human dataset as a reference, as this was previously reported to have the highest median neuronal gene detection[26], and used $knn_k = 50$. For Louvain community detection, we used a resolution of 0.25.

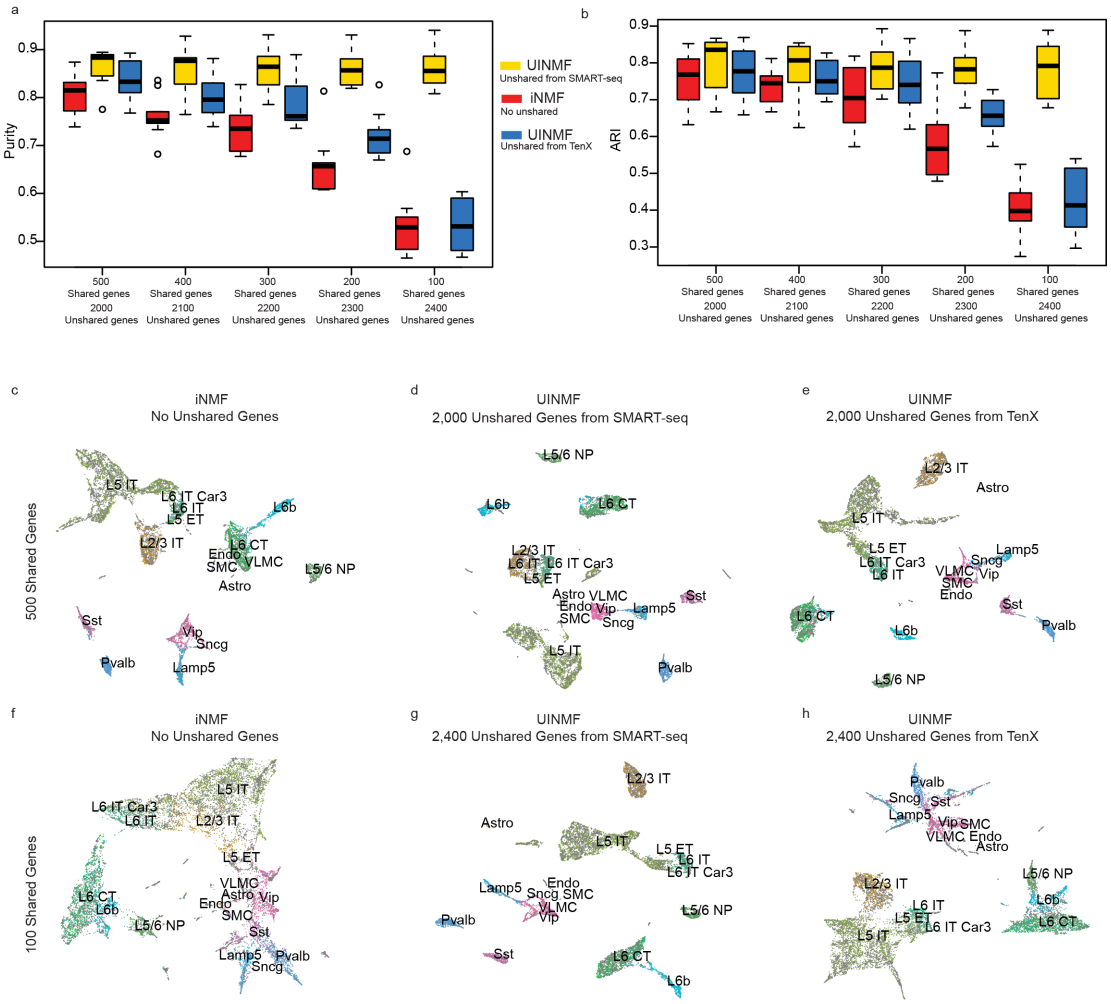
2.5.12 Supplementary Tables and Figures

Figure 2.9: Addition of unshared features is not time or memory prohibitive



Incorporating unshared features does not require prohibitive increases in time or memory usage. Using the STARmap and Dropviz datasets, which have 282 shared genes, we sampled equal numbers of cells from each dataset, taking additional cells from Dropviz once all 32,845 cells from STARmap were sampled. We ran UINMF with an additional 2,775 unshared genes, as shown in Fig. 4, while the other approaches used only 28 features. We benchmarked the increase in peak memory usage (**a**) and computational time (**b**) for UINMF, as well as other dataset integration methods that are incapable of including additional features, across an increasing number of cells. Data are presented as mean values +/- SEM.

Figure 2.10: Incorporating unshared features is especially advantegous when there are few shared features



Incorporating unshared genes significantly improves integration performance, particularly as number of shared genes decreases. In terms of both Purity (**a**) and ARI (**b**), including the unshared features from the SMART-seq and TenX data preserves higher dataset integration quality as the number of shared features declines. The bold line indicates the median data value, and the boundaries of each box are defined by the first and third data quartiles (25% and 75%, respectively). The upper (lower) whiskers extend from to highest(lowest) point within 1.5 of the interquartile range. Outliers beyond the whiskers are plotted as points.Including the unshared features from the SMART-seq dataset has the most added benefit, perhaps a result driven by the higher quality of the SMART-seq data. The difference in dataset integration is also visually apparent, as there are clear, distinguishable clusters for all three algorithms using 500 shared features (**c,d,e**), but with only 100 shared features, iNMF produces a lower resolution UMAP (**f**). The UINMF algorithm used to include the high quality SMART-seq unshared features maintains substantial integrity, with distinct populations of *Lamp5*, *Pvalb*, and *Sst* cells (**g**). Lastly, UINMF used to include unshared features from TenX generally maintains higher quality integrations, but is subject to a substantial decline in purity and ARI using only 100 shared features (**h**).

Figure 2.11: Unshared gene loadings largely contribute to nonneuronal type reconstructions

a

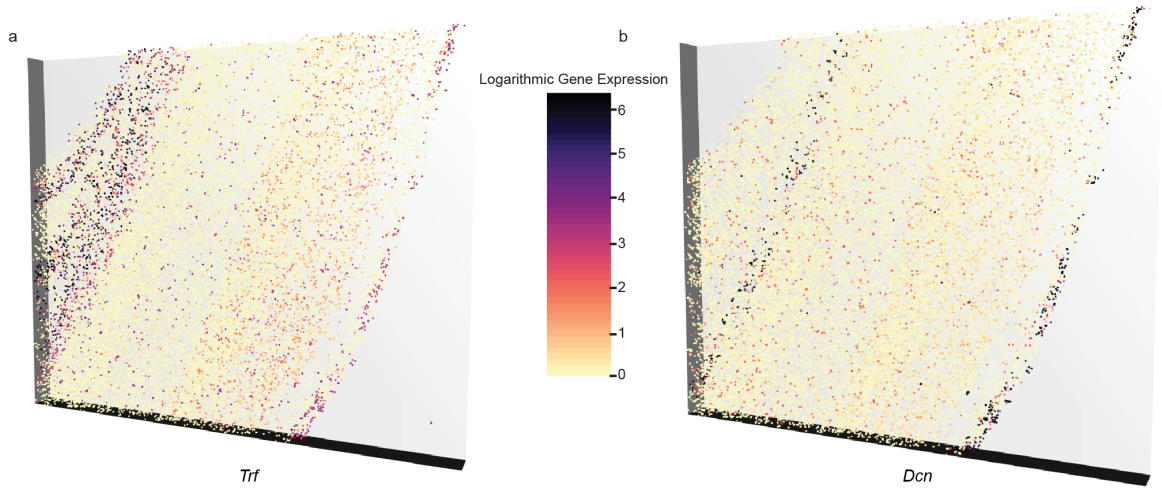
Factor	W Loadings	V^1 Loadings	V^2 Loadings	U Loadings
1	9.832	0.001	0.115	0.490
2	6.867	0.000	0.685	0.075
3	10.572	0.001	0.085	0.071
4	5.808	8.698	0.000	0.296
5	10.993	0.050	0.679	0.130
6	10.218	0.000	0.000	0.347
7	8.535	0.000	0.402	0.139
8	3.076	44.167	0.000	0.138
9	3.236	43.604	0.001	0.198
10	8.558	0.000	0.051	0.133
11	1.408	0.001	1.092	0.245
12	14.251	0.000	0.556	0.061
13	8.851	0.000	0.022	0.461
14	3.911	0.000	0.436	0.493
15	1.579	0.000	1.130	0.360
16	6.416	249.953	0.014	0.088
17	4.249	0.005	0.179	0.158
18	3.575	0.000	0.026	0.288
19	10.437	0.002	0.062	0.036
20	2.435	0.001	0.546	0.094
21	0.622	399.544	0.000	0.254
22	11.129	0.013	0.194	0.115
23	2.954	0.001	1.221	0.784
24	11.161	0.092	0.479	0.055
25	11.418	75.898	0.068	0.138
26	1.975	120.449	0.000	0.227
27	15.958	0.000	1.568	0.132
28	2.930	0.000	0.000	0.543
29	2.846	0.000	0.219	0.307
30	1.883	0.000	0.939	0.466
31	9.085	0.000	1.766	0.253
32	5.280	0.000	0.000	0.190
33	3.886	0.000	0.105	0.188
34	8.090	0.005	0.000	0.303
35	3.715	0.000	0.403	0.267
36	0.776	17.815	0.000	0.161
37	10.299	0.002	0.010	0.111
38	5.102	0.000	0.226	0.152
39	3.872	0.000	0.015	0.287
40	2.856	0.000	0.264	0.265

b

Factor	U Loadings	Annotation
23	0.7839	Microglia/Macrophage
28	0.5429	Multiple Clusters
14	0.4933	Polydendrocytes
1	0.4902	Endothelial Stalk
30	0.4664	Mural
13	0.4607	Fibroblasts
15	0.3605	Endothelial Stalk
6	0.3472	Macrophage
29	0.3072	Noise
34	0.3034	Astrocytes

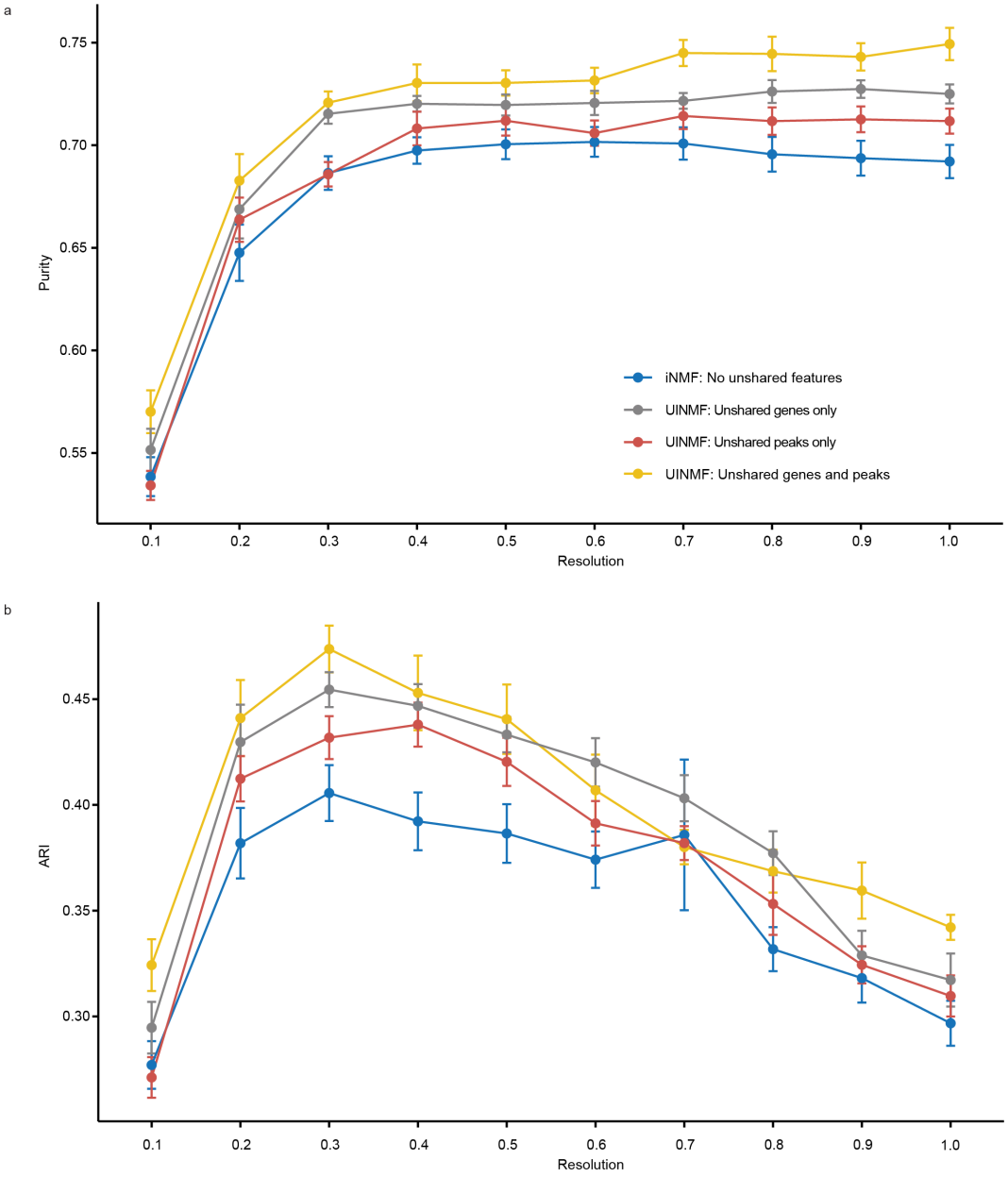
Relative contributions of metagenes to cell-type-specific reconstructions. Each matrix (W, V^1, V^2, and U) has a distinct contribution to the final matrix reconstruction (**a**), with the shared features, W, generally contributing the most, and the dataset-specific metagenes (V^1, V^2) contributing the least. The highest ten factor loadings, and their associated cell type, are shown for U matrix decomposition (**b**). The associated cell types are primarily nonneuronal.

Figure 2.12: KNN imputation for missing genes reflect expected distributions



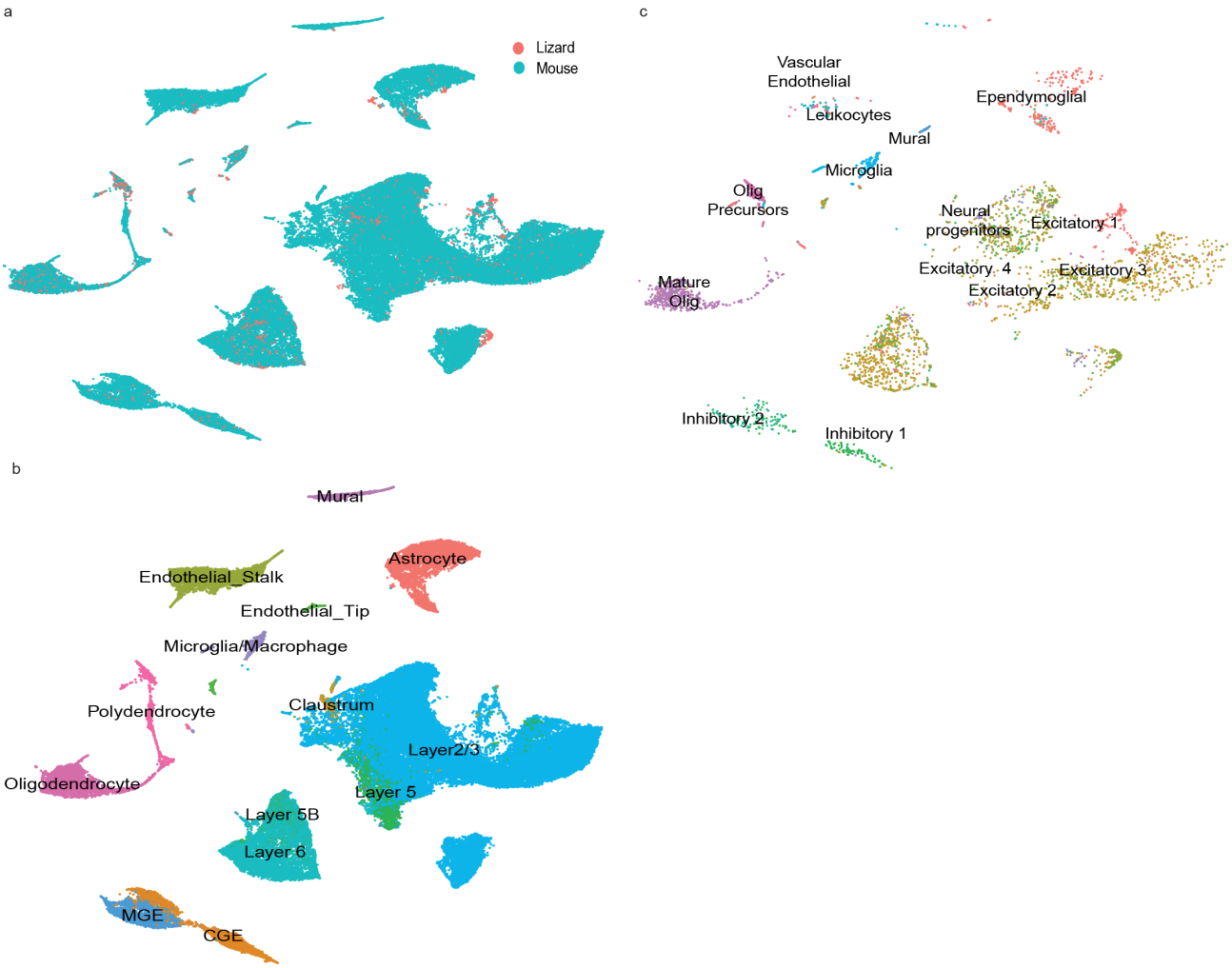
Spatial expression patterns imputed for *Trf* and *Dcn*. After performing KNN imputation, we are able to visualize the imputed expression profiles of the *Trf* (**a**) and *Dcn* (**b**) genes. *Trf*, a recognized marker gene of oligodendrocyte cells[38], consequently shows the greatest expression in the region composed of white matter. *Dcn* has previously been found in three types of vascular leptomeningeal cells (VLMCs)[178], which are known to comprise vascular structures. Therefore, we expected to see cells with highest imputed expression of *Dcn* expression near the blood brain barrier.

Figure 2.13: Using unshared genes has a greater improvement on the integration than use of chromatin features



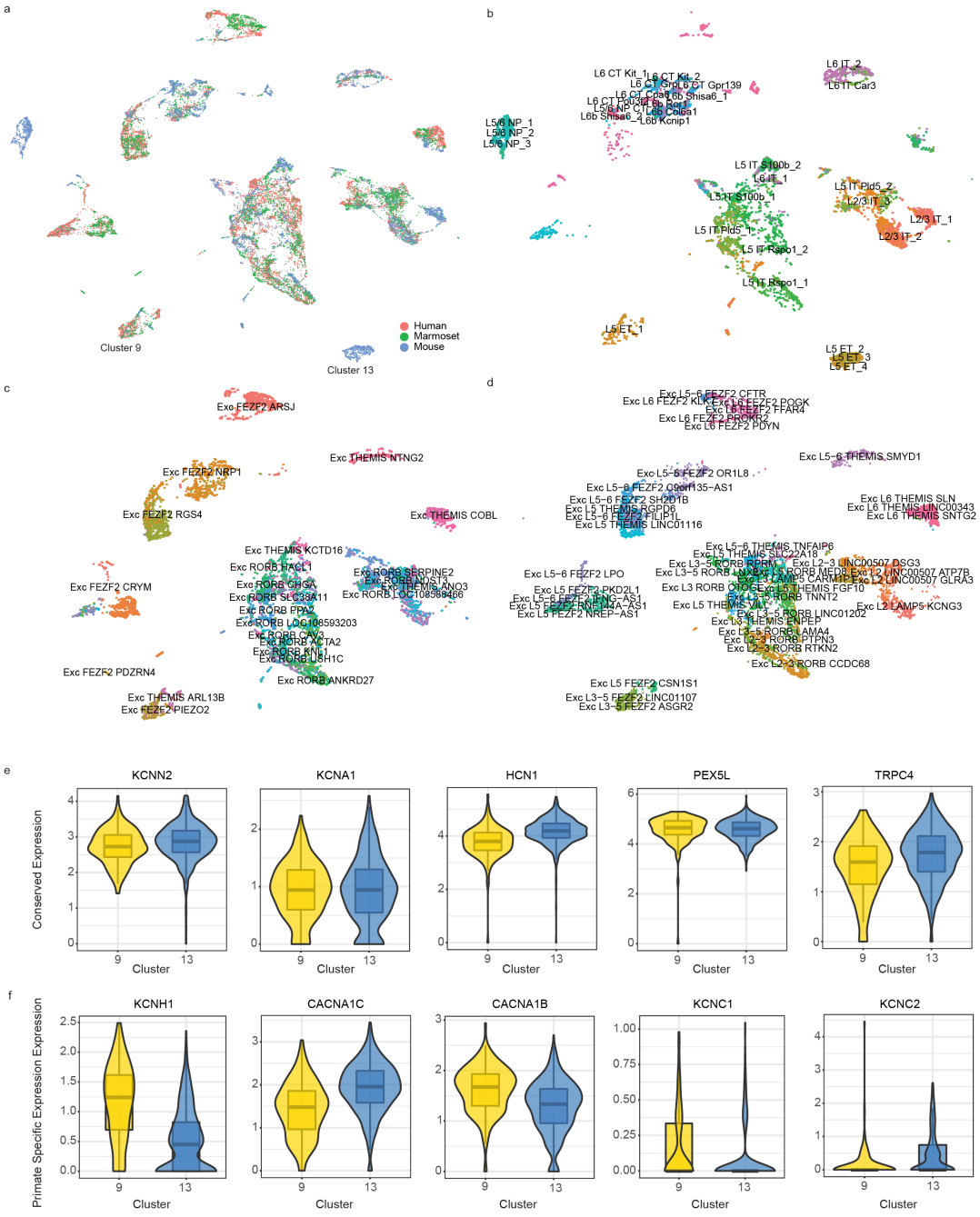
Unshared genes and unshared chromatin peaks each improve integration performance. To quantify the individual benefit of using the unshared genes and unshared peaks, we perform the SNARE-seq and STARmap integration with no unshared features, only the unshared genes as unshared features, only the unshared chromatin peaks as unshared features, and both unshared genes and unshared chromatin regions as unshared features. To demonstrate how the inclusion of each set of features changes the integration quality, we show the difference in purity (**a**) and ARI (**b**) metrics. Data are presented as mean values +/- SEM.

Figure 2.14: Cross-Species analysis using only orthologous genes



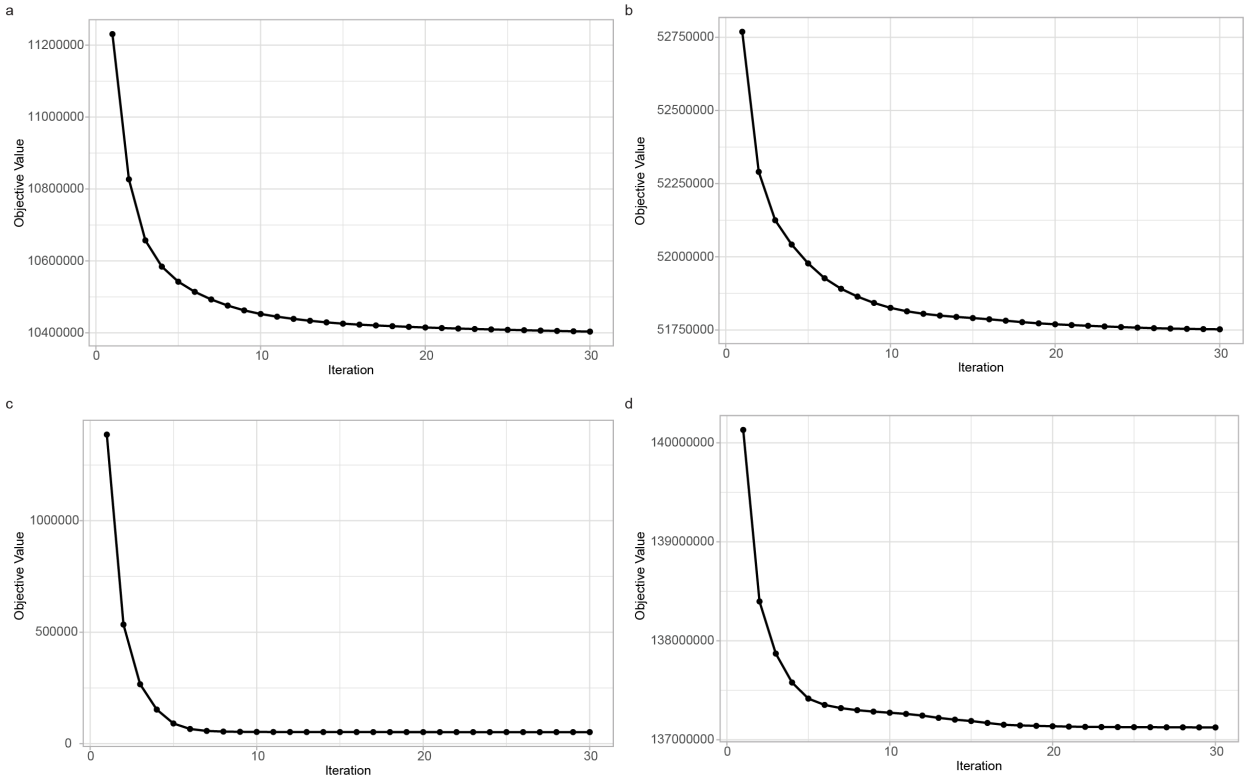
Integrating the lizard and mouse datasets using only orthologous genes, we still show adequate alignment between the two datasets (**a**). To examine cell type correspondence between the two datasets, we examined the mouse (**b**) and lizard (**c**) cells separately, labeled with their originally published labels.

Figure 2.15: UINMF integration of mouse, marmoset, and human datasets



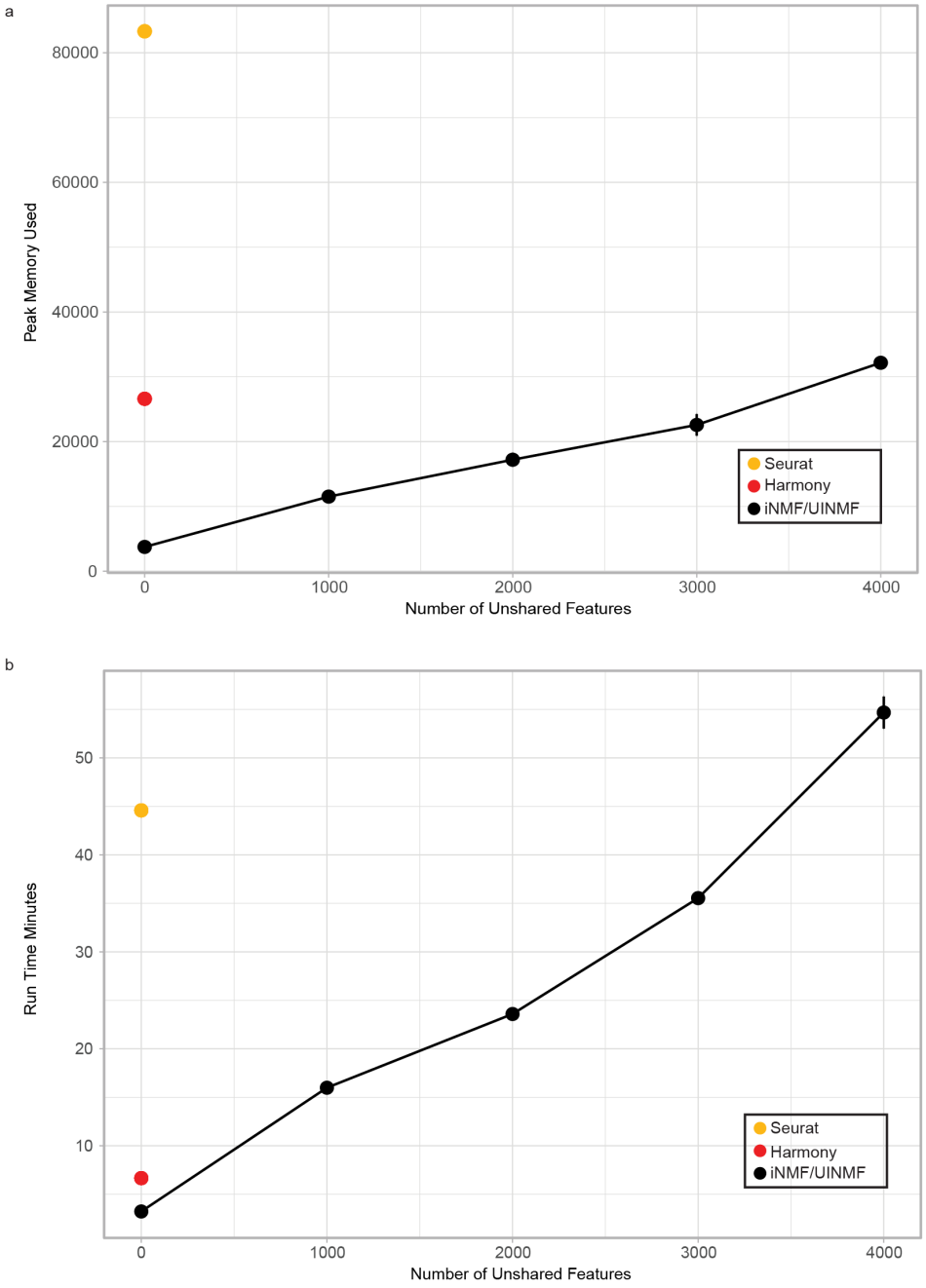
UINMF integrates excitatory neurons from mouse, marmoset, and human using non-orthologous genes. After confirming that the cells from each species are generally well mixed within the latent space (**a**), we examine the correspondence between the originally published labels for the mouse (**b**), marmoset (**c**), and human (**d**). We further investigate why only a single mouse L5 ET cluster aligns with the primate L5 ET neurons.

Figure 2.16: BCD Algorithm for UINMF Converges Consistently



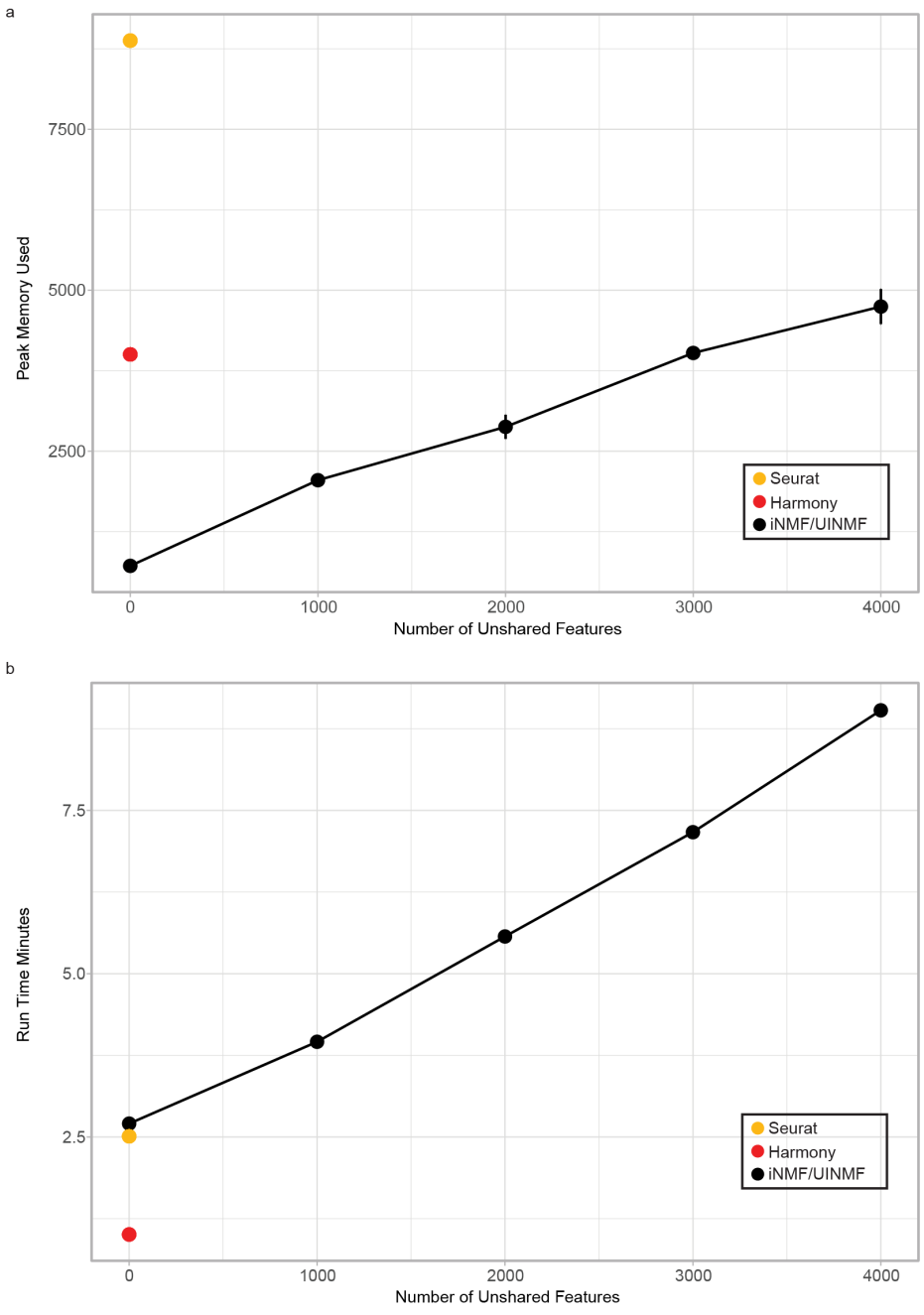
The objective function values for iNMF (**a**) and UINMF (**b**) across progressive iterations for the osmFISH and scRNA-seq analysis. The objective function values for iNMF (**c**) and UINMF (**d**) across progressive iterations for the SNARE-seq and STARmap data set integration. The first initialization is not shown, as it was large enough that including it resulted in the loss of the ability to visually identify the convergence of the successive objective values.

Figure 2.17: Time and memory usage of UINMF is competitive with existing methods when integrating STARmap and Dropviz datasets



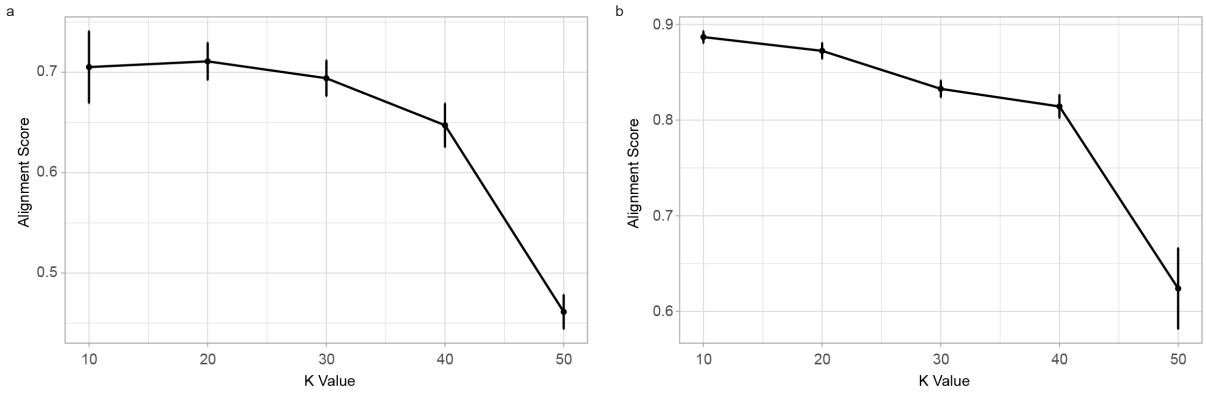
The increase in time and memory usage resulting from the addition of extra features is manageable. We compare the difference in memory usage (**a**) and run time (**b**) over 5 initializations for each set of unshared features when integrating the STARmap data and scRNA-seq data (28 shared genes). Note that performing UINMF with zero additional features is the equivalent of performing iNMF. Data are presented as mean values +/- SEM.

Figure 2.18: Time and memory usage of UINMF is competitive with existing methods when integrating STARmap and SNARE-seq datasets



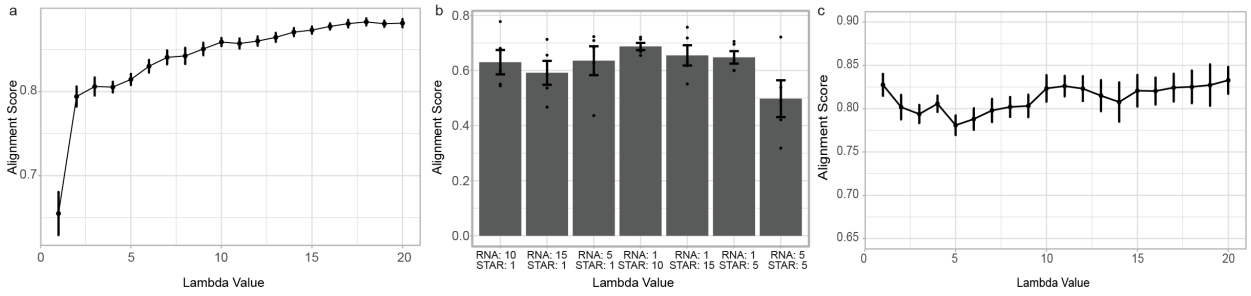
The increased time and memory usage from using UINMF is appropriate for the increase in the number of features used. Using the SNARE-seq and STARmap data, we analyze the memory usage (**a**) and runtime (**b**) of UINMF, Seurat v3, and Harmony. There are 944 shared features. It should be noted that including 0 unshared features for UINMF is the equivalent of performing iNMF. Data are presented as mean values +/- SEM.

Figure 2.19: The choice of K does not severely impact the alignment score



UINMF Alignment Score Is Relatively Robust to K. The alignment scores for the STARMAP (**a**) and osmFISH (**b**) are not severely impacted by the choice of K until an unusually large value of $K(K > 40)$ is chosen. Alignment scores were averaged over ten initializations. Data are presented as mean values +/- SEM.

Figure 2.20: UINMF Alignment Score Is Relatively Robust to Lambda



We calculated the alignment score of the algorithm across varying values of lambda for the SNAREseq and STARmap data integration. The results of ten initializations are shown (**a**). For the STARmap data, we found the penalty 10,1, where the higher penalty is assessed to the STARmap dataset, yielded the highest alignment over 5 random initializations (**b**). Likewise, we calculated the alignment score over 10 random initializations for a variety of lambdas for the osmFISH integration (**c**). Data are presented as mean values +/- SEM.

Figure 2.21: Summary of shared and unshared features used for integration

Dataset 1	Dataset 2	Dataset 1 Cells	Dataset 2 Cells	Shared Features	Unshared Features	Source of Unshared Features	Type of Unshared Features
SNARE-seq snATAC[2]	SNARE-seq scRNA[2]	10,309	10,309	2,589	2,000	Dataset 1	Intergenic Peaks
STARmap 3D[10]	DROPviz[29]	32,845	70,514	28	2,775	Dataset 2	Genes
osmFISH[12]	DROPviz[29]	6,471	70,020	33	2,000	Dataset 2	Genes
SNARE-seq[2]	STARmap 2D[10]	10,309	2,522	944	4,119	Dataset 1	Genes (2,668), Peaks (1,431)
Lizard[37]	DROPviz[29]	4,187	71,639	1,979	166	Dataset 1	Genes

CHAPTER 3

Constructing a Whole-Brain Cross-Modality Molecular Atlas with Spatial Resolution

3.1 Preamble

The following work has been submitted to a peer-reviewed journal for publication in conjunction with the work described in Chapter 4. For the work described in the following chapter, Joshua D. Welch conceived analysis ideas. Gaurav Gadhvi performed pre-processing on molecular samples (ATAC) for quality control. Joshua Sodicoff and Yichen Wang derived and implemented SiNMFiD. April R. Kriebel and Joshua D. Welch implemented RCTD. April R. Kriebel performed the molecular data integrations. April R. Kriebel, Joshua Sodicoff, Ming-Ching Wen, Albert Hung, Jingyi Liu annotated molecular cell clusters. Joshua D. Welch and April R. Kriebel designed the figures and drafted the manuscript.

3.2 Introduction

The fundamental goal of neuroscience is to understand the molecular cell types that compose neural circuits and their dynamic activity in cognition and behavior. The BRAIN 2025 report described a series of steps necessary to achieve this goal [2]. First, they emphasize that it is necessary to characterize the molecules that each type of cell uses. Second, we must discern how the cells are spatially arranged into anatomical structures.

The field of neuroscience has made huge strides on both problems. To understand the molecules used by each type of cell, a growing number of molecular whole-brain mouse atlases have been generated. For example, Yao et al. and Langlieb et al. have generated single nucleus and single cell datasets, capturing transcriptomic profiles of individual cells across the brain [94, 205]. To capture the epigenetic landscape, Zu et al. and Liu et al. have used snmC-seq and snATAC-seq sequencing assays, respectively [104, 219]. The cumulative

results of these endeavors are robust, single-cell level datasets that provide whole-brain coverage from the perspective of a single modality. The clear next step is the integration of these datasets to create multi-modal cell type profiles characterized by measurements from multiple modalities. In an initial effort by Liu et al., they separately integrated their snmC-seq data with sc10Xv2/3 [205], snATAC-seq [219], and SMART-seq [205] data, yielding three distinct integration spaces. Such an approach fails to represent the distinct datasets in a single, shared latent space. Furthermore, including sn10Xv3 samples, in addition to single-cell transcription measures, in atlas generation attempts is critical, since scRNA and snRNA have been shown to differ in cell type assignment on the basis of marker gene assignment [141].

Here, we describe the creation of our whole-brain, integrated atlas. Using online iNMF [60], we simultaneously integrate cells from six distinct modalities into a single, shared latent space. The iNMF framework accommodates for dataset specific variance, and constructs the latent space using biological interpretable metagene loadings [197]. Using our integrated cell profiles, we then investigate the spatial arrangement of such cell populations within the brain. Leveraging two distinct spatial transcriptomics datasets [94, 97], we estimate the spatial distribution of our defined cell types using two distinct spatial deconvolution methods. By providing estimated distributions for our molecular profiles, we address the second task set out by the BRAIN 2025 report [2].

3.3 Results

3.3.1 Brain-Wide Integration and Spatial Mapping of Molecular Cell Types from Single-Cell Transcriptome and Epigenome Data

To construct a single-cell, multi-modal atlas from transcriptome and epigenome profiles, we gathered whole-brain mouse datasets sequenced using SMART-seq [205], sn10Xv3 [94], sc10Xv2/3 [205], snmC-seq [104], and snATAC-seq [219]. All of these datasets were generated from dissociated tissue, resulting in the detailed spatial information for each cell being lost. In addition to using different sequencing modalities, the data generating labs also used different regional divisions to dissect the brain (Fig. 3.1a). Based on the overlaps among dissection regions, we categorized the brain into 18 distinct regions that are as small as possible while maintaining correspondence among dissections used to generate all data types. The final regions, as well as the cell counts for each region, before and after quality control, can be found in Supplemental Tables 3.1-2. After separating the data by region,

we grouped the cells for each region into non-neuronal cells, glutamatergic neurons, and GABAergic/other neurons using the high-level labels from the individual data-generating labs. We then performed online integrative nonnegative matrix factorization (online iNMF) using our previously published LIGERv1.0.0 [197, 106, 59] software to integrate the cells from the differing modalities into the same latent space (Fig. 3.1b). Online iNMF decouples time and memory usage from dataset size by operating on mini-batches of cells, making it particularly useful for this large-scale analysis. After performing factorization with online iNMF, we used the latent factors to perform Leiden community detection, identifying joint clusters across all data types (Fig. 3.1c). We integrated a total of 10,453,583 single-cell profiles (after quality control) across 54 integration analyses (three sub-analyses for each of the 18 regions) and manually annotated the resulting clusters. Our integration results are publicly available on the NeMO Analytics portal, and UMAP plots for each analysis can also be found in the Supplementary Figures (3.3-56).

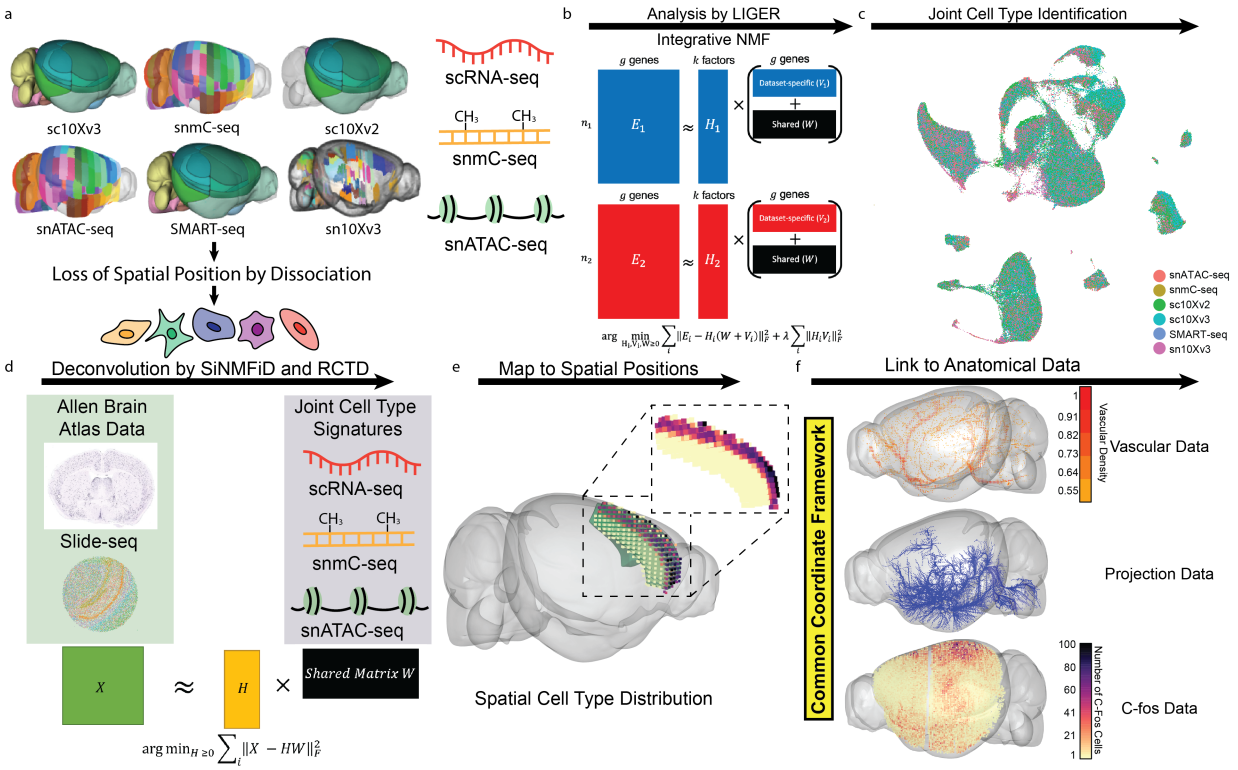
We next used these joint cell type assignments in concert with brain-wide spatial transcriptomic data to map the spatial distribution of each cell type within the Allen Institute's common coordinate framework (CCF) [194]. We used two different types of spatial data: the Allen Brain Atlas in situ hybridization data, which contains measurements for targeted genes quantified in 200 μm voxels across the mouse brain and a recently published Slide-Seq dataset, which measured transcriptome-wide expression across 101 coronal sections (spaced 100 μm apart) at 10 μm resolution within each section [94]. Note that the Slide-Seq data was registered to the Allen Brain Atlas CCF. We used two different computational approaches to map cell types in space: a novel deconvolution algorithm called supervised iNMF-informed deconvolution (siNMFiD) and the previously published RCTD algorithm (Fig. 3.1d) [37]. We developed siNMFiD for the unique challenge of deconvolving the Allen Brain Atlas ISH data [97] from multi-omic cell types and also used RCTD to deconvolve the Slide-seq dataset [94]. As a preliminary confirmation of the deconvolution results using our integrated data, we examined the spatial distribution of the cortical layers across four regions of the cortex SS, MOp, MOs, and VIS (Fig. 3.57-60). Each cortical region exhibits the appropriate assignment of the cortical layers, such that each cortical neuron type has an identified spatial distribution that corresponds to its expected area of occupancy within the cortex, and the expected layer ordering of cortical neurons is observed. By completing this deconvolution, we effectively mapped the spatial distributions of our molecular cell types (Fig. 3.1e) in the same CCF as previously published brain-wide anatomical data, including vascular density, mesoscale connectivity, and neuron activity in response to footshock (Fig. 3.1f).

3.3.2 Neuronal Cell Types Exhibit Highly Characteristic Spatial Distributions

The spatial distributions of neurons provide a potential bridge between molecular and anatomical measurements. For example, if a neuron type exclusively occupies a small volume of space, then imaging measurements of that position will largely reflect the properties of that neuron type. The more distinctive or characteristic the spatial distribution of a cell type, the easier it will be to identify from spatial position alone.

We examined the spatial distributions of the neuronal cell types in our analysis and visualized the broad class, spatial centroid, and volume for each type (Fig. 3.2a). Many of the known aspects of neuronal spatial arrangement are apparent from this visualization. For example, GABAergic cortical neurons are known to be molecularly similar across cortical areas; in contrast, the neurons of the Interbrain are composed of small, distinct nuclei that

Figure 3.1: Overview of approach for linking molecular and anatomical data types across the whole mouse brain



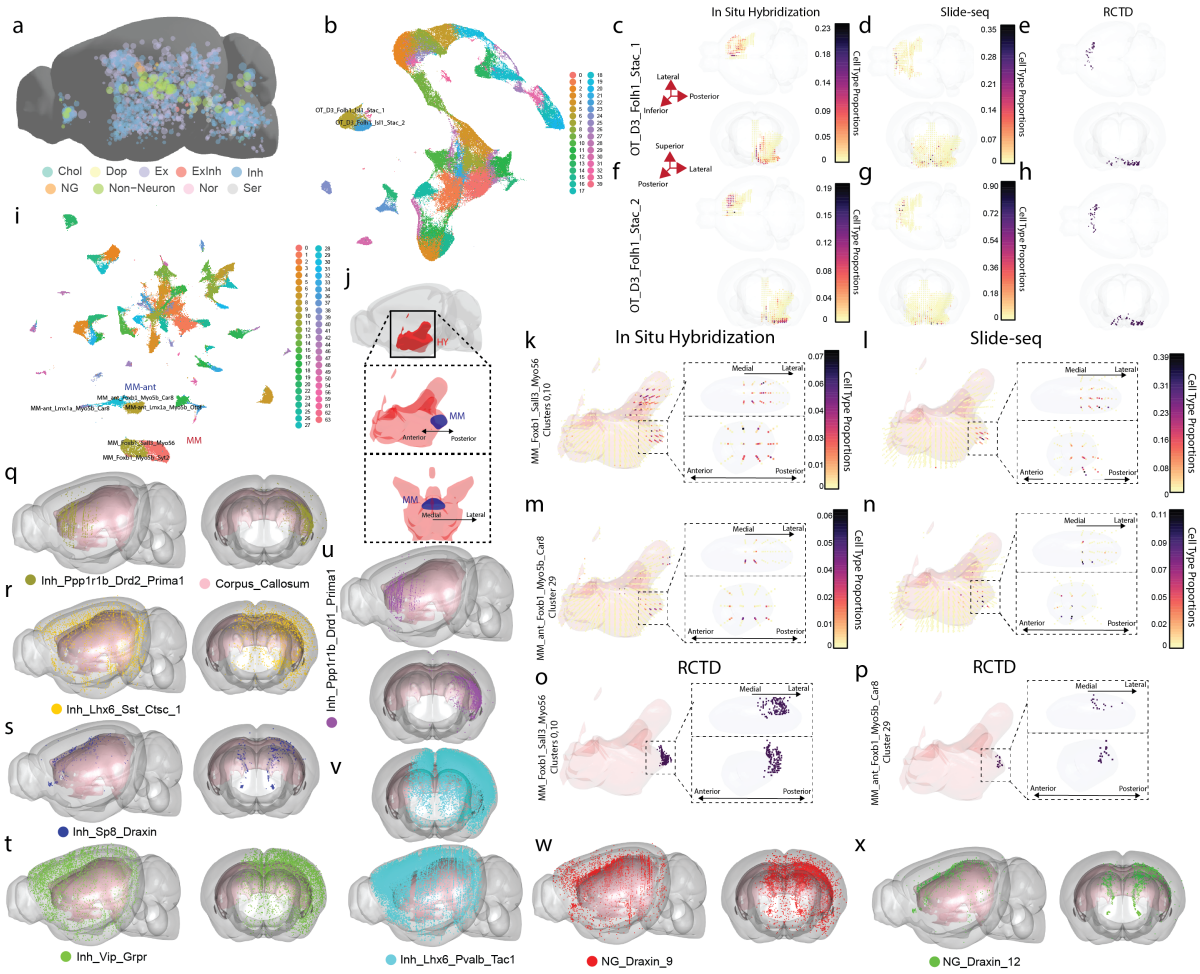
(**a**) Dissection regions for each modality, with different colors indicating different dissection regions used by the data generating labs. Areas that are gray and without color indicate that no data was captured from this region. (**b**) Diagram of LIGER integrative non-matrix factorization approach used to integrate single-cell transcriptome and epigenome datasets. Colored rectangles indicate matrices: (E_i) is a dataset of (n_i) cells by g genes that is decomposed into cell factor loadings (H_i), metagene loadings shared between datasets (W), and metagene loadings that are unshared between datasets (V_i) . (**c**) Example LIGER integration result: UMAP plot resulting from the integration of the primary motor cortex GABAergic neurons is shown with cells colored by their sequencing modality.(**d**) Diagram of deconvolution approach to map spatial distributions of molecular cell types. Allen Brain Atlas ISH data (X) was deconvolved with siNMFiD, using the shared metagenes matrix (W) inferred during the LIGER integration. The proportion of each cell type within a voxel was estimated by calculating cell factor loadings (H) for each spatial voxel. (**e**) After completing the deconvolution, the derived cell type distributions were mapped into the Allen Brain Atlas Common Coordinate framework (CCF), where each dot represents a voxel, and the shade of the dot is determined by the estimated cell type proportion for a specific cell type. (**f**) The spatial distribution of each cell type was used to link to anatomical data registered to the same CCF, such as vascular density, mesoscale connectivity, or brain-wide neuron activity.

demonstrate much greater molecular heterogeneity. Thus, the visualization shows that the volume occupied by each neuron type is largest toward the superior portion of the cerebrum, where the centroids of the cortical layers would lie, and decreases in the more inferior portions of the Cerebrum, in the Interbrain regions. These restricted spatial distributions are what may allow links between molecular profiles and anatomical features based on spatial position in a common coordinate system.

To quantify how distinctive the spatial distributions of neuronal cell types are, we registered Slide-seq beads into 100 μm voxels throughout the brain and calculated the number of voxels that contained at least one confidently assigned bead for each cell type. We refer to the total number of 100 μm voxels that a cell type can be found in as the cell type volume; for example, a cell type with a volume of 5 is present within 5 voxels across the brain. To examine the spatial distribution and variety of cell types within the brain, we plotted each cell type by its centroid, calculated as the centroid of all the voxels where a cell type was confidently assigned (Fig. 3.2a). We stratified the cell types by their major cell type distinctions, and the size of each dot reflects the volume occupied by that particular molecularly defined cell type. Overall, 95% of neuronal cell types within the brain occupy 892 or fewer voxels, corresponding to a total volume of less than .892 mm^3. We calculate the volume occupied by the majority of cell types in two ways. First, we use the whole brain volume of 415 mm^3, as reported by Xu et al [201]. Using this volume, most neuronal cell types occupy less than 0.430% of the brain. Next, we use the ABA coordinate framework's [194] 505,359 registered voxels to estimate that most neuronal cell types occupy less than 0.353 % of the brain. Both reported numbers indicate that the spatial distributions of neuronal cell types tend to be highly characteristic and localized.

We further identified several examples of neuron types with particularly unusual and distinctive spatial distributions. One example is the cells present in the Islands of Calleja, which are spatially distributed in disconnected clusters within the pallidum and striatum. These groups of cells exist in two locations, the olfactory tubercle (in the striatum), and along the border of the septum (pallidum), nucleus accumbens (striatum), and the nucleus of the diagonal band (pallidum) [56]. The cells are highly molecularly and histologically distinct from the D1 and D2 neurons of the striatum [211] and express the D3 Dopamine receptor. While little is known about the role of the Islands of Calleja, recent reports suggest that grooming is one function in which they play a regulatory role [211]. We identified 2 molecularly distinct D3-expressing clusters within the integration of the pallidum inhibitory and other neuronal cell types (Fig. 3.2b), and we used our molecular profiles to deconvolve both the ABA ISH data and Slide-Seq data (Fig. 3.2c,d,f,g). We observed a clear increase

Figure 3.2: Neuronal cell types exhibit highly distinctive spatial distributions



(**a**) Each cell type is plotted by its centroid within the gray outline of the brain, where the size of the dot reflects the number of voxels that cell type is found within, and the color of the dot corresponds to the primary cell type assignment for that cell. (**b**) UMAP plot of the LIGER integration of the GABAergic and other non-glutamatergic neurons from the Pallidum, where the colors signify distinct clusters. (**c-h**) The colored cubes set within the gray outline of the brain are voxels within the pallidum. The color of each voxel reflects the cell type proportion of a cell type within that particular voxel, such that a darker color indicates that a higher proportion of the voxel is occupied by a particular cell type. The cell type proportions of OT_D3_Folh1_Stac_1 and OT_D3_Folh1_Stac_2 are shown. The cell type proportions were determined using siNMFiD on ISH data and Slide-seq data, and by using RCTD on Slide-seq data. Cell type proportions are shown using 200 µm voxels. (**i**) UMAP plot of the LIGER integration of Glutamatergic neuronal types of the Hypothalamus, where the colors represent distinct clusters. An anatomical diagram showing the hypothalamus, colored red, within the brain, shown in gray. (**j**) The medial mammillary nucleus (MM), one of the small nuclei found within the hypothalamus, is colored blue. The small colored cubes represent individual voxels within the hypothalamus, and the color of each voxel is determined by the cell type proportion of a specific cell type.

Figure 3.2 *(previous page)*: **(k-p)** The voxels are shown within the hypothalamus, colored red. The voxels were subset to those within the MM, shown in blue, and provide both medial/lateral as well as anterior/posterior viewing windows. The cell type proportions for the MM_Foxb1_Sall3_Myo56 and the MM_ant_Foxb1_Myo5b_Car8 are shown. SiNMFiD was used to deconvolve the ISH and Slide-seq data, and also deconvolve the Slide-seq data using RCTD. **(q-x)** The outline of the brain is shown in gray, with the corpus callosum shown in pink. Each Slide-seq bead confidently assigned the selected cell type is plotted within each brain, the color of the dots indicating the cell type.

in cell type proportions for the D3 cell type populations in the inferior portion of the pallidum, corresponding to the expected distribution of the Islands of Calleja [211]. We used the same molecular clusters to run RCTD on the Slide-seq data and obtained very similar distributions (Fig.3. 2e,h).

Another interesting example of a neuron type with highly distinct spatial distribution is the Foxb1+ cells of the medial mammillary (MM) nucleus of the hypothalamus. These neurons form a multi-dimensional spatial gradient, with density that varies simultaneously along the anterior-posterior and medial-lateral axes. Known for its role in recollective memory [136], the MM may be composed of even smaller nuclei [123]. Recent investigations into the molecular profiles of MM neurons suggest both posterior/anterior and medial/lateral axes of transcriptional variance [123]. Examining the spatial distribution of the molecularly defined cell types MM_Foxb1_Sall3_Myo56 and MM_ant_Foxb1_Myo5b_Car8 within the MM indicates this dual-axis gradient variation (Fig. 3.2i). The hypothalamus is located within the interbrain, and the MM is located at the posterior end of the hypothalamus (Fig. 3.2j). Using our SiNMFiD algorithm, we deconvolved both the in situ hybridization data (Fig. 3.2k,m) and the Slide-seq data (Fig. 3.2l,n), and examined the spatial distribution of the MM_Foxb1_Sall3_Myo56 and MM_ant_Foxb1_Myo5b_Car8 cell types. From visual inspection, it is clear that MM_Foxb1_Sall3_Myo56 has a more posterior/lateral distribution, while MM_ant_Foxb1_Myo5b_Car8 has a more anterior/medial distribution. While the general spatial distribution of the two cell types is similar when either the ISH or the Slide-seq data is used, the higher resolution of the Slide-seq data does yield more clearly defined spatial compartmentalization of the cell types. However, our deconvolution algorithm is still able to accurately assign a distinct spatial distribution profile from data with lower resolution. We also observed similar cell type gradients when using RCTD on the Slide-Seq data (Fig. 3.2o-p).

Surprisingly, we additionally observed several neuron types whose spatial distribution overlapped with fiber tracts. While it is well-established that the white matter is composed

of axons, oligodendrocytes, and other glia, the presence of neurons in the white matter has also been reported [190, 171]. The molecular characterization of these white matter interstitial neurons remains incomplete [171], and our analysis provides a new opportunity to characterize them. To avoid artifacts from imperfect registration, we focused on neurons within the corpus callosum, whose large volume is well above the registration error of Slide-Seq data (estimated at 22–94 μm by Langlieb et al.[94]). To build confidence in the spatial location of neuron types within the fiber tracts, we manually examined each cell types that mapped within fiber tracts, the expression of their marker genes within the Allen Brain Atlas ISH data, and previous literature.

The majority of white matter interstitial neuron types we identified were GABAergic, consistent with previous reports [171]. We found neurons with molecular similarity to striatal spiny projection neurons (Inh_Ppp1r1b_Drd1_Prima1, Inh_Ppp1r1b_Drd2_Prima1), Sp8-expressing neurons (Inh_Sp8_Draxin), and neurons with molecular similarity to cortical interneurons (Inh_Lhx6_Sst_Ctsc_1, Inh_Lhx6_Pvalb_Tac1, Inh_Vip_Grpr) (Fig. 3.2q,r,s,t,u,v). We further validated these findings by looking for the expression of key marker genes for these cell types in soma that occur within fiber tracts using ABA ISH data (Fig. 3.62). Expression of parvalbumin (Pvalb) and somatostatin (Sst) has been previously reported within white matter [154], further supporting our findings. Additionally, because Slide-Seq beads may overlap with 0, 1, or more cells, we confirmed our findings using multiple replicates of MERFISH data from a recent publication (Fig. 3.61); this indicates that expression of these marker genes occurs within regions identified as individual segmented cells.

We also identified two cell types within fiber tracts that express markers of neurogenesis, NG_Draxin_9 (Fig. 3.2w) and NG_Draxn_12 (Fig. 3.2x) We confirmed sparse expression of the Draxin marker gene in somas within the white matter using the ABA ISH data (Fig. 3.62). The spatial distribution of these types parallels the known location of the rostral migratory stream (RMS) and the subventricular zones (SVZ) [39]. The RMS is responsible for conveying newly formed neurons from the subventricular zones to the olfactory bulb [14], and is found in close physical proximity to the white matter. However, the cells we identified are located well within the corpus callosum, above the RMS, and thus may represent cells from the RMS that have aberrantly migrated into the fiber tract.

3.4 Discussion

Herein, we describe the generation of a single-cell whole brain molecular atlas that is jointly constructed by leveraging 6 distinct modalities, including both transcriptomic and epigenomic measures. By performing our integration analyses in 18 distinct regions, further

separated into non-neuronal cells, excitatory neurons, and inhibitory and other neurons, we establish clusters that meaningfully separate cell types within a region and broad cell class. In well-studied regions of the brain, we identify cell types consistent with expectation, and in regions less well characterized, we find robust clusters composed of multiple modalities. To disseminate our atlas as a reference tool for future research, we collaborated with NeMO analytics to share our resulting clusters. Our data can be accessed on the NeMO Analytics Portal.

To define the spatial distribution of our defined molecular cell types, we deconvolved two distinct spatial transcriptomics datasets, using both the established RCTD algorithm [37], as well as debuting our novel deconvolution algorithm, SiNMFiD. We validate the resulting cell type distributions using molecular cell types with particularly distinct spatial characteristics. The estimated spatial distributions, provided by RCTD and SiNMFiD, are widely accessible to the scientific community on the Welch Lab Neuroglancer Portal.

To continue to advance our understanding of the molecular cell types within the brain, and their corresponding spatial distributions, resources will need to be devoted to manually curating brain cell types. A common, standardized nomenclature will need to be established, such that all established datasets, as well as future datasets, share common cell labels, enabling easy comparisons and analyses. A direct corollary to establishing common nomenclature will be the identification of marker genes that consistently identify cell types, regardless of the modality of origin. Optimally, the establishment of a common nomenclature and standardized marker genes will allow annotation to be semi-automated.

Semi-automated annotations would allow for newly-generated or unannotated datasets to be projected into the same latent space as previously annotated data, and label-transfer to occur. While this process will likely need manual curation, the notion of automatically generated draft annotations is an attractive one that would dramatically reduce the time needed for cell type annotation, which is typically a time consuming and error-prone process. After establishing a robust annotation pipeline that provides consensus annotations at the molecular level, further exploration would be needed to define consensus spatial distributions.

Defining a consensus spatial distribution for specific molecular cell types is difficult because of sample variance and current technological limitations. While whole-brain datasets with, or near, single-cell resolution have been generated [209, 94], these are typically captured from a single specimen. To identify a generalized spatial distribution per cell type, multiple brains will need to be sampled, a costly endeavor. Computational methods for analyzing spatial transcriptomics data may also need to be re-evaluated. The lack of ground truth for validating cell segmentation and spatial deconvolution pipelines have left ambiguity in how well these methods are performing, and benchmarking datasets and metrics need to be

established to accurately evaluate these methods. After optimizing how spatial transcriptomics data is processed, datasets from multiple animals can be successfully integrated into a consensus distribution of cell type distribution, and typical between animal variance can be quantified. To deeply understand the brain, there is an immense amount of work that remains, but we envision our whole-brain molecular atlas with spatial resolution as being a critical stepping stone in defining both the molecules that characterize each cell type, as well as the spatial arrangement of these cell types.

3.5 Methods

3.5.1 Defining Anatomical Regions Across Datasets

The anatomical dissection regions across datasets (i.e. SMARTseq, sc10Xv2/3, sn10Xv3, snmC-seq, and snATAC-seq) were not standardized between labs. To assign anatomical regions, we established common regional annotations between datasets, such that as much regional granularity as possible was retained, while still encompassing several modalities in each defined region. Ultimately, we define 18 distinct regions across the mouse brain, and assign each cell to a common anatomical region on the basis of the data-generating lab's annotation. Similarly, using the data-generating lab's original annotations, we perform an initial subsetting of the data for each region into three distinct categories: non-neuronal cell, excitatory neurons, and inhibitory and other neurons. For the SMART-seq data, which we did not have initial annotations for, we performed an initial integration of only the SMART-seq cells in order to separate the SMART-seq data into excitatory neurons, inhibitory neurons, and non-neuronal cell populations. The cell counts for each region, and each analysis, for each modality can be found in Supplementary Tables 3.1-2.

3.5.2 Online Integration of Single-Cell Datasets

To integrate single-cell profiles across modalities, we use online integrative non-negative matrix factorization, a component of the LIGER package. The online capabilities of online iNMF were instrumental in integrating datasets of such scale. We use LIGER to perform data normalization, scaling, feature selection, and integration. Clustering was performed using the Leiden algorithm, and final plots were generated using the UMAP functionality of the LIGER package. For explicit details on the parameters used in each analysis, please see the Supplementary Table 3.3. We manually annotate each cluster.

3.5.3 Preprocessing for the ABA In Situ Hybridization Data

In 2007, Lien et al. performed in situ hybridization staining and imaging for 20,000 genes in 160 sagittal sections at 25 μm [97]. The total number of genes were slightly variable across regions. The resulting images were registered into the Allen Brain Atlas common coordinate framework [169], and the expression energy for each gene was calculated by taking the average intensity across all pixels. Ultimately, the ISH ABA data provides whole-brain gene expression in 200 μm^3 voxels. We use the anatomical distinctions within the ABA to subset the voxelized gene expression data into the 18 anatomical regions used in integration. Prior to deconvolution, we performed any translations necessary to ensure proper registration with the ABA anatomical region. To account for the sparsity of ISH data, we calculated the fraction of missing values by gene. Then, a Poisson distribution is fit, and all genes with more than a standard deviation above the mean of missing values are removed. Voxels with less than 150 reads for the selected variable genes are removed. Each gene is scaled by its variance, and the resulting matrix, $S_{spatial}$, is $g \times v$, where g is the number of genes retained after filtering, and v is the number of quality voxels. We impose a non-negativity constraint on the data since we scale without centering.

3.5.4 Preprocessing for the Whole-Brain Slide-seq Data

We utilize the whole-brain Slide-seq data generated by Langlieb et. al [94]. Briefly, the Slide-seq whole-brain mouse atlas is composed of sequenced 10 μm beads from 101 pucks, created using coronal slices across the brain. We separated the beads by region using metadata provided by the original publication [94]. Any beads outside the 18 designated regions of interest were not deconvolved.

Prior to deconvolution, we scale the data, and perform any translations necessary to ensure proper registration with the ABA anatomical region. We impose a non-negativity constraint on the data since we scale without centering. We summarize the Slide-seq data into 100 μm^3 and 200 μm^3 voxels, where we sum the mRNA expression for all beads that fall within each voxel. Voxels with less than 150 reads for the selected variable genes are removed.

After the completion of preprocessing and quality control, we have 18 region-specific spatial transcriptomics matrices, $S_{spatial}$, that are z voxels by g genes for both the Slide-seq and ISH data.

3.5.5 SiNMFiD Reference Dataset Downsampling and Gene Selection

For deconvolution, we leverage the cell type annotations derived from our regional integration of the transcriptomic and epigenomic datasets. For each region, the number of unique cell types equals K. We select all transcriptomic datasets available for a region as reference datasets, or E_j, where j is the index of each distinct reference dataset used. For the majority of regions, we were able to leverage transcriptomic profiles from four reference datasets: sc10Xv3, sc10Xv2, sn10Xv3, and SMART-seq. Therefore, for our deconvolution analyses, $j \leq 4$. For specific parameters used for each deconvolution analysis, please see Supplementary Table 3.4.

Within our reference datasets, E_j, the number of cells assigned to each cell type, K, is not equivalent across K. Let $t_{k,j}$ equal the number of cells assigned to cell type K in reference set j. To ensure that each molecular cell type is well-represented during deconvolution, we downsample the reference datasets, $E_{1,\ldots,j}$, such that $\sum_{j=1}^{j} t_{k,j} \leq q$. For all deconvolution analyses, we set $q = 500$.

To select genes that are highly variable between cell types, we use the Kruskal Wallis one-way nonparametric ANOVA for each gene for each cell type K. Using the resulting χ^2 statistic, we select the top 2000 genes, such that $g = 2000$ for all SiNMFiD deconvolution analyses. We consequently subset each reference dataset E_j for the selected g variable genes and $t_{k,j}$ cells. We remove all genes not shared between reference datasets, and normalize each gene by dividing the gene expression vector by its root mean sum of squares. Any non-numeric or negative entries are set to 0.

3.5.6 Leveraging ANLS to Derive a Shared W Among Reference Datasets

Prior to deconvolving the spatial transcriptomics datasets, we define gene signatures for each cell type K using all available reference datasets $E_{1,\ldots,j}$. For each reference dataset E_j, let C_j be a N by K matrix, where N is the number cells retained after downsampling and K is the number of molecular cell types. In previous publications, we have decomposed each dataset E_j into dataset-specific metagenes (V_j), shared metagenes (W), and cell factor loadings (C_j)[197], such that V_j and W are both K cell types by g genes. Now, we populate C_j with a one hot encoding of the cell type assignments, and solve for V_j and W with the minimizing the following objective function:

$$argmin_{W\geq 0, V\geq 0}\sum_{i=1}^{j}\left\|E_i - C_i(W+V_i)\right\|_F^2 + \lambda\left\|C_iV_i\right\|_F^2$$

$$W = argmin_{W\geq 0}\left\| \begin{pmatrix} C_1 \\ \vdots \\ C_j \end{pmatrix} W - \begin{pmatrix} E_1 & - & C_1V_1 \\ & \vdots & \\ E_j & - & C_jV_j \end{pmatrix} \right\|_F^2$$

$$V_i = argmin_{V_i\geq 0}\left\| \begin{pmatrix} C_i \\ \sqrt{\lambda}C_i \end{pmatrix} V_i - \begin{pmatrix} E_1 & - & C_iW \\ & 0_{N_i\times g} & \end{pmatrix} \right\|_F^2$$

The minimization problem is iteratively solved until convergence or a maximum number of iterations is reached. We define convergence as a decrease in the reconstruction error of less than 10^{-8} between iterations. We set the maximum number of iterations equal to 100; however, all analyses converged within 100 iterations. For all deconvolution analyses, we set $\lambda = 1$, where λ is a regularization term used to balance reconstruction error with the magnitude of the expected dataset specific expression.

To account for variation in library size by cell type when determining cell type composition during deconvolution, the mean library size for all cells of a given type is calculated. Let the vector of these values be s. W_s, the scaled gene signature matrix, is

$$W_s = \text{diag}(s)W$$

3.5.7 SiNMFiD Implementation for Deconvolution of Spatial Transcriptomics Data

After solving for the shared gene signatures of each cell type, given in W_s, we can construct an optimization problem to deconvolve the spatial transcriptomics data, represented as $S_{spatial}$, into a final $H_{Decon.}$ matrix. $H_{Decon.}$ is z voxels by K cell types, and provides the estimated cell type composition of each voxel z. We solve:

$$argmin_{H\geq 0}\|S_{spatial} - HW_s\|_F^2$$

To generate a matrix of the cell type proportions of each cell type K in each voxel z for each sample j, we divide each row of the $H_{Decon.}$ matrix by its sum.

3.5.8 Advantages of supervised iNMF-informed deconvolution

We present a novel spatial deconvolution algorithm, supervised iNMF-informed deconvolution (siNMFiD). While the development of a novel method was necessitated by the inability

of current methods to deconvolve the large 200 μm tissue volumes of the ISH ABA data, SiNMFiD has several attractive attributes. First, siNMFiD is not designed to be constrained to operating on spatial transcriptomics data generated from a single modality. Many current deconvolution algorithms operate on specific data types, such as Vizgen or Slide-seq. siNMFiD can perform deconvolutions on any count or intensity-based modality, where the accuracy of the deconvolution varies in direct proportion to read depth. As such, siNMFiD is capable of deconvolving voxelized Slide-seq data as well as 200 μm ISH data, although the quality of the deconvolution results is still susceptible to being impacted by poor quality spatial transcriptomic data. As such, poorer quality spatial transcriptomics data will likely yield less robust estimations of spatial distribution for specific molecular cell types in comparison to spatial transcriptomic data of higher quality and read count. Unlike most currently available methods, which assume 2-dimensional data types, siNMFiD is capable of accommodating 2- and 3-dimensional coordinates.

SiNMFiD is not limited to the selection of a single reference dataset. In such cases as atlas generation, there may be several transcriptomic datasets generated with multiple technologies. siNMFiD is capable of leveraging multiple single cell and nucleus sequencing references to deconvolve several modalities of spatial transcriptomic data with similar accuracy. This multimodality feature also supports the correction of technical effects. The removal of technical artifacts from the cell type signatures prior to deconvolution ensures that the most robust signatures are used in determining cell type proportions. Since spatial transcriptomic data carries technological biases of its own, it becomes especially important to perform deconvolution analyses with as generalizable signatures as possible. SiNMFiD fulfills a unique and necessary niche for spatial deconvolution needs not met by other existing algorithms.

3.5.9 Spatial Deconvolution of the Whole-Brain Slide-seq Data Using RCTD

Analysis of the Slide-seq data [94] was completed using the RCTD methodology from the spacexr R package proposed in Cable et. al. [37]. In essence, RCTD fits a linear combination of cell types derived from a single cell reference dataset with expected Poisson-distributed gene expression plus bead and technology-specific covariates to sample.

For application of the method, annotations generated during the region-specific single cell analysis were used and beads were analyzed together by the provided corresponding dissection regions. Labels with fewer than 25 cells assigned were removed for compliance with the constraints of the methodology. Input coordinates for the beads were "flattened" i.e. the z-coordinate was removed for compliance, though because the method does not account

for spatial localization no impact on results was expected. A minimum of 250 UMIs was set for beads. The method itself was run on doublet mode, allowing for the assignment of a bead to two or more cell types to account for technical effects. For final cell type assignments, RCTD returns 4 distinct labels: reject, singlet, doublet_uncertain, and doublet_certain. For our downstream analyses, we remove all beads labeled reject. For beads labeled singlet or doublet_uncertain, we assign these beads their most highly probable cell type identity. All doublet_certain beads we annotate with the top two most probable cell type annotations.

To confirm the integrity of our manual cluster annotations, we examined the spatial distribution of cortical layer neurons in key isocortical areas, i.e. MOp, MOs, SS, VIS. We observe appropriate patterning of the cortical neurons within space, with the L2/3 IT neurons lying most superior, and the increasing layers (L4, L4/5, L5, L6, and L6b) increasing in inferiority (Supplementary Fig. 3.57-60). We also compare the resulting cell type assignments with the ABA cortical layer distinctions. For each RCTD bead assigned a cortical layer type, we assess what ABA cortical layer it falls within anatomically. There is imperfect registration between the RCTD and ABA; consequently, perfect cell type assignments are not anticipated. However, the majority of cortical neurons annotated with a cortical layer are assigned to their corresponding primary layer in the ABA (Supplementary Fig. 3.57-60).

3.5.10 Determining Cell Type Radius

To ensure that each cell type has the opportunity to be represented evenly, we filter the Slide-seq beads such that we retain only beads assigned to the left hemisphere. We downsample the beads into 100 µm voxels, and count the number of voxels that express a cell type confidently. Confidence is defined as a score of > 0.3. We calculate the centroid of each cell type by finding the centroid of all the voxels in which it is confidently expressed. In Figure 3.2a, color is assigned by the major cell type label, and the size of each circle is a reflection of its cell type volume, where we set the dot size to the log normalization of the number of voxels a cell occupies.

3.5.11 Performing QC for Filtering Low Quality Fragments from snATAC-seq Data

In addition to the cells annotated as low-quality by the data-generating lab, we perform an additional level of quality control on the snATAC-seq data. All unique fragments per cell were extracted using snaptools [8] from the processed snap files based on the known cell barcodes. Using these unique read fragments as input, fragments overlapping the peak regions were identified by calling peaks using the MACS2 [5] peak-caller with default parameters. We

used pycisTopic [32, 7] for Fragments in Peaks calculations and filtered out all low-quality cells with a FRiP score smaller than 0.4, less than 1000 unique fragments, and a TSS enrichment score less than 4. This is in accordance with the parameters utilized by the original publication [32].

3.5.12 Division of SMART-seq Cells into Primary Cell Classes

Annotations from the data-generating labs made dividing the cells of each region into Excitatory Neurons, Inhibitory and Other Neurons, and Non-neuronal cells straight forward. However, we did not have a complete set of annotations for the SMART-seq cells. To determine primary cell classes for the SMART-seq cells, we used LIGER to perform dimensionality reduction on the SMART-seq dataset. Using all SMART-seq cells in a single matrix, we normalized the matrix, and selected the top 2,546 variable genes. We scaled, but did not center, the data. We performed online_iNMF with $K = 30$, $\lambda = 5$, $max.epochs = 20$. After running online_iNMF, we performed quantile normalization and louvain clustering, with a $resolution = 0.5$ and $k = 200$. Finally, we create a final UMAP using the runUMAP with parameters $n_neighbors = 30$, $min_dist = 0.30$, and $distance = "cosine"$. We then used marker genes to determine whether each SMART-seq clusters was non-neuronal, inhibitory or other neurons, or an excitatory neuron, and divided the SMART-seq data by region and cell class accordingly.

3.5.13 Supplementary Tables and Figures

Table 3.1: Cell Counts by Dataset and Region Prior to Quality Control

Region	Analysis	sc10Xv3	sc10Xv2	snATAC	Meth	sn10Xv3	Smartseq	Total
AI	2	24140		13145				37285
AUD	2	11251	2166	39414		36003		88834
CB	2	15419		6144		78341		99904
CTXsp	2	45588	1407					46995
HB	2	214411		121108		595391		930910
HPFENT	2	100823	3521	86619		157504		348467
HY	2	67142	12944	34330		60344		174760
MB	2	169143	7824	134676		389278		700921
MOp	2	3395	1400	18282		62320		85397
MOs	2	19505		16155				35660
OLF	2	25806	21340	57108		49179		153433
ORBACA	2	37777	1918	25534		38829		104058
PAL	2	50438		39241		139108		228787
RSP	2	20114	1218	40932		54760		117024
SS	2	18546	5116	50405				74067
STR	2	98480		68453		65333		232266
TH	2	61696	33311	80159		249814		424980
VIS	2	16540	2885	23681		30228		73334
AI	4	6213	13878	3896				23987
AUD	4	2925	21979	11686	1153	14156		51899
CB	4	51818		4601	6014	63854	6452	132739
CTXsp	4	11277	11306		2780		2902	28265
HB	4	51316		34193	11087	116801	9203	222600
HPFENT	4	17206	25238	12952	2584	39468	4337	101785
HY	4	48605	46058	33533	6047	20178	21765	176186
MB	4	68101	12964	31720	11388	93294	11934	229401
MOp	4	13190	27421	5378	1522	17414	2152	67077
MOs	4	3164	6260	3754	1095		4648	18921
OLF	4	49762	65059	23701	5210	128589	3799	276120
ORBACA	4	7734	23696	6629	1664	12763	1696	54182
PAL	4	51907		25008	9842	112534	6351	205642
RSP	4	2344	6926	7549	1138	12111		30068
SS	4	2823	12498	13148	3907		1621	33997
STR	4	160409		100160	15659	69278	9802	355308
TH	4	11151	14704	13644	4952	50348	4465	99264
VIS	4	4272	49011	7608		11265	8214	80370
AI	5	16039	38558	22978	1889		1139	80603
AUD	5	12868	106176	66602	10167	81295	2961	280069
CB	5	114866		91630	6046	490360	5627	708529
CTXsp	5	21647	31617		2176		3847	59287
HB	5	72221		14486	11925	150438		249070
HPFENT	5	63785	179751	139057	27578	199541	8943	618655
HY	5	47513	41847	2461	5590	32463		129874
MB	5	101678	9212	51953	18180	200854		381877
MOp	5	108221	94678	37149	7668	87893	4641	340250
MOs	5	14164	28225	26695	5668		4881	79633
OLF	5	13750	107814	79659	18931	15039	4759	239952
ORBACA	5	35316	103262	53085	12704	69624	6471	280462
PAL	5	6495			1927	22115		30537
RSP	5	22352	59882	70684	12287	69983	1533	236721
SS	5	16101	125041	124023	23309		7402	295876
STR	5	27032		13522	2807	42665		86026
TH	5	58190	83961	28425	12057	168212	5950	356795
VIS	5	37752	267600	63386	9621	72807	13406	464572

Table 3.2: Cell Counts by Dataset and Region After Quality Control

Region	Analysis	sc10Xv3 QC	sc10Xv2 QC	snATAC QC	Meth QC	sn10Xv3 QC	Smartseq QC	Total
AI	2	24068		10412				34480
AUD	2	11163	2063	33462		32054		78742
CB	2	14291		5861		70651		90803
CTXsp	2	45315	1407					46722
HB	2	214411		121107		595362		930880
HPFENT	2	99927	3476	84932		155514		343849
HY	2	66664	12526	34111		59856		173157
MB	2	168919	7820	121624		389207		687570
MOp	2	3231	1273	16381		56879		77764
MOs	2	19502		12703				32205
OLF	2	25806	21340	57108		49179		153433
ORBACA	2	37777	1918	25534		38829		104058
PAL	2	50400		38126		138235		226761
RSP	2	19356	1196	38610		51966		111128
SS	2	18538	5107	47653				71298
STR	2	98480		63409		65333		227222
TH	2	58971	29888	75098		245053		409010
VIS	2	16174	2813	21770		28933		69690
AI	4	6102	13721	3698				23521
AUD	4	2789	20922	10720	1038	12691		48160
CB	4	50711		3978	4293	48817	5910	113709
CTXsp	4	5737	4907		713		1738	13095
HB	4	50418		31640	9977	109265	9153	210453
HPFENT	4	16927	24601	12050	2383	34236	4291	94488
HY	4	44297	43102	29072	5487	16357	19877	158192
MB	4	67314	12873	31091	11171	85013	11573	219035
MOp	4	13142	27160	4915	1491	15353	2134	64195
MOs	4	3089	6249	3559	1059		4562	18518
OLF	4	49054	62993	21873	4779	124812	3214	266725
ORBACA	4	7608	22725	6058	1635	12414	1691	52131
PAL	4	51077		24262	9351	109009	6249	199948
RSP	4	2298	6656	6831	966	10257		27008
SS	4	2783	12414	12415	3684		1607	32903
STR	4	158747		99362	15050	68826	9772	351757
TH	4	6617	8646	8275	1564	27786	1332	54220
VIS	4	3893	44801	6123		8431	7582	70830
AI	5	13993	33208	19239	1637		1035	69112
AUD	5	12333	102076	61113	9296	77360	2761	264939
CB	5	114155		89770	5999	487791	4858	702573
CTXsp	5	20070	29485		1646		3610	54811
HB	5	52566		11490	10362	104793		179211
HPFENT	5	62664	176630	135087	22977	190950	8791	597099
HY	5	46435	40828	1816	5281	27369		121729
MB	5	94562	8585	40670	16095	144113		304025
MOp	5	108002	94428	37108	7626	86990	4612	338766
MOs	5	14089	27926	24944	5340		4861	77160
OLF	5	12906	97556	76965	16320	13608	4662	222017
ORBACA	5	35235	102638	51330	11904	67167	6469	274743
PAL	5	6088			1574	18226		25888
RSP	5	21841	58941	65466	12149	67740	1356	227493
SS	5	15929	124025	116560	23204		7392	287110
STR	5	16557		5572	1158	21140		44427
TH	5	56431	79304	27776	8998	146516	5607	324632
VIS	5	37217	259858	60798	9536	69581	13198	450188

Table 3.3: Parameters for Online iNMF Analyses

Region	Analysis	QuantNorm Ref	Leiden	minDist UMAP	K	Genes	Lambda	Batch Size	Seed
AI	2	sc10Xv3	0.5	0.1	30	2553	5	5000	123
AUD	2	sc10Xv3	0.5	0.1	30	2542	5	5000	123
CB	2	sc10Xv3	0.1	0.1	30	2512	5	5000	123
CTXsp	2	sc10Xv3	0.5	0.1	30	2523	5	5000	123
HB	2	sc10Xv3	0.1	0.1	30	2560	5	5000	123
HPFENT	2	sc10Xv3	0.5	0.1	30	2489	5	5000	123
HY	2	sc10Xv3	0.1	0.1	30	2506	5	5000	123
MB	2	sc10Xv3	0.1	0.1	30	2520	5	5000	123
MOp	2	sc10Xv3	0.5	0.1	30	2458	5	5000	123
MOs	2	sc10Xv3	0.5	0.1	30	2453	5	5000	123
OLF	2	sc10Xv3	0.5	0.1	30	2517	5	5000	123
ORBACA	2	sc10Xv3	0.5	0.1	30	2465	5	5000	123
PAL	2	sc10Xv3	0.1	0.1	30	2577	5	5000	123
RSP	2	sc10Xv3	0.5	0.1	30	2499	5	5000	123
SS	2	sc10Xv3	0.5	0.1	30	2507	5	5000	123
STR	2	sc10Xv3	0.1	0.1	30	2524	5	5000	123
TH	2	sc10Xv3	0.1	0.1	30	2573	5	5000	123
VIS	2	sc10Xv3	0.5	0.1	30	2462	5	5000	123
AI	4	sc10Xv3	0.5	0.1	30	2486	5	5000	123
AUD	4	sc10Xv3	0.5	0.1	30	2423	5	5000	123
CB	4	sc10Xv3	0.5	0.1	80	2569	5	5000	123
CTXsp	4	sc10Xv3	0.5	0.1	30	2438	5	5000	123
HB	4	sc10Xv3	1	0.1	80	2564	5	5000	123
HPFENT	4	sc10Xv3	0.5	0.1	30	2418	5	5000	123
HY	4	sc10Xv3	1	0.1	80	2528	5	5000	123
MB	4	sc10Xv3	1	0.1	80	2457	5	5000	123
MOp	4	sc10Xv3	0.5	0.1	30	2456	5	5000	123
MOs	4	sc10Xv3	0.5	0.1	30	2424	5	5000	123
OLF	4	sc10Xv3	0.5	0.1	30	2463	5	5000	123
ORBACA	4	sc10Xv3	0.5	0.1	30	2565	5	5000	123
PAL	4	sc10Xv3	1	0.1	30	2457	5	5000	123
RSP	4	sc10Xv3	0.5	0.1	30	2545	5	5000	123
SS	4	sc10Xv3	0.5	0.1	30	2436	5	5000	123
STR	4	sc10Xv3	1	0.1	30	2482	5	5000	123
TH	4	sc10Xv3	1	0.1	80	2512	5	5000	123
VIS	4	sc10Xv3	0.5	0.1	30	2493	5	5000	123
AI	5	sc10Xv3	0.5	0.1	30	2494	5	5000	123
AUD	5	sc10Xv3	0.5	0.1	30	2448	5	5000	123
CB	5	sc10Xv3	1	0.1	30	2518	5	5000	123
CTXsp	5	sc10Xv3	0.5	0.1	30	2397	5	5000	123
HB	5	sc10Xv3	1	0.1	80	2510	5	5000	123
HPFENT	5	sc10Xv3	0.5	0.1	30	2459	5	5000	123
HY	5	sc10Xv3	1	0.1	80	2461	5	5000	123
MB	5	sc10Xv3	1	0.1	80	2381	5	5000	123
MOp	5	sc10Xv3	0.5	0.1	30	2532	5	5000	123
MOs	5	sc10Xv3	0.5	0.1	30	2429	5	5000	123
OLF	5	sc10Xv3	0.5	0.1	30	2495	5	5000	123
ORBACA	5	sc10Xv3	0.5	0.1	30	2459	5	5000	123
PAL	5	sc10Xv3	1	0.1	30	2522	5	5000	123
RSP	5	sc10Xv3	0.5	0.1	30	2580	5	5000	123
SS	5	sc10Xv3	0.5	0.1	30	2517	5	5000	123
STR	5	sc10Xv3	1	0.1	30	2508	5	5000	123
TH	5	sc10Xv3	1	0.1	80	2475	5	5000	123
VIS	5	sc10Xv3	0.5	0.1	30	2429	5	5000	123

Table 3.4: Parameters for SiNMFiD Deconvolution Analyses

Region	Genes	QC Genes	Voxels	sc10Xv2	sc10Xv3	sn10Xv3	Smartseq	Seed	Lambda
AI	2009	1830	748	TRUE	TRUE	FALSE	TRUE	123	1
AUD	2013	1887	909	TRUE	TRUE	TRUE	TRUE	123	1
CTXsp	1976	1709	855	TRUE	TRUE	FALSE	TRUE	123	1
MOp	1977	1673	927	TRUE	TRUE	TRUE	TRUE	123	1
MOs	1950	1633	1139	TRUE	TRUE	FALSE	TRUE	123	1
OLF	2002	1839	4060	TRUE	TRUE	TRUE	TRUE	123	1
ORBACA	2022	1673	1571	TRUE	TRUE	TRUE	TRUE	123	1
RSP	2026	1747	1031	TRUE	TRUE	TRUE	TRUE	123	1
SS	1996	1826	2825	TRUE	TRUE	FALSE	TRUE	123	1
TH	2025	1746	1771	TRUE	TRUE	TRUE	TRUE	123	1
VIS	1981	1828	1283	TRUE	TRUE	TRUE	TRUE	123	1
MB	1958	1669	3222	TRUE	TRUE	TRUE	TRUE	123	1
HY	1997	1649	1534	TRUE	TRUE	TRUE	TRUE	123	1
HPFENT	1996	1862	3266	TRUE	TRUE	TRUE	TRUE	123	1
PAL	2041	1772	1062	FALSE	TRUE	TRUE	TRUE	123	1
HB	1995	1698	4174	FALSE	TRUE	TRUE	TRUE	123	1
CB	2000	1836	4699	FALSE	TRUE	TRUE	TRUE	123	1
STR	2035	1824	3756	FALSE	TRUE	TRUE	TRUE	123	1

Figure 3.3: UMAP integration of the AI Non-neuronal cells

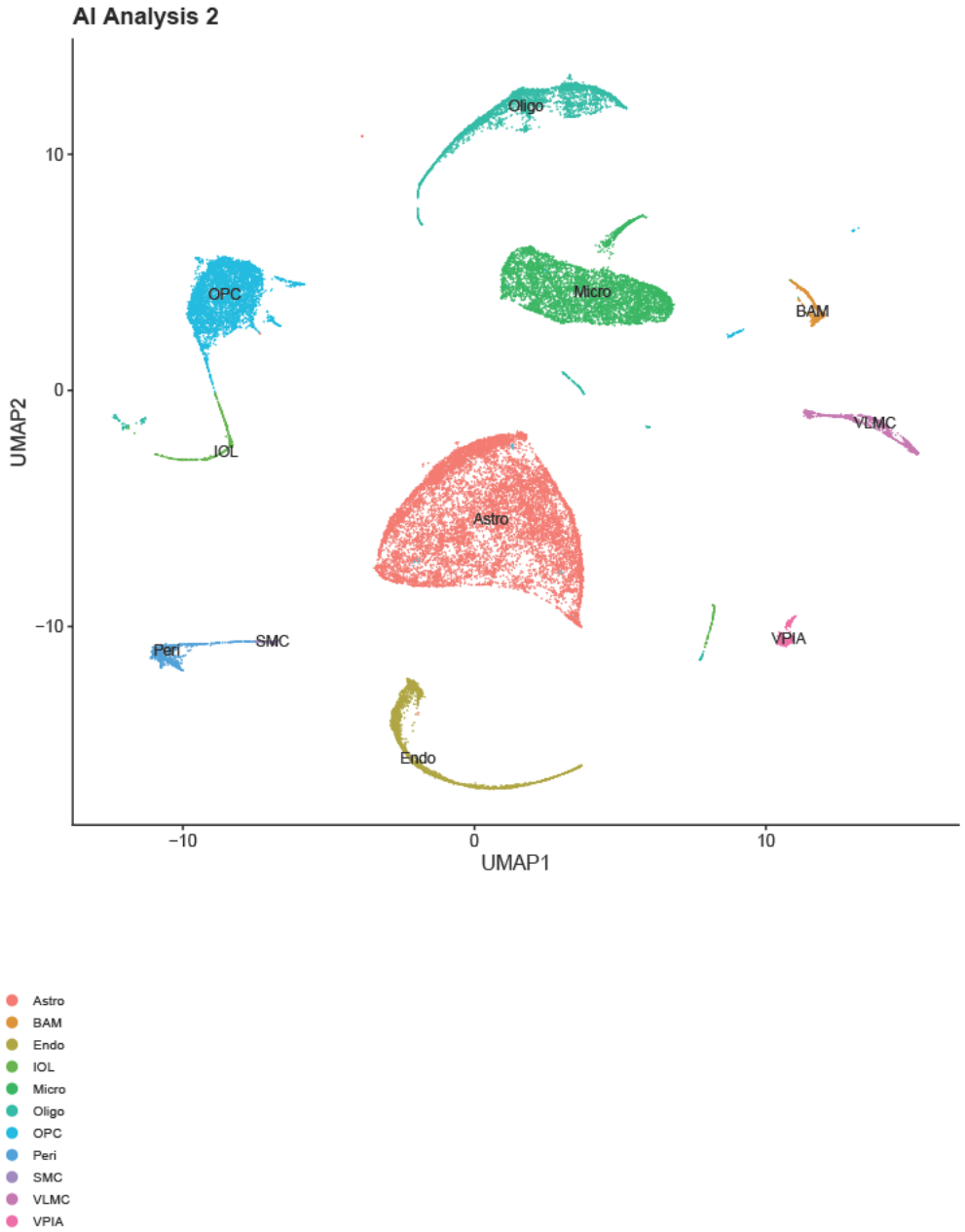


Figure 3.4: UMAP integration of the AI Inhibitory and Other Neurons

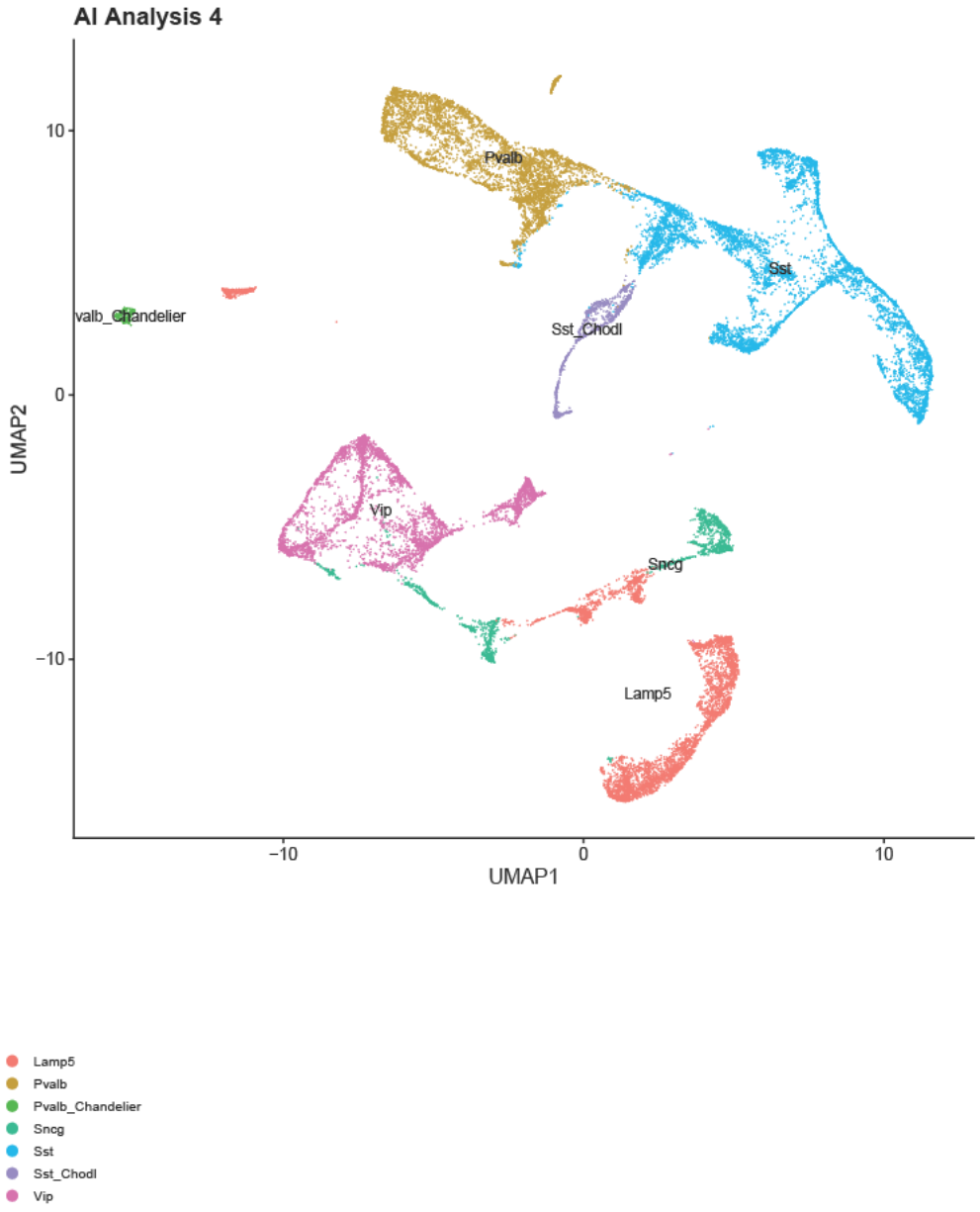


Figure 3.5: UMAP integration of the AI Excitatory Neurons

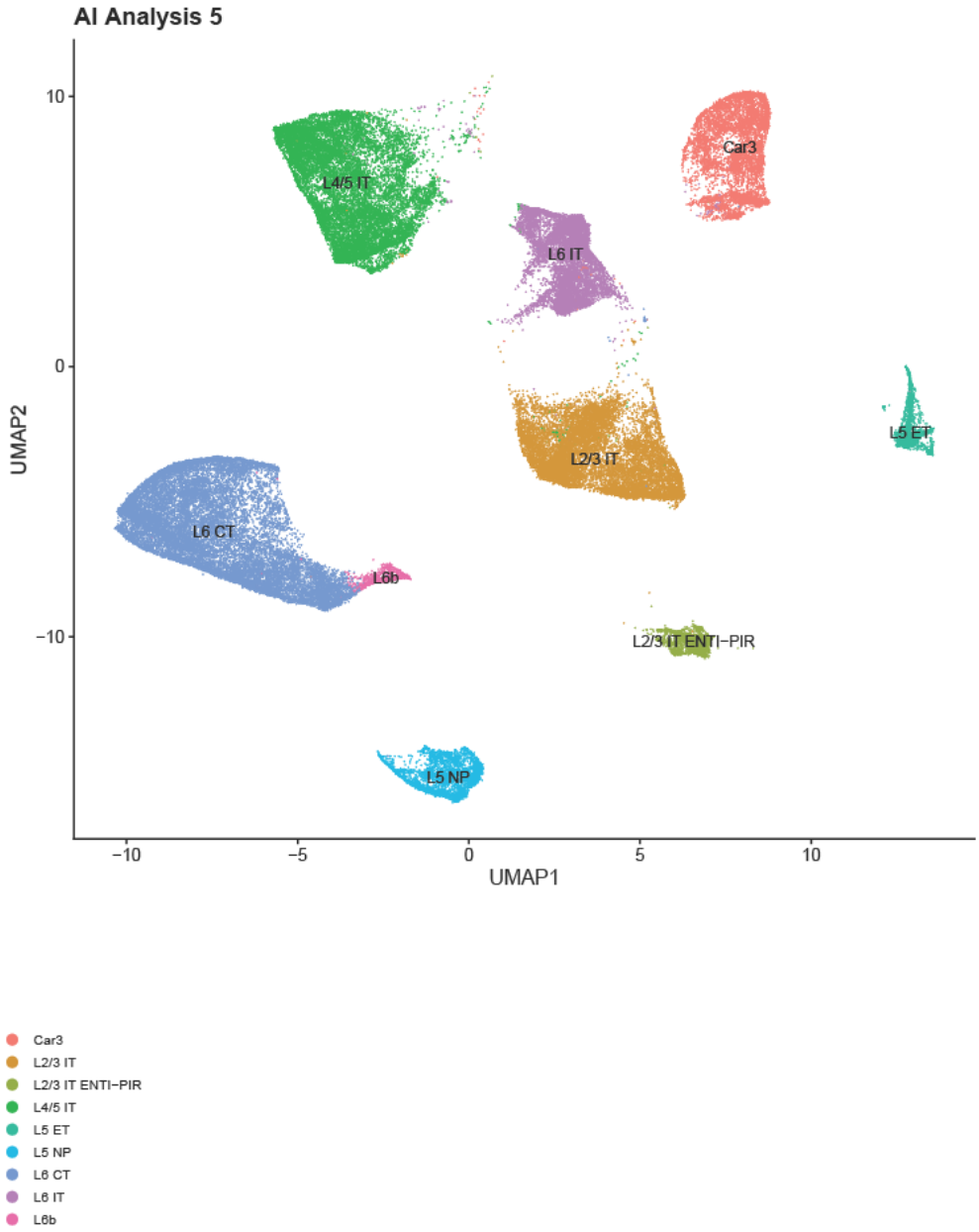


Figure 3.6: UMAP integration of the AUD Non-neuronal cells

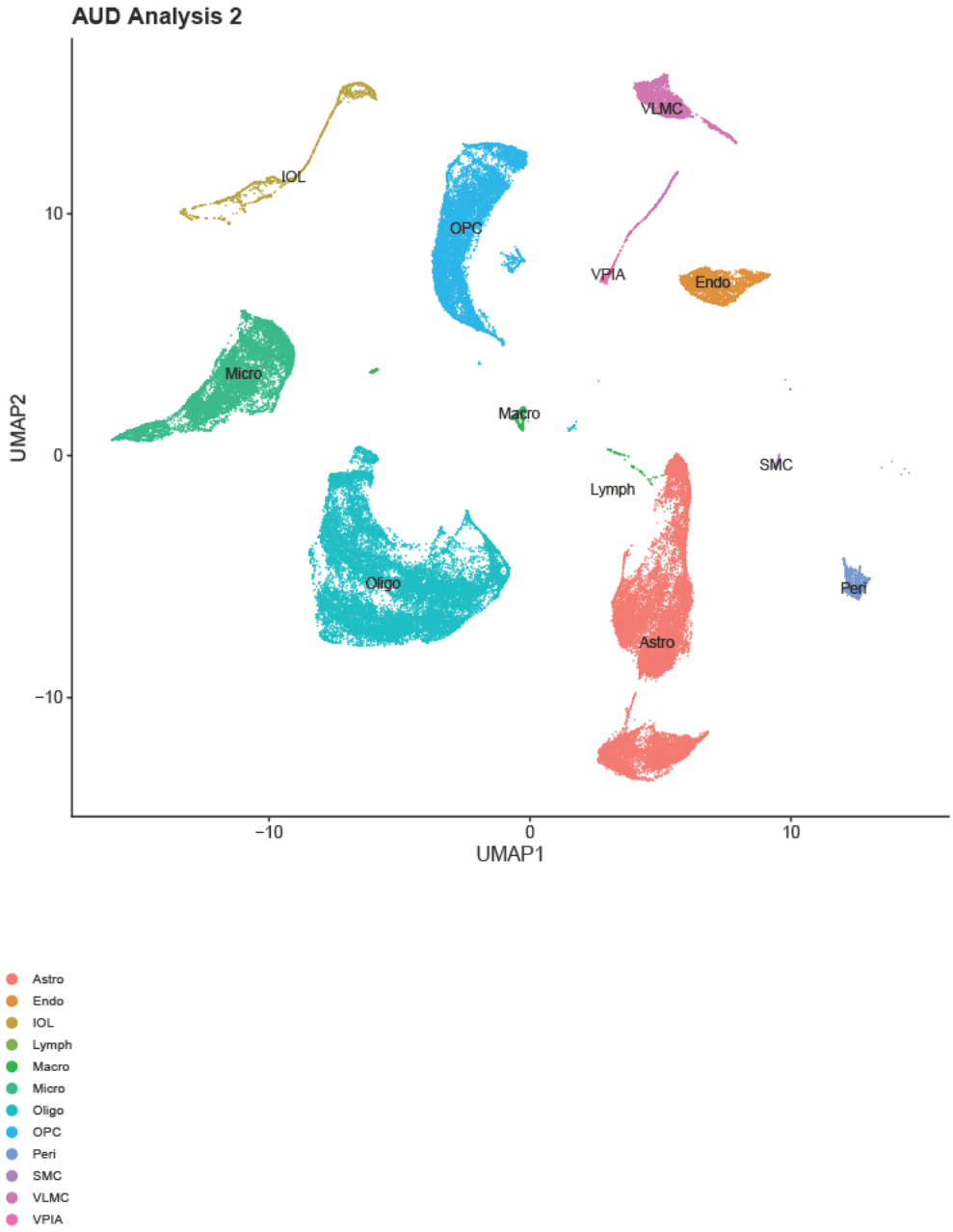


Figure 3.7: UMAP integration of the AUD Inhibitory and Other Neurons

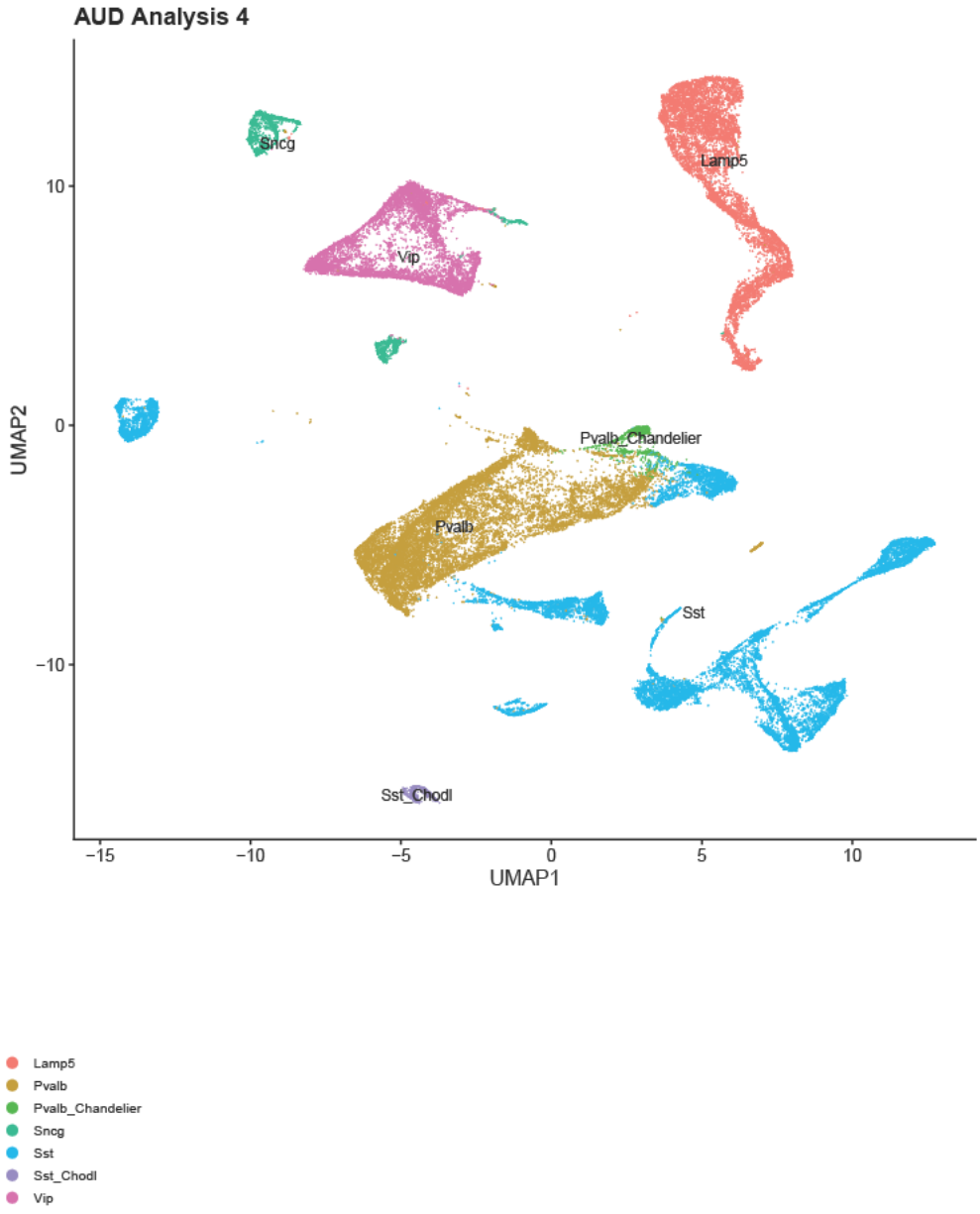


Figure 3.8: UMAP integration of the AUD Excitatory Neurons

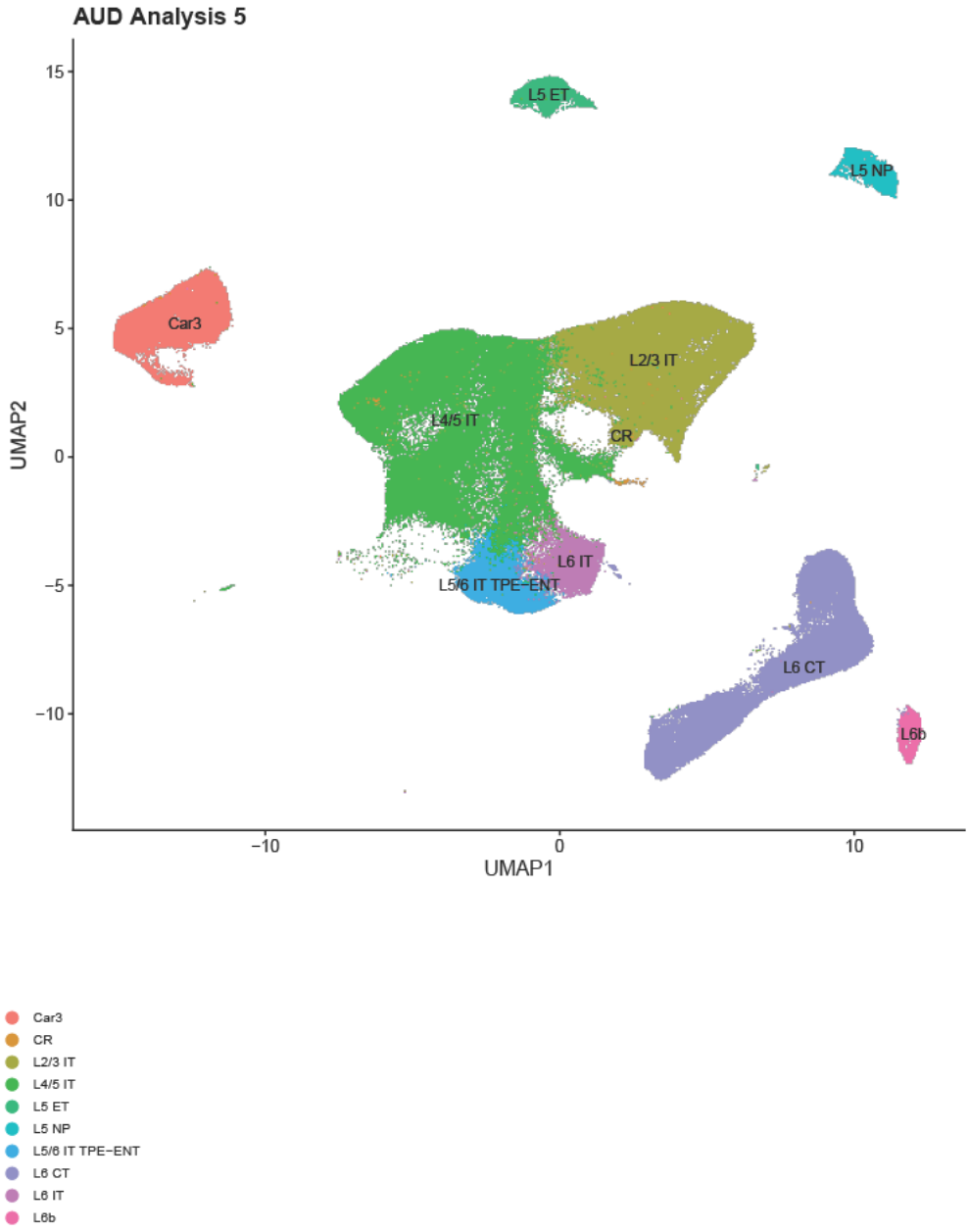


Figure 3.9: UMAP integration of the CTXsp Non-neuronal cells

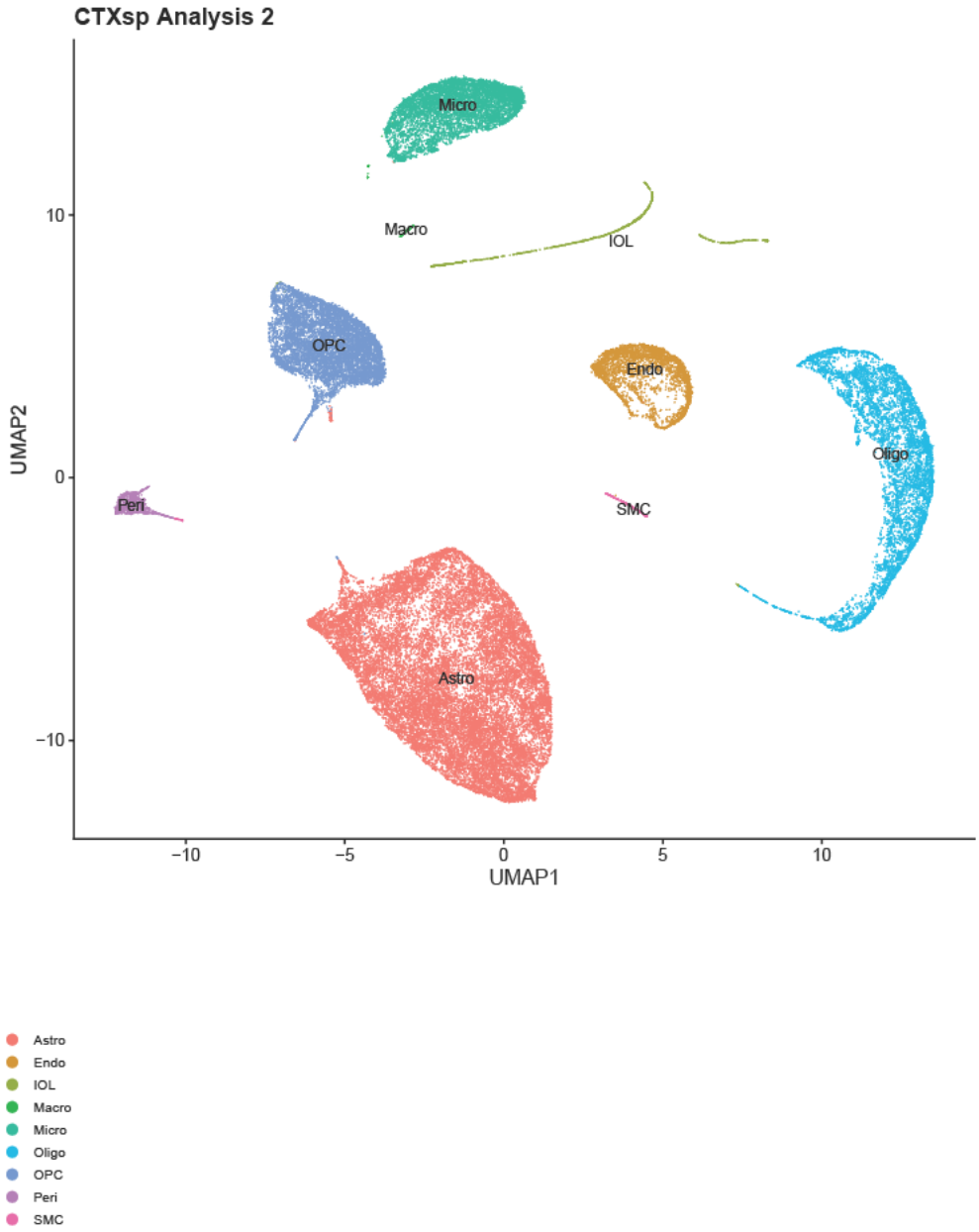


Figure 3.10: UMAP integration of the CTXsp Inhibitory and Other Neurons

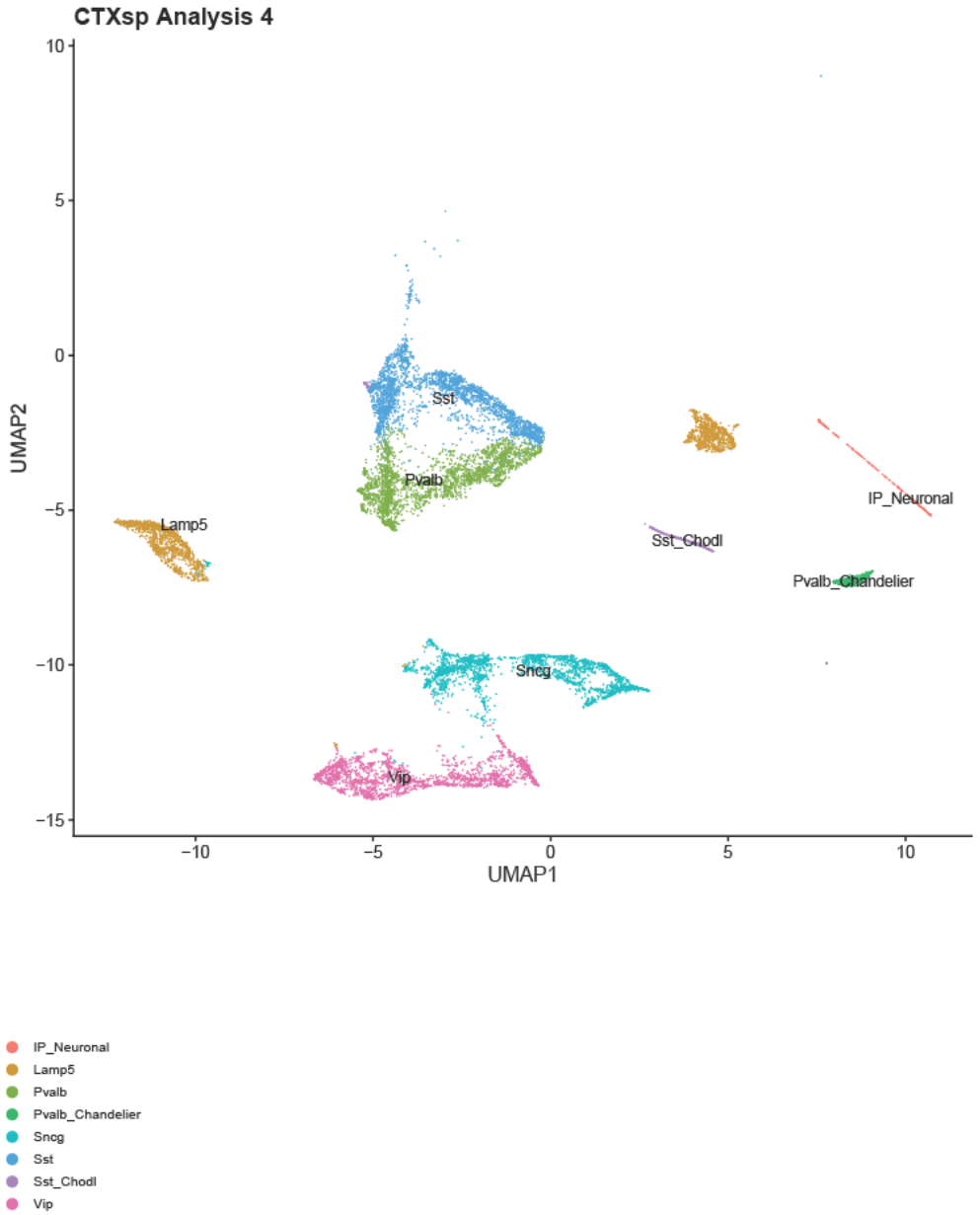


Figure 3.11: UMAP integration of the CTXsp Excitatory Neurons

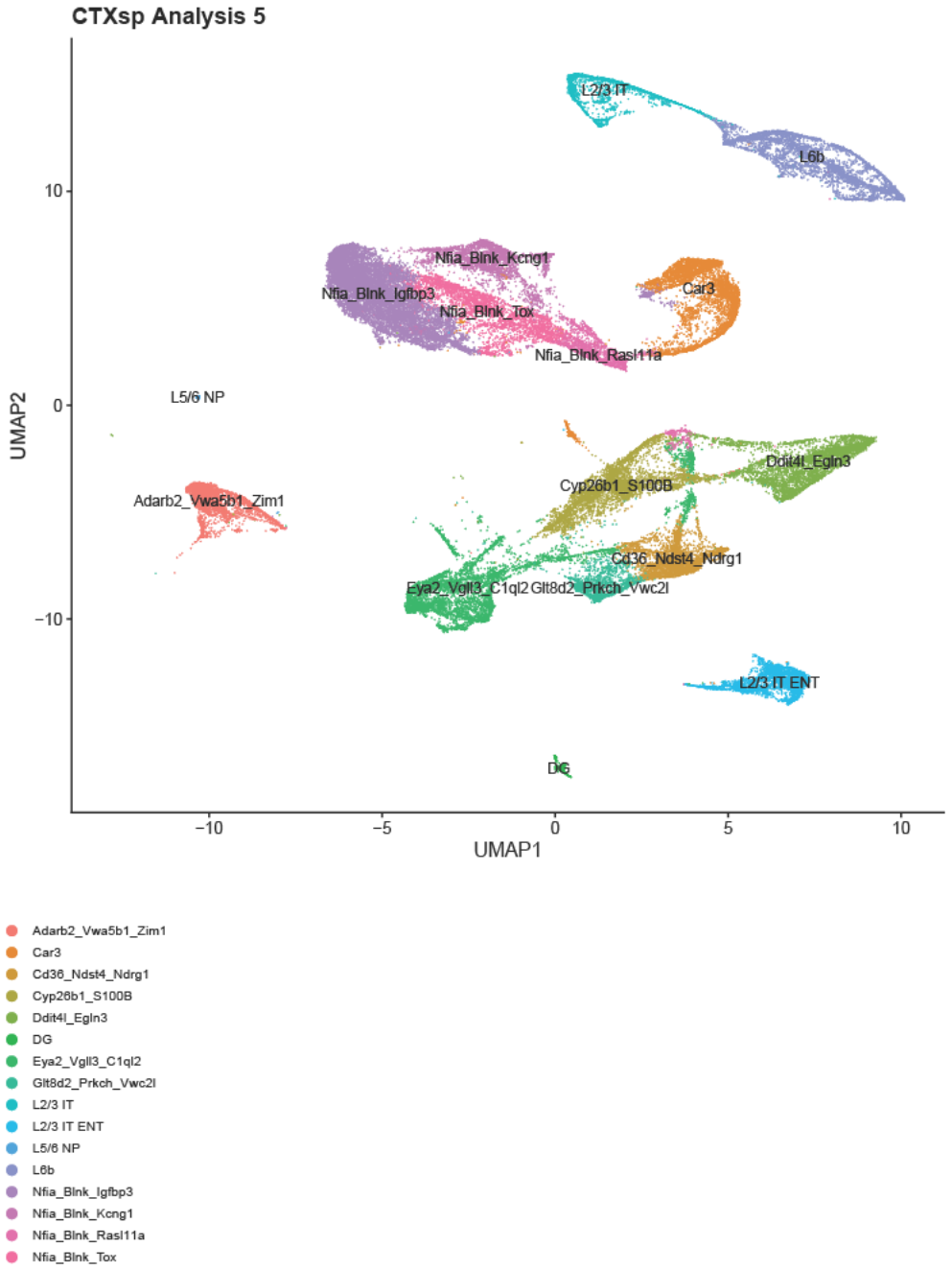


Figure 3.12: UMAP integration of the CB Non-neuronal cells

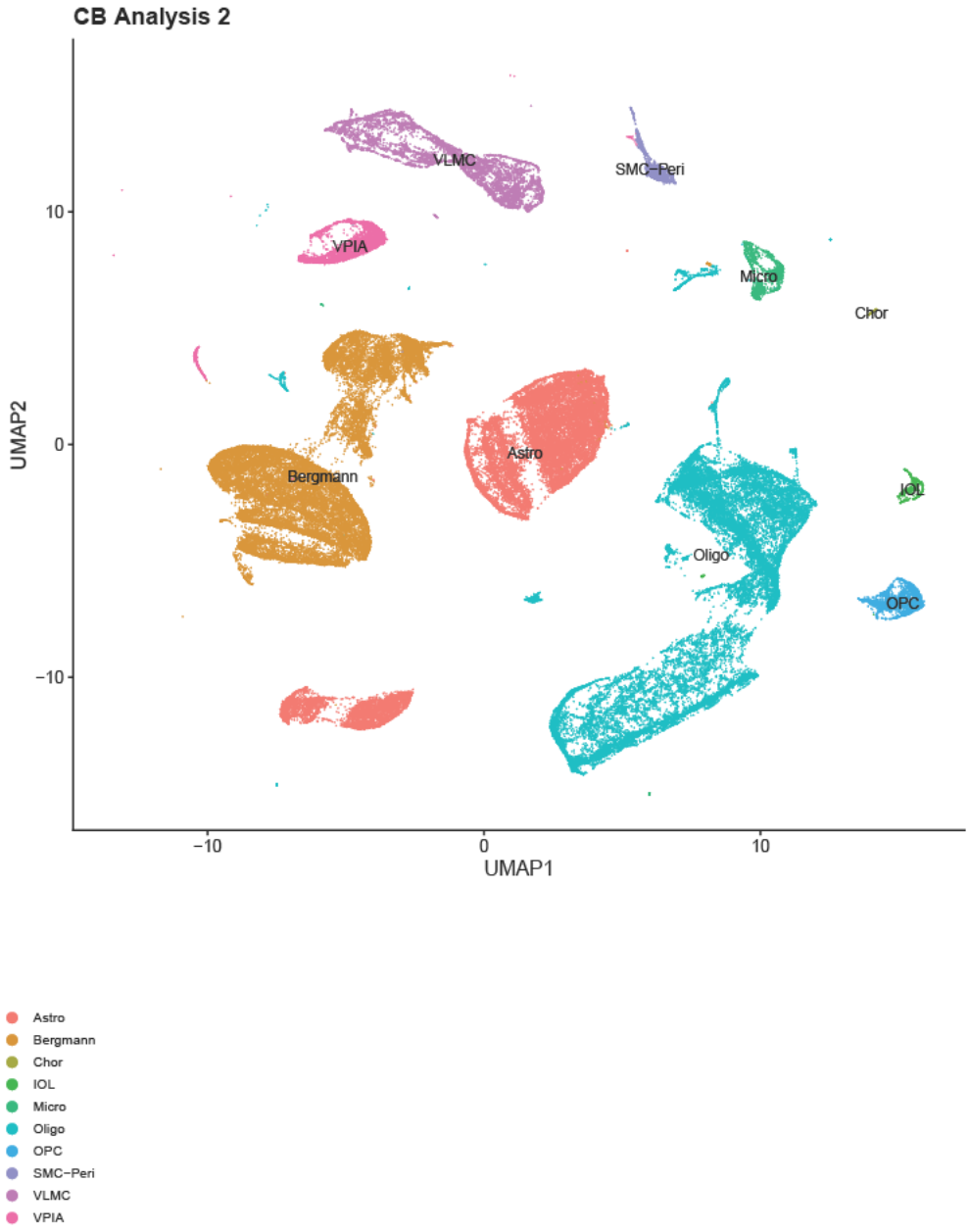


Figure 3.13: UMAP integration of the CB Inhibitory and Other Neurons

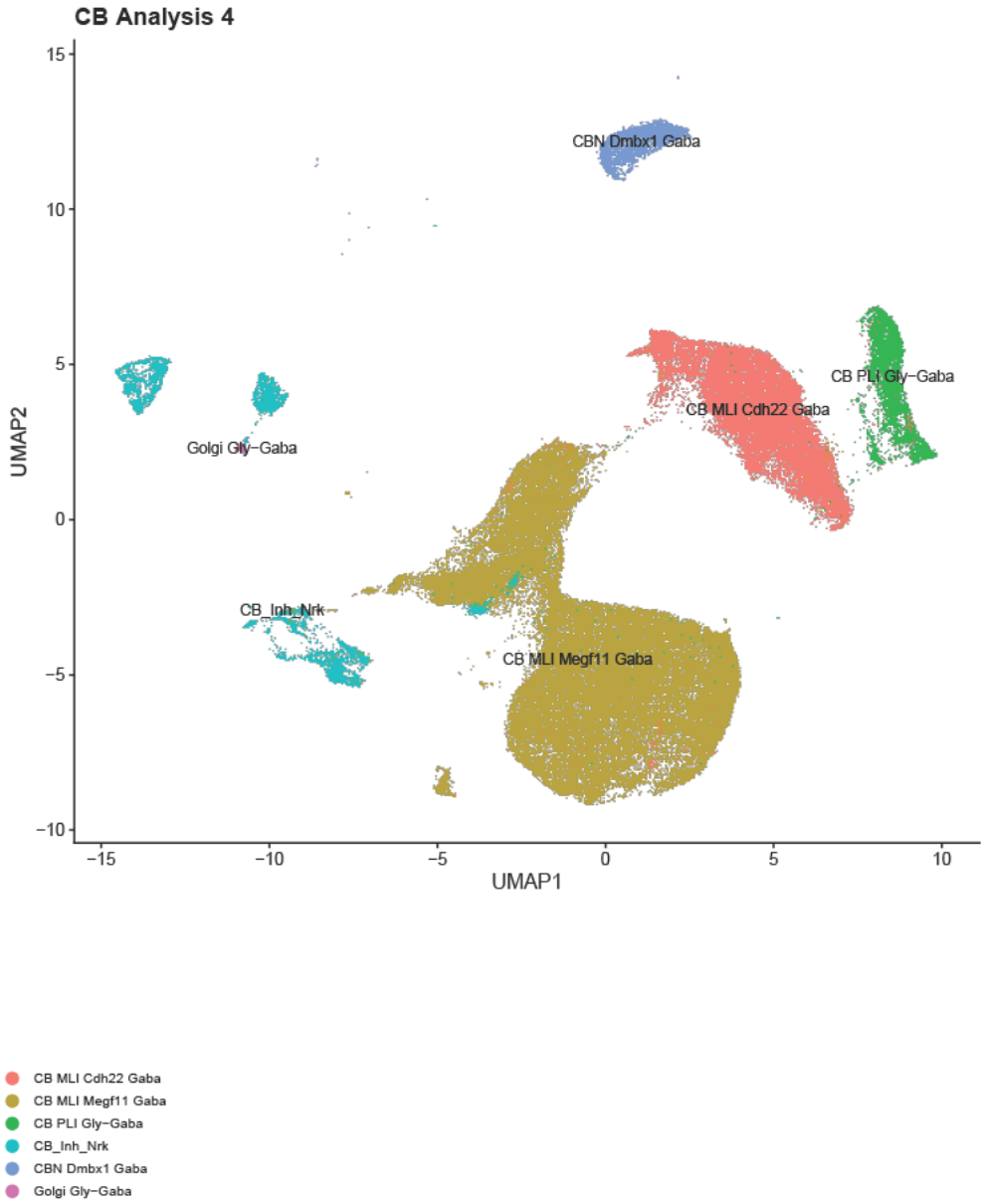


Figure 3.14: UMAP integration of the CB Excitatory Neurons

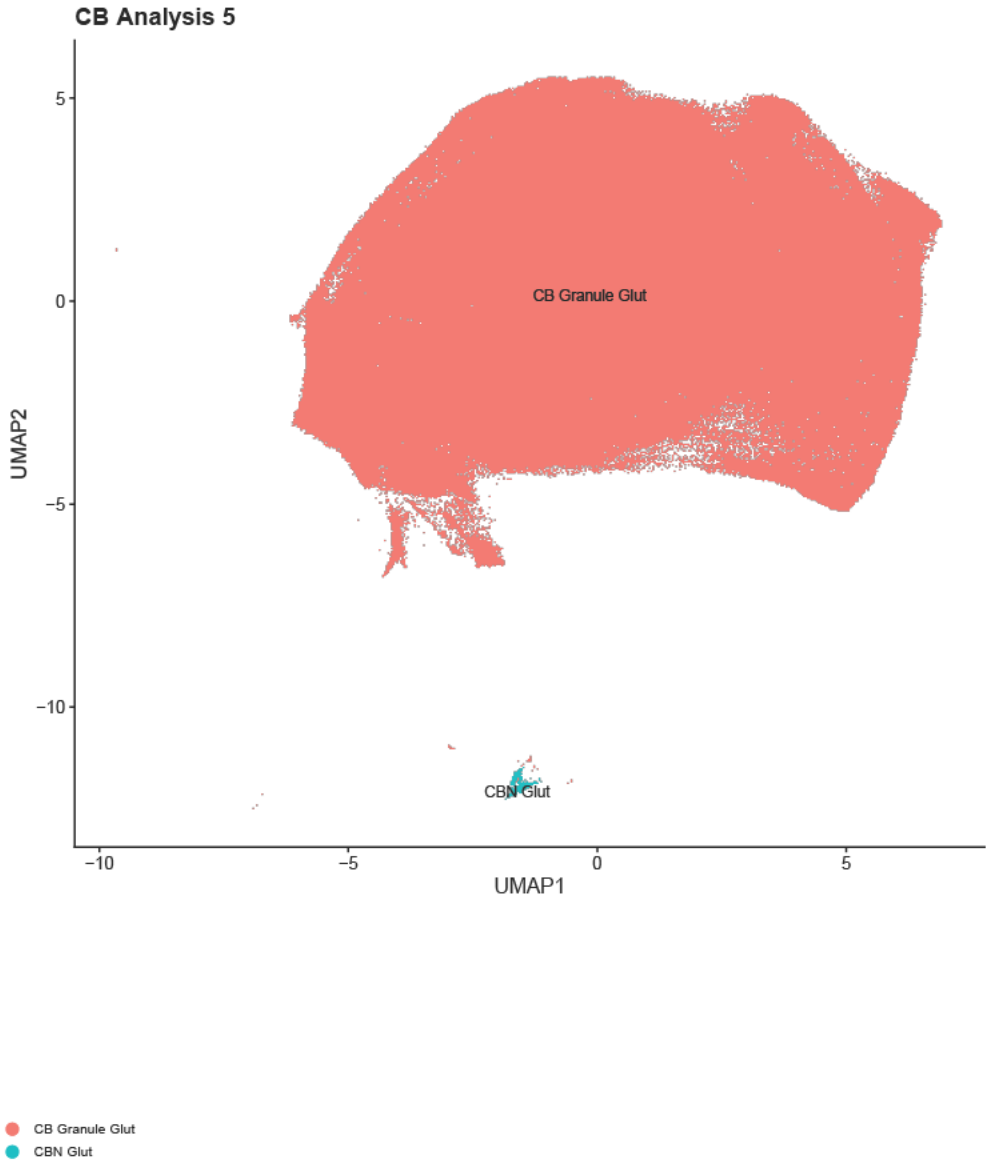


Figure 3.15: UMAP integration of the HB Non-neuronal cells

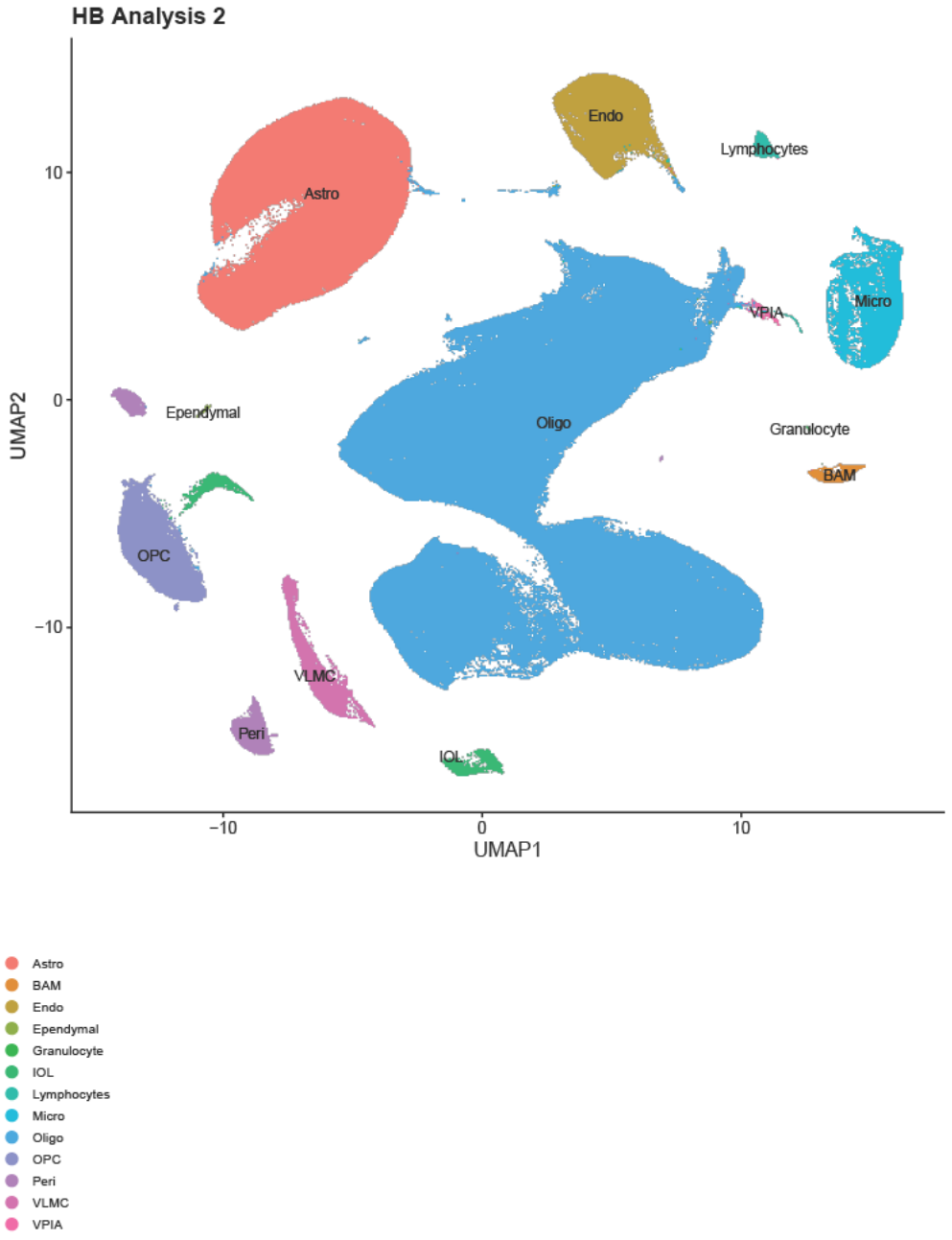


Figure 3.16: UMAP integration of the HB Inhibitory and Other Neurons

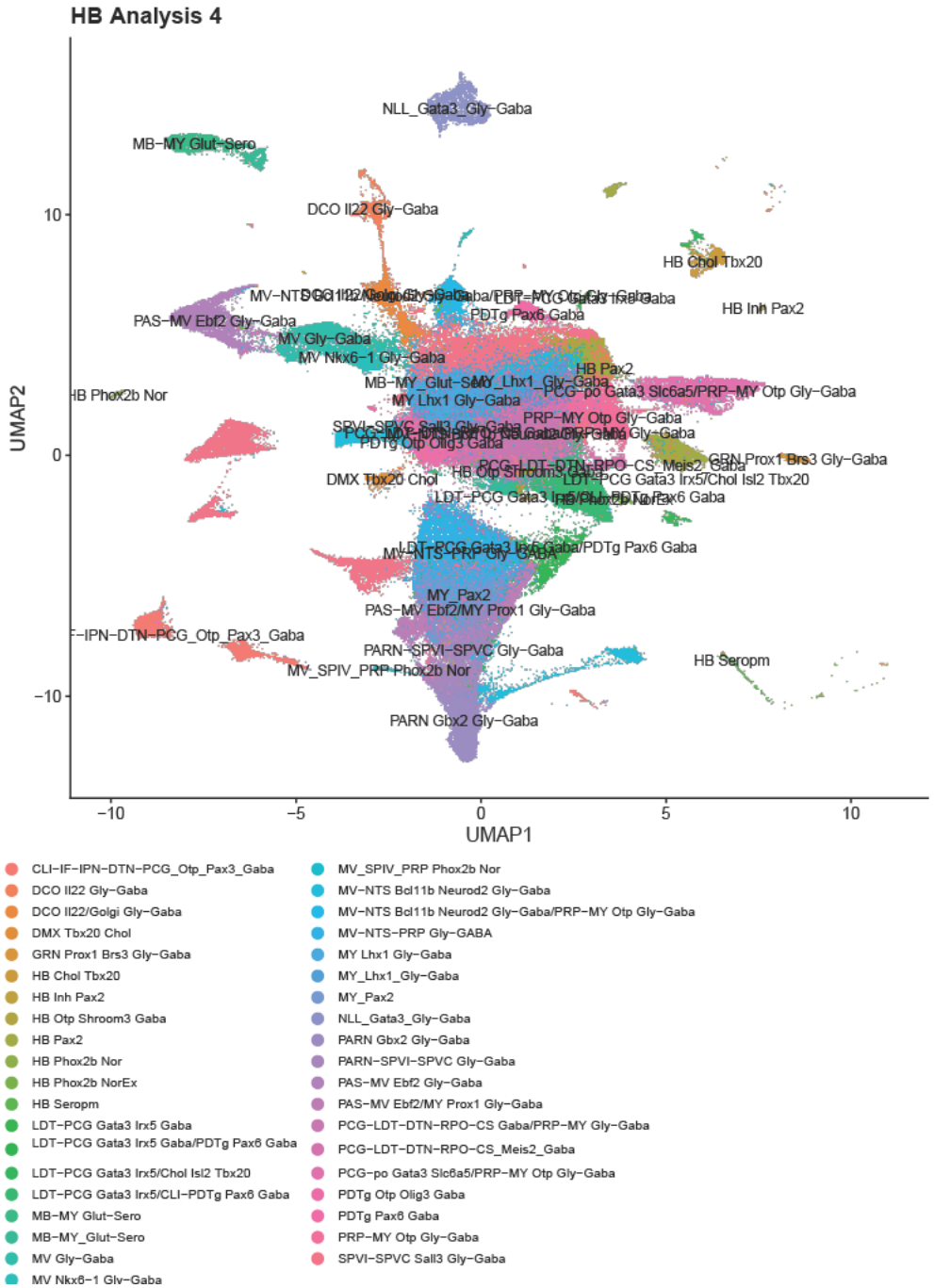


Figure 3.17: UMAP integration of the HB Excitatory Neurons

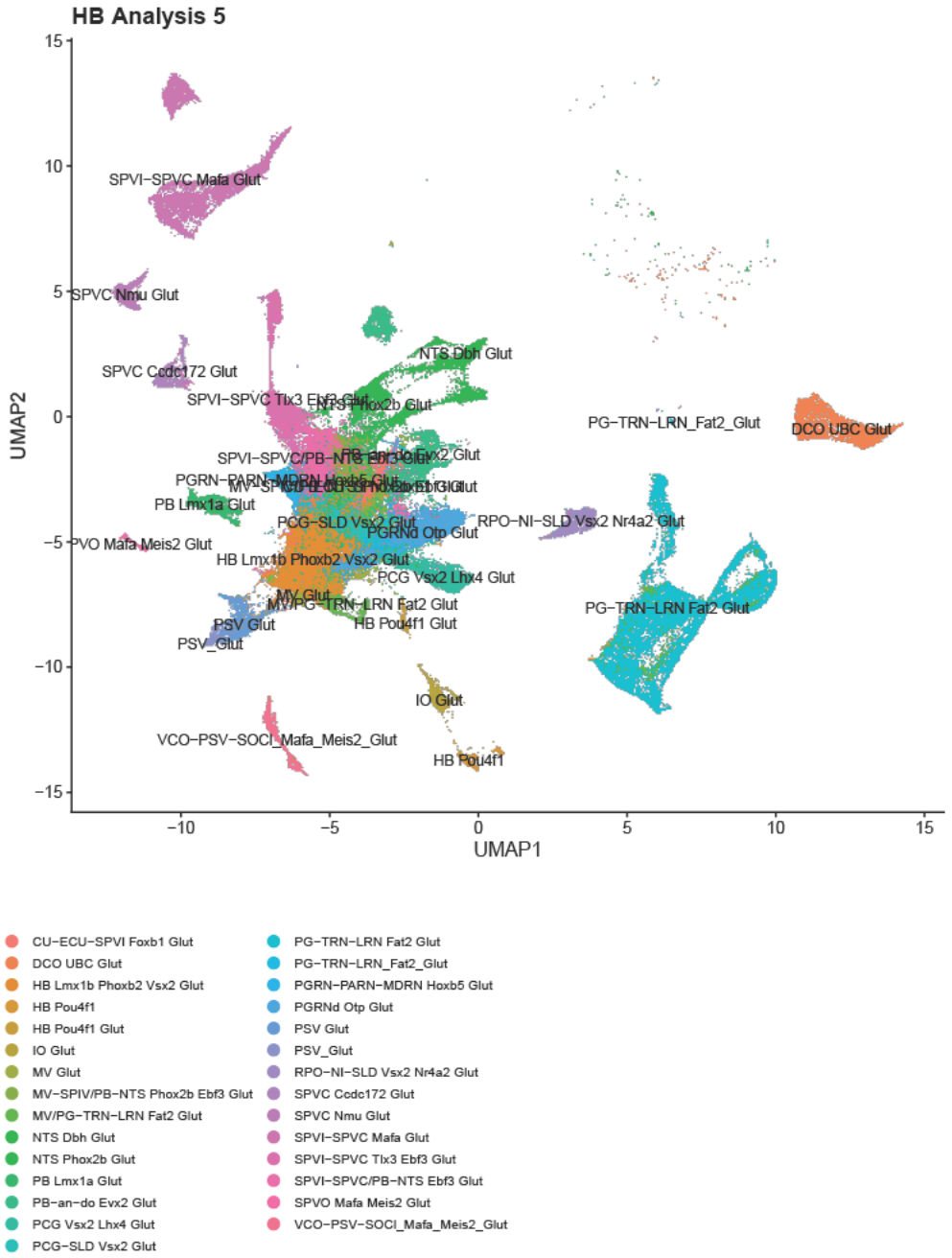


Figure 3.18: UMAP integration of the HPFENT Non-neuronal cells

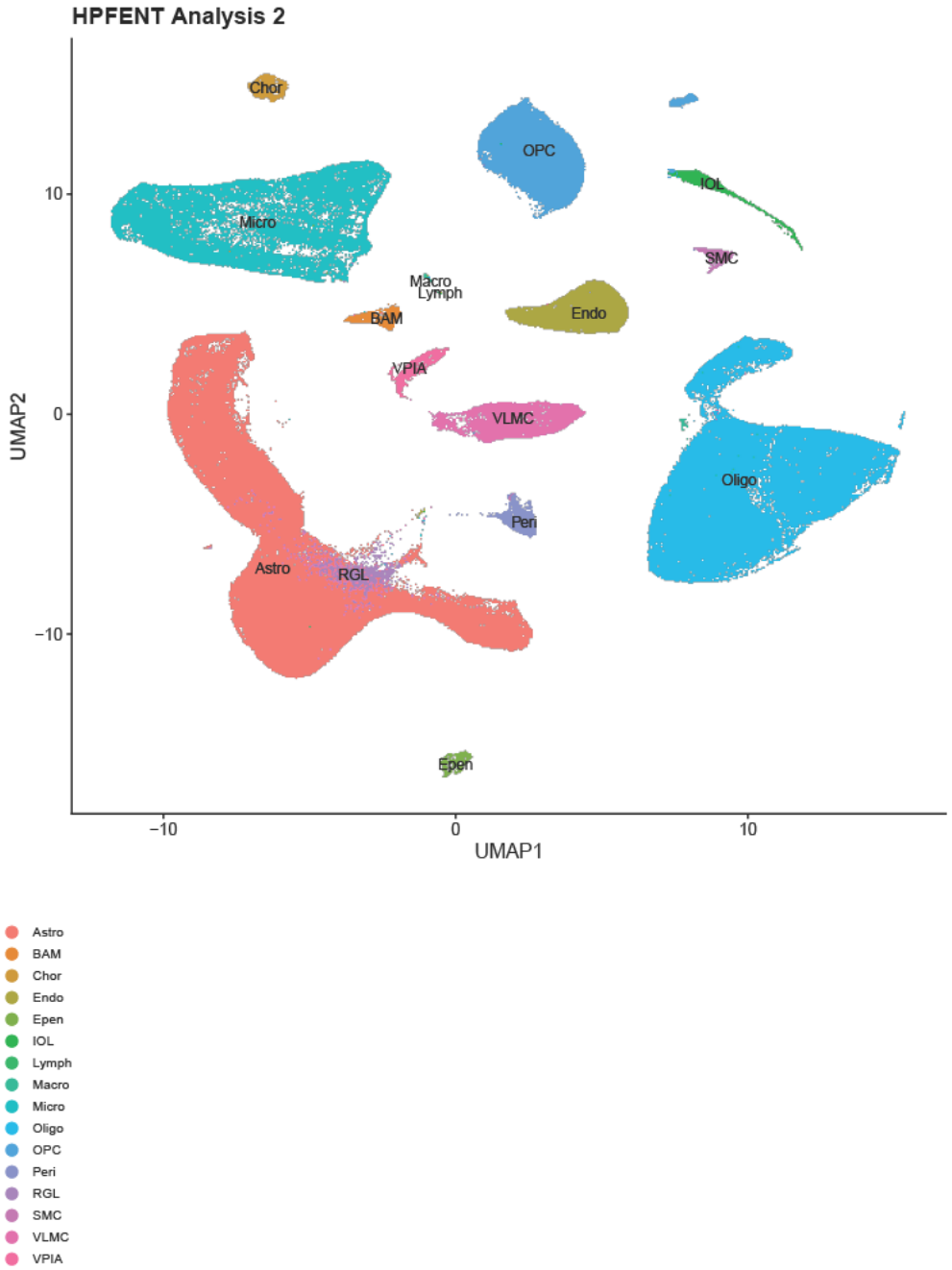


Figure 3.19: UMAP integration of the HPFENT Inhibitory and Other Neurons

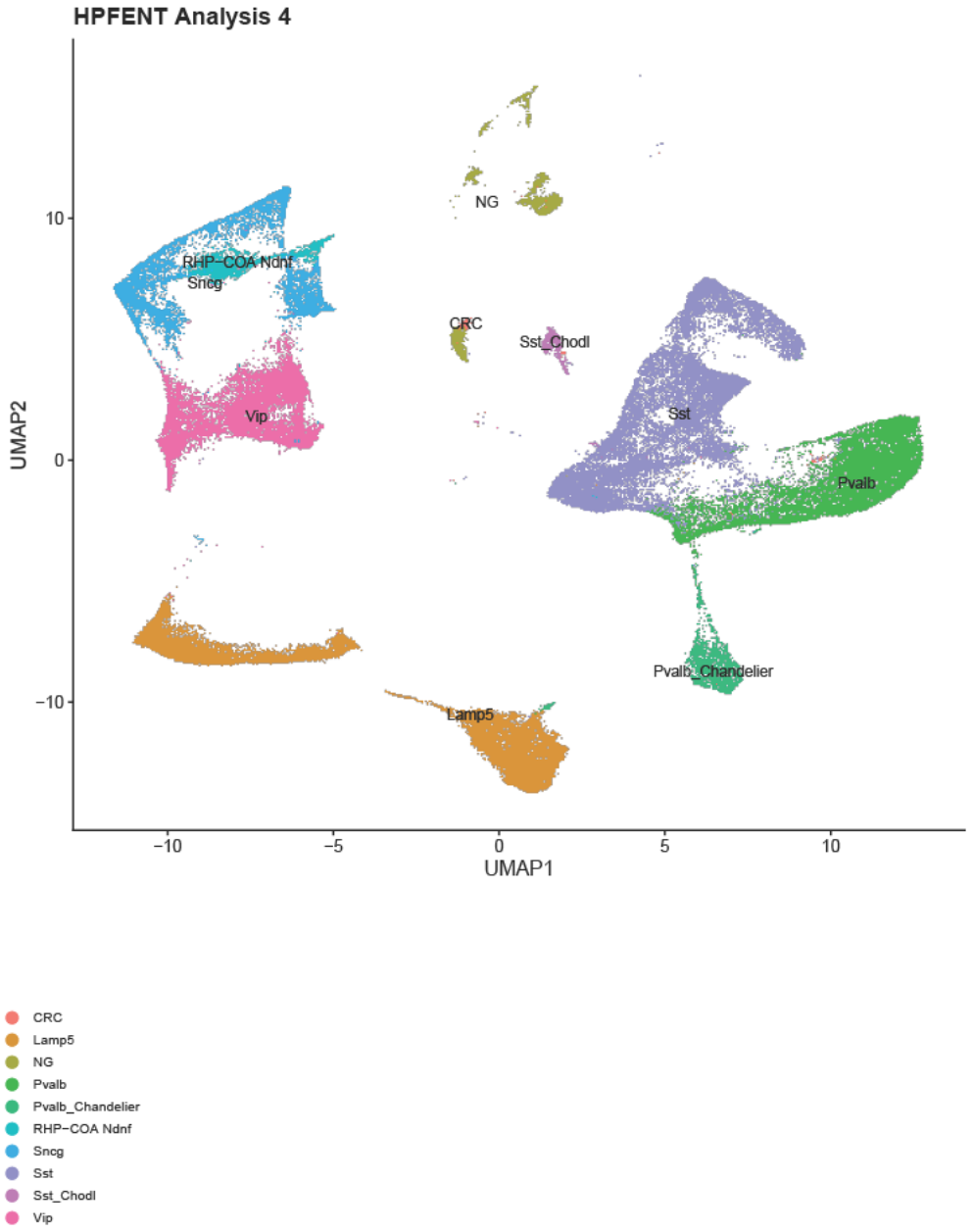


Figure 3.20: UMAP integration of the HPFENT Excitatory Neurons

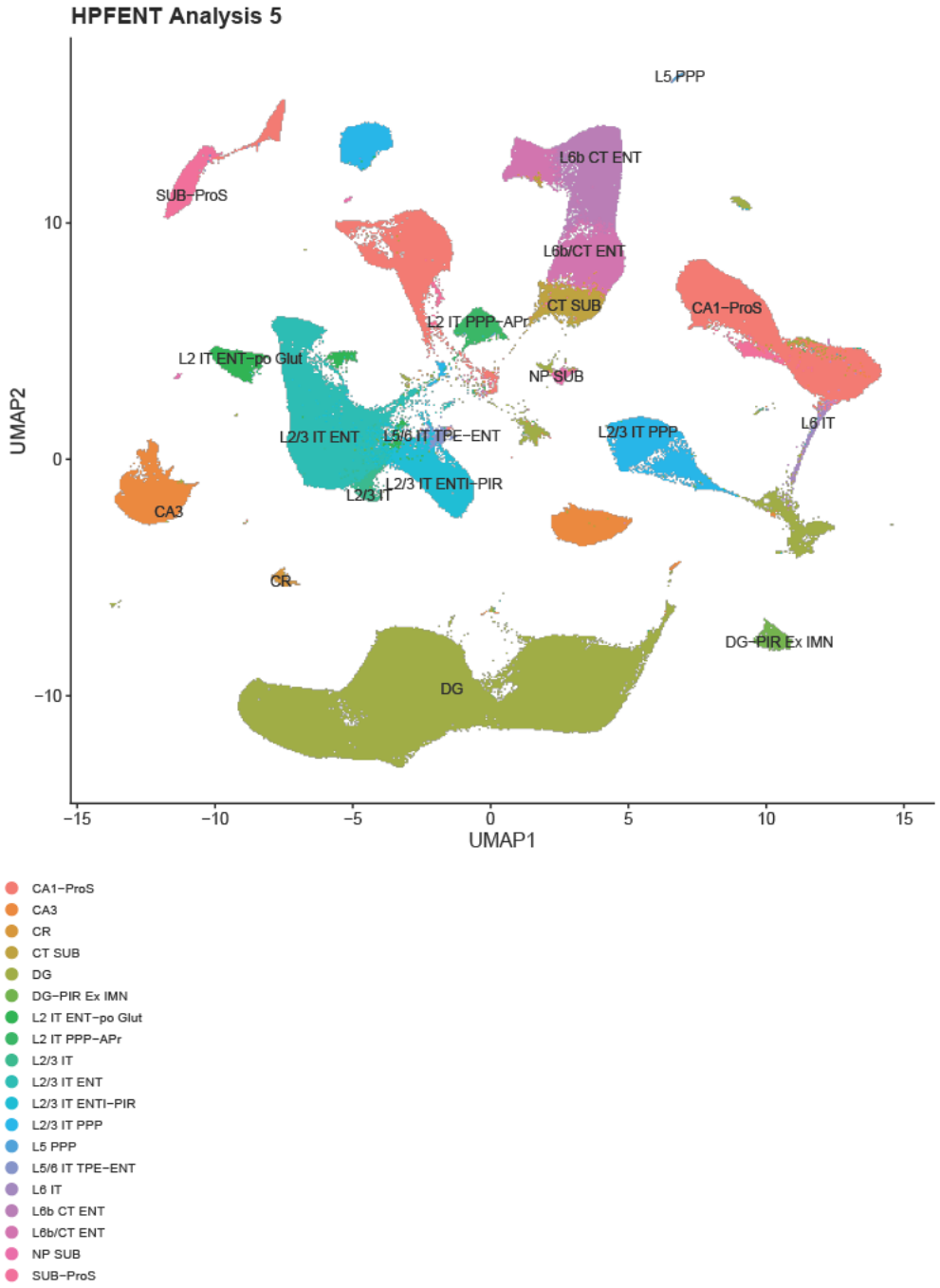


Figure 3.21: UMAP integration of the HY Non-neuronal cells

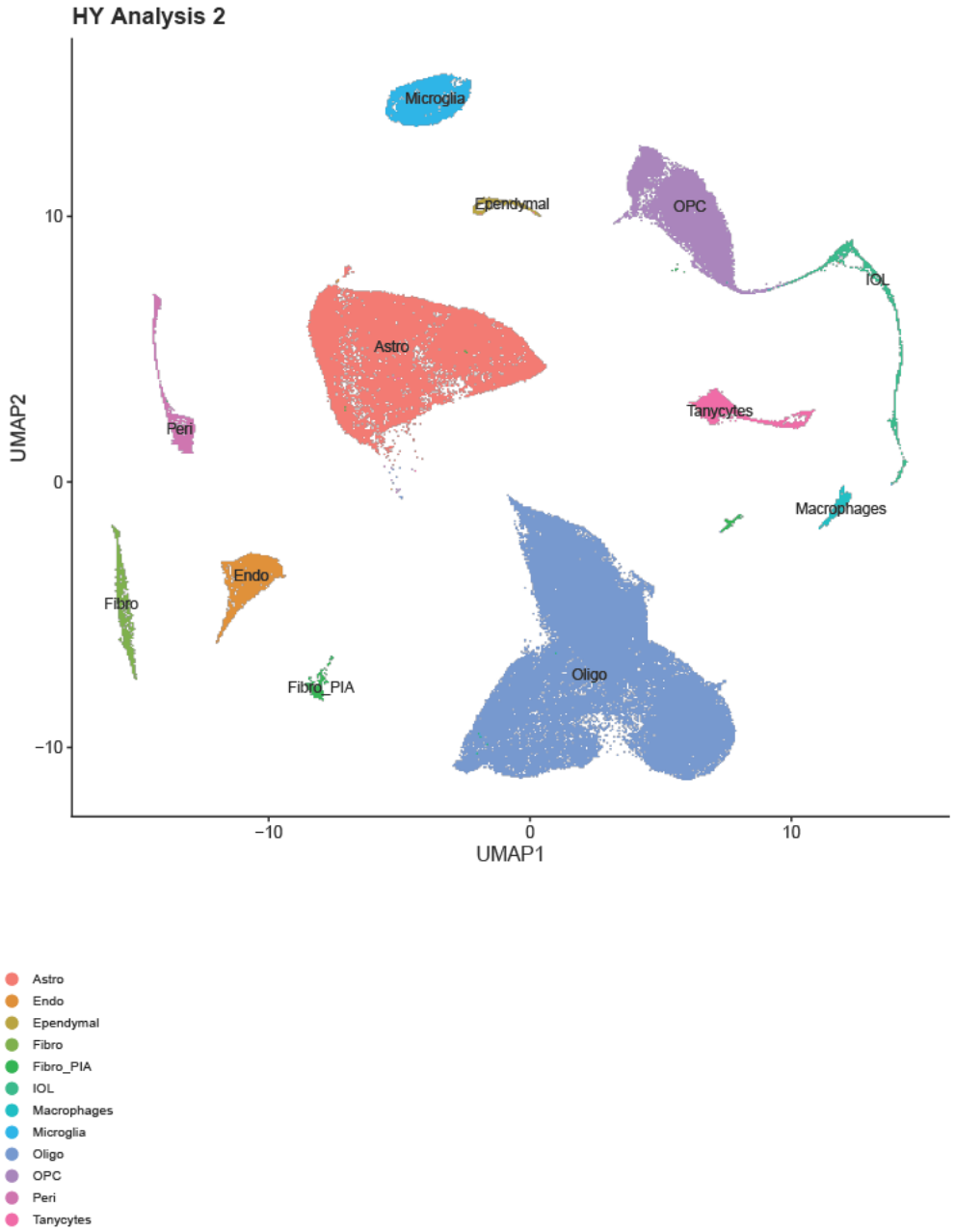


Figure 3.22: UMAP integration of the HY Inhibitory and Other Neurons

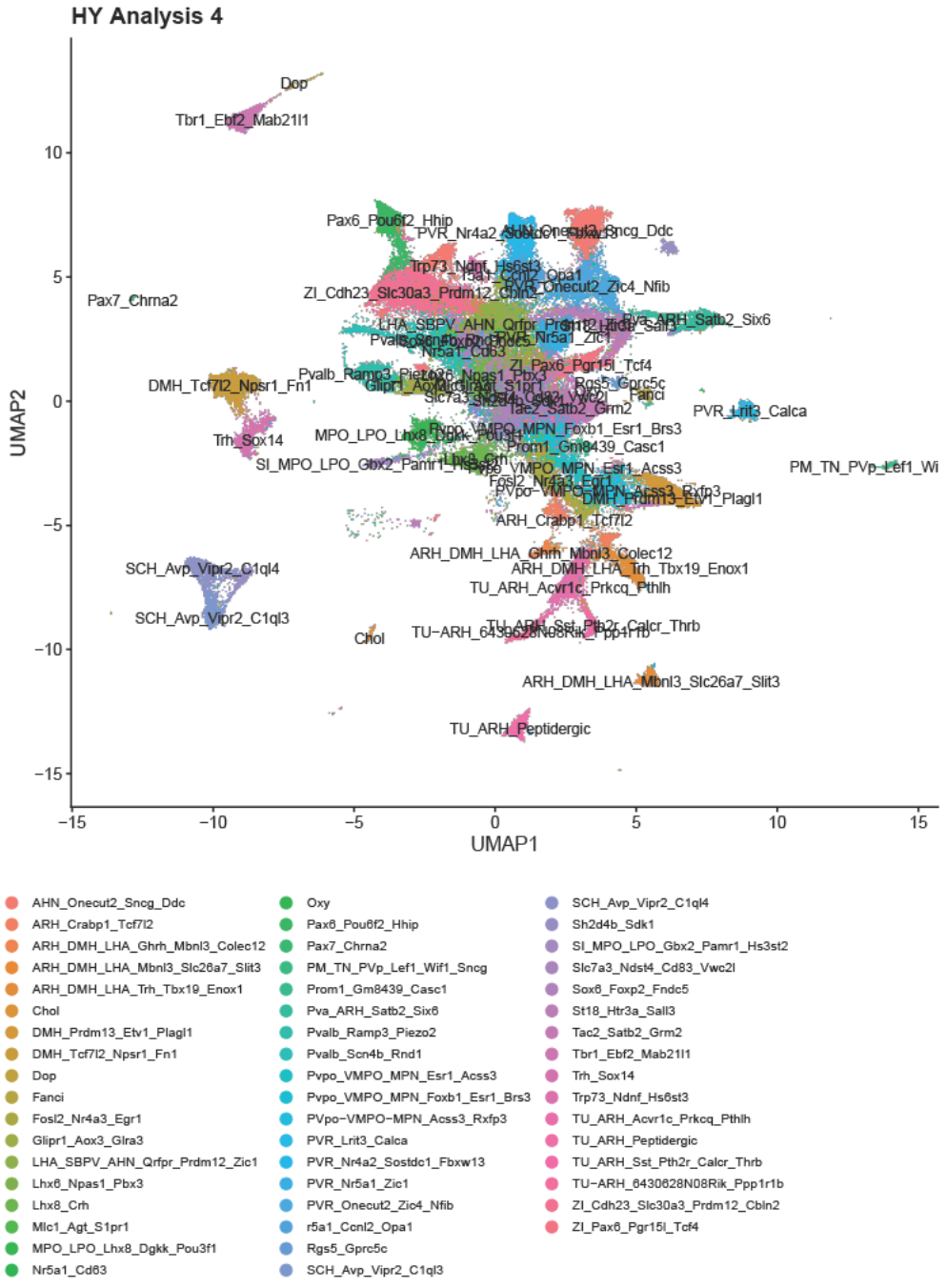
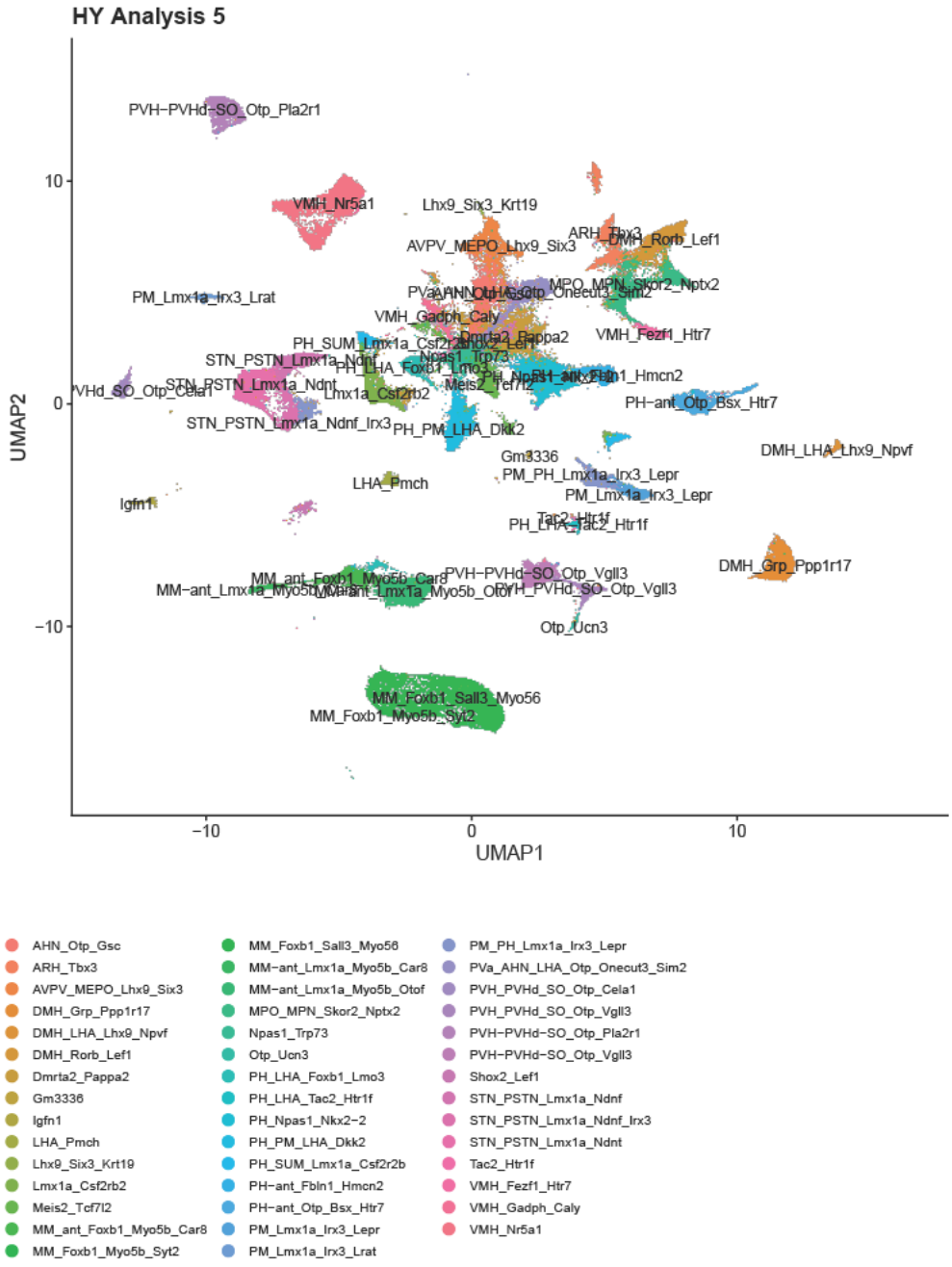


Figure 3.23: UMAP integration of the HY Excitatory Neurons



- AHN_Otp_Gsc
- ARH_Tbx3
- AVPV_MEPO_Lhx9_Six3
- DMH_Grp_Ppp1r17
- DMH_LHA_Lhx9_Npvf
- DMH_Rorb_Lef1
- Dmrta2_Pappa2
- Gm3336
- Igfn1
- LHA_Pmch
- Lhx9_Six3_Krt19
- Lmx1a_Csf2rb2
- Meis2_Tcf7l2
- MM_ant_Foxb1_Myo5b_Car8
- MM_Foxb1_Myo5b_Syt2
- MM_Foxb1_Sall3_Myo56
- MM−ant_Lmx1a_Myo5b_Car8
- MM−ant_Lmx1a_Myo5b_Otof
- MPO_MPN_Skor2_Nptx2
- Npas1_Trp73
- Otp_Ucn3
- PH_LHA_Foxb1_Lmo3
- PH_LHA_Tac2_Htr1f
- PH_Npas1_Nkx2−2
- PH_PM_LHA_Dkk2
- PH_SUM_Lmx1a_Csf2r2b
- PH−ant_Fbln1_Hmcn2
- PH−ant_Otp_Bsx_Htr7
- PM_Lmx1a_Irx3_Lepr
- PM_Lmx1a_Irx3_Lrat
- PM_PH_Lmx1a_Irx3_Lepr
- PVa_AHN_LHA_Otp_Onecut3_Sim2
- PVH_PVHd_SO_Otp_Cela1
- PVH_PVHd_SO_Otp_Vgll3
- PVH−PVHd−SO_Otp_Pla2r1
- PVH−PVHd−SO_Otp_Vgll3
- Shox2_Lef1
- STN_PSTN_Lmx1a_Ndnf
- STN_PSTN_Lmx1a_Ndnf_Irx3
- STN_PSTN_Lmx1a_Ndnt
- Tac2_Htr1f
- VMH_Fezf1_Htr7
- VMH_Gadph_Caly
- VMH_Nr5a1

Figure 3.24: UMAP integration of the MB Non-neuronal cells

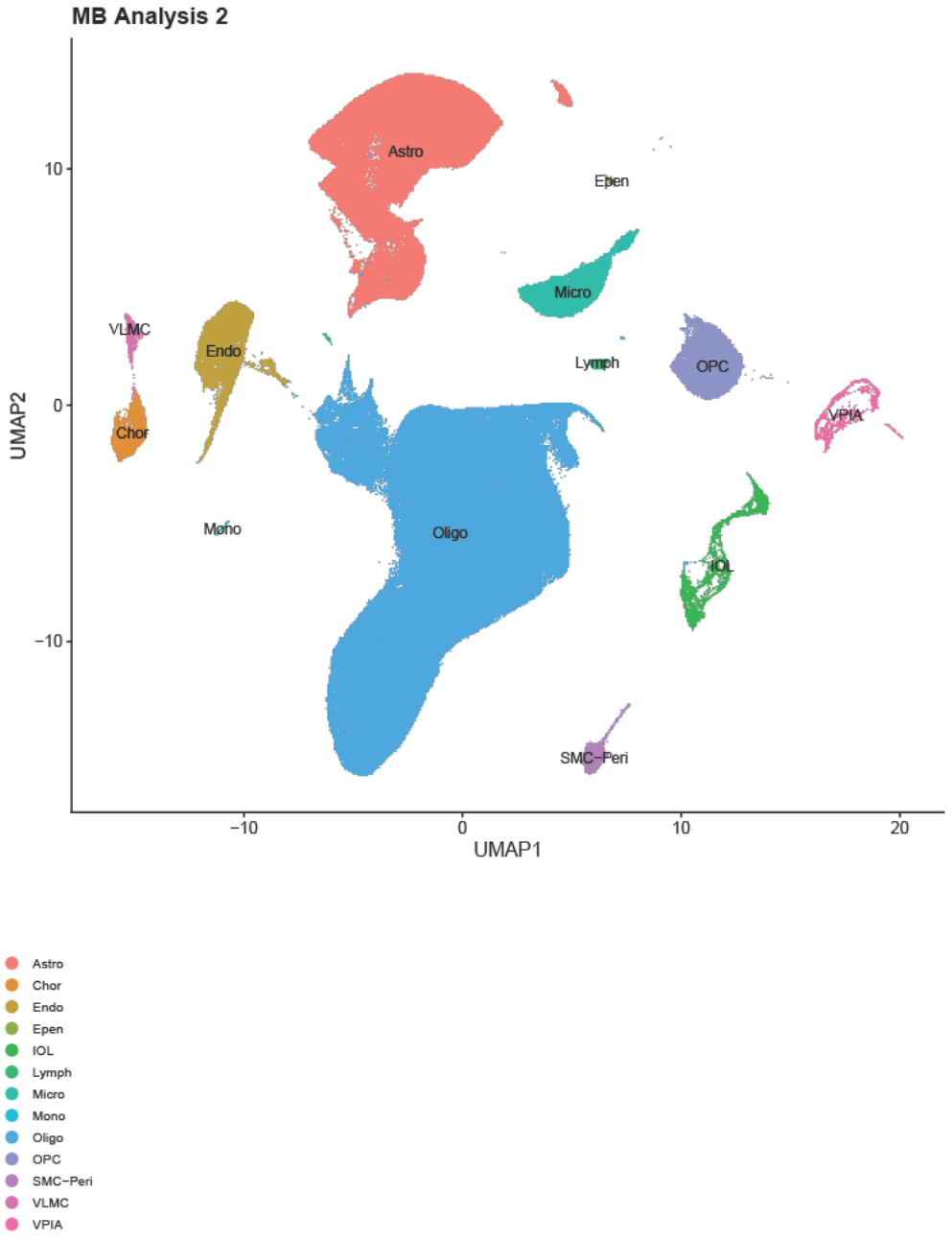


Figure 3.25: UMAP integration of the MB Inhibitory and Other Neurons

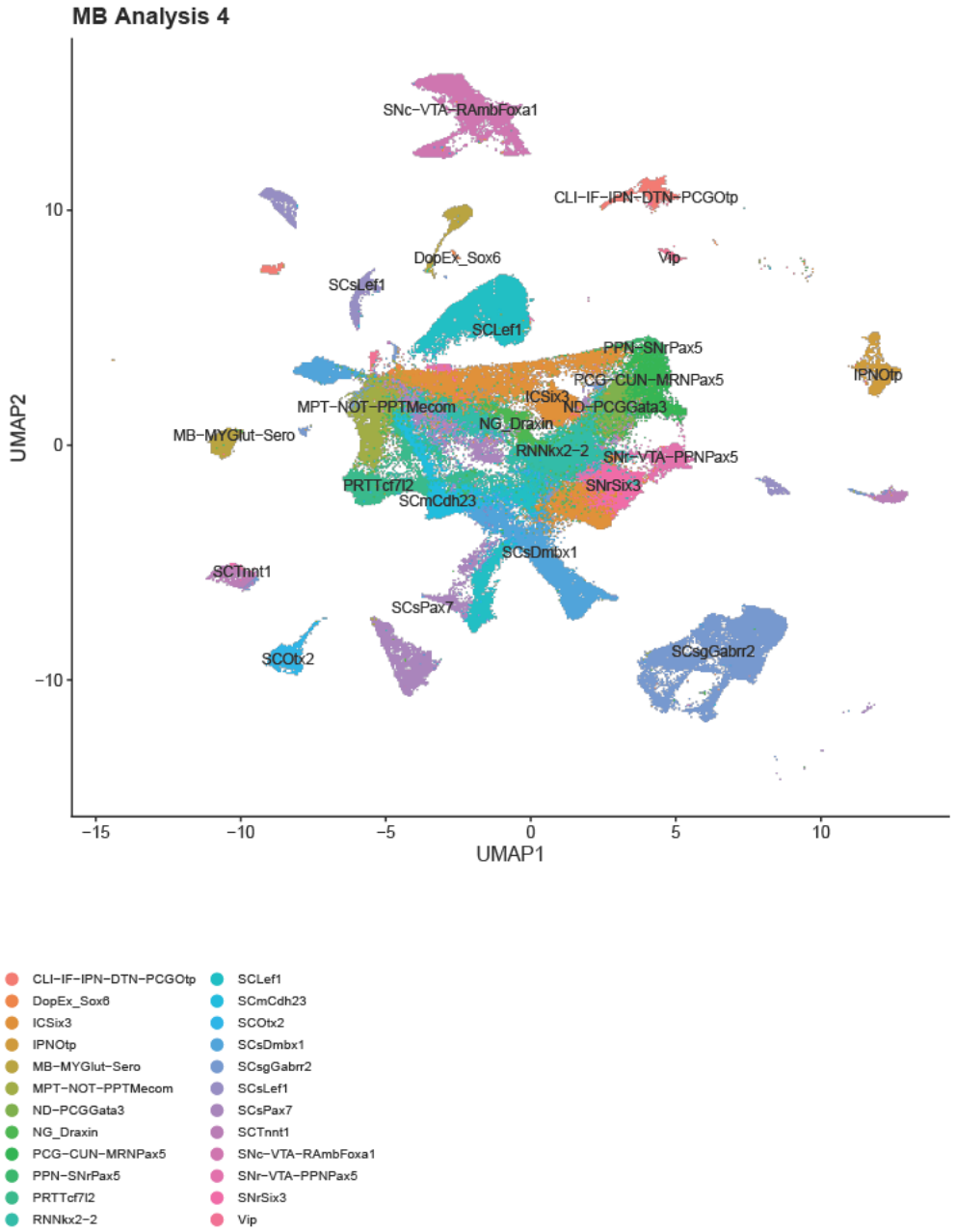


Figure 3.26: UMAP integration of the MB Excitatory Neurons

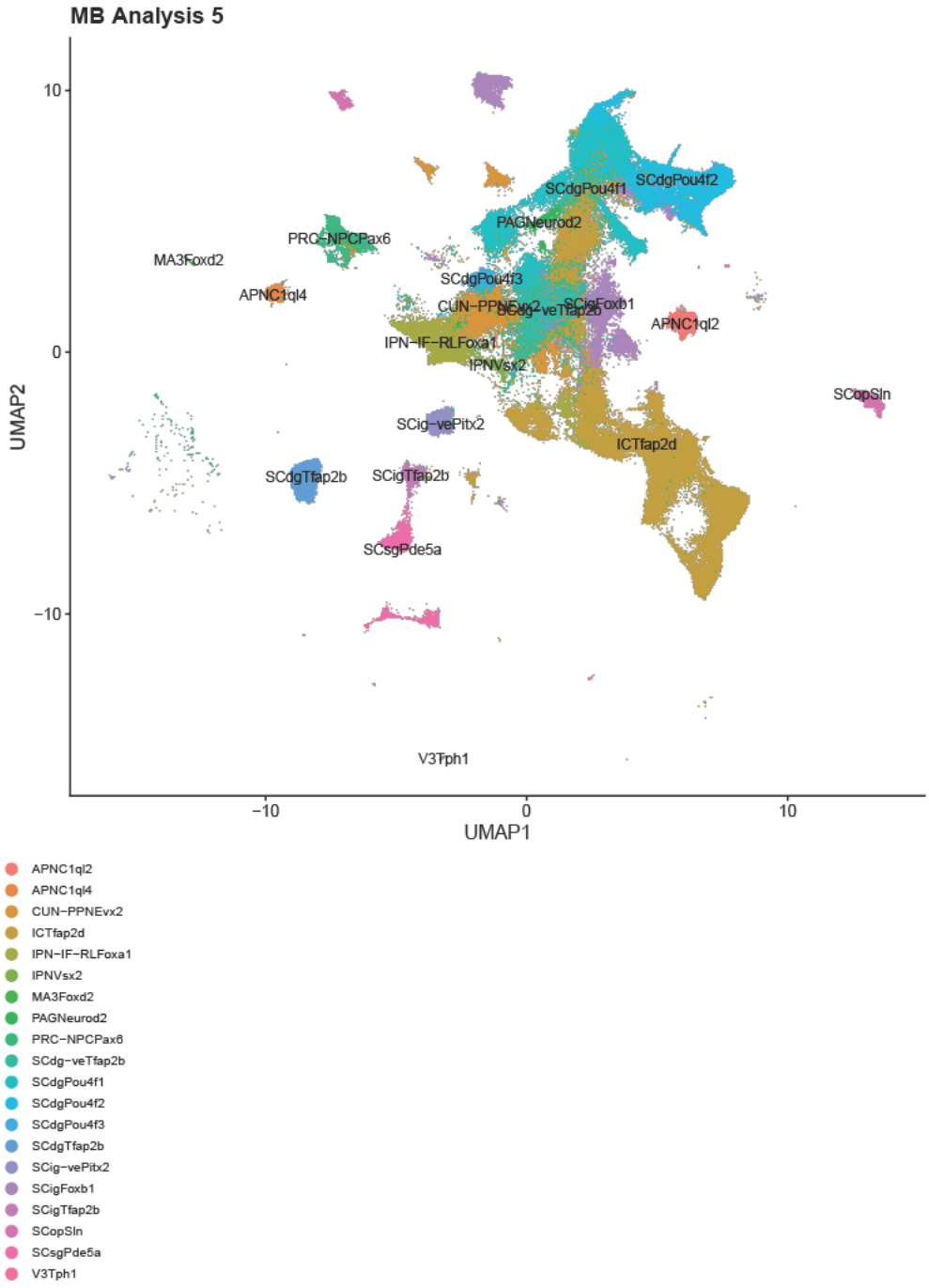


Figure 3.27: UMAP integration of the MOp Non-neuronal cells

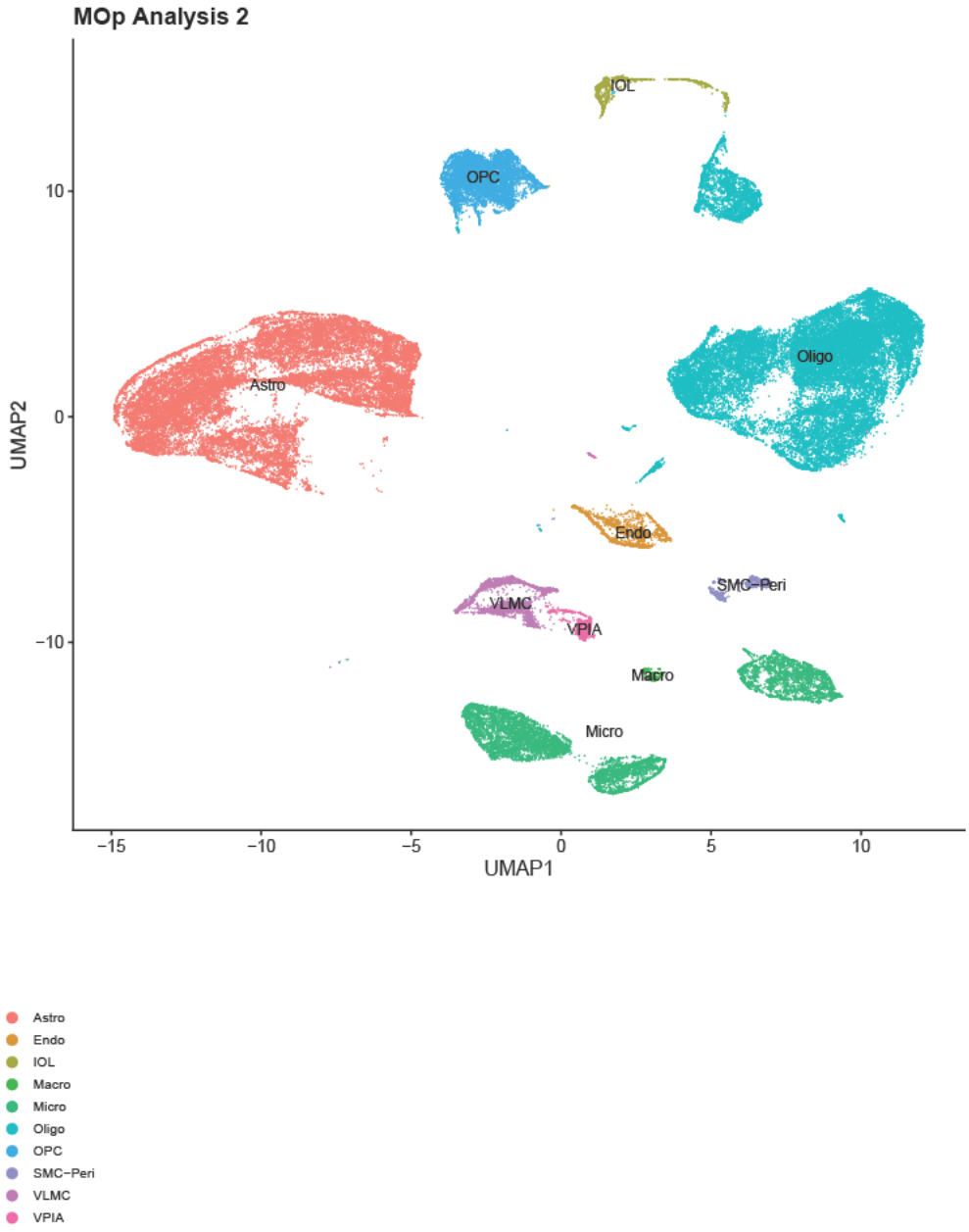


Figure 3.28: UMAP integration of the MOp Inhibitory and Other Neurons

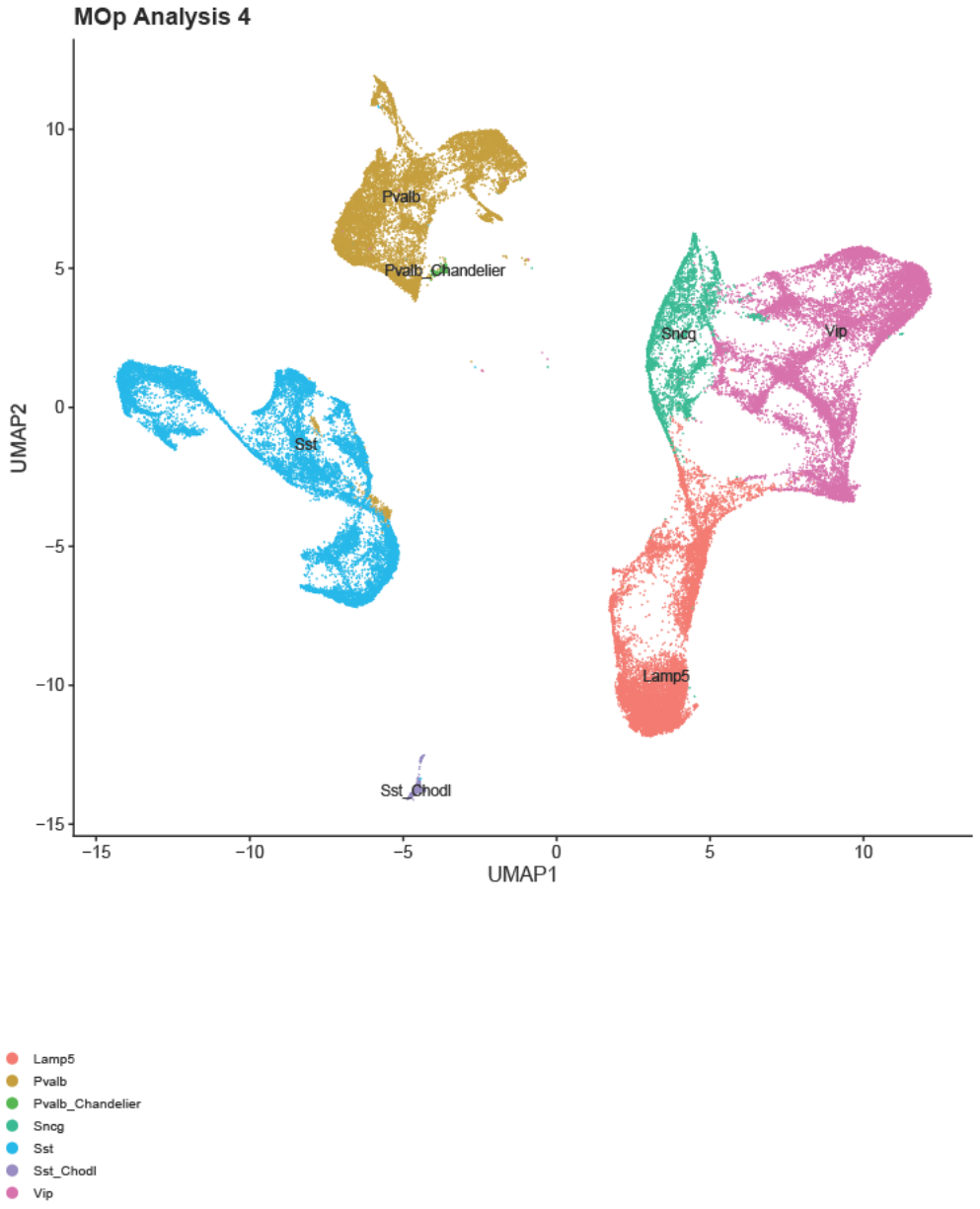


Figure 3.29: UMAP integration of the MOp Excitatory Neurons

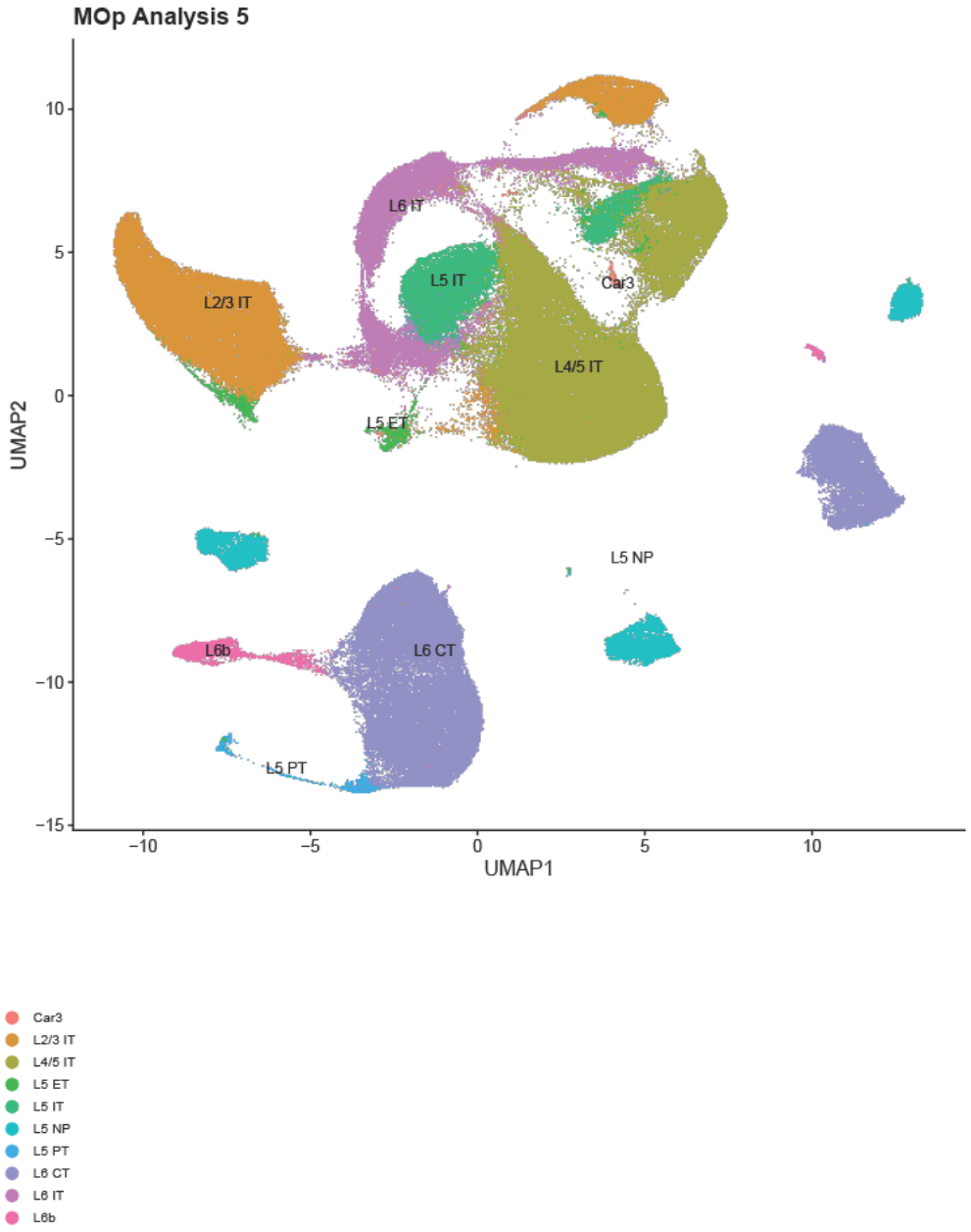


Figure 3.30: UMAP integration of the MOs Non-neuronal cells

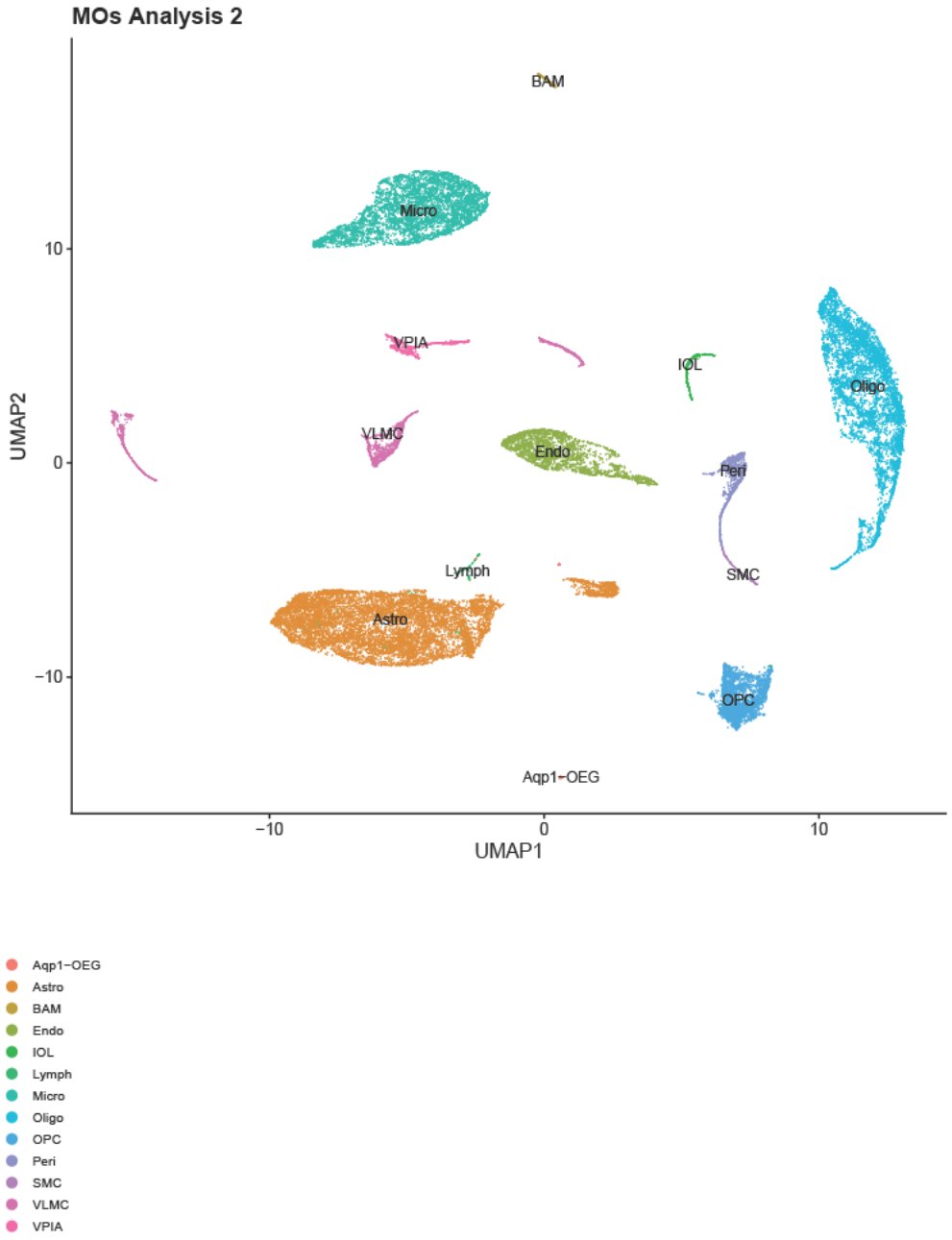


Figure 3.31: UMAP integration of the MOs Inhibitory and Other Neurons

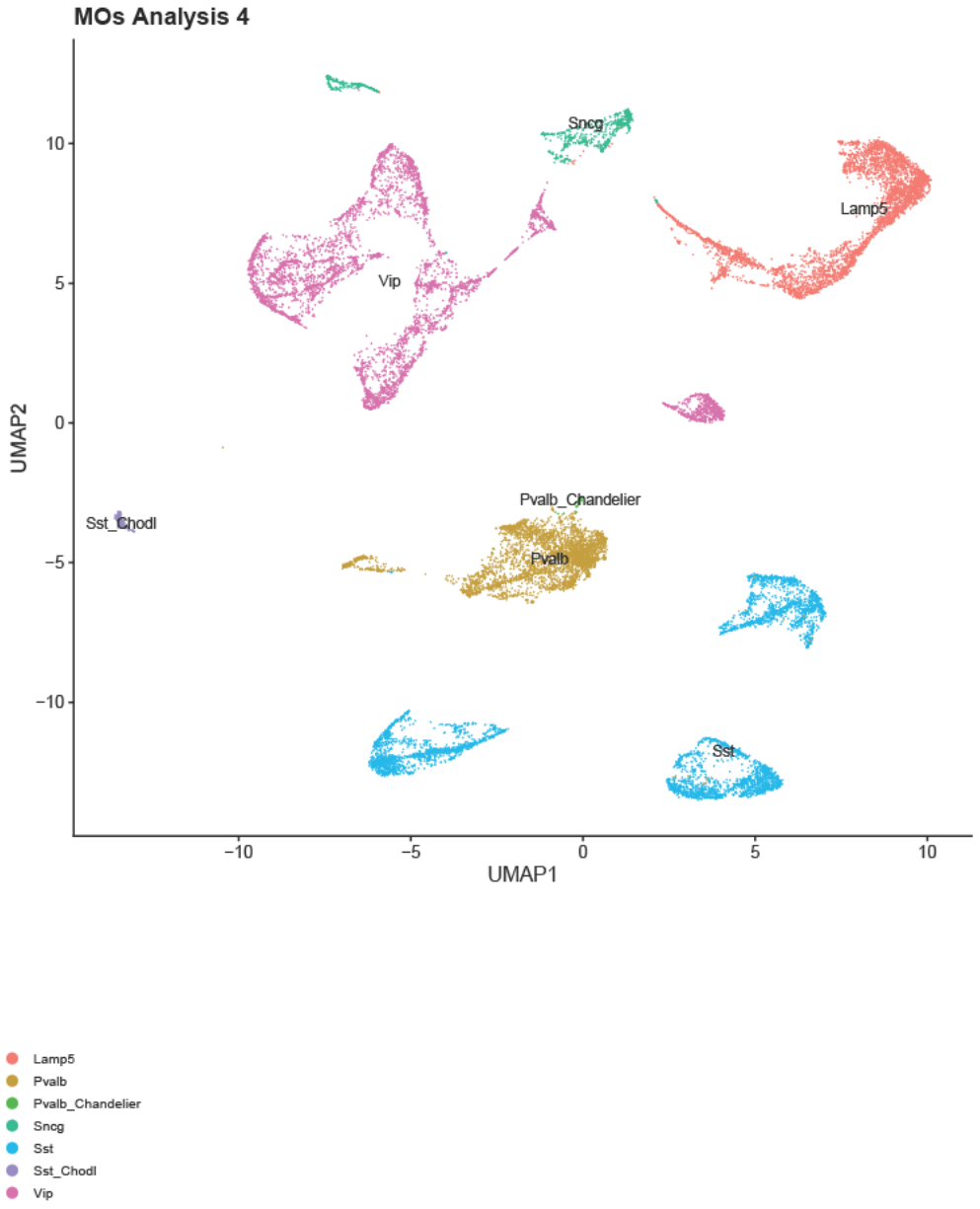


Figure 3.32: UMAP integration of the MOs Excitatory Neurons

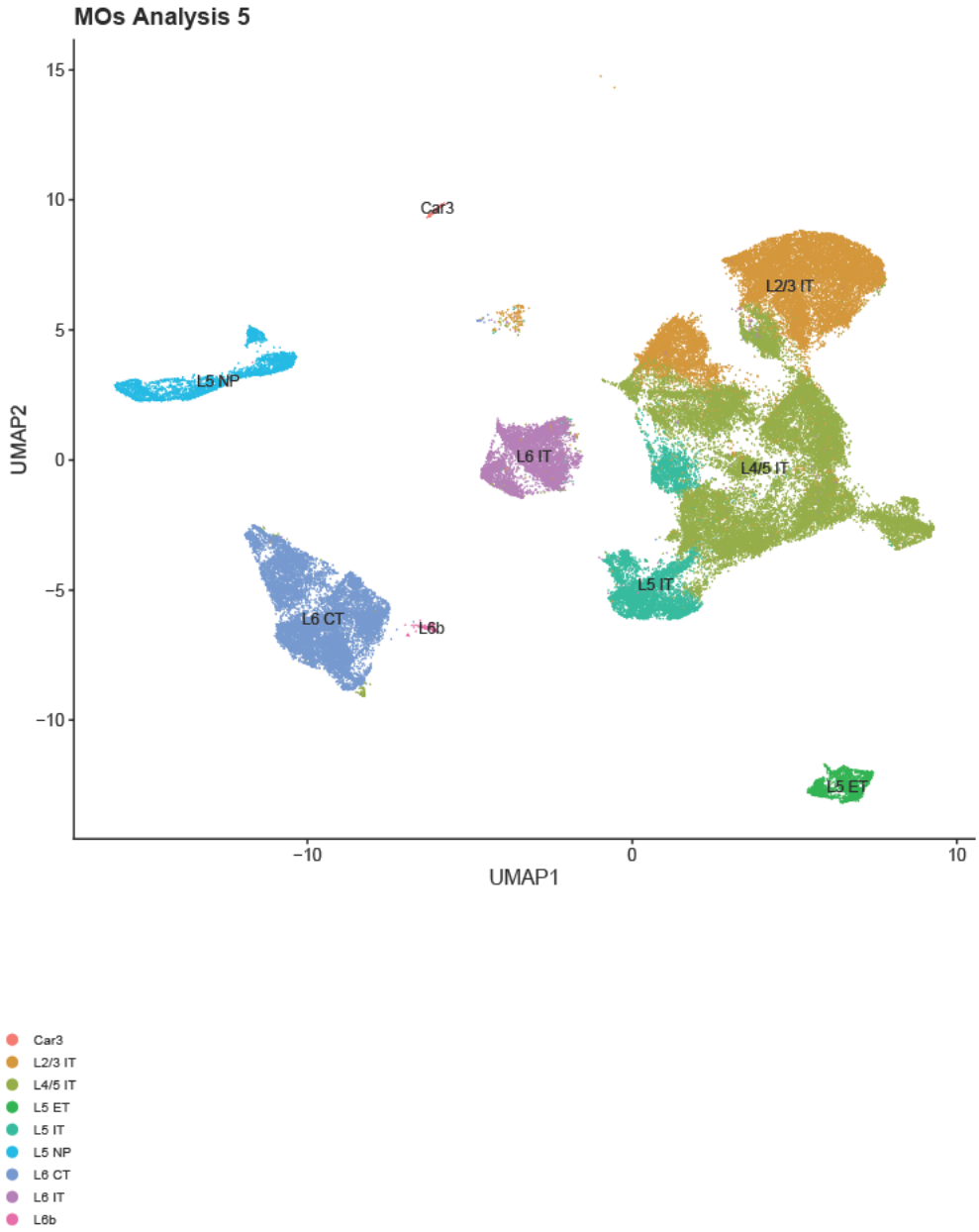


Figure 3.33: UMAP integration of the OLF Non-neuronal cells

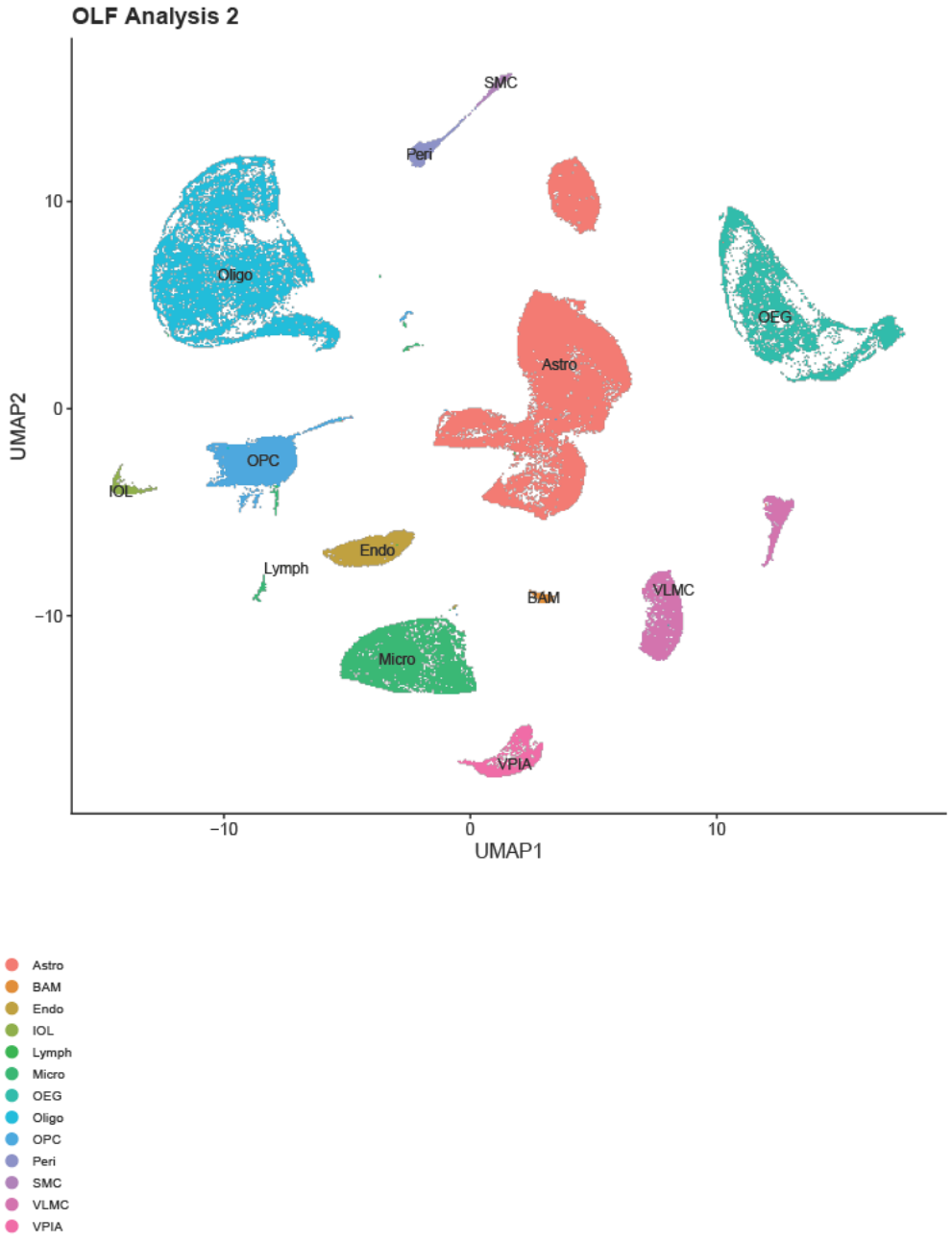


Figure 3.34: UMAP integration of the OLF Inhibitory and Other Neurons

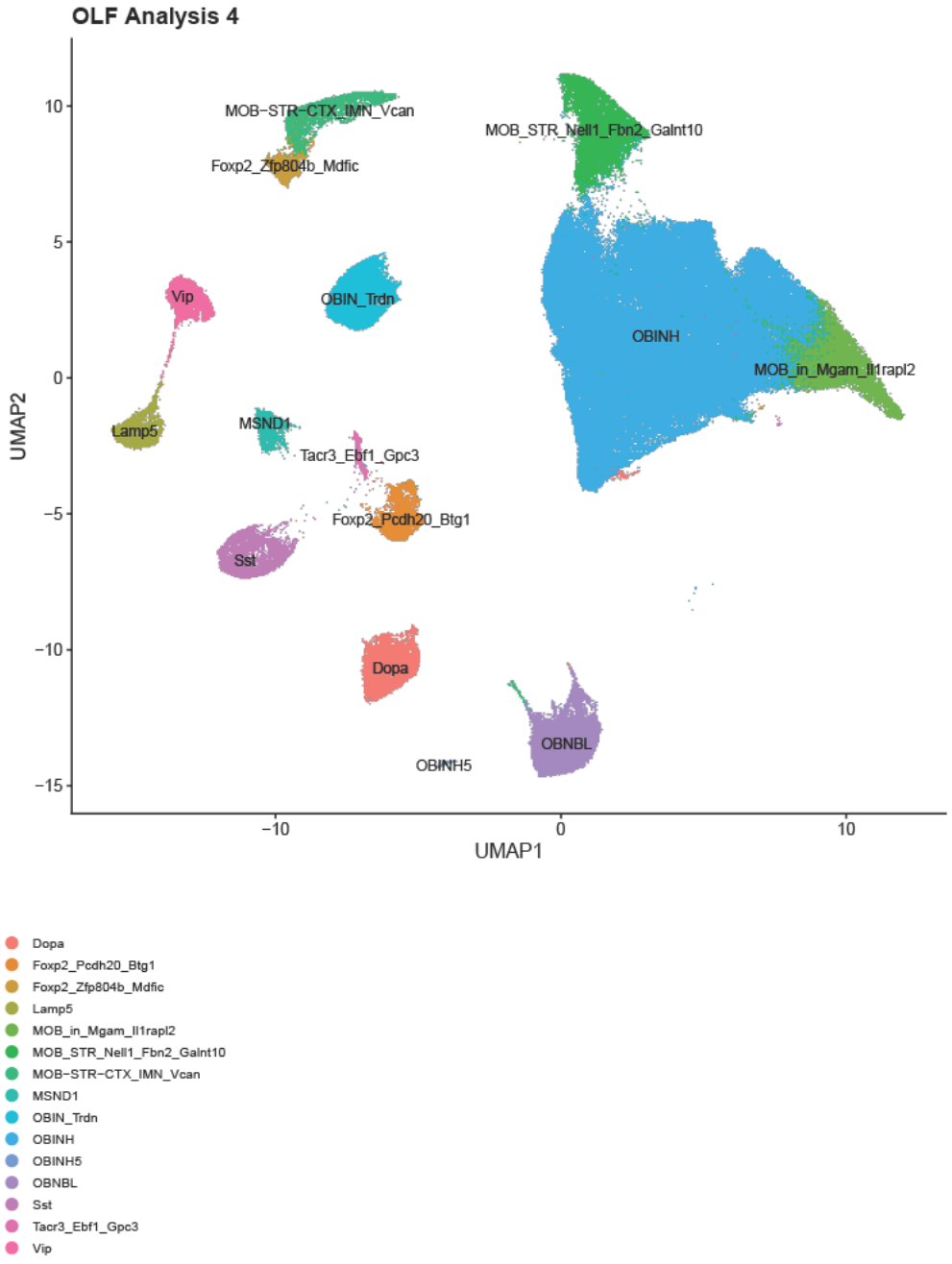


Figure 3.35: UMAP integration of the OLF Excitatory Neurons

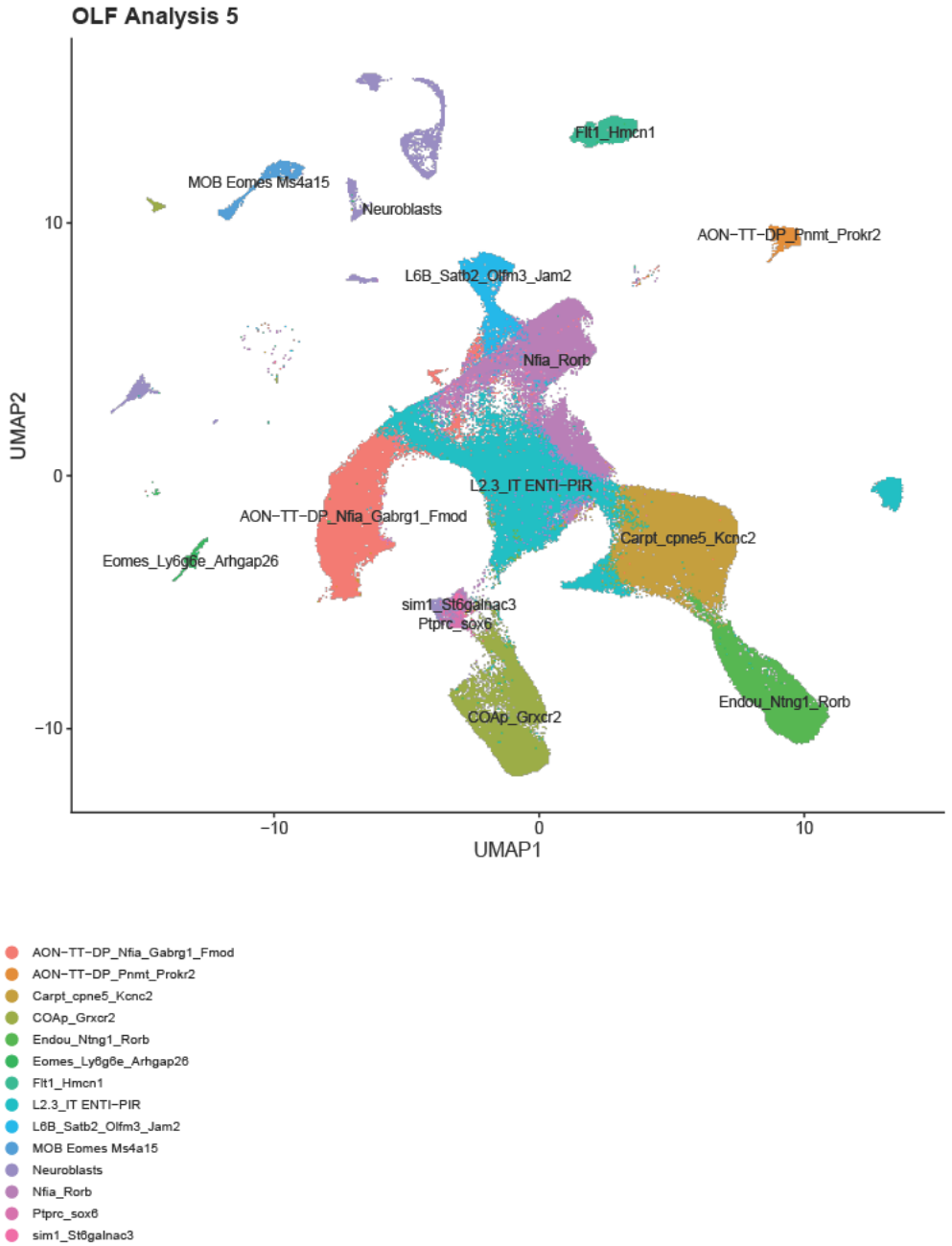


Figure 3.36: UMAP integration of the ORB_ACA Non-neuronal cells

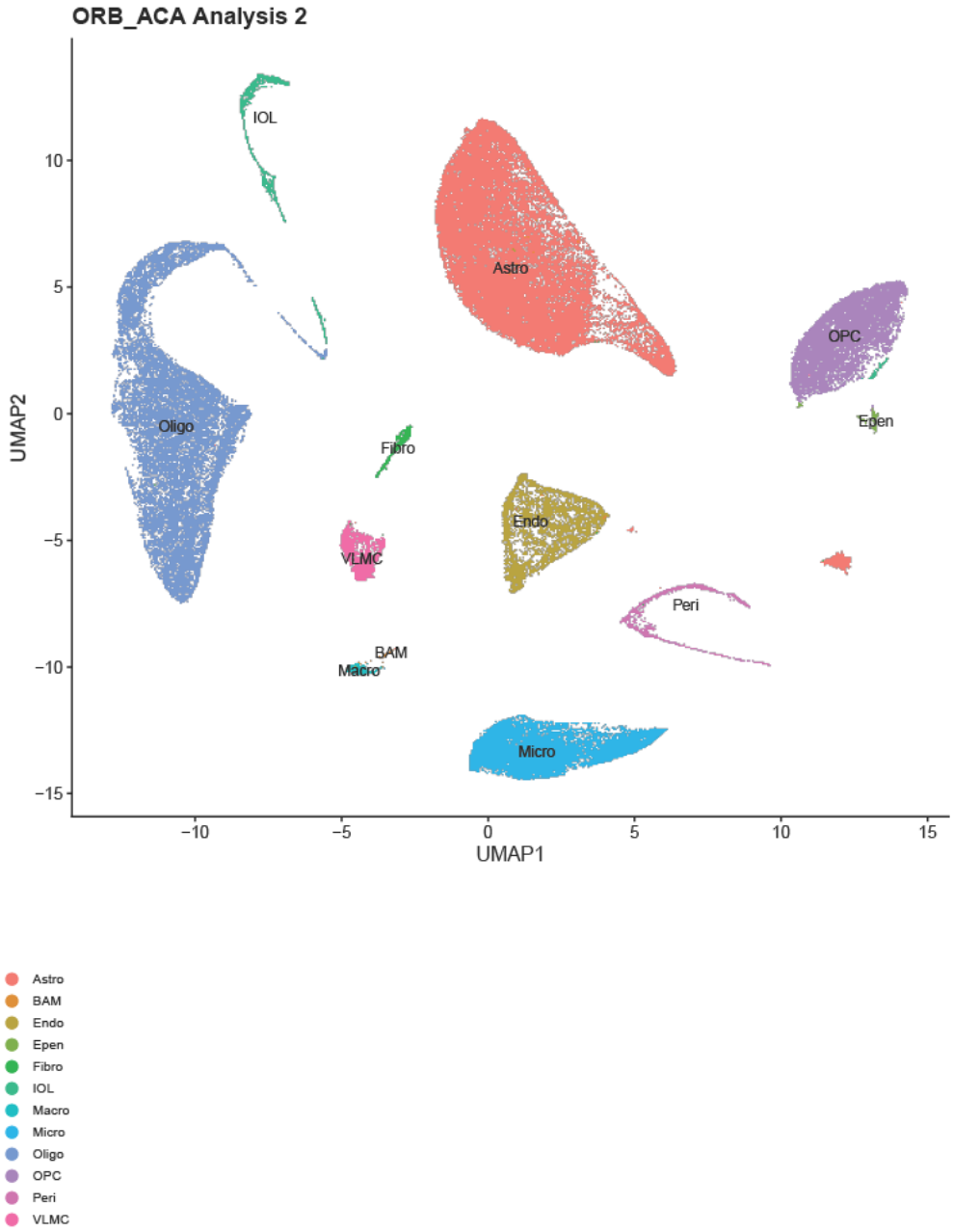


Figure 3.37: UMAP integration of the ORB_ACA Inhibitory and Other Neurons

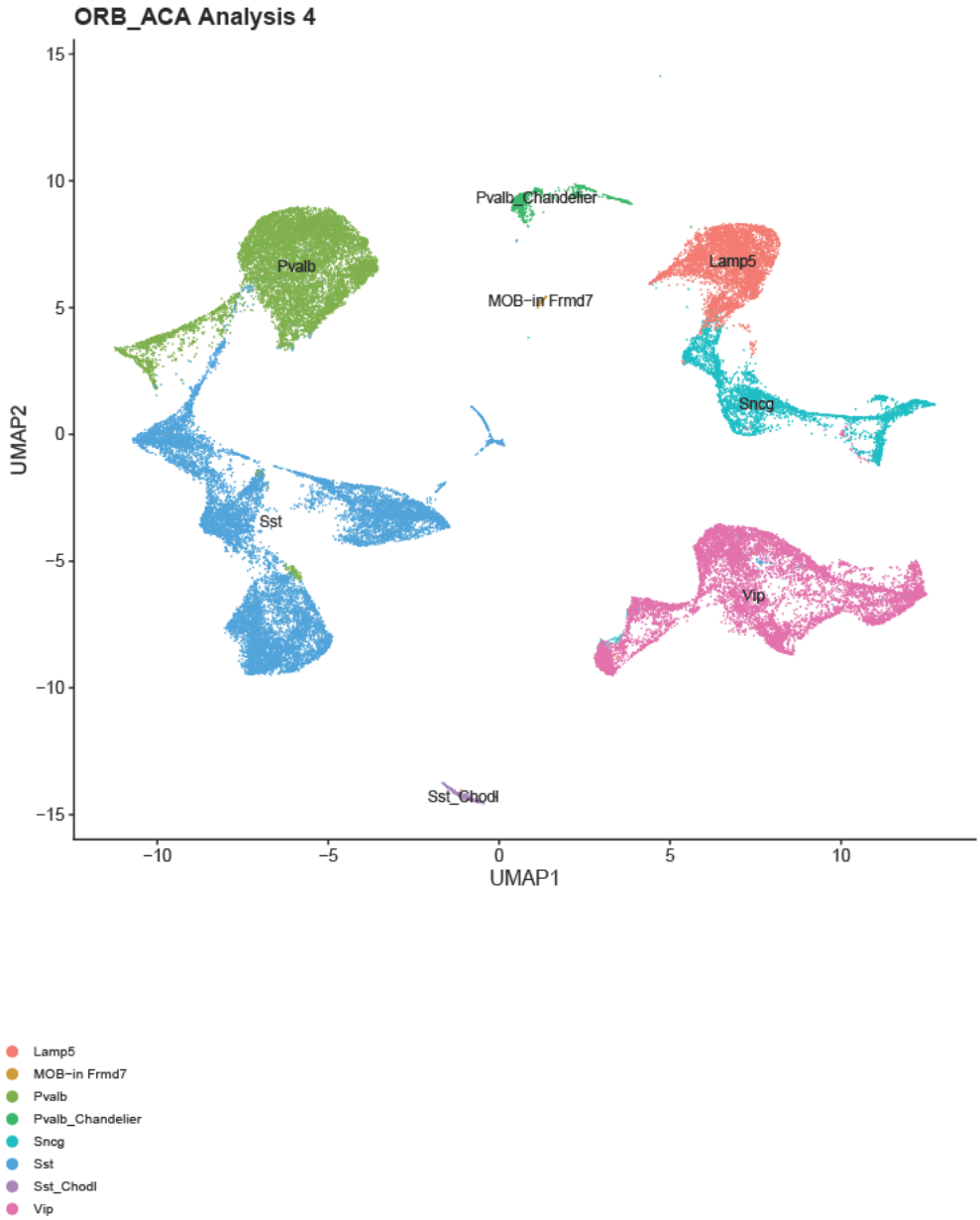


Figure 3.38: UMAP integration of the ORB_ACA Excitatory Neurons

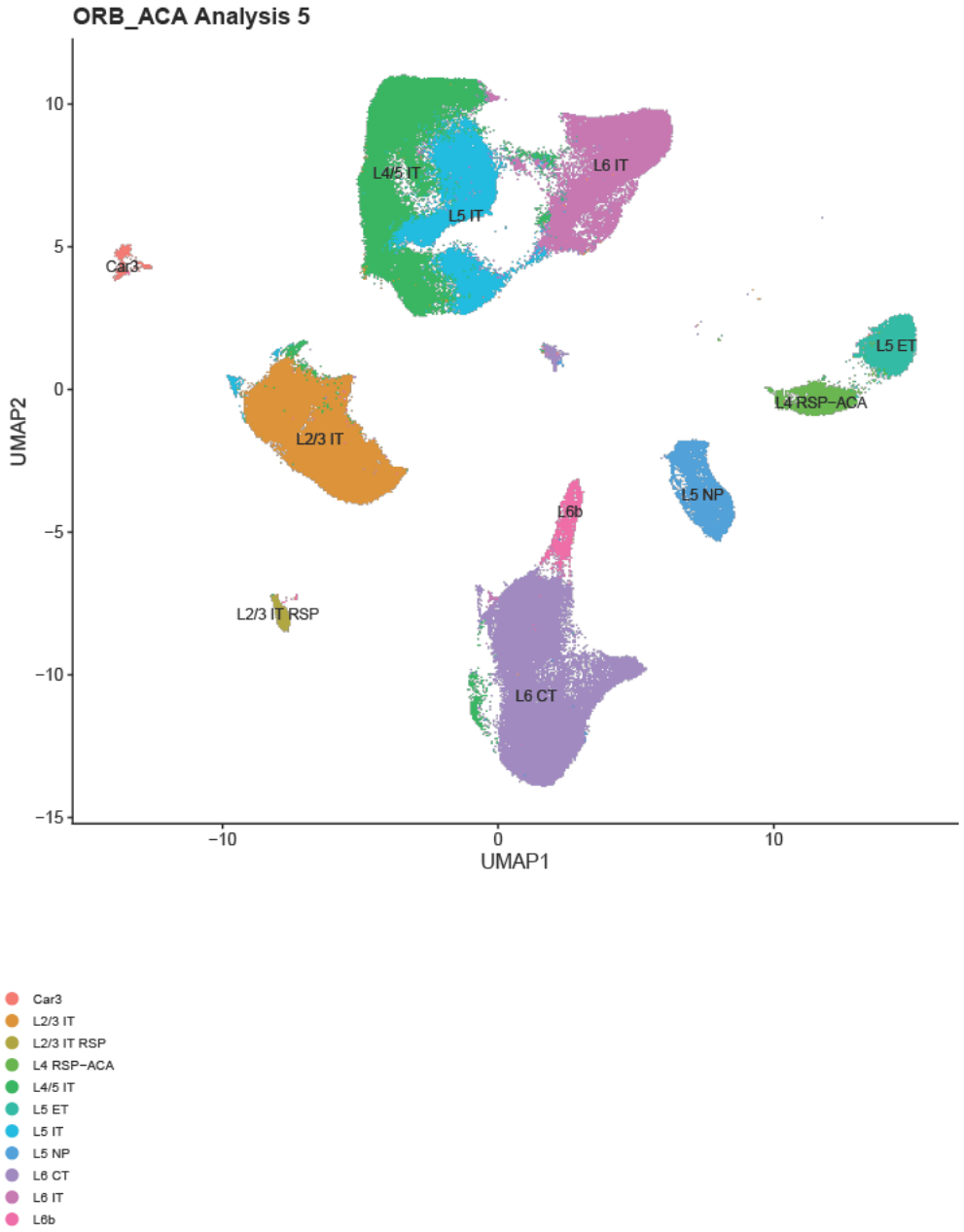


Figure 3.39: UMAP integration of the PAL Non-neuronal cells

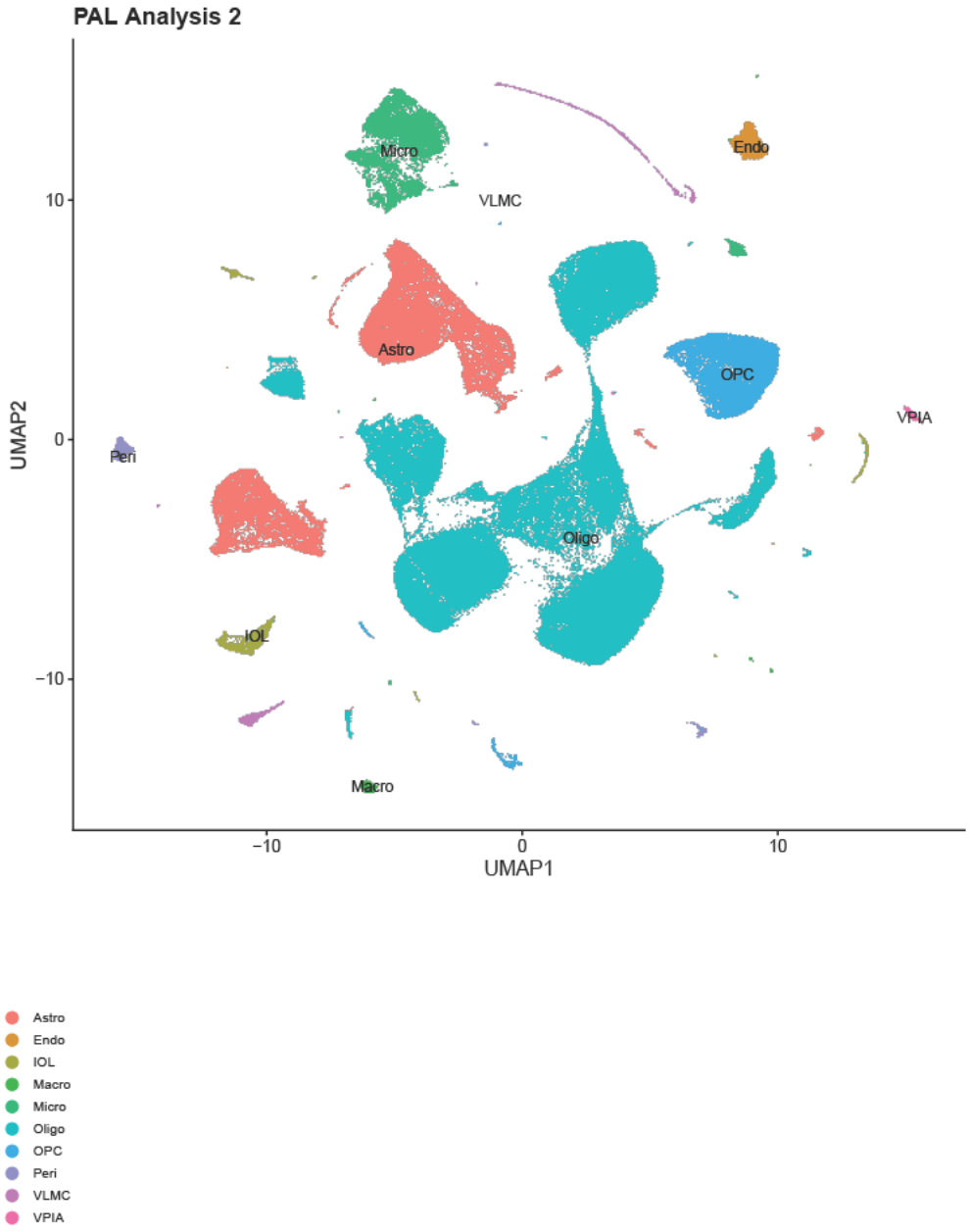


Figure 3.40: UMAP integration of the PAL Inhibitory and Other Neurons

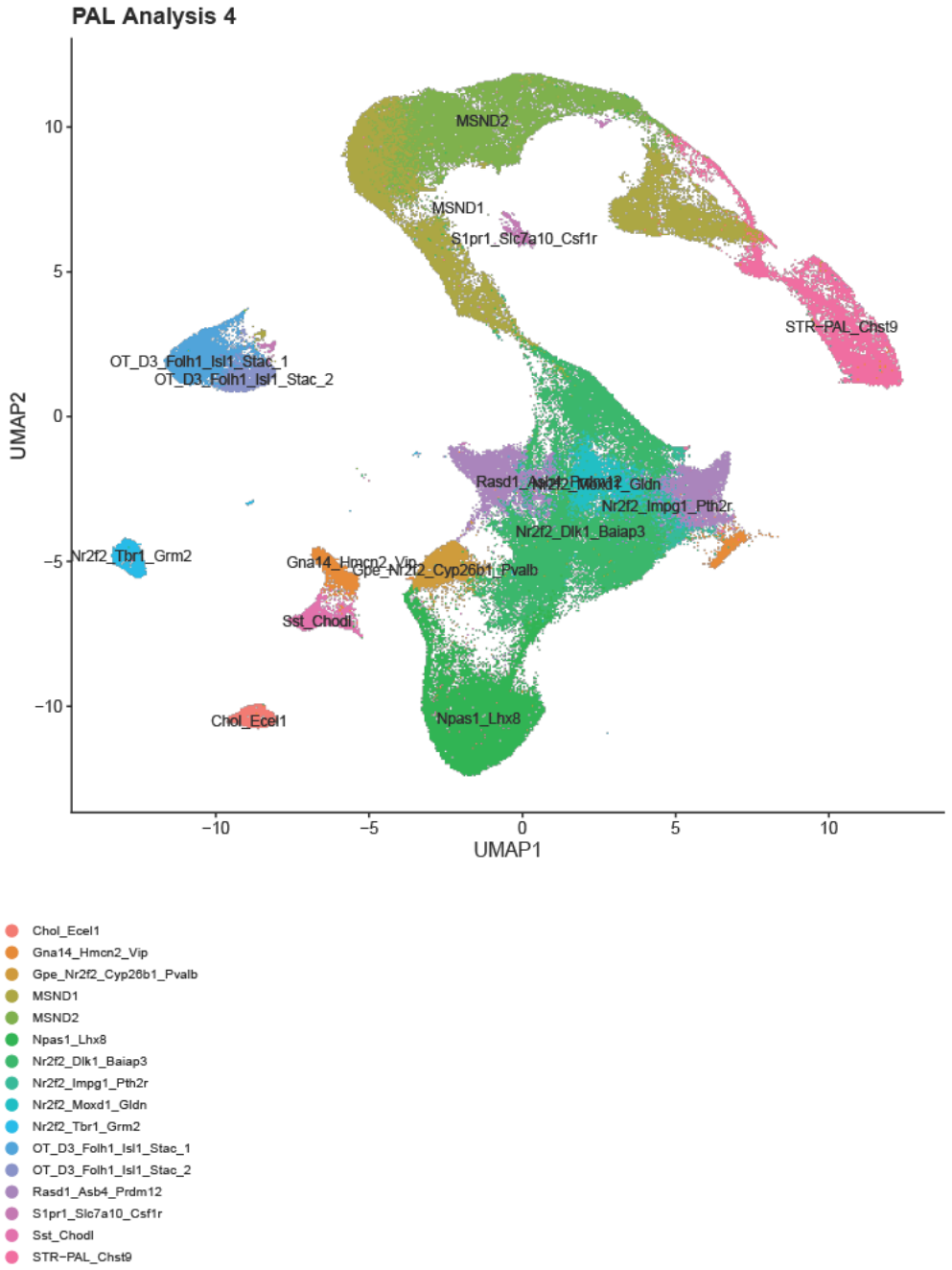


Figure 3.41: UMAP integration of the PAL Excitatory Neurons

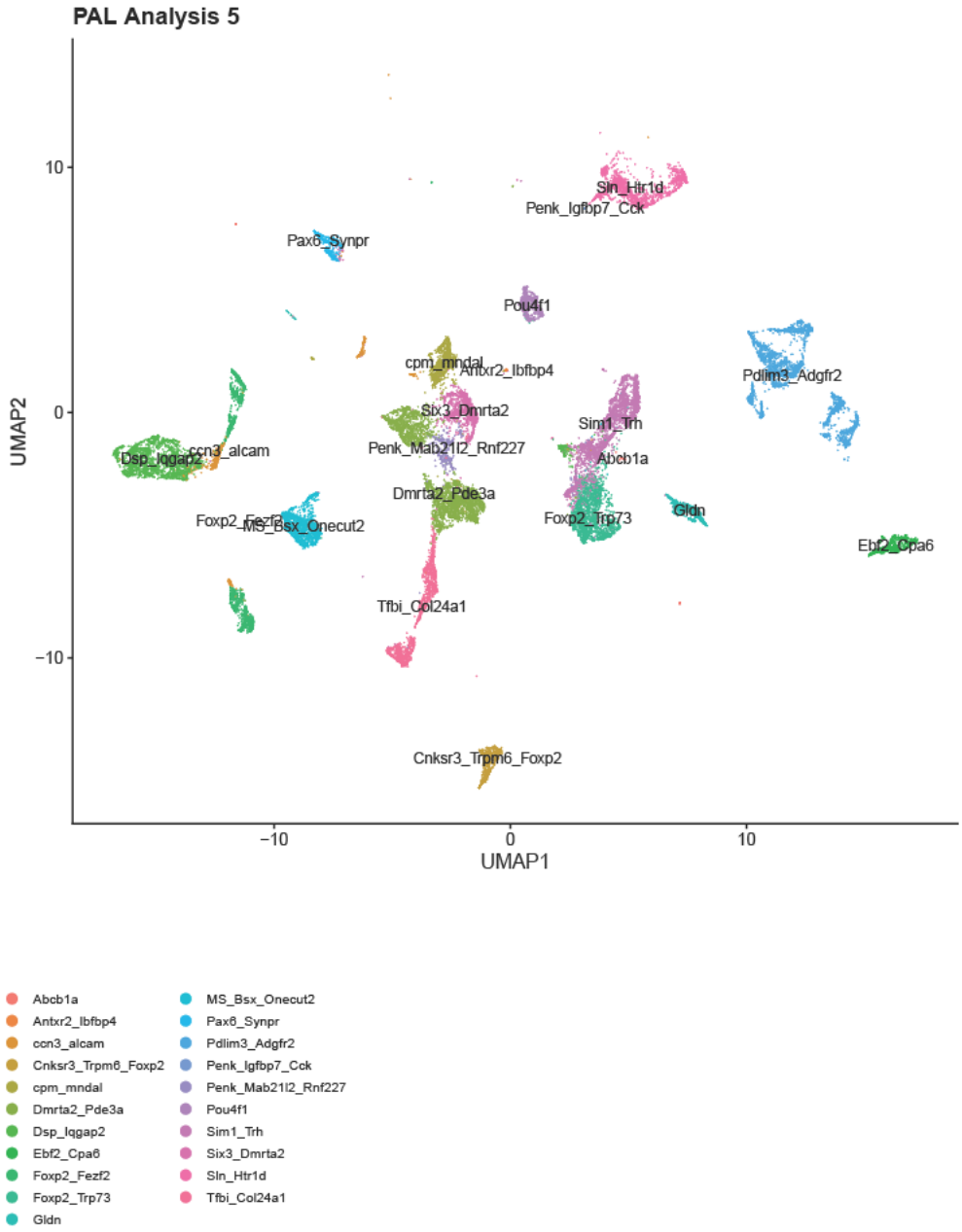


Figure 3.42: UMAP integration of the RSP Non-neuronal cells

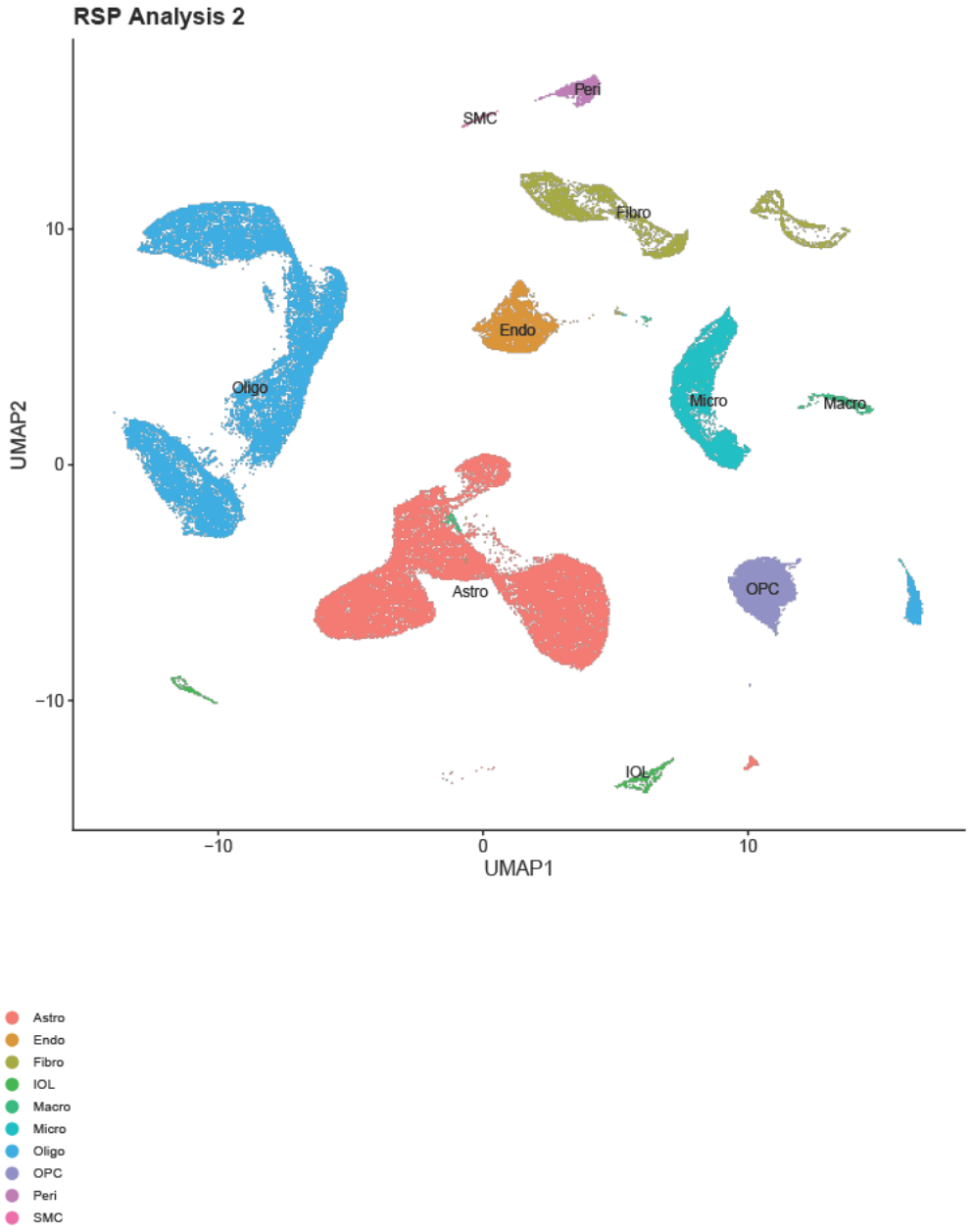


Figure 3.43: UMAP integration of the RSP Inhibitory and Other Neurons

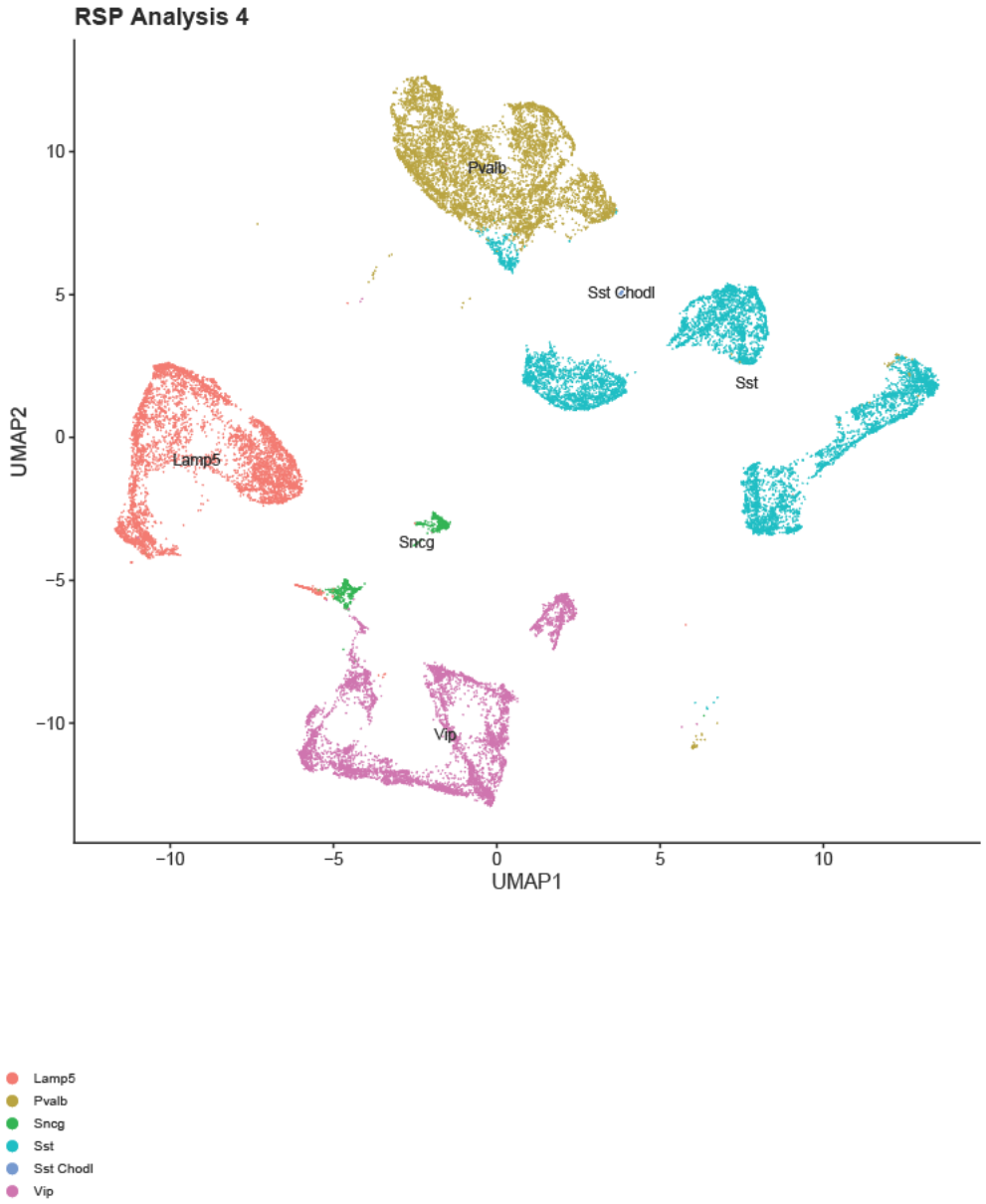


Figure 3.44: UMAP integration of the RSP Excitatory Neurons

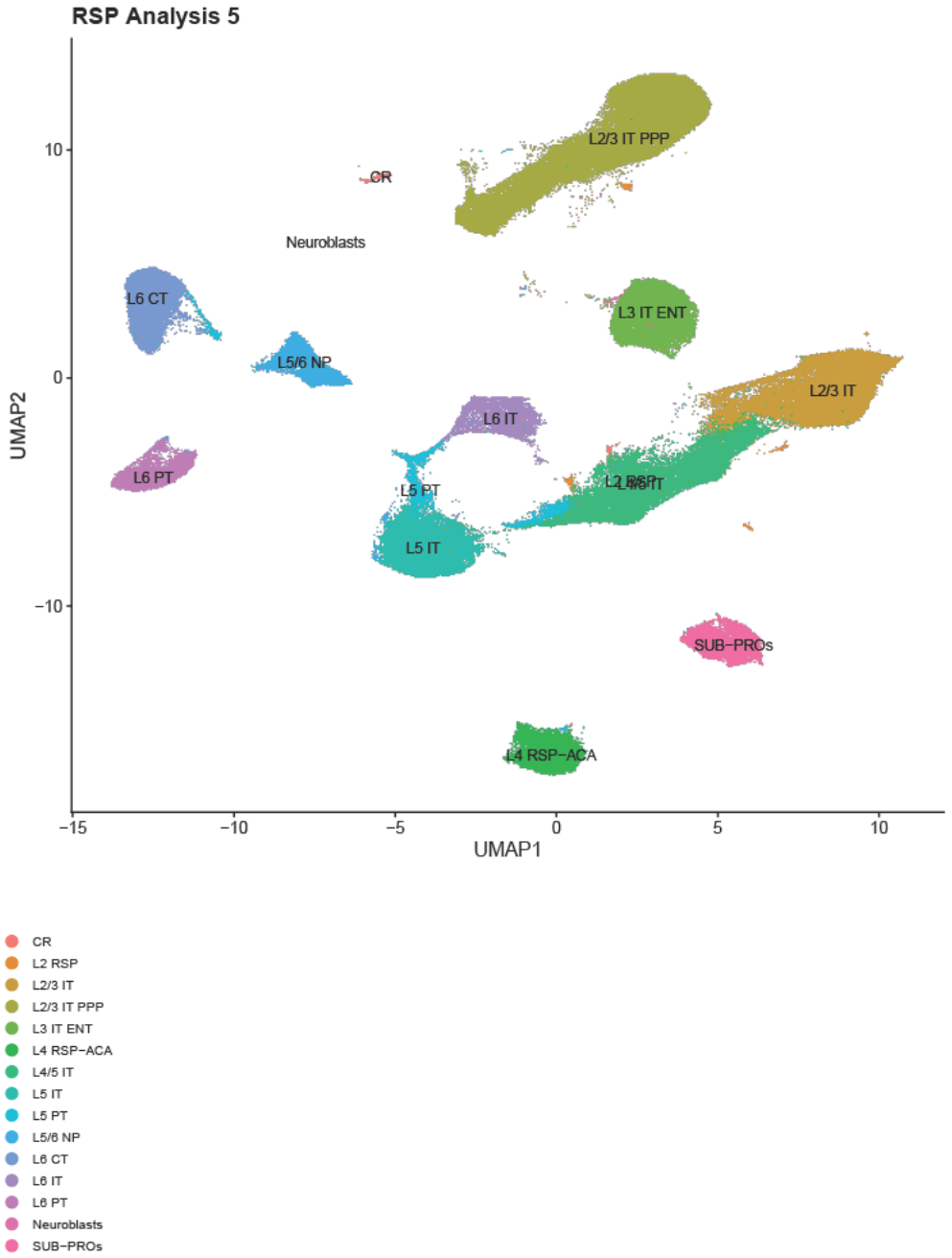


Figure 3.45: UMAP integration of the SS Non-neuronal cells

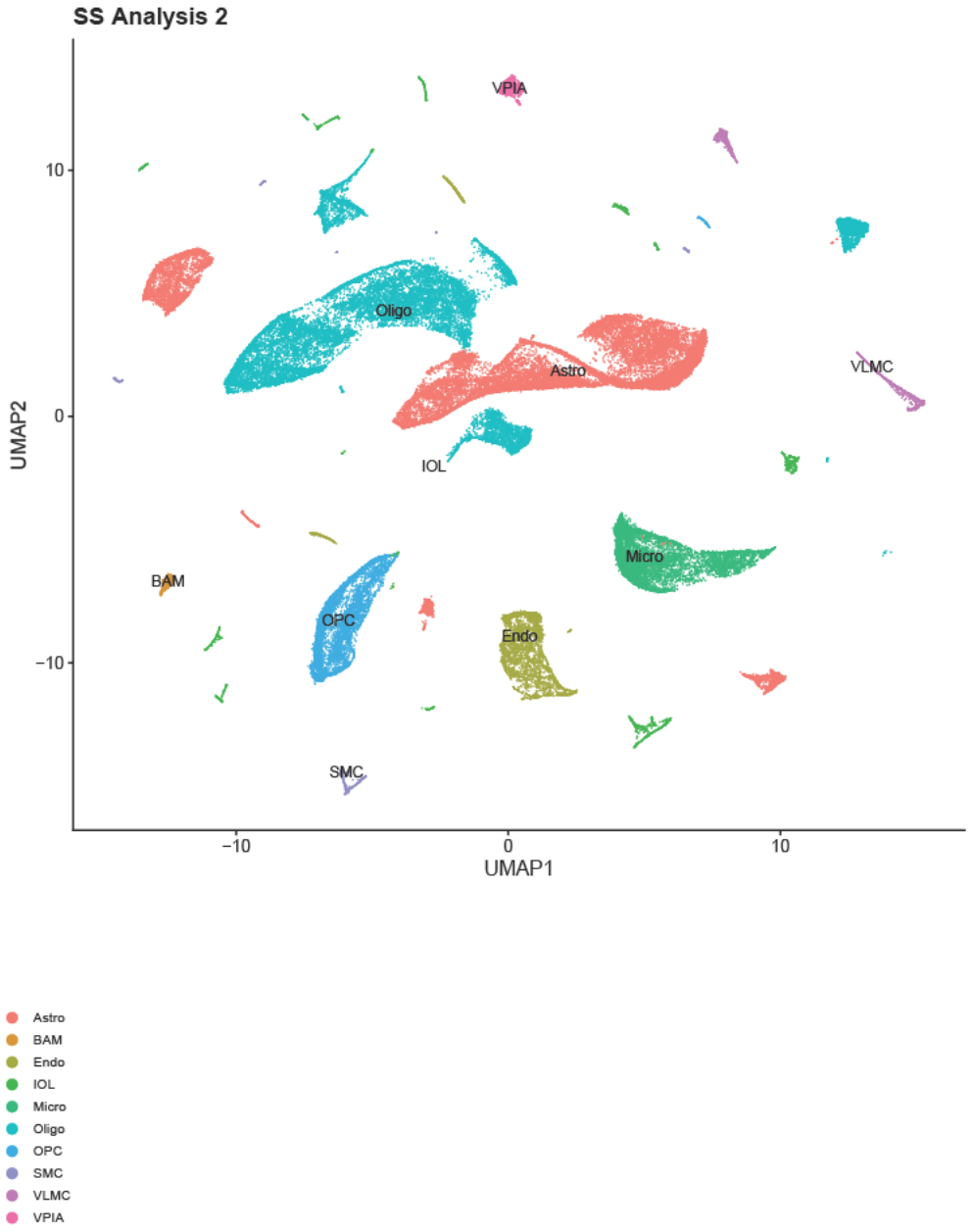


Figure 3.46: UMAP integration of the SS Inhibitory and Other Neurons

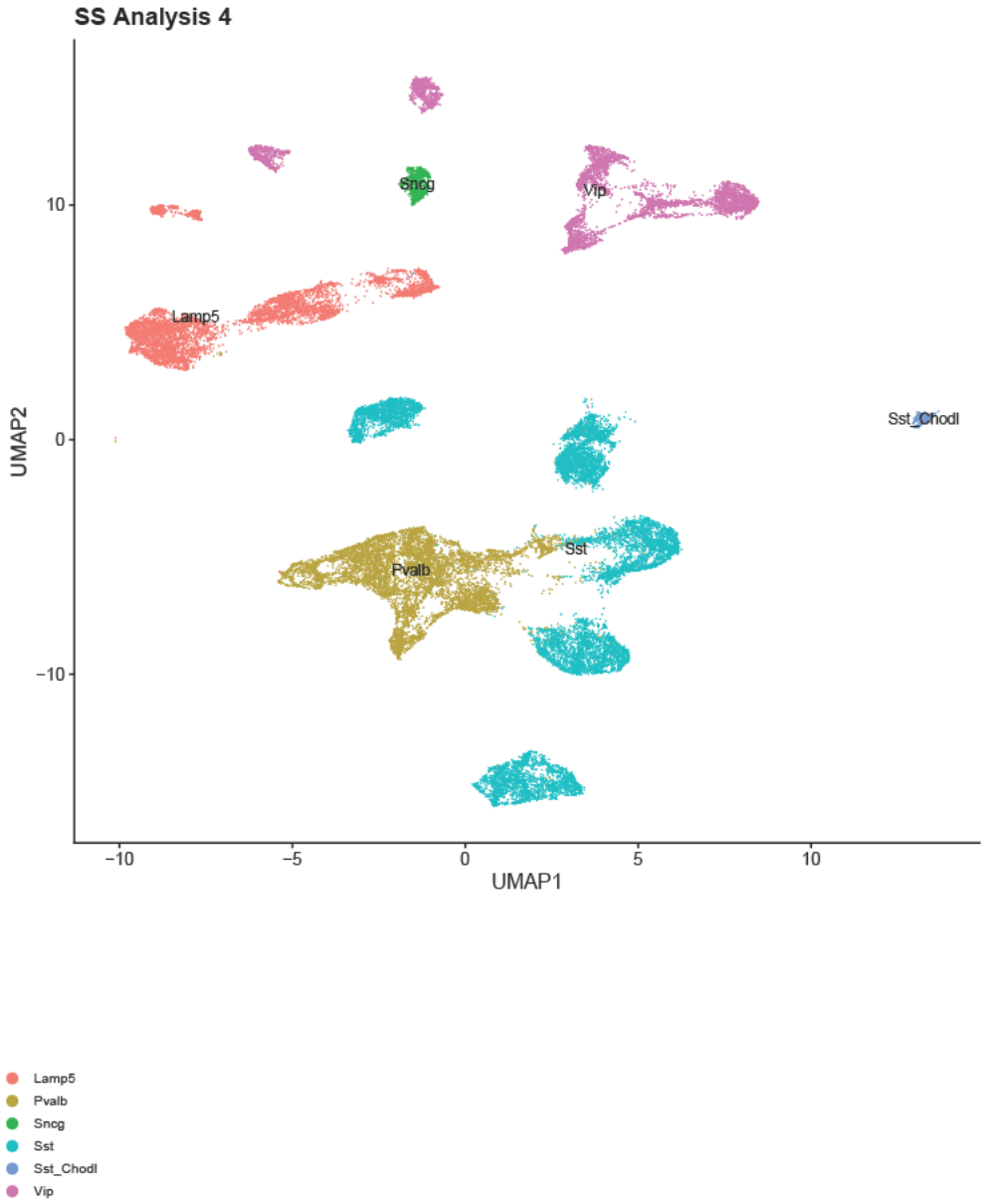


Figure 3.47: UMAP integration of the SS Excitatory Neurons

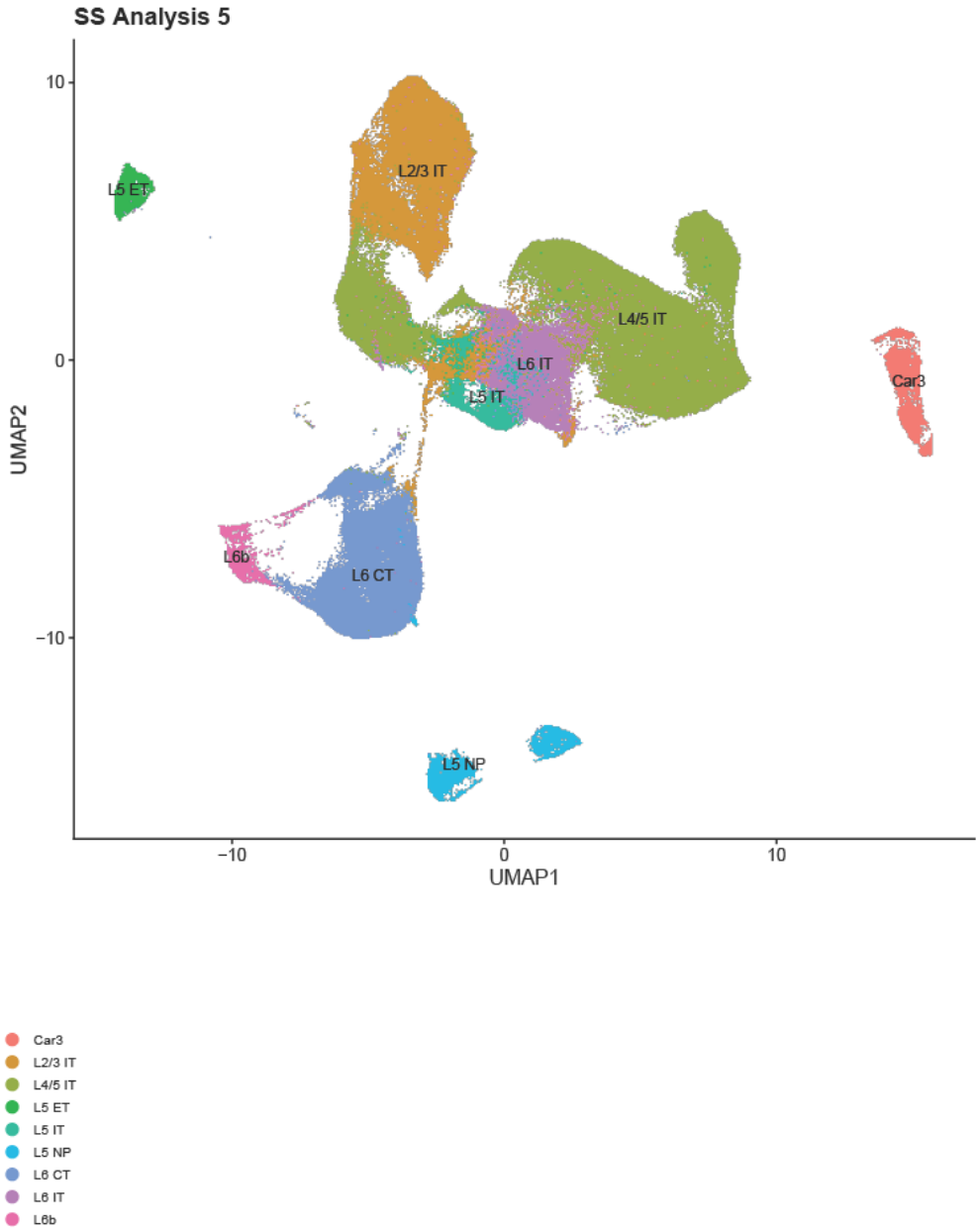


Figure 3.48: UMAP integration of the STR Non-neuronal cells

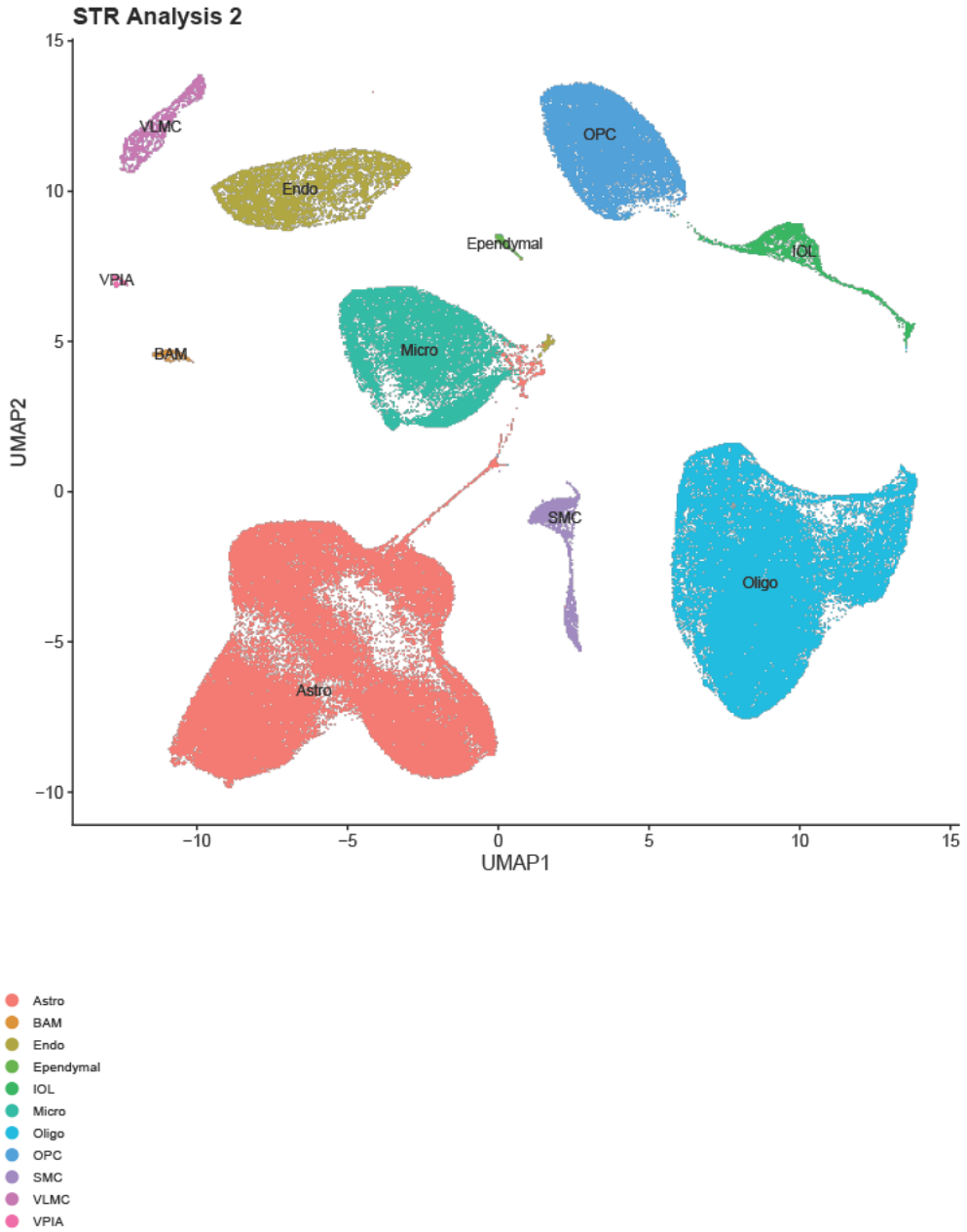


Figure 3.49: UMAP integration of the STR Inhibitory and Other Neurons

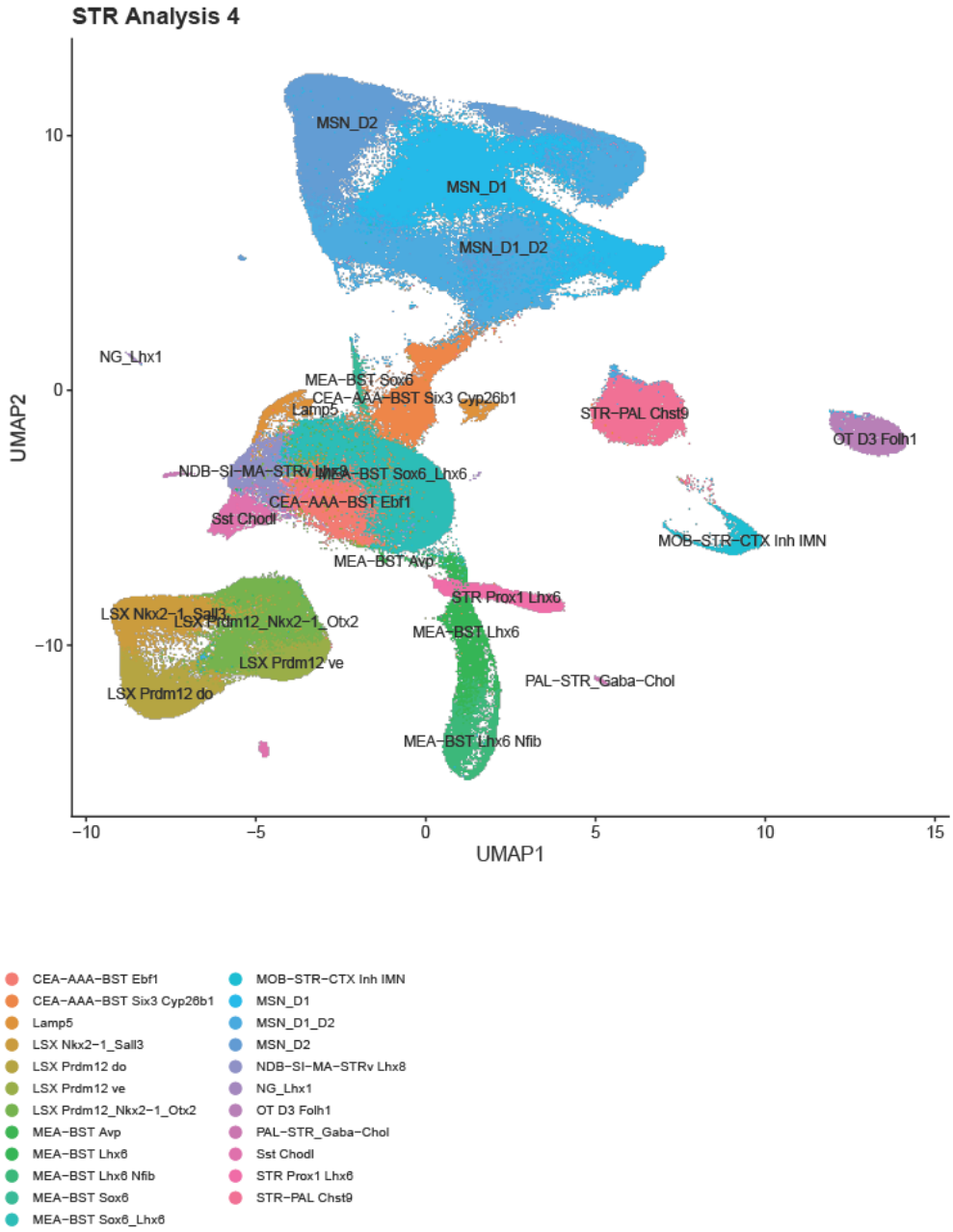


Figure 3.50: UMAP integration of the STR Excitatory Neurons

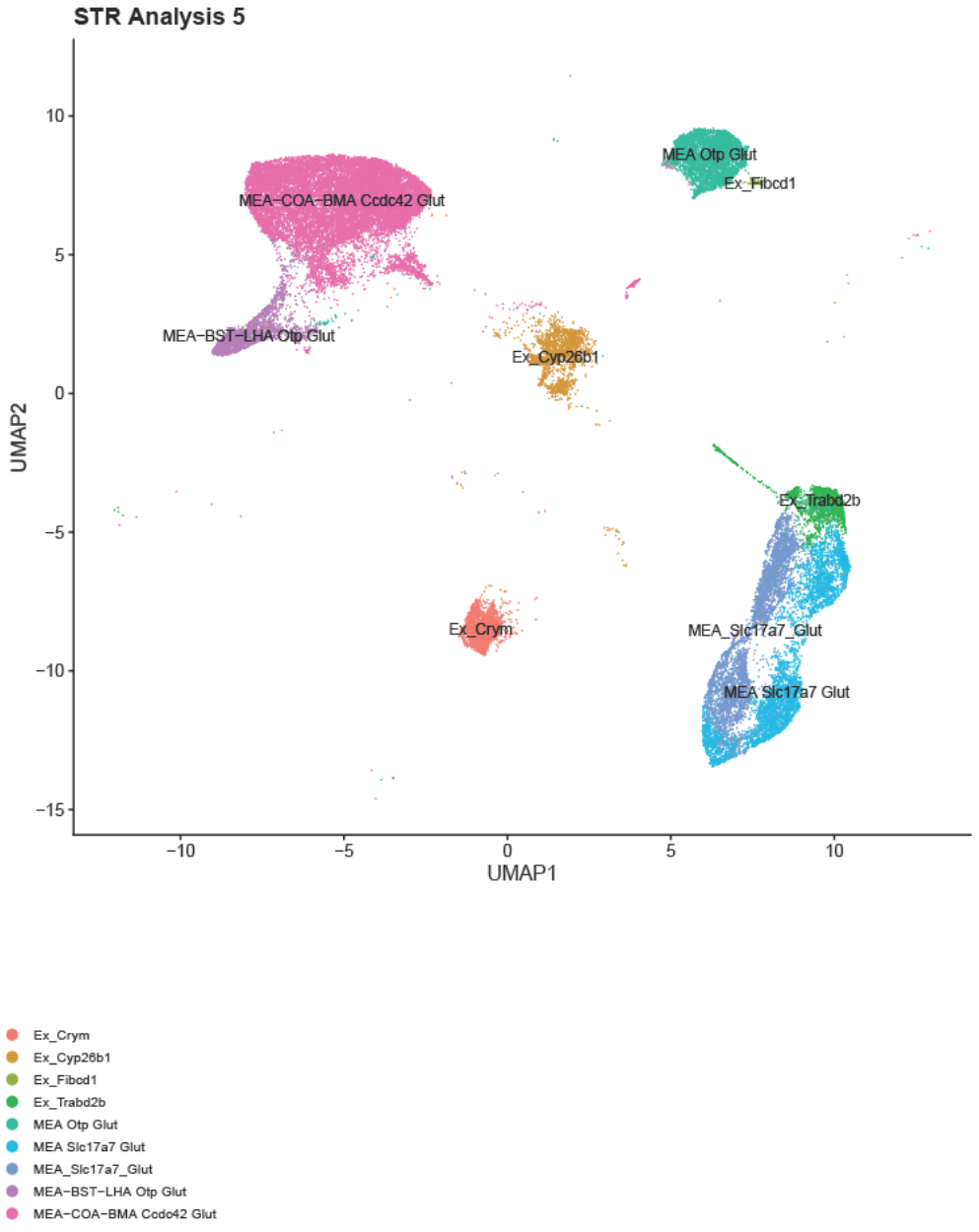


Figure 3.51: UMAP integration of the TH Non-neuronal cells

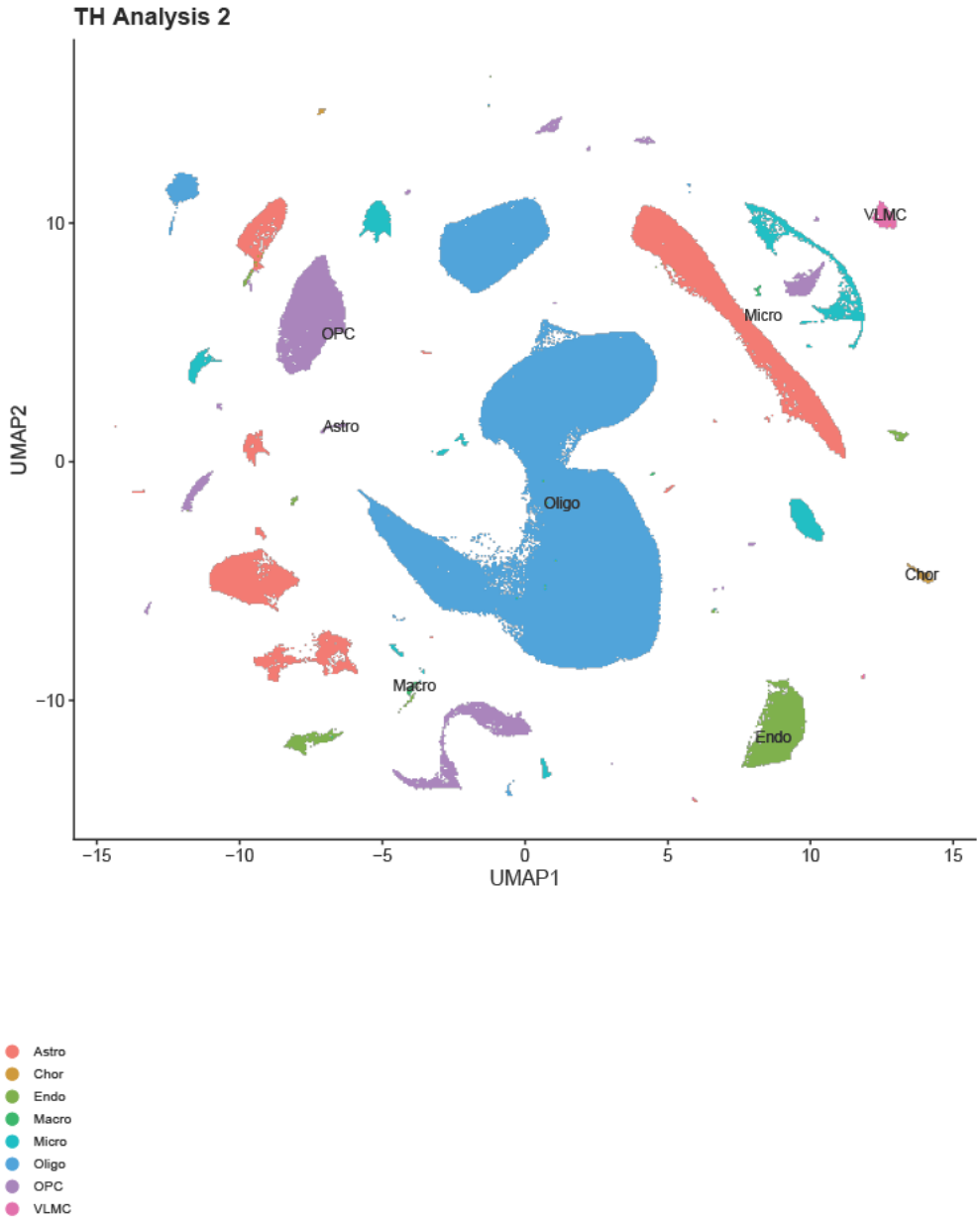


Figure 3.52: UMAP integration of the TH Inhibitory and Other Neurons

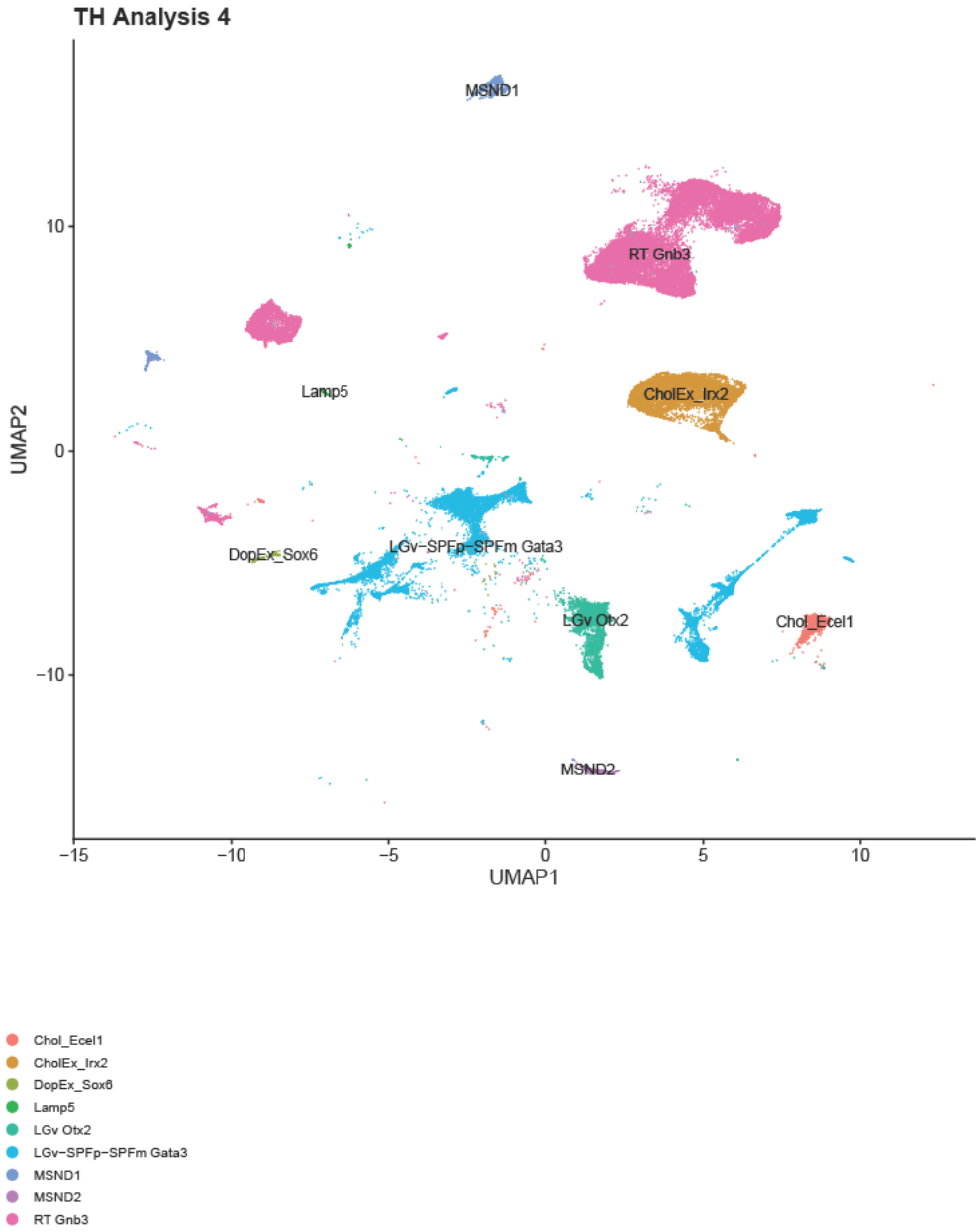


Figure 3.53: UMAP integration of the TH Inhibitory and Other Neurons

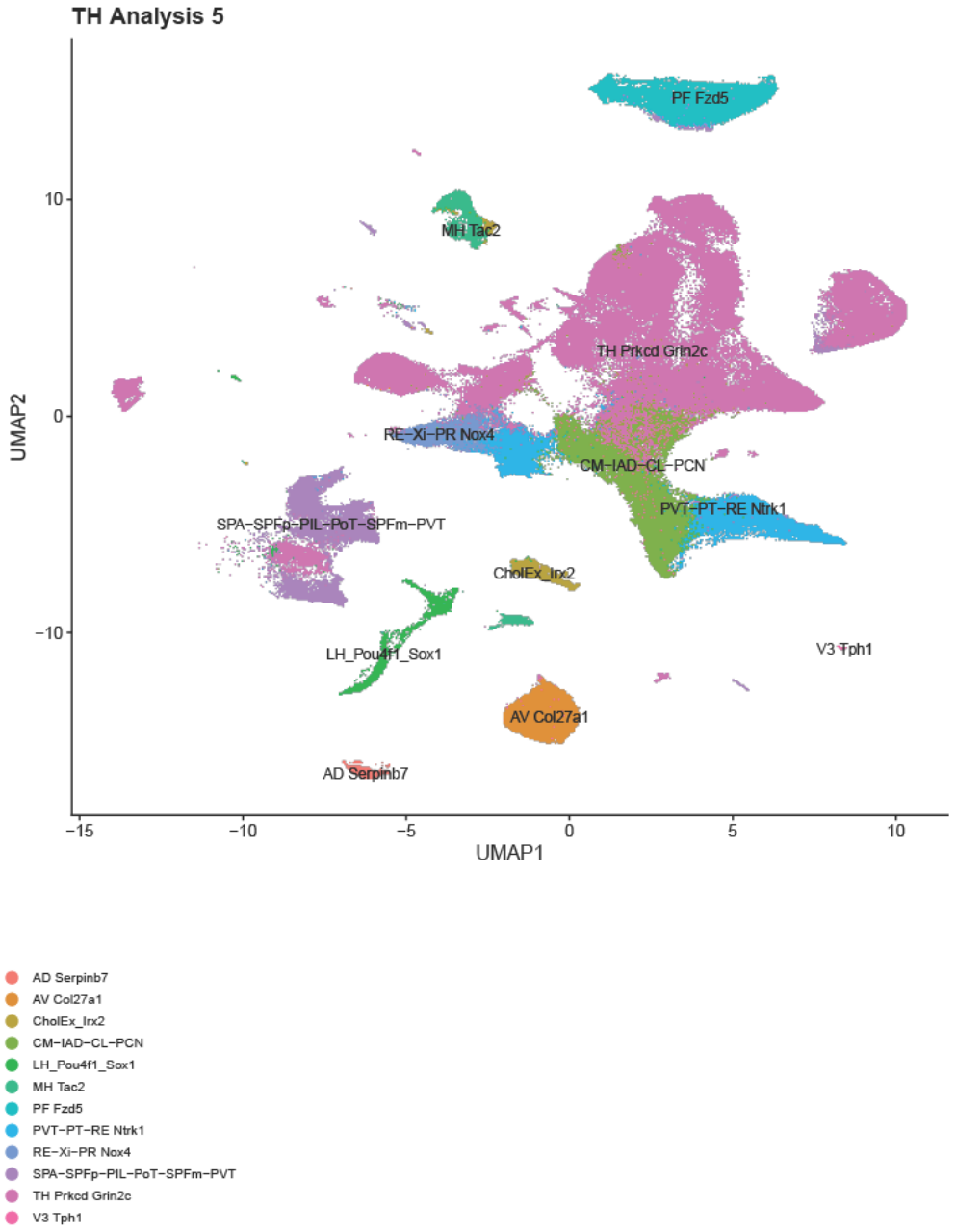


Figure 3.54: UMAP integration of the VIS Non-neuronal cells

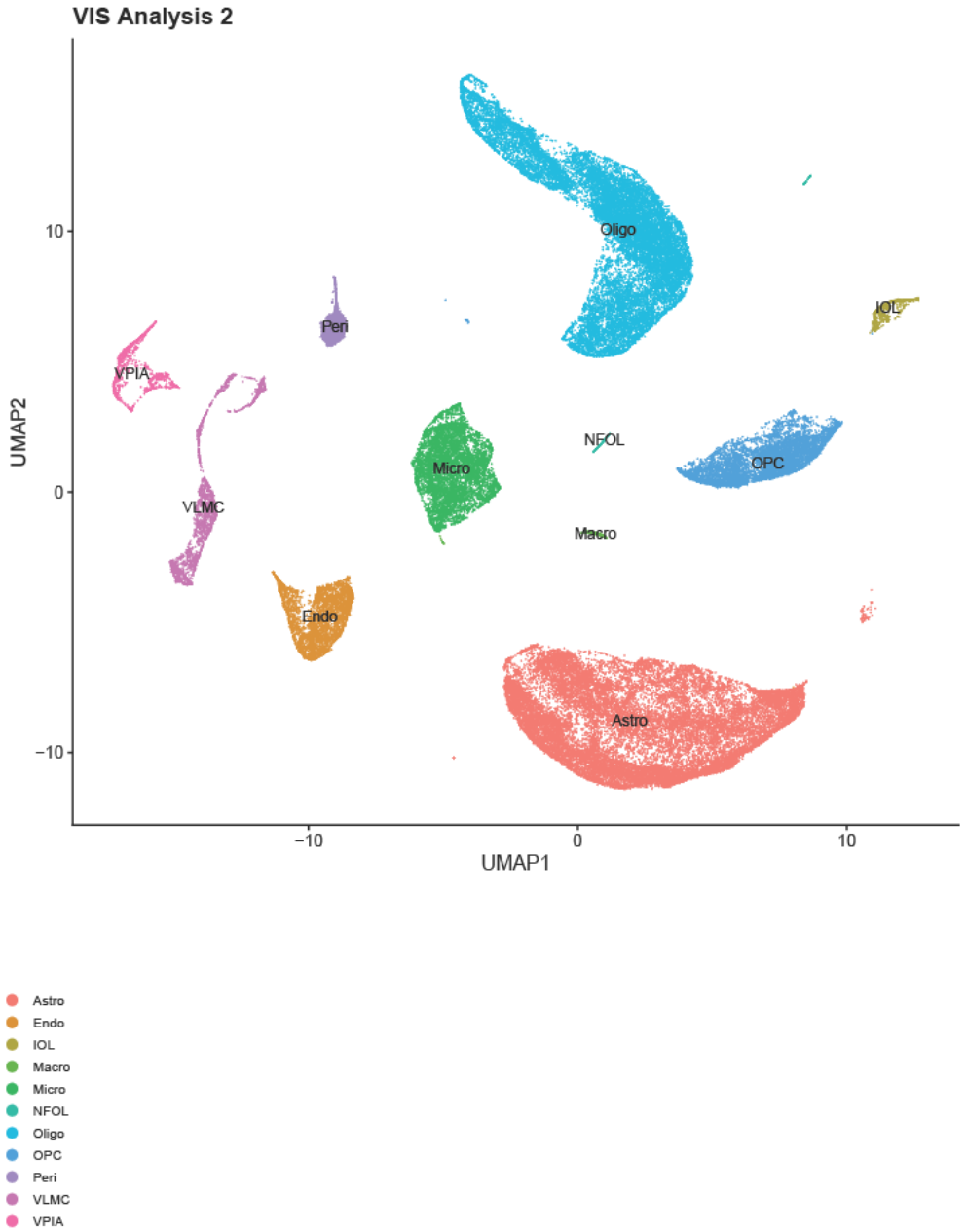


Figure 3.55: UMAP integration of the VIS Inhibitory and Other Neurons

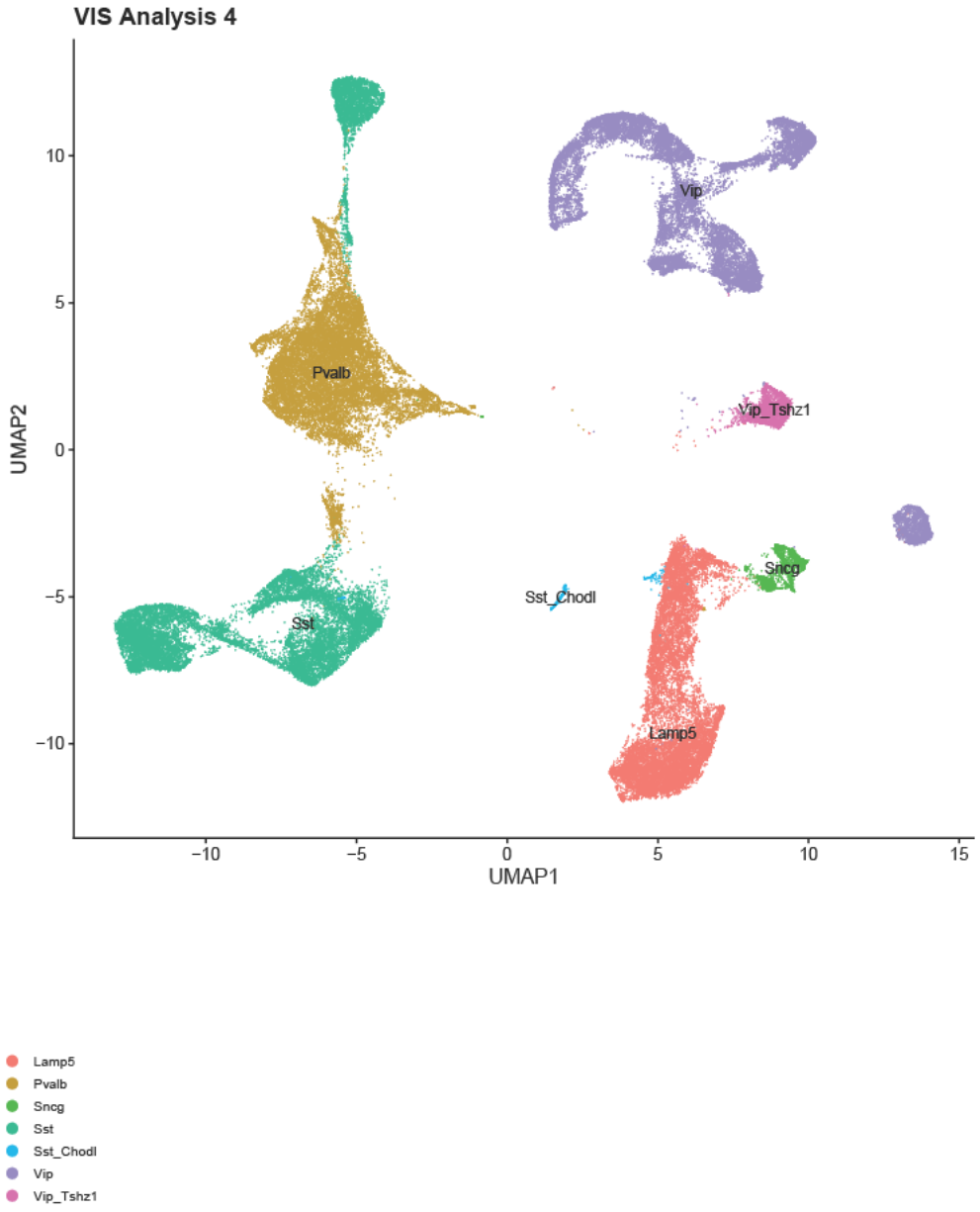


Figure 3.56: UMAP integration of the VIS Excitatory Neurons

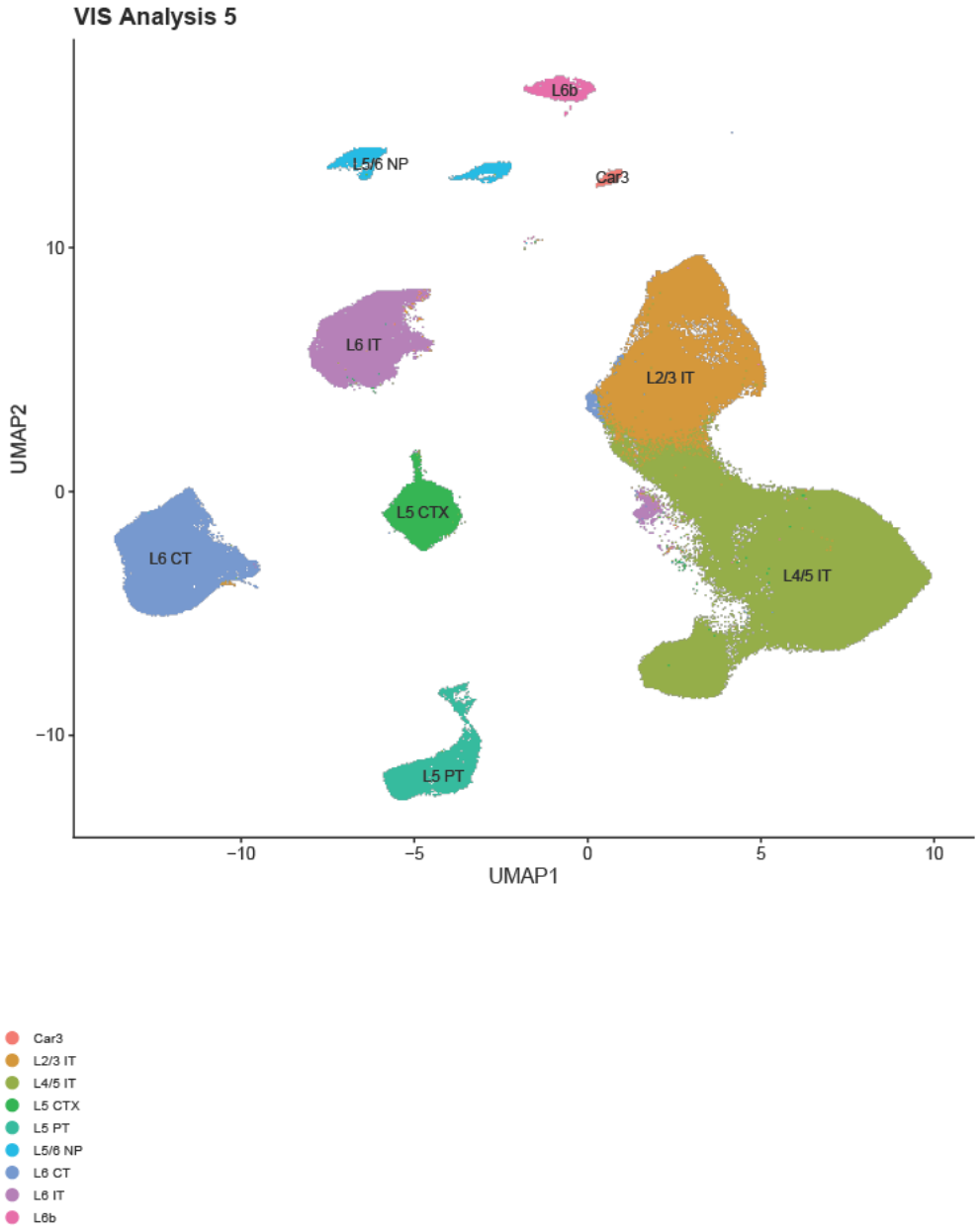
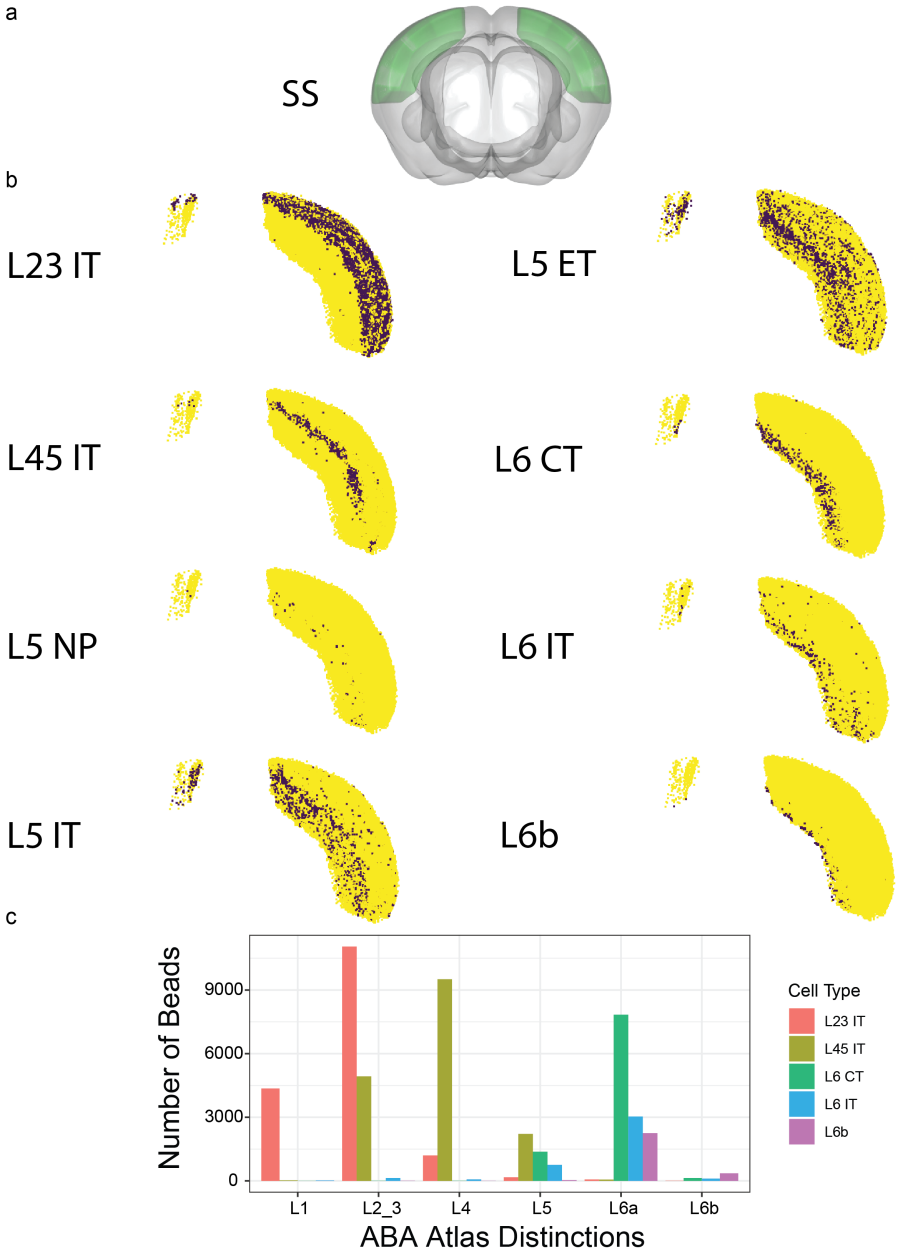
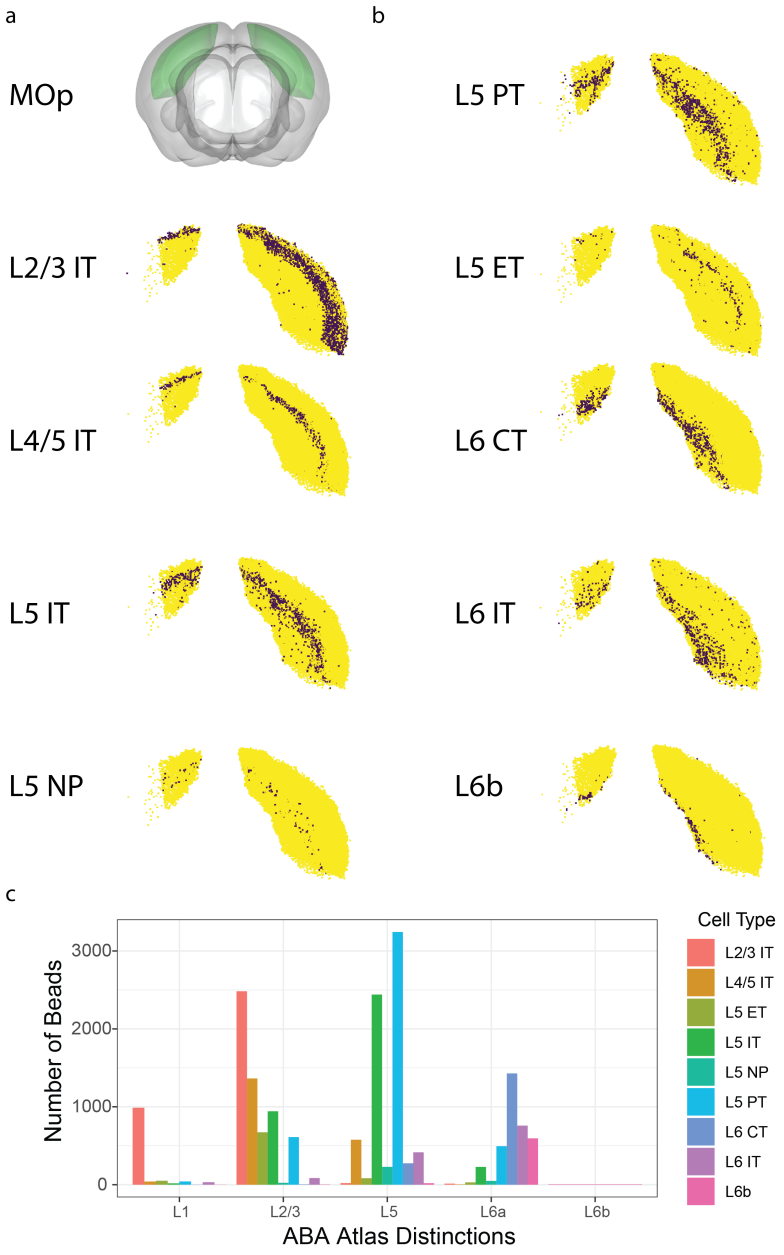


Figure 3.57: Identified Cortical Layer Distribution in the SS



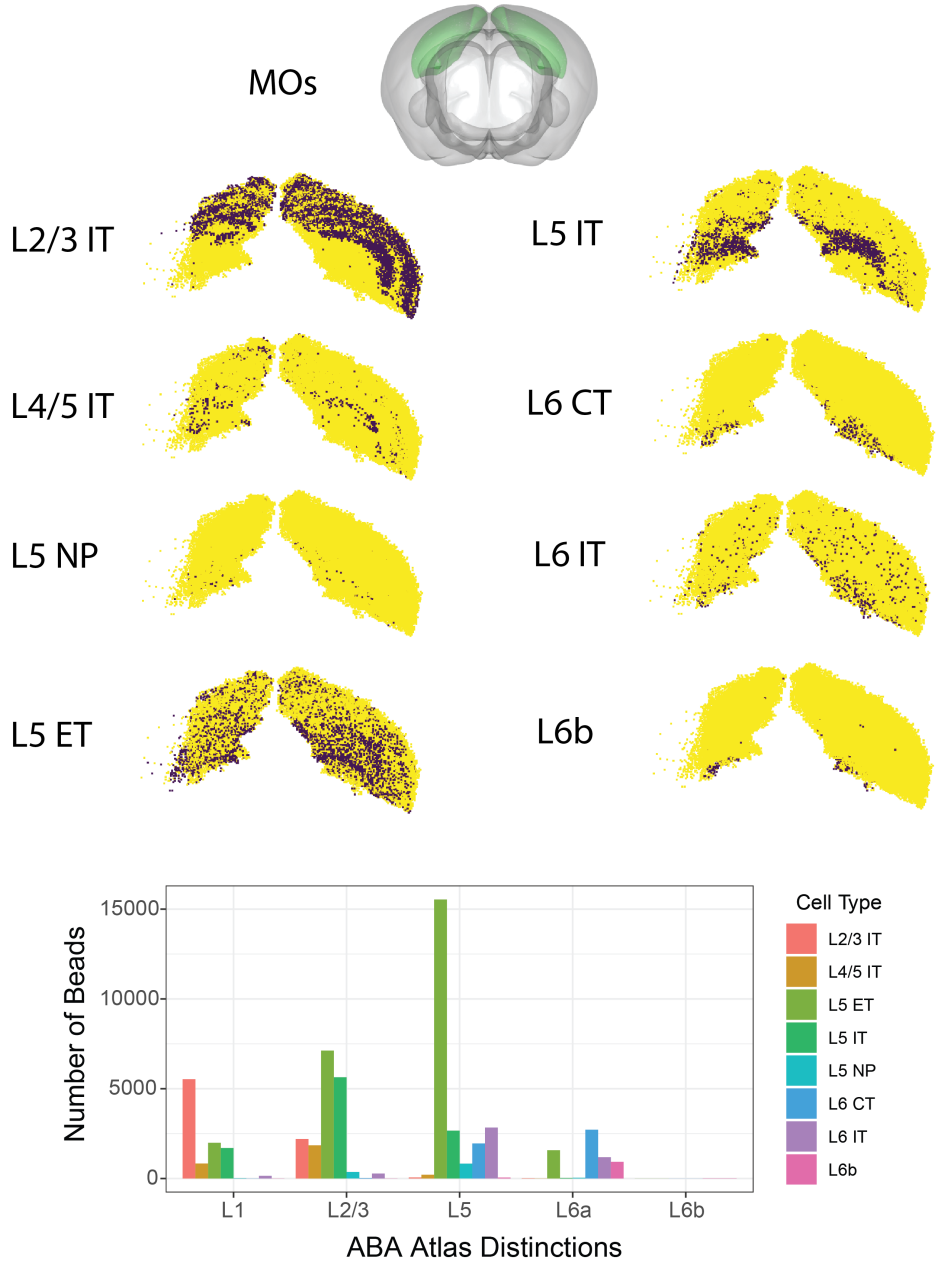
(**a**) The ABA brain is shown with a coronal view in gray, where the somatosensory region (SS) is shown in green. (**b**) For each cortical neuron cluster assignment that resulted from the SS glutamatergic integration, we color each bead confidently assigned that cell type within the SS to be purple, and each cell type not confidently assigned that cell type yellow. We take each bead assigned a glutamatergic cortical layer annotation, and look at the anatomical layer this bead location corresponds to within the ABA. (**c**) Along the y-axis, we see the number of Slide-seq beads confidently assigned by RCTD to a particular cell type. The cell type assigned can be distinguished by the color of each bar within the graph. The x-axis denotes the cortical region that the bead's location would be assigned to within the ABA.

Figure 3.58: Identified Cortical Layer Distribution in the MOp



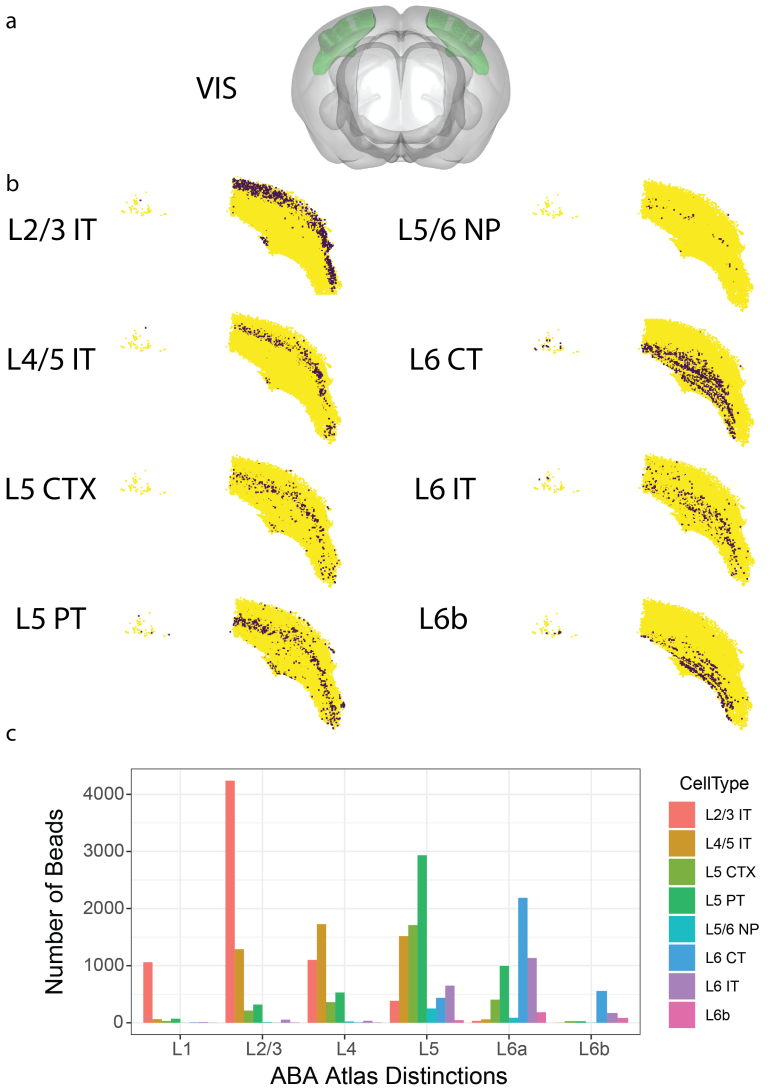
(**a**) The ABA brain is shown with a coronal view in gray, where the primary motor area (MOp) is shown in green. (**b**) For each cortical neuron cluster assignment that resulted from the MOp glutamatergic integration, we color each bead confidently assigned that cell type within the MOp to be purple, and each cell type not confidently assigned that cell type yellow. (**c**) We take each bead assigned a glutamatergic cortical layer annotation, and look at the anatomical layer this bead location corresponds to within the ABA. Along the y-axis , we see the number of Slide-seq beads confidently assigned by RCTD to a particular cell type. The cell type assigned can be distinguished by the color of each bar within the graph. The x-axis denotes the cortical region that the bead's location would be assigned to within the ABA.

Figure 3.59: Identified Cortical Layer Distribution in the MOs



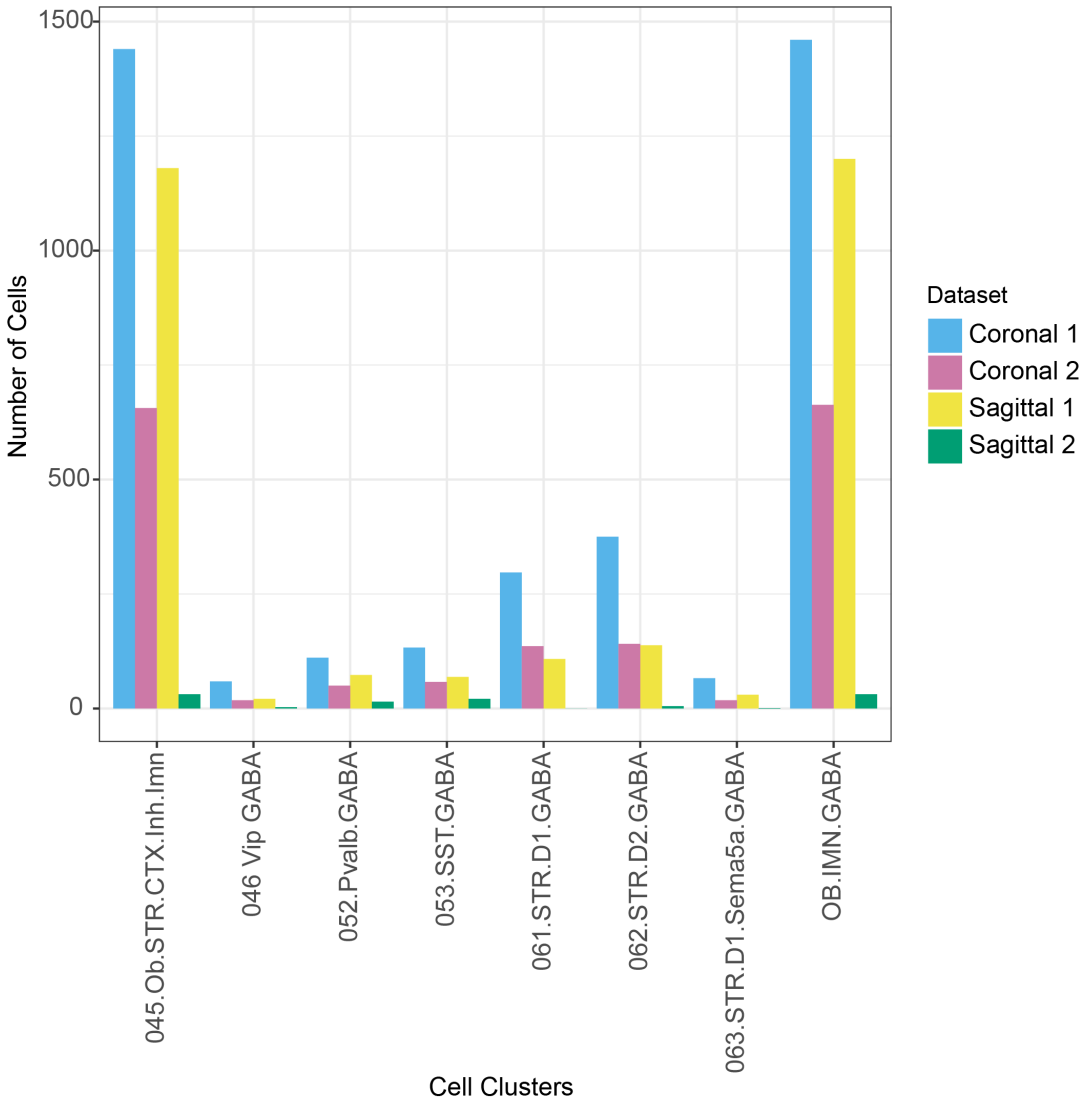
(**a**) The ABA brain is shown with a coronal view in gray, where the secondary motor area (MOs) is shown in green. (**b**) For each cortical neuron cluster assignment that resulted from the MOs glutamatergic integration, we color each bead confidently assigned that cell type within the MOs to be purple, and each cell type not confidently assigned that cell type yellow. We take each bead assigned a glutamatergic cortical layer annotation, and look at the anatomical layer this bead location corresponds to within the ABA. (**c**)Along the y-axis, we see the number of Slide-seq beads confidently assigned by RCTD to a particular cell type. The cell type assigned can be distinguished by the color of each bar within the graph. The x-axis denotes the cortical region that the bead's location would be assigned to within the ABA.

Figure 3.60: Identified Cortical Layer Distribution in the VIS



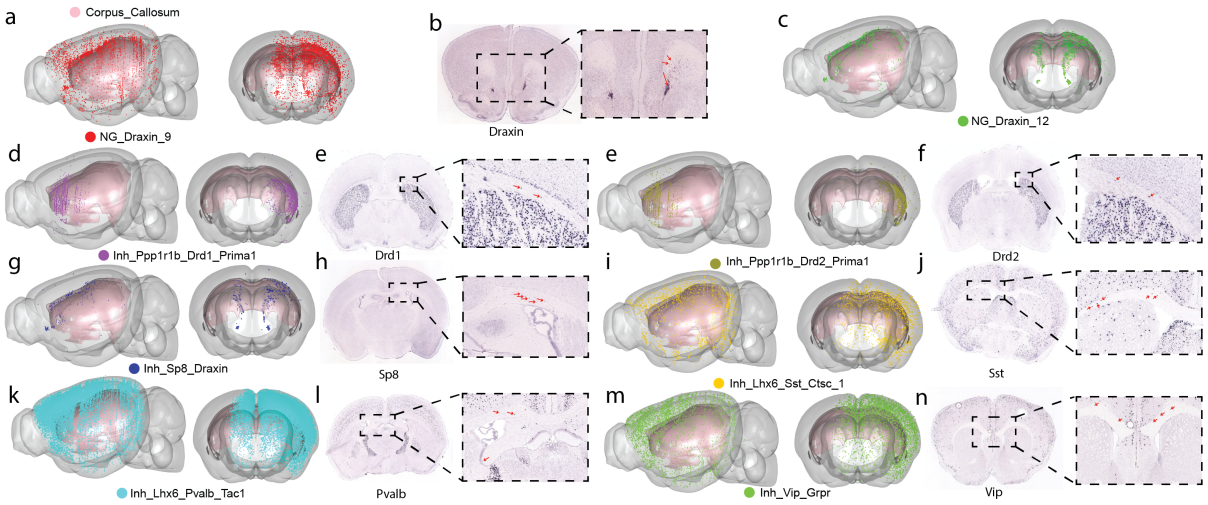
(**a**) The ABA brain is shown with a coronal view in gray, where the visual area (VIS) is shown in green.(**b**) For each cortical neuron cluster assignment that resulted from the VIS glutamatergic integration, we color each bead confidently assigned that cell type within the VIS to be purple, and each cell type not confidently assigned that cell type yellow. We take each bead assigned a glutamatergic cortical layer annotation, and look at the anatomical layer this bead location corresponds to within the ABA. (**c**) Along the y-axis , we see the number of Slide-seq beads confidently assigned by RCTD to a particular cell type. The cell type assigned can be distinguished by the color of each bar within the graph. The x-axis denotes the cortical region that the bead's location would be assigned to within the ABA.

Figure 3.61: Interneuron and Neurogenesis Cell Types of the Corpus Callosum in the MERFISH Dataset



We display the number of cells present within the corpus callosum for each cell type within each of the four distinct cross sectional datasets from the MERFISH ABA portal. The Y-axis details the number of cells, while the bars are grouped across the X-axis by their molecular cell types. The bars are colored by their anatomical dataset distinction.

Figure 3.62: Interneuron and Neurogenesis Cell Types of the Corpus Callosum in the ABA ISH Data



(**a**) Each cell type is plotted by its centroid within the gray outline of the brain, where the size of the dot reflects the number of voxels that cell type is found within, and the color of the dot corresponds to the primary cell type assignment for that cell. The outline of the brain is shown in gray, with the corpus callosum shown in pink. (**a,c,d,e,g,i,k,m**) Each Slide-seq bead confidently assigned the selected cell type is plotted within each brain, the color of the dots indicating the cell type. (**b,e,f,h,j,l,n**) Nissl stains from the Allen Brain Institute for specific genes are shown, with red arrows drawing attention to expression that seems to be present within the fiber tracts.

CHAPTER 4

Leveraging Spatial Coordinates to Correlate Anatomical and Physiological Features with Molecular Cell Type Profiles

4.1 Preamble

The following work has been submitted to a peer-reviewed journal for publication in conjunction with the work described in Chapter 3. For the work described in the following chapter, Joshua D. Welch conceived analysis ideas. April R. Kriebel performed the neuronal projection and c-Fos analyses. Jingyi Liu and April R. Kriebel performed vascular analyses. Andrew Robbins built the data portal. Joshua D. Welch and April R. Kriebel designed the figures and drafted the manuscript.

4.2 Introduction

In addition to solidifying molecular cell type profiles and spatial distributions, the roadmap provided by the BRAIN 2025 report [2] details two additional major milestones pivotal to the advancement of neuroscience; (1) which neurons are synaptically connected to each other; and (2) how patterns of neuron activity vary over time.

The field has advanced significantly in its ability to measure neuron morphology and connectivity, driven by advances in tissue clearing, viral labeling, imaging, and computational segmentation [133, 107, 134, 214]. Remarkable progress has also been made in measuring patterns of neuron activity over time: live-cell activity recording with calcium imaging [68], Neuropixel probes [83, 167], and optogenetic perturbation [217, 175, 180] allow detailed studies of neuron activity dynamics and stimulus-dependent activity.

Nevertheless, efforts to map molecular cell types and locate their spatial positions have proceeded largely independently from efforts to probe their synaptic connections and dy-

namic activity patterns. Molecular and anatomical experiments measure such different neuronal features that the insights from analyzing them have remained largely disjoint. The immense challenges of measuring morphology, connectivity, and electrophysiology make it difficult to identify the molecular features of the cells profiled. Additionally, the different types of measurement operate on different spatial scales, ranging from subcellular to regional resolution.

However, spatial location within the brain is a common feature measured by both spatial transcriptomic and imaging experiments, providing a potential bridge between molecular and anatomical data types. Here, we show that using spatial transcriptomic data to estimate a spatial distribution for each cell type makes it possible to relate molecular cell types to anatomical data types measured in the same common coordinate framework (CCF) [194]. We integrate our spatially resolved cell types (details can be found in Chapter III) with brain-wide vascular density [182], mesoscale connectome data [133], and stimulus-dependent neuron activity [31] to infer properties of molecular cell types at mesoscale spatial resolution.

4.3 Results

4.3.1 Joint Registration of Whole Brain Vasculature and Cell Type Location Reveals Differences in Local Vascular Density of Neuron Types

Blood flow is a critical component of homeostasis, but it is of particular import within the brain. Regulation of cerebral blood flow is necessary to monitor the exchange of metabolites and cells through the blood brain barrier [52], and to ensure that the metabolic needs of the tissue are met. Endothelial cells that line the walls of the blood vessels in the central nervous system modulate the interchange between the brain and the blood, and these endothelial cells are influenced by vascular, immune, and neural cells [52]. This regulation is necessary, in part, because the vascular demands within the brain are not stagnant, as increased neuron activity results in increased blood flow, a relationship referred to as neurovascular coupling [65]. In order to achieve the correct modulation of cerebral blood flow, the cells within the neurovascular unit have to work in harmony. The neurons and glia give the signals for vasodilation and the vascular cells carry out the desired operation [65, 22]. Thus, cerebral blood flow is intimately linked to neuronal activity. Yet, our understanding of what neuronal cell types are most likely to occupy regions of high vascular density is limited. Generally, broad cell types, such as astrocytes, have been associated with the regulation of blood flow [75], but knowledge about more refined molecular cell types and their relationships to

vascular density have not been thoroughly investigated brain-wide.

We reasoned that a brain-wide map of molecular cell type distributions provides an unprecedented opportunity to investigate differences in the local vascular density of neurons. Although brain vasculature at the capillary scale varies significantly among individuals, the position of large vessels and overall vascular density should show less inter-individual variation. By co-registering the Slide-seq molecular profiles and vascular density information within the same anatomical coordinate system, we enable the assessment of the correlation of well-defined, molecular cell type profiles with vascular density information. The identification of neuron types that are surrounded by high vascular density, and those conversely surrounded by low vascular density, may lead to insights about which neuron types need larger amounts of blood flow, have the capacity to exert a greater influence on the blood flow volumes within the brain, or impact the formation of vascular networks to the greatest extent.

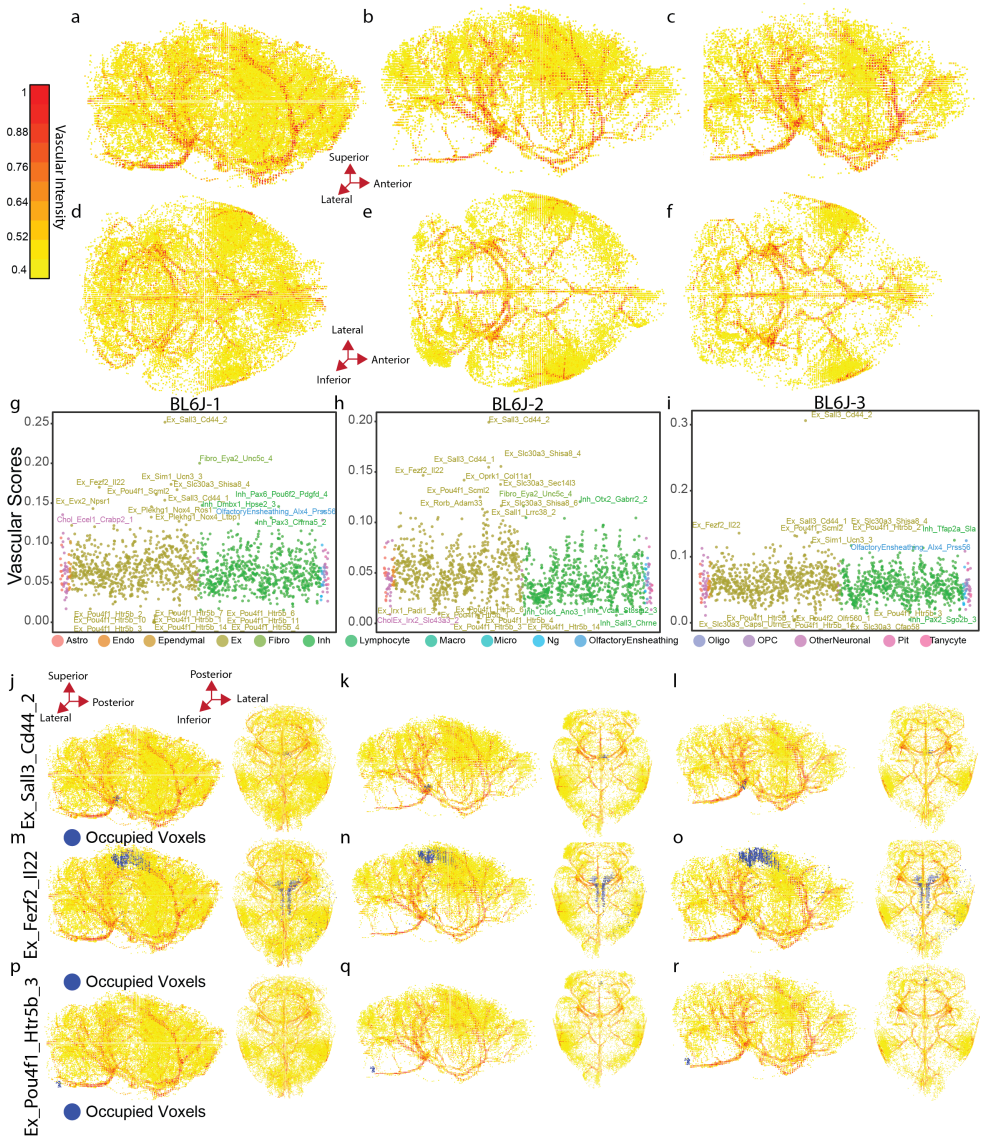
To investigate the vascular density of molecular cell types across the mouse brain, we used data from Todorov et al., who reconstructed the whole brain vasculature of three distinct BL6J mice [182]. We first summarized the vascular density within 100 μm voxels, then co-registered the vascular data [182] and Slide-seq data [94] within the Allen CCF (Fig. 4.1 a-f). While variability in vascular density among the three mice can be observed, major vascular structures and mesoscale vascular density patterns are quite consistent, as reported by Todorov et al. To quantify the local vascular density of each cell type, we calculated a vascular score by summing the local vascular density of each voxel that contains at least one confidently mapped cell, weighted by the number of cells of that type in the voxel (see methods) for each of the three samples (Fig. 4.1g-i). The vascular scores reflect the association between a specific molecular cell type and vascular density, with a higher score indicating a higher correspondence with high vascular density, and a lower score reflecting a low correspondence with vascular density.

Despite the between-sample variance present within the vascular density data, most clearly emphasized by the different highest magnitude of each sample's vascular scores (Fig. 4.1 g-i, see Y-axis), the themes and key identified cell types remain largely consistent across replicates. To ensure that our results were not confounded by each cell type's prevalence across the brain, we compared the vascular score and number of voxels each neuronal cell type was present in, confirming that vascular score was not a function of cell type prevalence (Supplementary Fig. 4.5). Of the broad cell type categories, such as excitatory neurons, inhibitory neurons, and non-neuronal cell types, indicated by the color of the dots, no one category is substantially higher or lower in vascular scores. Rather, within each broad category, there is a wide range of fluctuation between the vascular scores of each cell type. The

indication that there is no single broad cell type with consistently higher, or lower, vascular scores than all other cell types fits with the current understanding of the neurovascular unit encompassing both neuronal and non-neuronal cells [65].

While no single broad neuronal category shows evidence of elevated or depressed vascular density across all of its cell types, some individual cell types consistently demonstrate particularly high or low vascular scores. For example, Ex_Sall3_Cd44_2 has the highest vascular score across all three brains (Fig. 4.1g). This cell type occurs in a small spatial volume that appears to cluster tightly around the basilar artery [200]. Despite the inter-individual variance that is known to exist in the vascular structures within the brain [90], the cell type

Figure 4.1: Joint Registration of Spatially Resolved Molecular Profiles with Whole-Brain Vasculature Reveals Differences in Local Vascular Density of Neuron Types



(**a-f**) Vascular density across three mouse brains from a sagittal plane view and a horizontal plane view. Each dot represents a 100 μm voxel and is colored by the average vascular density of that voxel (yellow is low, red is high). Voxels with vascular density lower than that of 0.4 are not shown for clarity, thus creating some areas of white space where vascular density is particularly low.(**g-i**) Scatter plot of vascular scores per cell type. Each dot is a cell type, the y-axis is the vascular score, and the x-axis organizes the cell types by broad cell class. The dots are colored by broad cell type classifications, and cell type names of the cell types with high and low vascular scores are shown.

Figure 4.1 *(previous page)*: (**j-r**) Spatial plot of vascular density and cell type location for the Ex_Sall3_Cd44_2 (**j-l**), Ex_Fexf2_Il22 (**m-o**), and Ex_Pou4f1_Htr5b_3 (**p-r**) cell types. Blue dots are individual Slide-seq beads confidently assigned to each cell type, and these beads are shown registered within the vascular density of the three sample brains, again using sagittal and horizontal plane views.

identified as having the highest association with increased vascular density clusters in tight spatial proximity to the same major artery in all three samples (Fig. 4.1j-l). The Ex_Sall3_Cd44_2 cell type exists primarily within the medial mammillary nucleus (MM) of the hypothalamus, consistent with a previous report by Wu et. al. that the MM is an area of elevated vascular density [199].

Another cell type with a high vascular score across all three brains is Ex_Fezf2_Il22, a cortical excitatory neuron type with a wider radius of spatial distribution (Fig 4.2m-o) than Ex_Sal3_Cd44_2. While this cell type is not clustered specifically around a singular artery, it occupies a region of the brain encompassing the retrosplenial artery and posterior internal frontal artery [200], critical contributors to cerebral blood flow in the most superior portions of the brain. This cell type is primarily located within the RSPv6, a region previously noted to have high vascular density in comparison with other cortical areas [199].

In contrast, Ex_Pou4f1_Htr5b_3 has low vascular scores across all three samples. Spatially, Ex_Pou4f1_Htr5b_3 occupies the region of low vascular density superior to the vertebral artery (Fig. 4.1p-r). Along the ventral edge of the hindbrain runs the vertebral artery, which eventually becomes the basilar artery. At the posterior end of the hindbrain, the paraolivary artery branches off of the vertebral artery and runs in the superior-anterior direction from its branch point [200]. As a result, the posterior-dorsal portion of the hindbrain is not transected by any major vasculature, which can be seen visually in the lack of high density voxels above the vertebral artery. Ex_Pou4f1_Htr5b_3 is a cell type spatially constrained to the inferior olivary complex, which has a strong anterior/posterior gradient to its vascular density [199]. While the more anterior regions of the inferior olivary complex have increased vascular density, the posterior sections have a reduced vascular density. As would be expected for a cell type with a low vascular score, Ex_Pou4f1_Htr5b_3 is found in the posterior portions of the inferior olivary complex. Ex_Pou4f1_Htr5b_3 has one of the lowest vascular scores of all molecular cell types across all three brains, and the lack of vascular transection within its occupied voxels confirms that cell types with lowest vascular scores most likely occupy an area characterized by few, if any, major vascular features, a finding confirmed by previous vascular density maps [199]. To ensure that the association between vascular density and molecular cell types can be further explored by the scientific community at large, we develop

the Welch Lab Neuroglancer portal online Neuroglancer portal, in which we upload the spatially resolved molecular cell type distributions in conjunction with the vascular density data.

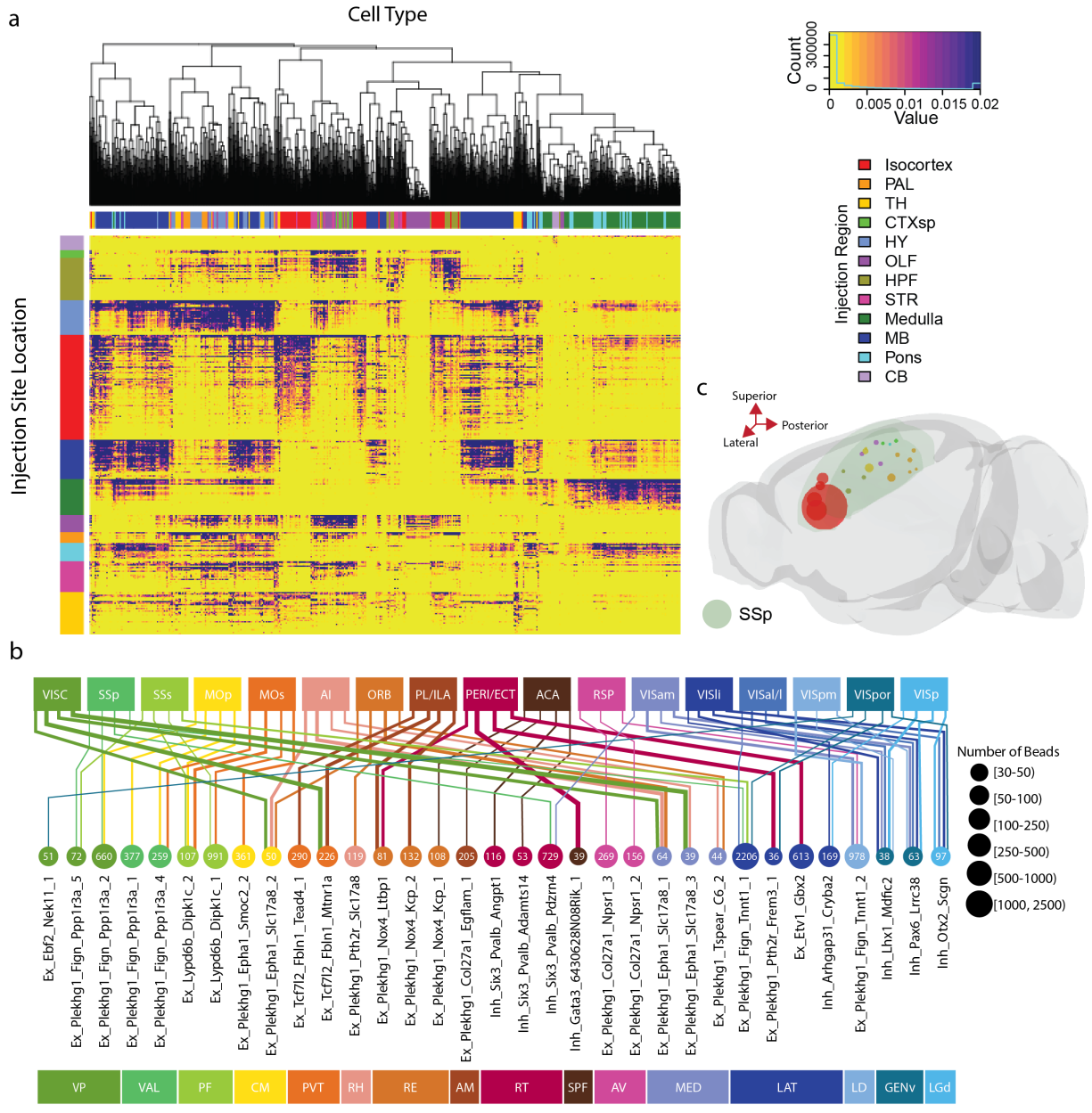
The co-registration of the vascular density data in conjunction with molecular cell type profiles also allows us to examine the correlation between individual neuronal activity and vascular density. We calculate the neural activity of each neuronal cell type in two distinct ways: by the log (counts per million) of genes that are annotated as immediate early genes and by using NeuroEstimator [24] (See Methods). We observe no correlation between vascular density and either measure of neural activity (Fig. 4.6). The lack of correlation between neural activity and vascular density most likely reflects the sampling differences between the two data types, as the vascular score for each voxel is fixed, and we expect neural activity at the cell type level to be dynamic in nature.

4.3.2 Joint Registration with Mesoscale Connectivity Data Reveals Regional Connectivity Differences Among Neuron Types

Determining how molecularly defined neuron types are connected in neural circuits remains a key challenge in neuroscience. Leveraging recombinant adeno-associated virus (AAV) expressing enhanced green fluorescent protein combined with two-photon microscopy, Oh et. al. generated an atlas of neuron connectivity across the whole mouse brain [133]. This atlas provides connectivity at the mesoscale, capturing both long-range and local connections by imaging neurons across the brain. While this resource provides incredible insights about connectivity, the only link to molecular cell types comes from Cre lines, which do not provide the same level of detailed molecular profile as single-cell sequencing experiments. Some emerging molecular assays such as MAPseq [85], BARseq [47], and epi-retro-seq [216], combine projection mapping with single-cell sequencing, but the available datasets are not at the same scale as the Allen Connectivity Atlas.

We reasoned that integrating the spatial distributions of molecular cell types and mesoscale connectivity data could enable the investigation of the regional projection patterns of molecularly defined neuron types. To this end, we co-registered the Allen Connectivity Atlas with the spatially resolved transcriptomic profiles captured in the whole mouse brain Slide-seq experiments. We then used the 498 experiments from wild-type mice (no Cre lines) to calculate a projection score for each injection site and each molecular cell type identified within the Slide-seq data. The projection score quantifies how much projection energy (quantified by the Allen Connectivity Atlas) for each distinct injection site occurs within

Figure 4.2: Co-Registration of Connectivity and Cell Types Reveals Cell-Type-Specific Connectivity



(**a**) The heatmap shows the projection score for each projection (y-axis) and for each cell type (x-axis), where the y-axis is organized by each projection's injection site region, and the x-axis is allowed to cluster by profile similarity, as indicated by the dendrogram. The colored bar next to the y-axis indicates the region of each projection's injection site location, and the colored bar above the x-axis indicates the region in which the majority of the cell type beads are located. (**b**) Within the wiring diagram, the square boxes across the top indicate a cortical injection region, where each region is indicated by a different color. The circles below each represent a molecular cell type, the name of which is given vertically below the dot.

Figure 4.2 *(previous page)*: The size of these circles, and the number inside the circle, indicates the number of voxels within the thalamus that have been confidently assigned this cell type. Each cell type is assigned a thalamic region on the basis of where the highest number of beads confidently assigned that cell type are located, and the molecular cell types are grouped and colored by this assignment. The lines between the cortical region boxes and the circular molecular cell types indicate the average highest scoring molecular cell types associated with the projections from each cortical injection region, where the thickness of the line indicates the magnitude of the average projection score. (**c**) The outline of the brain is shown in gray, and the primary somatosensory area (SSp) is shown in green. Each of the colored circles represents a projection that has an injection site location within the SSp and that projects into the thalamus. The dots are placed according to their injection site location, and colored by the subregion of the SSp in which the injection site lies. The radius of these dots indicates the projection score between the projection energy of an injection and the molecular cell profile Ex_Lypd6b_Dipk1c_2.

100 μm voxels containing each molecular cell type (Methods).

We summarized the projection scores for all injection sites and cell types across the whole brain using a heatmap, where rows are injection sites and columns are cell types (Fig. 4.2a). We additionally colored the columns by the brain region in which each cell type is predominantly located. Our analysis reflects many of the regional connectivity relationships noted by Oh et al., but with the addition of molecular cell types defined by single-cell and spatial transcriptomics [133]. Each injection occupies a substantial number of voxels within its injection site region, and this is reflected by each region having a strong association with cell types that lie within its boundaries. For example, the cerebellum has little projection energy outside of the bounds of the cerebellum, and correspondingly the molecular cell types associated with cerebellum injection sites are primarily localized within the cerebellum. Other, more complex, regional connections have also been maintained. The cell types of the pons and medulla receive few projections from the hypothalamus and isocortex, more from the midbrain, and the strongest signal is generated between the pons and medulla themselves, a pattern that is recapitulated in the magnitude of the projection scores of the molecular cell types. Similarly, the hypothalamus projections are highly associated with cell types from the hypothalamus, striatum, pallidum, thalamus, and midbrain, regions that are known recipients of hypothalamic projections. The correspondence between expected projection energies and the primary location of high scoring molecular cell types suggests that the projection score is correctly implicating molecular cell types associated with specific injections. Further explorations of the association between neural projection patterns and molecular cell types is facilitated by the Welch Lab Neuroglancer portal, where we provide the spatially resolved molecular cell type distributions in conjunction with the neuronal

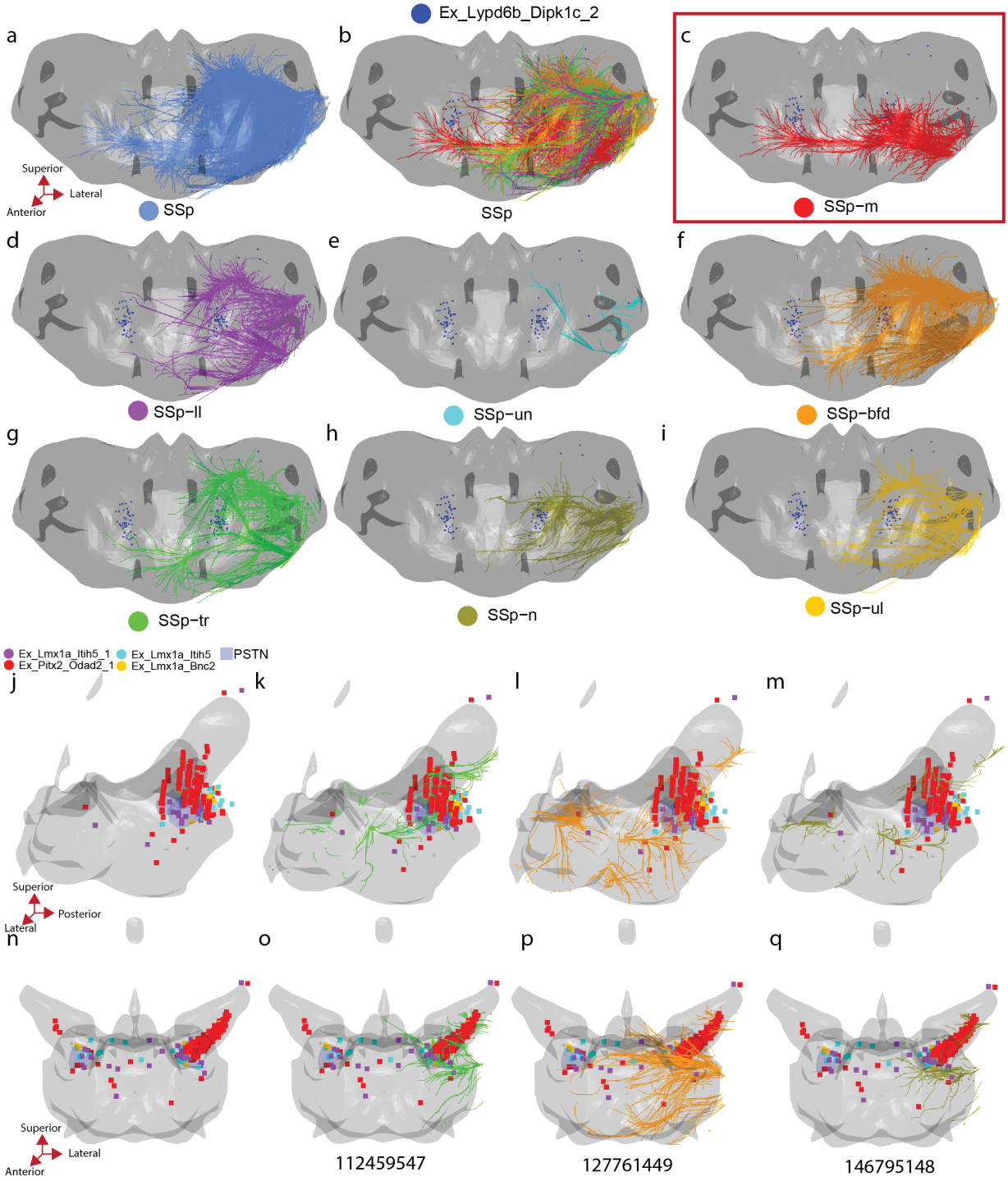
projection data.

We further used our projection scores to create a wiring diagram of cell types receiving cortico-thalamic projections (Fig. 4.2b). Previously, Oh et al. created a wiring diagram reflecting the connections between regions of the cortex and the thalamus [133], but our cell-type-specific projection scores allow us to further elaborate this diagram with the addition of molecularly defined cell types occurring in each region of the thalamus. The wiring diagram derived from our projection scores, with no manual curation, is highly consistent with the regional connectivity patterns reported by Oh et al. Previous studies have noted the large prevalence of glutamatergic neurons, and the scarcity of GABAergic neurons within the thalamus[216]. Consistent with these reports, we note that inhibitory cell types rank within the top projection scores only for thalamic nuclei that have been identified as containing GABAergic neuron populations [94]. Our wiring diagram highlights interesting connections between the visual (VIS) and restrosplenial (RSP) areas and specific thalamic cell types. The cell types with highest projection scores for injection sites in the visual areas are Ex_Plekhg1_Fign_Tnnt1_2, Inh_Lhx1_Mdfic2, Inh_Pax6_Lrrc38, and Inh_Otx2_Scgn. These cell types are located in the LD, LGd, and Genv, a region which contains the LGv [133]. The lateral geniculate nucleus is an established component of the visual pathways in the brain, both sending and receiving signals from the visual areas [198]. Similarly, a recent morphological reconstruction of individual IT neurons confirmed that non-IT VISpm neurons project to the LD [166]. We also observe the RSP primarily connected with cell types Ex_Plekhg1_Col27a1_Npsr1_2 and Ex_Plekhg1_Col27a1_Npsr1_3 in the AV, consistent with known reciprocal connections between the RSP and AV [108, 160].

We observed particularly region-specific cortico-thalamic projections from the somatosensory cortex, in which small differences in injection site origin result in noticeable differences in projection scores among thalamic cell types (Fig. 4.2c). The primary somatosensory area (SSp) is comprised of several smaller regions, including the nose (SSp-n), barrel field (SSp-bfd), lower limb (SSp-ll), mouth (SSp-m), upper limb (SSp-ul), trunk (SSp-tr) and unassigned (SSp-un). Each injection site within the SSp has a distinct projection score within the thalamus for the molecular cell type Ex_Lypd6b_Dipk1c_2, as indicated by the magnitude of the injection site dots in Fig 4.2c. The higher projection scores of projections with injection sites within the SSp-m implicate Ex_Lypd6b_Dipk1c_2 as a potential target for the SSp-m projections.

To further investigate this hypothesis, we examined the spatial organization of the projections with injection sites within the SSp as they project into the thalamus (Fig. 4.3a). We colored each projection by its injection site region and graphed the Slide-seq beads which contain the cell type Ex_Lypd6b_Dipk1c_2, examining all the injection regions cumulatively

Figure 4.3: Neuronal Projections Identify Probable Molecular Cell Type Associations within Distinct Brain Regions



(**a**) The thalamus is shown in gray, and the physical projection patterns of projections that possess an injection site located within the SSp are shown in blue. (**b**) Each of the SSp originating projections is colored by their subregion of origin . The navy blue dots indicate Slide-seq beads which have confidently been assigned the cell type Ex_Lypd6b_Dipk1c_2.

Figure 4.3 *(previous page)*: (**c-i**) The SSp injection regions are each graphed separately by their subregion of origin, where the gray shape is the thalamus, the blue dots are Slide-seq beads that captured the Ex_Lypd6b_Dipk1c_2 molecular cell type, and the color of the neuronal projections indicates their subregion of origin. (**j-q**)The hypothalamus is shown in gray, and the blue shaded area indicates the parasubthalamic nucleus (PSTN) . The colored dots indicate Slide-seq beads that have been confidently assigned a specific molecular profile. Three neuronal projections with injection site origins within the central amygdalar nucleus (CEA) are shown, both sagittal and coronal planes are shown for each projection, and each projection is traced by a unique color.

(Fig. 4.3b) and separately (Fig. 4.3c-i). Visual inspection confirms that projections with injection sites in the SSp-m have a much greater spatial proximity to the beads containing the Ex_Lypd6b_Dipk1c_2 cell type.

The projections from the SSp mouth (SSp-m) region are more highly associated with the Ex_Lypd6b_Dipk1c_2 molecular cell type than projections from other sub-regions within the SSp. Beads containing a confident assignment of the Ex_Lypd6b_Dipk1c_2 cell type are primarily spatially constrained within two regions of the thalamus: the parafascicular thalamic nucleus and the ventral posteromedial nucleus (VPM). The VPM is where the three main tracts of the trigeminal system (spinal trigeminal, ventral trigeminothalamic, and dorsal trigeminothalamic tracts) synapse with third-order neurons and then send that information to the primary somatosensory cortex [140]. The trigeminal nerve transmits sensory input from the face such as pain, temperature, touch, and proprioception. It also captures mouthfeel, the physical or textural sensations in the mouse caused by foods and liquids [161]. Therefore, it is plausible that the molecular cell type Ex_Lypd6b_Dipk1c_2 may be a type of third-order neuron within the thalamus, relaying sensory information from the mouth to the portion of the somatosensory areas that processes mouth stimuli.

We also identified several interesting cell-type-specific projection patterns not involving the isocortex. We identified three WT projections from the Allen Brain Atlas that had injection sites within the central amygdalar nucleus (CEA) that projected substantially into the hypothalamus. We then used projection scores to find the 10 highest scoring molecular cell types for these three injection sites. Five of these cell types have projections from all CEA injections, and four (Ex_Pitx2_Odad2_1, Ex_Lmx1a_Itih5_1, Ex_Lmx1a_Itih5, Ex_Lmx1a_Bnc2) occur within the parasubthalamic nuclei (PSTN) (Fig. 4.3j-q). The CEA regulates food intake, causing food inhibitions learned from aversion, and also affects refeeding behavior after a fast [218]. Zséli et al. found the CEA to be highly interconnected with refeeding-activated brain areas, including the PSTN [218]. Kim et. al identified a subpopulation of the PSTN that expresses *Tac1*. When inhibiting the activity of these PSTN

Tac1 cells, they saw an increase in meal frequency, while increasing expression significantly reduced mouse feeding patterns, suggesting neuronal types found within the PSTN are capable of demonstrating anorexigenic effects on feeding patterns [87]. In summary, the CEA's intimate connection with the PSTN confirms that the projection score identifies molecular cell types with plausible connections between these regions.

4.3.3 Joint Registration with c-Fos Imaging Implicates Neuronal Subtypes in Stress-Induced Neuron Activity

Characterizing the dynamic activity patterns of specific molecular cell types over time and in response to stimuli is a crucial step toward understanding the cellular basis of cognition and behavior. However, determining the molecular types of neurons observed in temporal activity measurements has remained challenging because high-throughput molecular measurement requires invasive procedures such as fixation, tissue dissociation, or tissue sectioning.

We reasoned that the spatial distributions of molecularly defined cells could be used to address this limitation, implicating specific neuron types in dynamic neuron activity patterns based on their location within the same CCF. To demonstrate this, we used a recently published dataset of time-course neuron activity imaging. Bonapersona et al. assessed brain activity in response to stress–a one-time foot shock experiment [31]. State-of-the-art imaging and tissue clearing approaches enabled whole-brain quantification of c-Fos protein levels within individual neurons at multiple time points after foot shock. While such experiments reliably indicate which neurons were active at the time of imaging, c-Fos imaging does not simultaneously capture detailed gene expression information to identify the molecular subtypes of the active neurons. Thus, this sort of experiment can identify brain regions with increased neuron activity, but cannot reveal what subtypes of neurons are activating.

We co-registered the c-Fos data with the spatial distributions of cell types from Slide-Seq (Fig. 5.4a), summed the number of c-Fos expressing cells within each 100 μm voxel, and adjusted for baseline c-Fos expression. (Bonapersona et al. established baseline neuron activity levels at time 0 by placing the animals within the foot-shock chamber but not shocking the animals.) We then calculated an activity score for each of our molecular cell types, similar to the vascular density and projection scores used in the previous sections. For each cell type, we summed the number of c-Fos+ cells within all voxels containing that cell type, weighted by the number of Slide-seq beads containing the cell type in that voxel. Because Bonapersona measured c-Fos expression at multiple time points (0, 30, 90, and 180 minutes after foot shock), we used the change in number of c-Fos+ cells compared to T0 in later time points. Thus, we obtained an activity score for each cell type at each time

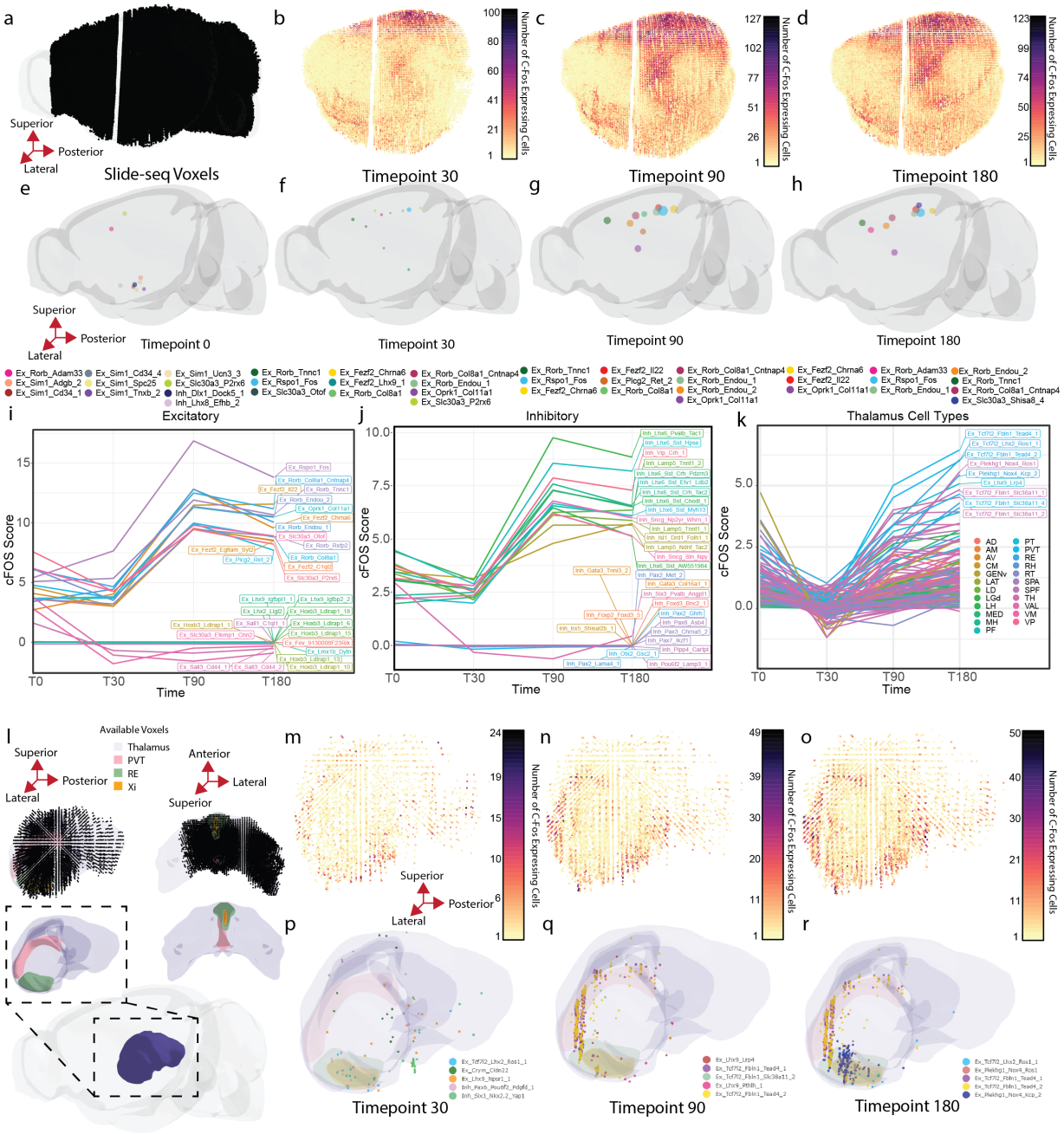
point, with scores at T30, T90, and T180 indicating the change in activity after foot shock and scores at T0 indicating baseline activity. We present the intensity of neuronal activity as the change in c-Fos expressing cells within each voxel, removing all voxels with a change equal to 0 for clearer visualization (Fig. 5.4b-d). There is a clear shift in c-Fos expression across time, with the greatest number of c-Fos expressing cells within a single voxel being measured at 90 minutes after footshock (n = 127). However, it is not only the number of c-Fos expressing cells that shifts along the temporal axis, as the breadth of regions with a high count of c-Fos expressing cells is wider in the later time points than at 30 minutes after foot shock.

To examine the spatial distribution of the molecular cell types with the highest c-Fos activity scores for each time point, we ranked the neuronal types that appear in more than 20 voxels brain-wide by their c-Fos activity scores. Graphing the ten highest scoring cell types from each time point by their centroids, with the size of the dot corresponding to the magnitude of their c-Fos score shows shifting spatial distributions of c-Fos activity over time (Fig. 5.4e-h). Prior to foot-shock stress, the cells most associated with c-Fos expression lie in the hypothalamus. Yet, upon foot-shock induction, we see a dramatic shift of c-Fos activity to the cortex, with especially prominent activity in the motor, somatosensory, and visual areas, implicating neurons activated by the experimental condition.

We identified cell types whose activity score was elevated at a single experimental time point, as well as those that demonstrated more complex temporal trends. Ex_Slc30a3_Otof, for example, is within the top ten expressing cell types at time point 30, but is not as highly elevated, in comparison to other cell types, at T90 or T180. Similarly, Ex_Rorb_Col8a1 is one of the most highly associated molecular cell types in T30 and T90, but not in T180 (Fig. 5.4e-h). This hints at sequential patterns of neuron activation, in which different molecular cell types are activated over time. For example, the sensation of pain would need to be registered, processed, and then potentially encoded into memory.

We next looked at which cell types had the highest and lowest cFos scores across all brain regions and across post-foot-shock time points. To do this, we averaged the scores for each cell type across T30, T90, and T180 and selected the 15 highest and 15 lowest ranking Glutamatergic and GABAergic neuron types (Fig. 5.4i,j). We confirmed that the c-Fos score is not a function of the number of voxels a cell type occupies by examining the relationship between the c-Fos score and number of voxels occupied by each cell type for each timepoint (Supplementary Fig. 4.7). Glutamatergic and GABAergic neurons with high-ranking c-FOS scores are located almost entirely within the isocortex, particularly within the MO, SS, and VIS. For both inhibitory and excitatory neurons, the cell types with the smallest response to the experimental stimulus, measured using the c-Fos score, are primarily located in the

Figure 4.4: Joint Registration with c-Fos Imaging Data Implicates Neuronal Subtypes in Stress-Induced Neuron Activity



(**a**) The spatial distribution of the Slide-seq dataset, registered into voxels, is shown, where each black dot represents a voxel where at least one Slide-seq bead is present. The outline of the mouse brain is shown in gray.(**b-d**) The change of c-Fos expressing cells from time point 0 to time point 30, 90, and 180 is shown, where each dot is a 100 μm voxel, and the color of the dot reflects the magnitude of change in c-Fos expressing cells within that voxel. c-Fos activity is calculated by counting the number of c-Fos expressing cells within each voxel.(**e-h**) For each time point, the molecular cell types with the ten highest c-Fos activity scores are selected, and graph the centroid of each cell type within the brain, shown in gray.

Figure 4.4 *(previous page)*: The dots represent each centroid, and are colored by molecular cell type. The diameter of the dots indicates the magnitude of the c-Fos score for that cell type. (**i-j**) For both the inhibitory and excitatory molecular cell types, the cell types with the fifteen highest and fifteen lowest average c-Fos scores across the experimental condition (T30, T90, T180) are shown. The X-axis shows the c-Fos scores across timepoints, and Y-axis reflects the magnitude of the c-Fos score. Each line is colored uniquely and represents a distinct molecular cell type. (**k**) All neuronal cell types that have been confidently assigned to more than 10 voxels within the thalamus are shown, plotted as their c-Fos scores across time. Each line represents a molecular cell type, and the lines are colored by the thalamic nucleus that the cell type primarily occupies. Cell types that have a substantial presence in more than one nuclei have been labeled with a spatial location of thalamus, or TH. (**l**) The distribution of the Slide-seq data within the thalamus, shown in violet, is shown as black dots, where each dot is a voxel occupied by one or more Slide-seq beads . Different internal structures of the thalamus, including the paraventricular nucleus of the thalamus (PVT), nucleus of reuniens (RE), and xiphoid thalamic nucleus (Xi) as pink, green, and yellow, respectively, and the thalamus, shown in dark violet, can be visualized within the larger context of the brain, shown in gray.(**m-o**) The change in the number of c-Fos expressing cells in each voxel within the thalamus is shown for time point 30, 90, and 180 minutes after footshock. Each dot represents a voxel, and the color reflects the magnitude of change in c-Fos expressing cells from time point 0 to the time point shown.(**p-r**)For each time point, the five molecular cell types with the highest c-Fos activity score within the thalamus are selected. Each bead confidently assigned one of those molecular cell types within the thalamus is graphed, where the color of the bead reflects the molecular cell type assigned to that bead.

hindbrain, although molecular cell types from the midbrain, interbrain and olfactory bulb are also identified as lowly associated with the neural activity in response to foot-shock stimuli.

The finding that molecular cell types Ex_Sall3_Cd44_2 and Ex_Sall3_Cd44_1 within the MM of the hypothalamus had low c-Fos activation scores was somewhat surprising, as Bonapersona et. al noted that the mammillary bodies of the hypothalamus (MBO), a brain region predominantly associated with recollective memory [136], showed c-Fos response after foot shock [31]. However, we think that our findings are consistent with theirs, and indicate fine-grained regional differences in foot-shock response within the highly localized and specialized nuclei of the MBO. The MBO encompasses the lateral mammillary nucleus (LM), the medial mammillary nucleus (MM), supramammillary nucleus (SUM), and the tuberomammillary nucleus (TM). Our ranking of high and low scoring c-FOS scores identified two distinct MM molecular cell types with little to no increased c-Fos expression in response to foot-shock: Ex_Sall3_Cd44_2 and Ex_Sall3_Cd44_1. In other regions of the MBO, such as the LM and SUM, we identify several molecular cell types have positive average c-Fos scores across the experimental timepoints. Although understanding of the MBO and its nuclei is

still incomplete, recent findings indicate that the MM and LM operate in distinct circuits, with the MM functioning in spatial memory [189]. Additionally, projections from the MM reach to the ventral tegmental area, which is characterized by a large population of dopamine neurons and is consequently associated with reward behaviors [136]. Lesions within the TM, however, were found to enhance long term memory during fear conditioning, implying that the TM is involved with memory encoding during a fear-based exercise [155]. In short, this analysis underscores the fine-grained regional and molecular differences in c-Fos activity that we can identify. We upload the co-registered c-Fos response data and molecular cell type distributions in our online Welch Lab Neuroglancer portal to encourage further hypothesis generation and dataset exploration.

After examining the neuronal types with highest average activity across all time points, we looked at the temporal dynamics of cFos scores. Broadly, we observed a clear shift in the neuronal activity after footshock (T30, T90, T180) in comparison to baseline T0 c-Fos scores. At T30, we see a lower c-Fos activity score than at T90 and T180, observing the largest c-Fos scores at T90. The Fos protein has previously been found to be at maximal levels 1-2 hrs after administration of c-Fos mRNA via hypertonic saline, and disappears 4-8 hours after administration [156]. This previous finding is in consensus with our observed peak change in c-Fos being observed at T90. The cell type Ex_Rspo1_Fos, an excitatory neuron cluster differentiated from its surrounding clusters by its increased expression of Fos, demonstrates the greatest change from baseline at all experimental time points. The calculated c-Fos activity score correctly identifies molecular cell types associated with the expression of c-Fos, and differentially scores cell types after the onset of the experimental condition.

Having examined the c-Fos response at the whole-brain level, we narrowed our focus to a single region, the thalamus (Fig. 4.4l). We chose the thalamus because it contains highly regionally specialized cell types and Bonapersona reported strong thalamic response after foot shock. The thalamus, located in the interbrain, is composed of nuclei that function to relay information between brain and body. Every sensory system, excluding olfaction, uses a thalamic nucleus to receive, process, and distribute signals to the appropriate cortical area [183]. Interestingly, while the cell types of the thalamus and cortical areas show a similarly increasing trend in c-Fos expressing cells from T30 to T90, c-Fos activity scores for thalamic cell types continue to increase at T180 (Fig. 4.4m-o). This is in contrast to the whole-brain trends we observed, where T90 had the highest c-Fos activity. To investigate whether this phenomenon was driven by one or many molecular cell types, we calculated thalamus-specific c-Fos activity scores, re-calculating the c-Fos activity score as described above but using only voxels within the thalamus. We then plotted the thalamus-specific c-Fos scores across time

points (Fig. 4.4k). The heightened c-Fos activity score at T180 is the result of increased activity of many of the molecular cell types present within the thalamus, but is especially apparent in molecular cell types primarily located in the PVT, RE, and PT, as well as a few more widely distributed molecular cell types (TH).

To examine the spatial distribution of the top-ranking cell types for each time point, we ranked the cell types by the magnitude of their thalamic c-Fos activity score and then graphed the Slide-seq beads that contain each high-ranking cell type (Fig. 4.4p-r). We observe a clear shift in the spatial distribution of highly active neurons across time points. In T30, we see activation scattered throughout the thalamic nuclei, with Ex_Tcf7l2_Lhx2_Ros1_1 being the only neuronal type that has a marked presence in the RE. By T90, we see that the most highly active neuron populations, especially Ex_Tcf7l2_Fbln1_Tead4_1 and Ex_Tcf7l2_Fbln1_Tead4_2, occur within the paraventricular nucleus of the thalamus (PVT). This result is consistent with the known role of the PVT as a critical component of the brain anxiety network, regulating fear processing by projections to both the lateral division of the central amygdala (CeL) and the basolateral amygdalar nucleus (BLA) [135, 89]. Specifically, the posterior PVT has neurons that project to both the BLA and CeL, and an increase in the number of c-Fos expressing neurons within the pPVT as a result of fear-conditioning via foot-shock has been recorded [135]. Similarly, Baisley et. al. found Fos expression within the PVT 90 minutes after foot shock [25] .

While the role of the PVT in fear processing is well-known, our analysis implicates specific molecular cell types, especially Ex_Tcf7l2_Fbln1_Tead4_1 and Ex_Tcf7l2_Fbln1_Tead4_2, with increased activity in the PVT, particularly the posterior PVT, as highly active at 90 minutes after experiencing foot-shock. At T180, we see a continuation of neuronal activity within the PVT, driven by Ex_Tcf7l2_Fbln1_Tead4_1 and Ex_Tcf7l2_Fbln1_Tead4_2, but we also see other highly active molecular cell types, such as Ex_Tcf7l2_Lhx2_Ros1_1, Ex_Plekhg1_Nox4_Ros1, and Ex_Plekhg1_Nox4_Kcp_2, shift the concentration of neuronal activity to include the RE and Xi nuclei of the thalamus.

The cell types we implicated using cFos activity scores agree well with existing knowledge of the functions of the RE and Xi. An established link in the chain of episodic memory encoding, the RE relays signals from the medial prefrontal cortex (mPFC) to the hippocampus (HPF) [142]. Reciprocal interactions between the mPFC and the HPF are critical for contextual memory formation, but while the HPF has direct projections to the mPFC, the mPFC first transmits signals to the RE, and the RE forwards these signals to the HPF. Inactivation of the RE prior to foot-shock impairs fear-conditioning [142], establishing the role of the RE in memory formation after trauma (i.e. foot-shock). The Xi, the small nuclei enveloped by the RE, has recently been found to be integral to the response to visual threats

[150]. The increased c-Fos activity score in Ex_Tcf7l2_Lhx2_Ros1_1, Ex_Plekhg1_Nox4_Ros1, and Ex_Plekhg1_Nox4_Kcp_2 at T180 demonstrates the ability of our c-Fos activity score to appropriately prioritize neuron populations in key regions known to be linked to fear-based response to foot-shock, highlighting the utility of co-registration of molecular cell types and physiological response data.

4.4 Discussion

We leveraged a spatial coordinate framework to integrate molecular profiles with their corresponding anatomical and physiological datasets. In doing so, we were able to identify molecular cell types with high and low correlations with vascular density, specific neuronal projections, and c-Fos activity in response to foot shock. Our generated correlations are intended to serve for hypothesis generation and formation, but we demonstrate the biological feasibility of our predictions for each anatomical and physiological dataset analyzed. To increase the accessibility of the integrated datasets, we have made them publicly available in our hosted Welch Lab Neuroglancer Portal. Here, users can select one or multiple anatomical and physiological features in conjunction with molecular cell type spatial distributions.

The methodology of coupling molecular profiles with anatomical and physiological ones via a common coordinate framework is highly extensible, and can be applied to any dataset registered into the CCF. For example, Brunner et al. used functional ultrasound imaging to measure changes in cerebral blood volume in response to neuronal activity, and performed whisker-stimulation experiments, visual-evoked activity experiments, and imaged the brain during an active sensory-motor task [33]. This data has been registered to the CCF, and, if voxelized, would allow for the exploration of molecular cell profiles with the previously mentioned experimental conditions. For datasets not already registered to the CCF, new registration tools, such as DeepMapi [129], offer streamlined registration pipelines. It is our hope that our experimental design will prompt more datasets to be processed into the CCF and analyzed in conjunction with their spatially corresponding molecular profiles. As the number of datasets registered to the CCF grows, and those datasets are made easily publicly accessible, multi-way interaction analyses can be considered. The interaction between multiple anatomical and physiological datasets can be assessed in conjunction with spatially relevant molecular cell types.

In addition to the generation of new hypotheses, experimental work exploring the proposed relationships will need to be conducted to validate findings. However, given the cost and time expenditure associated with wet lab experiments, our methodology can also function as a filter to de-prioritize unlikely connections and relationships for investigation.

Ultimately, we present a common coordinate framework pipeline for establishing connections between high resolution molecular cell profiles and anatomical and physiological data thus creating novel means for hypothesis exploration.

4.5 Methods

4.5.1 Integration of Molecular Profiles and Vascular Data

The original VesSAP publication registered the CCF to the vascular data, rather than vice versa. To obtain vascular density measurements across CCF voxels, we used elastix to fit an inverse transform using B-splines. We used identical parameters to the original registration with the exception of the displacement magnitude penalty, which we set as described in the elastix documentation for inverting transformations. We next downsampled the vascular data [182] to 100 μm resolution, and translated the coordinates so that they aligned with the coordinates used in for the Slide-seq registration. Notably, this step did not entail stretching or distorting the original image; we simply shifted the brains within 3D space to correctly align. Following the original publication's recommendation [94], we consider cell types with a confidence score of greater than 0.3 to be present on a bead. By necessity, we can only calculate vascular scores for voxels where there is at least one confidently assigned cell type present, and where there is a measure of vascular intensity. For each of the three BL6J mouse samples captured by the vascular dataset, we calculate a vascular score such that

$$VascularScore_A = \frac{\sum_{j=1}^{m} \left(I_j \times \frac{N_j^A}{N_j^B} \right)}{\sum_{j=1}^{m} N_j^A}$$

where N is the number of beads, m is the total number of voxels with a vascular intensity score and at least 1 bead with a confidentially assigned cell type, j is a particular voxel, A is the cell type for which we are calculating the vascular score, B is all cell types present within voxel j, I_j is the vascular intensity within voxel j. Consequently, we take the ratio of beads with cell type A in voxel j and weight it by the number of confidentially assigned cell types within j. We then multiply this by the vascular intensity within that voxel (I_j) and compute the sum of all weighted intensity scores across all voxels (m). We then weigh the final vascular score by the number of times a cell type is confidentially assigned.

For generating the 3D vascular brains shown in Figure 4.1, we filter to retain all voxels with a vascular intensity of greater than 0.4. We then bin the remaining intensity scores into 10 bins, and assign the gradient of colors accordingly. When showing the spatial distribution

of cell types Ex_Sall3_Cd44_2, Ex_Fezf2_Il22, and Ex_Pou4f1_Htr5b_3, we color the voxels that have at least one bead that confidently expresses this cell type.

For plotting a global view of vascular scores (Fig. 4.1 g-i), we filter to retain cell types that are present in more than 15 voxels that are shared in common with the vascular data brain-wide. Cells originally annotated as CholEx, CholInh, Chol, NorEx, SerEx, DopEx, ExInh, Nor, or Ser are re-annotated as OtherNeuronal.

4.5.2 Integration of Molecular Profiles and ABA Projection Data

Using the Allen Brain Atlas Projection data [133], we calculated the relationship between projection energy and molecular cell types across the whole mouse brain. The Allen Projection Data had previously been registered at 100 μm resolution, and we translate the coordinates such that the Slide-seq and projection data share the same coordinates within a 3D space. We find the intersection of all voxels with at least one confident cell type loading (score of > 0.3, per original publication), and all voxels for which the Allen Brain Atlas measured projection intensity. To calculate the association between each projection and molecular cell types within the brain, we calculated a projection score such that

$$ProjectionScore_A = \frac{\sum_{j=1}^{m} \left(I_j \times \frac{N_j^A}{N_j^B} \right)}{\sum_{j=1}^{m} N_j^A}$$

Where P is a single injection id, N is the number of beads, m is the number of voxels with a projection intensity score and at least 1 bead with a confidentially assigned cell type, j is a particular voxel, A is the cell type for which we are calculating the projection score, B is all cell types present within voxel j, I_j is the projection energy within voxel j. Consequently, we take the ratio of beads with cell type A in voxel j and weight it by the number of confidentially assigned cell types within j. We then multiply this by the projection energy within that voxel (I_j) and compute the sum of all weighted intensity scores across all voxels (m). We then weigh the final projection score by the number of times a cell type is confidentially assigned.

We present the projection scores for 498 Wild Type, non-CRE line injection experiments from the Allen Brain Atlas. For the heatmap visualization in Fig 4.2a, we filter out all cell types with 0 scores for all injection experiments, all cell types that occupy fewer than 10 voxels throughout the brain, and filter to retain only neuronal cell types. The maximum projection score was set to be 0.2, such that all values greater than or equal to 0.2 are set to 0.2. Region assignments for each injection (Y-axis) were found using the injection site coordinates provided by the Allen Brain Atlas. To assign each molecular cell type a region (X-axis), we calculate the number of beads in each region for each cell type. We exclude

regions annotated as "root", "Fiber tracts", "Ventricle", or "unassigned" within the Allen Brain Atlas, and assign each cell type to the region in which it has the highest number of beads.

Narrowing our focus to a specific sub-region of the brain, we recalculated the projection scores using only voxels annotated as being part of the thalamus. We select projections that project significantly into the thalamus, removing injections whose streamlines occupy 30 or fewer voxels within the thalamus. For the wiring diagram shown in Fig. 4.2b, we filter the projections such that they significantly project into thalamus and have an injection region within the cortical regions. We filter for cell types that are neuronal and occupy greater than or equal to 30 beads within the thalamus. We assign each cell type to a region by counting the number of beads of each cell type present in each section of the thalamus, and labeling the cell type with the region where it most frequently appears, i.e. has the most beads. To determine mesoscale projection associations between cortical regions and thalamic regions, we average the projection scores for each cell type amongst all the projections originating from a distinct cortical injection region. We rank the average projection scores for each cortical region, and select the four highest scoring cell types for the wiring diagram. Notably, we remove all cell types with a projection score of less than or equal to 0.03, as this low of a projection score may not be significant enough to reflect biological projection patterns. For this reason, the AUD and the RSP only have 3 selected cell types, and the FRP cortical region is not shown.

Similarly, we also calculate a hypothalamus specific projection score, constraining the voxel space to that occupied by the hypothalamus, and re-calculating the projection score, removing injections whose streamlines occupy 30 or fewer voxels within the hypothalamus. When identifying quality cell types within the hypothalamus, we filtered to retain neuronal cell types that had a projection score with at least one neuron in the hypothalamus, and had been confidently assigned to more than 10 beads within the hypothalamus.

4.5.3 Integration of Molecular Profiles and c-FOS Data

Bonapersona et. al measures c-Fos expression within the 3D mouse brain at 30, 90 and 180 minutes after foot-shock [31]. To adjust for c-Fos expression at baseline, they also measure c-Fos expression at time point 0, where they place the animal into the foot-shock equipped cage, but do not shock the animal. To co-register the c-Fos activity data with the Slide-seq data [94], we translate the c-Fos data, voxelized at 100 μm, onto the Slide-seq data, also downsampled to 100 μm voxels. We filter out all voxels that do not have at least one Slide-seq bead that has been confidently assigned (confidence score > 0.3, per original publication's recommendation [94]). For all voxels in which there is a confident cell type assignment, but

no c-Fos signal, we assign these voxels a c-Fos activity score of 0. The remaining Slide-seq beads are shown in Fig. 4.4a. We measure the intensity of the c-Fos expression by counting the number of c-Fos expressing cells captured within each voxel. To account for changes in c-Fos expression from baseline expression, we calculate the change in c-Fos expressing cells, such that $\Delta cFos^z_{Timepoint_x} = N^z_x - N^z_0$, where z is a 100 μm voxel, N is the number of c-Fos expressing cells, and x is a time point after foot shock. We visualize significant c-Fos voxels in each time point by filtering out all voxels in which $cFos^z_x = 0$.(Fig. 4.4b-d). For each molecular cell type in each time point, we calculate a c-Fos score, such that

$$cFosActivityScore_{X_A} = \frac{\sum_{j=1}^{m} \left(C_j \times \frac{N^A_j}{N^B_j} \right)}{\sum_{j=1}^{m} N^A_j}$$

where x is a time point, N is the number of beads, m is all voxels retained after previous filtering steps, j is a particular voxel, A is the cell type for which we are calculating the c-Fos score, and B is all cell types present within voxel . For time point $x = 0$, C_j is the number of c-Fos expressing cells within voxel j. For all other time points, $C_j = cFos^j_x$ within voxel j. Consequently, we take the ratio of beads with cell type A in voxel j and weight it by the number of confidentially assigned cell types within j. We then multiply this by the change in c-Fos activity within that voxel (C_j) and compute the sum of all weighted intensity scores across all voxels (m). We then weigh the final c-Fos activity score by the number of times a cell type is confidentially assigned.

4.5.4 Analyzing the Association Between c-Fos Score and Molecular Cell Types Across Timepoints and Within Specific Regions

Filtering to retain only neuronal cell types that are present in greater than 20 voxels throughout the entirety of the brain, we can visualize the neuronal cell types with highest association to c-Fos activity throughout the brain at time points 0, 30, 90, and 180 minutes after foot-shock (Fig. 4.4e-h). Notably, time point 0 shows the c-Fos score calculated on the number of c-Fos expressing cells within each voxel, and time points 30, 90, and 180 present the c-Fos score calculated on the change in c-Fos score adjusted for baseline (time point 0) values. For each time point, we filter out all cell types that are confidently assigned to ≤ 20 voxels brain-wide. We then rank the molecular cell types by c-Fos score, and select the ten highest scoring cell types for each time point. We graph these cell types such that their dot location is given by their cell type centroid (see cell type radius analysis), and the magnitude of the

dot size is given by the magnitude of their c-Fos score (Fig. 4.4e-h). To examine the onset of the experimental condition on c-Fos scores, we retain only cell types present in greater than 20 voxels throughout the brain. We then select only cell types that are excitatory, and average their c-Fos score across experimental timepoints (T30, T90, T180). We then rank the molecular cell types in order of their average experimental score, and select the top 15 highest ranking molecular cell types and the 15 lowest ranking molecular cell types. We then graph them by their c-Fos score across time, and color the resulting line plot by molecular cell type (Fig. 4.4i). We repeat this process for the inhibitory neurons as well (Fig. 4.4j). Next, we examine the onset of the experimental condition within a single distinct region. We filter the molecular cell types to only neuronal cells confidently assigned to greater than 10 voxels within the thalamus. For each molecular cell type, we examine the number of beads present in each thalamic nuclei. When a substantial population of the molecular cell type's beads is present in a single thalamic nuclei, we annotate that molecular cell type as belonging to that thalamic anatomical region. For molecular cell types substantially present in more than one thalamic nuclei, we annotate these molecular cell types with the anatomical region of the thalamus. We then graph the c-Fos scores of the molecular cell types of the thalamus across time, with each line representing a molecular cell type, and the lines colored by the assigned anatomical thalamic region (Fig. 4.4k). Continuing to restrict our interest to voxels located within the thalamus, we similarly limit the beads within our visualizations to those annotated as being located within the thalamus (Fig. 4.4l). For each time point, we show the change in c-Fos expressing cells for each voxel with at least one Slide-seq bead with at least one confident cell type assignment, filtering out all voxels for which there are no changes in c-Fos expression (Fig. 4.4m-o). We calculate a thalamus-specific c-Fos activity score by constraining the voxel space to those annotated as belonging to the thalamus. For time points 30, 90, and 180, we rank the cell types by thalamus-specific c-Fos activity score, selecting the 5 highest scoring cell types that are present in 10 or more voxels within the thalamus. We graph each bead that has been confidently assigned one of these cell types within the bounds of the thalamus (Fig. 4.4p-r).

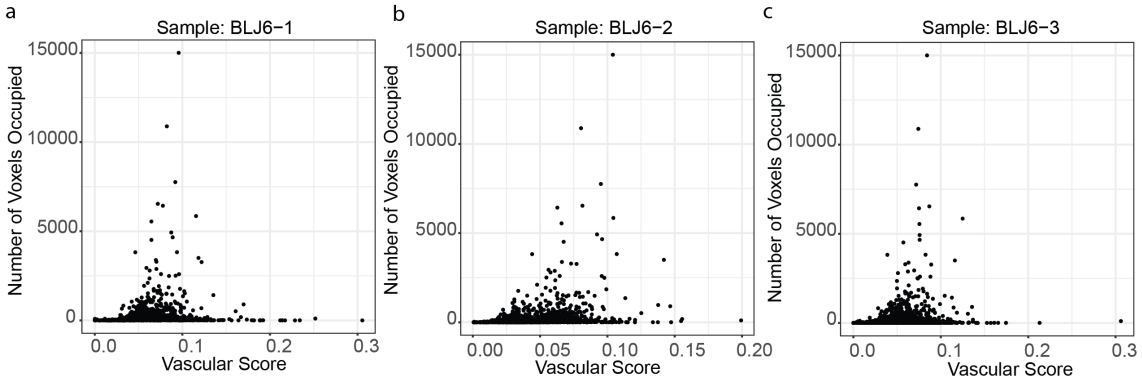
4.5.5 Assessing the Relationship between Neural Activity and Vascular Density

We also analyzed the relationship between neural activity and vascular score. To determine the neural activity for a given cell, we used two different methods. First, we calculated the neural activity by the log (counts per million) of genes that are annotated as immediate early genes. We refer to this score as the Neuronal Activity Score. We also calculated the

neural activity using NeuroEstimator [121]. After calculating the neural activity score for each cell using both of the above methods, we then found the average neuronal activity for a cell type by taking the average neuronal scores across all cells assigned to that cell type. This resulted in an average NeuroEstimator estimation of neuronal activity for each cell, as well as an average Neuronal Activity Score estimation of neuronal activity for each cell type. We filtered the cell types used such that each cell type was confidently assigned to ten or more voxels in the Slide-seq data [94], and that each cell type was classified as neuronal. Since the vascular scores had been calculated for three vascular samples (BLJ6-1, BLJ6-2, BLJ6-3), we also calculated the average vascular score for each cell type across all three samples. Next, we calculated the correlation between the Neuronal Activity Score and the vascular score, and found the correlation to be -0.036. The correlation between the NeuroEstimator score and the vascular score was -0.17 (Supplementary Figure 4.5). Because of this, we deemed no correlation between vascular density and neuronal cell types with elevated activity scores.

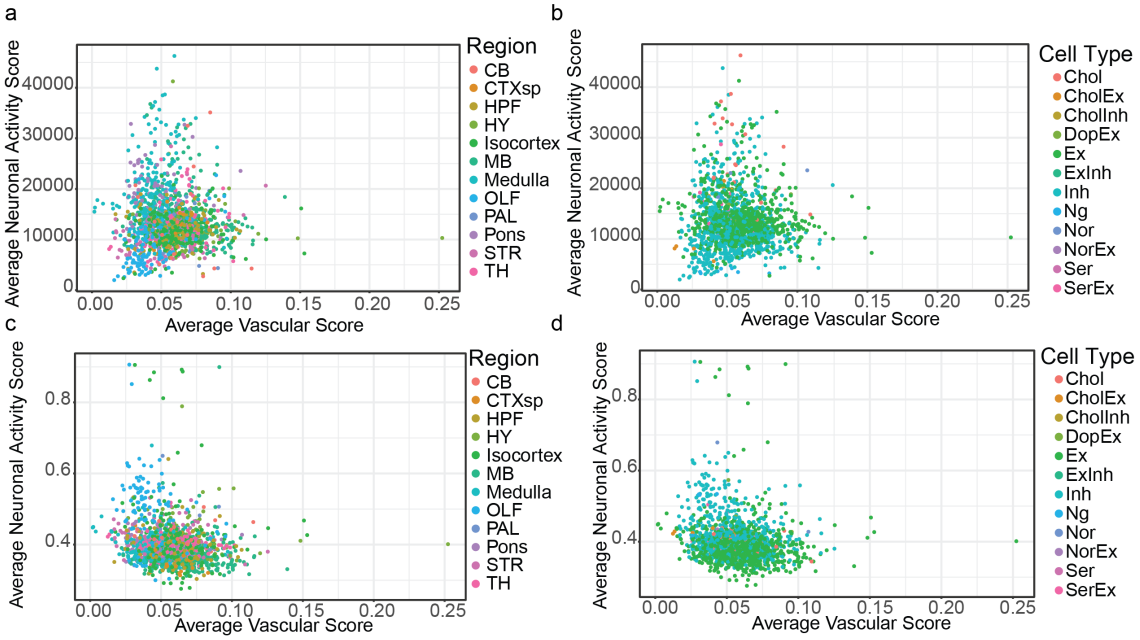
4.5.6 Supplementary Tables and Figures

Figure 4.5: Correlation between Vascular activity score and cell type prevalence



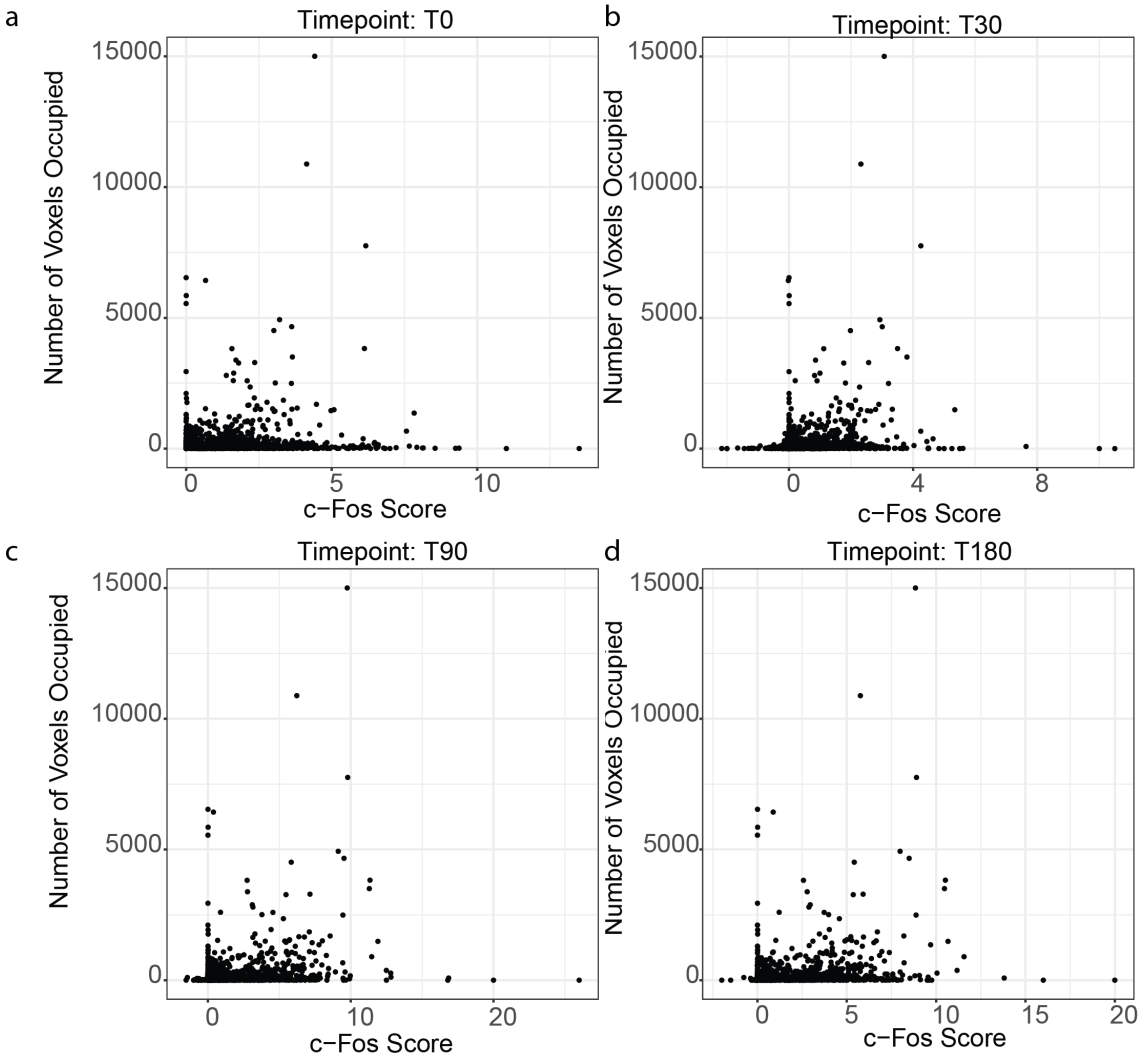
(**a**) Each molecular neuronal cell type is represented as a dot. The x-axis represents the vascular score for each cell type, and the number of voxels that each cell type is found in is presented by the y-axis. (**b-c**) The vascular scores for all samples used within the paper are shown.

Figure 4.6: Correlation between neuronal activity and vascular score



(**a-d**) We examine the relationship between neural activity and vascular density using both a Neuronal Activity Score and a NeuroEstimator Score. Each dot represents a cell type, where the position along the x-axis presents that cell type's averaged vascular score across 3 vascular samples, and the y-axis position represents the average neuronal score for that cell type. We color the cell types either by the region in which they most frequently appear, or by their broad cell class.

Figure 4.7: Correlation between c-Fos activity score and cell type prevalence



(**a**) Each dot represents a distinct neuronal cell type, with the x-axis reflecting the c-Fos score for each cell type, and the y-axis capturing the number of voxels in which a cell type is assigned. The c-Fos score for Time point 0 is calculated on the number of c-Fos expressing cells.(**b-d**) The c-Fos score for T30, T90, and T180 is calculated using the change in c-Fos expression from T0 to the time point of interest.

CHAPTER 5

Conclusion

5.1 Summary

The motivation for this dissertation is the construction of a more comprehensive cellular profile. In Chapter II, we refine cell type characterization by including additional, unshared features in data integrations. In Chapter III, we define an atlas of expected molecular cell types within the mouse brain and determine their spatial organization. In Chapter IV, we pair these molecularly defined cell types with anatomical and physiological features. To more deeply understand cell type profiles, we have derived two novel bioinformatics algorithms (UINMF, SiNMFiD), constructed a molecular atlas of the mouse brain with spatial resolution, and validated the leveraging of spatial coordinates to correlate molecular types with anatomical and physiological features. We have ensured that our work is accessible to the broader scientific community by utilizing Github, NeMO Analytics, and our own novel portal: Welch Lab Neuroglancer Portal. Below, we detail more fully the contribution of each chapter toward understanding cellular identity.

In Chapter II, we present our UINMF algorithm, which integrates molecular datasets while leveraging unshared features. We demonstrate the utility of our algorithm on cross-modality and cross-species integrations. By introducing unshared features, the UINMF algorithm generally improves the alignment and purity scores of the integrated datasets and is not prohibitive in either memory usage or run time. For datasets with few measured transcripts, such as the STARmap spatial transcriptomics dataset [195], including unshared features from a dataset with more measured transcripts provides greater structure within the shared latent space. Higher quality integrations allow for improved characterization of the molecular cell types with spatial resolution. Similarly, when observing cross-species integrations, including non-orthologous genes allows cell types that may be distinguished only by non-orthologous gene expression to be identified. A key advantage of building the UINMF algorithm on a

non-negative matrix factorization framework is the resulting biological interpretability of its factorization. Our algorithm is publicly available on Github.

In Chapter III, we describe the creation of our mouse whole-brain, spatially-resolved molecular atlas. To create the atlas, we integrate data from six distinct modalities, capturing single-cell and single-nucleus transcriptomic measures in conjunction with epigenomic profiles. To account for the regional heterogeneity that is observed within the brain, as well as the diverse number of cell types present in each region, we divide our data by anatomical dissection region and primary cell class (non-neuronal, excitatory neurons, inhibitory and other neurons) before integration. Ultimately, we perform 54 distinct integration analyses and manually curate and annotate the derived cellular clusters. Using the resulting clusters, we deconvolve two distinct spatial transcriptomics datasets, the ISH Allen Brain dataset [97] and the Slide-seq [94] dataset. For the deconvolution, we use an established algorithm, RCTD [37], to perform an initial deconvolution of the Slide-seq data using our derived cluster labels. Then, using our novel deconvolution method SiNMFiD, we deconvolve both the Slide-seq and ISH data. To confirm the accuracy of our cell type annotations, we examined the correctness of the distribution of the cortical layers using RCTD. As RCTD is not capable of deconvolving the large, 200 μm voxels of the ISH data, we compared the RCTD and SiNMFiD results on the Slide-seq data to ensure that SiNMFiD was deriving appropriate spatial distributions for the molecular cell types. Overall, we observe consistency between datasets and algorithms when examining the spatial distribution of a single molecular cell type. To ensure our findings are accessible to the general public, we upload our integration clustering results to the NeMO Analytics portal and publish our deconvolution results on our own portal Welch Lab Neuroglancer Portal.

In Chapter IV, we demonstrate how a spatial coordinate framework can be used to integrate molecular cell types with anatomical and physiological data. Specifically, after establishing spatial distributions for our molecular cell type profiles in Chapter IV, we then use the Common Coordinate Framework [194] to register the cell types into the same coordinate system as a vascular density dataset, a neuron projection dataset, and c-Fos activity after footshock dataset. After co-registration, we performed a correlation analysis. We determine distinct molecular cell types associated with elevated and depressed vascular density scores and identify cell types most associated with specific neuronal projections. Using c-Fos activity scores at three time points after footshock, we prioritize molecular cell types most likely to be involved in footshock response at particular time points along the event axis. An integrated interface of molecular cell profiles in conjunction with the vascular, neuronal projection, and c-Fos datasets can be found at Welch Lab Neuroglancer Portal.

5.2 Future Work

Efforts to understand cell types and their functions are a necessary and complex area of ongoing research. Refining cell types beyond what we have accomplished in this dissertation will likely be an ongoing and iterative process. UINMF, our whole-brain molecular atlas, and the co-registration of distinct datasets into a shared coordinate system can all be leveraged to improve our understanding of cellular identity.

5.2.1 Future Areas of Exploration for the UINMF Algorithm

One principal way we demonstrate the utility of UINMF is by illustrating the improved integration results from integrating single-cell RNA-seq data with many measured transcripts with spatial transcriptomics data with few measured transcripts. In 2023, Zhang et al. published a spatial transcriptomic atlas of the mouse brain utilizing the MERFISH technology to capture ∼1,100 genes [209]. In this study, the MERFISH data was integrated with the sc10Xv2/3 whole-brain data. An additional primary research question is whether using UINMF to integrate the sc10Xv2/3 data and the MERFISH data yields additional insights and better alignment. If so, redefining the MERFISH atlas annotations using UINMF integrations with the sc10Xv2/3 data would be beneficial. Potential future applications of UINMF extend past the integration types introduced in the original publication.

Multi-omic technologies have established a new era of defining a cell type by multiple facets of molecular identity. Commonly, assays aim to measure the transcriptomic features of a cell in conjunction with some other feature (i.e., epigenome, genome, proteome, etc.) [53]. UINMF can link different multi-omic datasets using the transcriptomic features as shared features. The other, non-transcriptomic features can still be used in defining cellular profiles within the latent space. As multi-omic technologies increase in popularity, UINMF can be used to explore multi-modal integration.

UINMF applications are not limited to molecular features. There are no constraints on the unshared feature space, meaning the unshared matrix can be defined in novel ways. Imagine we use Neuropixel probes to capture the activity of a neuron across time and also sequence that neuron to yield its transcriptomic measures. We can then build two matrices, one matrix defined as cells by genes and one matrix defined by cells by signal intensity across time. Utilizing the unshared features in new ways provides a means of algorithmically adding non-molecular features into our cell type profiles. We anticipate that UINMF will be beneficial for refining cell type identities in several distinct capacities.

5.2.2 Future Directions for Research in the Development of the Whole-Brain Cell Atlas

While an initial whole-brain molecular atlas was necessary to construct a foundation on which to build, such an atlas must be continuously improved. Initial iterations must focus on continued improvements to molecularly defined cellular profiles. Some of these improvements must encompass a standardization of nomenclature, ease of community interaction, and addition of non-healthy samples, as detailed below.

Creating a draft of the molecular cell-type atlas with spatial resolution is a critical initial step in furthering our current understanding of cell identity, but subsequent iterations of the whole-brain atlas should concentrate on deriving a standardized nomenclature for cellular classification. Currently, each of the data-generating labs that produced the six distinct modalities that we integrated for our atlas used a unique nomenclature and method for naming clusters and cell types. Ideally, a standard nomenclature between labs would allow for cell type profiles to be identified across modalities and labs, as well as promote continual solidification of the features of each cell type profile. For example, recent work on the motor cortex [206] identified cell types, assigned them spatial regions of origin, and provided marker genes that would allow the identification of these clusters in future analyses. Ideally, the whole-brain atlas nomenclature would also provide this information, allowing neuroscientists to gain intuition about the origin of each cell type and where it resides in the classification hierarchy. Creating a standardized nomenclature across the brain is a crucial step in reproducible neuroscience.

After the standardization of nomenclature, another critical component that would increase the impact of the whole-brain atlas would be the ease with which fellow scientists could utilize the constructed atlas to annotate and decipher their own data. Allowing users to project their data into the latent space defined by the atlas would provide users with an initial annotation draft of their dataset using standardized annotation labels. While manual curation of automatically generated labels will likely be recommended for several years to come, the ability to automatically generate an initial draft labeling would save countless hours of labor and provides a measure of rigor to a process currently limited by user expertise. The ease with which users can upload and interact with their data, as well as the atlas data, needs to be an area of focused and intentional improvement.

After increasing the standardization and useability of the whole-brain atlas, we can pivot our focus from healthy samples and cell types to those influenced by disease. Disease datasets, gathered from brains characterized by demylination [74], aging [17], Parkison's [213], and Alzheimer's [174] have already been sequenced. A detailed analysis integrating such diseased

datasets into the whole-brain atlas and identifying and categorizing perturbations between healthy and diseased states may point to global trends in the disease pathways. Cell states that remain unshared between healthy and diseased states may reflect cell types particularly influenced by the disease state. Understanding what specific pathways and cell types are perturbed during the onset of a disease condition is vital to understanding and eventually combating disease pathology.

It is critical to understand that the standardization of nomenclature, increasing the accessibility of the brain atlas, and interrogation of diseased cell state will be ongoing and interwoven tasks, as each iteration of the brain atlas will hopefully bring significant improvements on each of these fronts. Additionally, as technologies improve, our understanding of spatial distribution and anatomical and physiological features in conjunction with molecular profiles will likely result in continued curation of cellular profiles. Recent investigations suggest that significant cellular differences exist between transcriptionally similar cell types [134], and the continual addition of non-molecular features to cell type profiles will result in reconsideration of cell type divisions and definitions.

5.2.3 Continued Applications of Leveraging Co-Registration to Correlate Molecular and Non-molecular Features of Cellular Identity

Integrating molecular data into the same common coordinate system as anatomical and physiological data provides some of the first computationally derived associations between molecular cell types and anatomical and physiological data. Recent evidence suggests that cell types within the same molecular type can still have significant variations in terms of projection patterns [134]. Constructing comprehensive and informative cell-type profiles necessitates the addition of anatomical and physiological features. Our methodology of pairing molecular and other non-molecular features using spatial coordinates can lead to further discoveries and intuitions by the application to new datasets, the registration and improved accessibility of already generated data, and the development of more sophisticated association methods.

Expanding the spatial correlation framework to examine new datasets is a simple next step in understanding how molecular cell types respond physiologically and are influenced by their own anatomy (i.e., neuronal projection characteristics) and their surrounding anatomy (i.e., vascular density). The Slide-seq and ISH datasets used in our analyses [97, 94] are not single-cell resolution. Using spatial transcriptomics datasets captured at single-cell resolution, i.e., MERFISH [209], may provide a more accurate representation of the cell types within

each voxel. Higher resolution spatial transcriptomics data is one area of future exploration, as examining new anatomical and physiological data also promises exciting results. For example, Chen et al. used Neuropixel probe recordings to capture the flow of information within the brain throughout stimulus exposure, decision-making, movement response, and reward [45]. By integrating this data with molecular profiles, we can associate specific molecular cell types with each phase in the experimental process, including decision-making processes. Establishing such informative correlations is straightforward, requiring only that new data be registered to the CCF.

Many datasets have already been registered to the CCF prior to publication submission. However, in our experience, accessing these open-source datasets has been far from trivial. Author response times for data access requests have been delayed or absent entirely. Even when the publication data is accessible, it is often in its raw, unprocessed form. Frequently, the code and documentation for how the data-generating authors processed and quality-controlled their data is sparse and confusing. It is critical that a repository of CCF datasets is created, curated, and consistently maintained. Each CCF-registered dataset should have an accompanying tutorial for accessing and processing the data, ensuring that it is used in a rigorous and consistent manner. As the number of generated and accessible CCF-registered datasets grows, new ways of interacting with the repository data will naturally arise. In addition to looking for correlations between molecular profiles and non-molecular features, it would also be beneficial to explore associations between the non-molecular features themselves. Questions to explore include whether neuronal projection patterns are influenced by vascular density and if c-Fos expressing cells are more probably associated with specific neuronal projection patterns. As the number of well-curated datasets increases, so too will the number of explorable connections and interactions.

We anticipate that novel CCF-registered datasets will present their own unique challenges to dataset correlations. Our methods explore straightforward spatial correlations with success, but new approaches may need to be developed for more complex data. Data sets with signals captured across time necessitate the development of methods that correspondingly process signal intensity across time while accommodating the types of molecular cell types found at each spatial location. Developing strategies that account for transcriptionally similar, yet uniquely clustered, cell types would also be advantageous, as molecular differences can be subtle and can represent cell states instead of cell types [207]. Providing a weight to distance significantly differential cell types from similar ones may provide inferences for more global associations between molecular and non-molecular features. Developing more sophisticated methodologies for discovering associations between molecular cell types and complex datasets will be an area of intense future research.

This dissertation focuses on developing more comprehensive cell-type profiles. We begin by using UNIMF to discover more rigorous molecular cell type clusters, and we then expand our efforts to understand molecular cell types at the organ level, creating a molecular atlas of the brain. To investigate the spatial distribution of molecular cell types, we deconvolve our derived cell types within space and use spatial coordinates to draw associations between molecular subtypes and non-molecular features. Work aimed at deriving complete cell-type profiles is, by necessity, iterative. We provide public access to our data in the hopes that it will facilitate future revisions of the cell type profiles provided, and we created our 3D atlas portal to provide a hypothesis-generation tool for the community at large. We fervently hope these tools will continually facilitate the further development of robust and informative cell-type profiles.

BIBLIOGRAPHY

[1] Allen brain cell atlas. https://portal.brain-map.org/atlases-and-data/bkp/abc-atlas. Accessed: 2024-1-4.

[2] BRAIN 2025: A scientific vision. https://braininitiative.nih.gov/vision/nih-brain-initiative-reports/brain-2025-scientific-vision. Accessed: 2023-11-13.

[3] Chan zuckerberg CELLxGENE discover. https://cellxgene.cziscience.com/. Accessed: 2024-1-4.

[4] Home. https://www.braincelldata.org/. Accessed: 2024-1-4.

[5] MACS2.

[6] Muscle tissue. https://training.seer.cancer.gov/anatomy/cells_tissues_membranes/tissues/muscle.html. Accessed: 2023-12-13.

[7] pycistopic.

[8] SnapTools.

[9] Spatial gene expression. https://www.10xgenomics.com/products/spatial-gene-expression. Accessed: 2024-1-10.

[10] Visium HD spatial gene expression. https://www.10xgenomics.com/products/visium-hd-spatial-gene-expression. Accessed: 2024-1-10.

[11] Method of the year 2020: spatially resolved transcriptomics. *Nat. Methods*, 18(1):1, January 2021.

[12] Visium spatial gene expression protocol planner, document number CG000295 rev D. *10x Genomics*, 2021.

[13] What is a cell? In *Understanding Cancer*, pages 64–75. Cambridge University Press, May 2022.

[14] N David Åberg, Maria A I Åberg, and Peter S Eriksson. CHAPTER 11 - growth hormone and insulin-like growth Factor-I and cellular regeneration in the adult brain. In Fred J Nyberg, editor, *The Somatotrophic Axis in Brain Function*, pages 125–145. Academic Press, Burlington, January 2006.

[15] Sandra Ackerman. *Major Structures and Functions of the Brain*. National Academies Press (US), 1992.

[16] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Analyzing Protein Structure and Function*. Garland Science, 2002.

[17] William E Allen, Timothy R Blosser, Zuri A Sullivan, Catherine Dulac, and Xiaowei Zhuang. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell*, 186(1):194–208.e18, January 2023.

[18] Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. The ENCODE blacklist: Identification of problematic regions of the genome. *Sci. Rep.*, 9(1):1–5, June 2019.

[19] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.*, 21(1):111, May 2020.

[20] Ricard Argelaguet, Anna S E Cuomo, Oliver Stegle, and John C Marioni. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.*, 39:1202–1215, May 2021.

[21] Thomas L Athey, Daniel J Tward, Ulrich Mueller, Joshua T Vogelstein, and Michael I Miller. Hidden markov modeling for maximum probability neuron reconstruction. *Commun Biol*, 5(1):388, April 2022.

[22] David Attwell, Alastair M Buchan, Serge Charpak, Martin Lauritzen, Brian A Macvicar, and Eric A Newman. Glial and neuronal control of brain blood flow. *Nature*, 468(7321):232–243, November 2010.

[23] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.*, 11(1):5650, November 2020.

[24] Ethan Bahl, Snehajyoti Chatterjee, Muhammad Elsadany, Yann Vanrobaeys, Li-Chun Lin, K Peter Giese, Ted Abel, and Jacob J Michaelson. NEUROeSTIMator: Using deep learning to quantify neuronal activation from Single-Cell and spatial transcriptomic data. April 2022.

[25] Sarah K Baisley, Christina L Cloninger, and Vaishali P Bakshi. Fos expression following regimens of predator stress versus footshock that differentially affect prepulse inhibition in rats. *Physiol. Behav.*, 104(5):796–803, October 2011.

[26] Trygve E Bakken, Nikolas L Jorstad, Qiwen Hu, Blue B Lake, Wei Tian, Brian E Kalmbach, Megan Crow, Rebecca D Hodge, Fenna M Krienen, Staci A Sorensen, Jeroen Eggermont, Zizhen Yao, Brian D Aevermann, Andrew I Aldridge, Anna Bartlett, Darren Bertagnolli, Tamara Casper, Rosa G Castanon, Kirsten Crichton, Tanya L Daigle, Rachel Dalley, Nick Dee, Nikolai Dembrow, Dinh Diep, Song-Lin Ding, Weixiu Dong, Rongxin Fang, Stephan Fischer, Melissa Goldman, Jeff Goldy, Lucas T Graybuck,

Brian R Herb, Xiaomeng Hou, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Baldur van Lew, Yang Eric Li, Christine S Liu, Hanqing Liu, Jacinta D Lucero, Anup Mahurkar, Delissa McMillen, Jeremy A Miller, Marmar Moussa, Joseph R Nery, Philip R Nicovich, Joshua Orvis, Julia K Osteen, Scott Owen, Carter R Palmer, Thanh Pham, Nongluk Plongthongkum, Olivier Poirion, Nora M Reed, Christine Rimorin, Angeline Rivkin, William J Romanow, Adriana E Sedeño-Cortés, Kimberly Siletti, Saroja Somasundaram, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Xinxin Wang, Fangming Xie, Anna Marie Yanny, Renee Zhang, Seth A Ament, M Margarita Behrens, Hector Corrada Bravo, Jerold Chun, Alexander Dobin, Jesse Gillis, Ronna Hertzano, Patrick R Hof, Thomas Höllt, Gregory D Horwitz, C Dirk Keene, Peter V Kharchenko, Andrew L Ko, Boudewijn P Lelieveldt, Chongyuan Luo, Eran A Mukamel, Sebastian Preissl, Aviv Regev, Bing Ren, Richard H Scheuermann, Kimberly Smith, William J Spain, Owen R White, Christof Koch, Michael Hawrylycz, Bosiljka Tasic, Evan Z Macosko, Steven A McCarroll, Jonathan T Ting, Hongkui Zeng, Kun Zhang, Guoping Feng, Joseph R Ecker, Sten Linnarsson, and Ed S Lein. Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. April 2020.

[27] Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.*, 24(10):695–713, October 2023.

[28] Daphne W Bell. Northern blot. https://www.genome.gov/genetics-glossary/Northern-Blot. Accessed: 2024-1-26.

[29] Jon-Matthew Belton, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276, November 2012.

[30] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat. Methods*, 18(11):1352–1362, November 2021.

[31] Valeria Bonapersona, Heike Schuler, Ruth Damsteegt, Youri Adolfs, R Jeroen Pasterkamp, Martijn P van den Heuvel, Marian Joëls, and R Angela Sarabdjitsingh. The mouse brain after foot shock in four dimensions: Temporal dynamics at a single-cell resolution. *Proceedings of the National Academy of Sciences*, 119(8):e2114002119, 2022.

[32] Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, and Stein Aerts. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods*, 20(9):1355–1367, September 2023.

[33] Clément Brunner, Micheline Grillet, Arnau Sans-Dublanc, Karl Farrow, Théo Lambert, Emilie Macé, Gabriel Montaldo, and Alan Urban. A platform for brain-wide volumetric functional ultrasound imaging and analysis of circuit dynamics in awake mice. *Neuron*, 108(5):861–875.e7, December 2020.

[34] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12):1213–1218, December 2013.

[35] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, July 2015.

[36] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420, June 2018.

[37] Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.*, 40(4):517–526, April 2022.

[38] John D Cahoy, Ben Emery, Amit Kaushal, Lynette C Foo, Jennifer L Zamanian, Karen S Christopherson, Yi Xing, Jane L Lubischer, Paul A Krieg, Sergey A Krupenko, Wesley J Thompson, and Ben A Barres. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.*, 28(1):264–278, January 2008.

[39] Zayna Chaker, Paolo Codega, and Fiona Doetsch. A mosaic world: puzzles revealed by adult neural stem cell heterogeneity. *Wiley Interdiscip. Rev. Dev. Biol.*, 5(6):640–658, November 2016.

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

[41] Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E Vinyard, Sara P Garcia, Kendell Clement, Miguel A Andrade-Navarro, Jason D Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.*, 20(1):241, November 2019.

[42] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), April 2015.

[43] Lu Chen, Jaime M Tovar-Corona, and Araxi O Urrutia. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *Int. J. Evol. Biol.*, 2012:596274, July 2012.

[44] Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.*, 37(12):1452–1457, December 2019.

[45] Susu Chen, Yi Liu, Ziyue Aiden Wang, Jennifer Colonell, Liu D Liu, Han Hou, Nai-Wen Tien, Tim Wang, Timothy Harris, Shaul Druckmann, Nuo Li, and Karel Svoboda. Brain-wide neural activity underlying memory-guided movement. *Cell*, 187(3):676–691.e16, February 2024.

[46] Tsai-Ying Chen, Li You, Jose Angelito U Hardillo, and Miao-Ping Chien. Spatial transcriptomic technologies. *Cells*, 12(16), August 2023.

[47] Xiaoyin Chen, Yu-Chi Sun, Huiqing Zhan, Justus M Kebschull, Stephan Fischer, Katherine Matho, Z Josh Huang, Jesse Gillis, and Anthony M Zador. High-Throughput mapping of Long-Range neuronal projection using in situ sequencing. *Cell*, 179(3):772–786.e19, October 2019.

[48] Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, Oliver Stegle, and Wolf Reik. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, 9(1):781, February 2018.

[49] Simone Codeluppi, Lars E Borm, Amit Zeisel, Gioele La Manno, Josina A van Lunteren, Camilla I Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods*, 15(11):932–935, November 2018.

[50] Geoffrey M Cooper. *The Origin and Evolution of Cells*. Sinauer Associates, 2000.

[51] John A Cooper, Christopher Chow, Merton R Bernfield, Wilfred D Stein, Jonathan M W Slack, L Andrew Staehelin, Ronald A Laskey, Michael Cuffe, Bruce M Alberts, and Harvey F Lodish. *cell*. February 2024.

[52] Richard Daneman and Alexandre Prat. The blood-brain barrier. *Cold Spring Harb. Perspect. Biol.*, 7(1):a020412, January 2015.

[53] Maria A Dimitriu, Irina Lazar-Contes, Martin Roszkowski, and Isabelle M Mansuy. Single-Cell multiomics techniques: From conception to applications. *Front Cell Dev Biol*, 10:854317, March 2022.

[54] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable Single-Cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016.

[55] Rui Dong and Guo-Cheng Yuan. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol.*, 22(1):145, May 2021.

[56] J H Fallon, J N Riley, J C Sipe, and R Y Moore. The islands of calleja: organization and connections. *J. Comp. Neurol.*, 181(2):375–395, September 1978.

[57] Roozbeh Farhoodi, Benjamin James Lansdell, and Konrad Paul Kording. Quantifying how staining methods bias measurements of neuron morphologies. *Front. Neuroinform.*, 13:36, May 2019.

[58] Alejandra Freire-Regatillo, Pilar Argente-Arizón, Jesús Argente, Luis Miguel García-Segura, and Julie A Chowen. Non-Neuronal cells in the hypothalamic adaptation to metabolic signals. *Front. Endocrinol.*, 8:51, March 2017.

[59] Chao Gao, Jialin Liu, April R Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R Nery, Margarita M Behrens, Joseph R Ecker, Bing Ren, and Joshua D Welch. Iterative single-cell multi-omic integration using online learning. *Nat. Biotechnol.*, pages 1–8, April 2021.

[60] Chao Gao, Jialin Liu, April R Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R Nery, Margarita M Behrens, Joseph R Ecker, Bing Ren, and Joshua D Welch. Iterative single-cell multi-omic integration using online learning. *Nat. Biotechnol.*, April 2021.

[61] Pablo Garcia-Lopez, Virginia Garcia-Marin, and Miguel Freire. The histological slides and drawings of cajal. *Front. Neuroanat.*, 4:9, March 2010.

[62] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods*, February 2021.

[63] 10x Genomics. Chromium next GEM single cell multiome ATAC + gene expression reagent kits user guide. Technical Report Document Number CG000338 Rev B, December 2020.

[64] Paul G Giresi, Jonghwan Kim, Ryan M McDaniell, Vishwanath R Iyer, and Jason D Lieb. FAIRE (Formaldehyde-Assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, 17(6):877–885, June 2007.

[65] Helene Girouard and Costantino Iadecola. Neurovascular coupling in the normal brain and in hypertension, stroke, and alzheimer disease. *J. Appl. Physiol.*, 100(1):328–335, January 2006.

[66] Hui Gong, Shaoqun Zeng, Cheng Yan, Xiaohua Lv, Zhongqin Yang, Tonghui Xu, Zhao Feng, Wenxiang Ding, Xiaoli Qi, Anan Li, Jingpeng Wu, and Qingming Luo. Continuously tracing brain-wide long-distance axonal projections in mice at a one-micron voxel resolution. *Neuroimage*, 74:87–98, July 2013.

[67] Rajeshwar Govindarajan, Jeyapradha Duraiyan, Karunakaran Kaliyappan, and Murugesan Palanisamy. Microarray and its applications. *J. Pharm. Bioallied Sci.*, 4(Suppl 2):S310–2, August 2012.

[68] Christine Grienberger, Andrea Giovannucci, William Zeiger, and Carlos Portera-Cailliau. Two-photon calcium imaging of neuronal activity. *Nature Reviews Methods Primers*, 2(1):1–23, September 2022.

[69] Daniel Gyllborg, Christoffer Mattsson Langseth, Xiaoyan Qian, Eunkyoung Choi, Sergio Marco Salas, Markus M Hilscher, Ed S Lein, and Mats Nilsson. Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. *Nucleic Acids Res.*, 48(19):e112, November 2020.

[70] Gen Hamanaka, Ryo Ohtomo, Hajime Takase, Josephine Lok, and Ken Arai. White-matter repair: Interaction between oligodendrocytes and the neurovascular unit. *Brain Circ*, 4(3):118–123, July 2018.

[71] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[72] Chung-Chau Hon, Jay W Shin, Piero Carninci, and Michael J T Stubbington. The human cell atlas: Technical approaches and challenges. *Brief. Funct. Genomics*, 17(4):283–294, July 2018.

[73] Laura L Mays Hoopes. *Nucleic acid blotting: Southern and northern*. John Wiley & Sons, Inc., Hoboken, NJ, USA, September 2012.

[74] Jinchao Hou, Yingyue Zhou, Zhangying Cai, Marina Terekhova, Amanda Swain, Prabhakar S Andhey, Rafaela M Guimaraes, Alina Ulezko Antonova, Tian Qiu, Sanja Sviben, Gregory Strout, James A J Fitzpatrick, Yun Chen, Susan Gilfillan, Do-Hyun Kim, Steven J Van Dyken, Maxim N Artyomov, and Marco Colonna. Transcriptomic atlas and interaction networks of brain cells in mouse CNS demyelination and remyelination. *Cell Rep.*, 42(4):112293, April 2023.

[75] Clare Howarth, Anusha Mishra, and Catherine N Hall. More than just summed neuronal activity: how multiple cell types shape the BOLD response. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 376(1815):20190630, January 2021.

[76] Youjin Hu, Kevin Huang, Qin An, Guizhen Du, Ganlu Hu, Jinfeng Xue, Xianmin Zhu, Cun-Yu Wang, Zhigang Xue, and Guoping Fan. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.*, 17:88, May 2016.

[77] Hidetoshi Ikeno, Ajayrama Kumaraswamy, Kazuki Kai, Thomas Wachtler, and Hiroyuki Ai. A segmentation scheme for complex neuronal arbors and application to vibration sensitive neurons in the honeybee brain. *Front. Neuroinform.*, 12:61, September 2018.

[78] Laurent Jacob, Johann A Gagnon-Bartsch, and Terence P Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28, January 2016.

[79] Mika Sarkin Jain, Cecilia Dominguez Conde, Krzysztof Polanski, Xi Chen, Jongeun Park, Rachel A Botting, Emily Stephenson, Muzlifah Haniffa, Austen Lamacraft, Mirjana Efremova, and Sarah A Teichmann. MultiMAP: Dimensionality reduction and integration of multimodal data. February 2021.

[80] Sarah Jäkel and Leda Dimou. Glial cells and their function in the adult brain: A journey through the history of their ablation. *Front. Cell. Neurosci.*, 11:24, February 2017.

[81] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, January 2007.

[82] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.*, 12(3):e694, March 2022.

[83] James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydın, Mladen Barbic, Timothy J Blanche, Vincent Bonin, João Couto, Barundeb Dutta, Sergey L Gratiy, Diego A Gutnisky, Michael Häusser, Bill Karsh, Peter Ledochowitsch, Carolina Mora Lopez, Catalin Mitelut, Silke Musa, Michael Okun, Marius Pachitariu, Jan Putzeys, P Dylan Rich, Cyrille Rossant, Wei-Lung Sun, Karel Svoboda, Matteo Carandini, Kenneth D Harris, Christof Koch, John O'Keefe, and Timothy D Harris. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, November 2017.

[84] Hee Won Kang, Ho Kyu Kim, Bae Hun Moon, Seo Jun Lee, Se Jung Lee, and Im Joo Rhyu. Comprehensive review of golgi staining methods for nervous tissue. *Appl. Microsc.*, 47(2):63–69, June 2017.

[85] Justus M Kebschull, Pedro Garcia da Silva, Ashlan P Reid, Ian D Peikon, Dinu F Albeanu, and Anthony M Zador. High-Throughput mapping of Single-Neuron projections by sequencing of barcoded RNA. *Neuron*, 91(5):975–987, September 2016.

[86] J Kim and H Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *2008 Eighth IEEE International Conference on Data Mining*, pages 353–362, December 2008.

[87] Jessica H Kim, Grace H Kromm, Olivia K Barnhill, Jacob Sperber, Lauren B Heuer, Sierra Loomis, Matthew C Newman, Kenneth Han, Faris F Gulamali, Theresa B Legan, Katharine E Jensen, Samuel C Funderburk, Michael J Krashes, and Matthew E Carter. A discrete parasubthalamic nucleus subpopulation plays a critical role in appetite suppression. *Elife*, 11, May 2022.

[88] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *J. Global Optimiz.*, 58(2):285–319, February 2014.

[89] Gilbert J Kirouac. The paraventricular nucleus of the thalamus as an integrating and relay node in the brain anxiety network. *Front. Behav. Neurosci.*, 15:627633, February 2021.

[90] Christoph Kirst, Sophie Skriabine, Alba Vieites-Prado, Thomas Topilko, Paul Bertin, Gaspard Gerschenfeld, Florine Verny, Piotr Topilko, Nicolas Michalski, Marc Tessier-Lavigne, and Nicolas Renier. Mapping the Fine-Scale organization and plasticity of the brain vasculature. *Cell*, 180(4):780–795.e25, February 2020.

[91] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, 20(4):207–220, April 2019.

[92] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.*, 40(5):661–671, May 2022.

[93] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, 16(12):1289–1296, December 2019.

[94] Jonah Langlieb, Nina S Sachdev, Karol S Balderrama, Naeem M Nadaf, Mukund Raj, Evan Murray, James T Webber, Charles Vanderburg, Vahid Gazestani, Daniel Tward, Chris Mezias, Xu Li, Katelyn Flowers, Dylan M Cable, Tabitha Norton, Partha Mitra, Fei Chen, and Evan Z Macosko. The molecular cytoarchitecture of the adult mouse brain. *Nature*, 624(7991):333–342, December 2023.

[95] Lauren Alfiler, Kaitlyn Casimo, Clare Gamlin, Staci Soresen. Neurons: Beyond the textbook -educator webinar.

[96] Jeongwoo Lee, Do Young Hyeon, and Daehee Hwang. Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.*, 52(9):1428–1442, September 2020.

[97] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, Lin Chen, Li Chen, Tsuey-Ming Chen, Mei Chi Chin, Jimmy Chong, Brian E Crook, Aneta Czaplinska, Chinh N Dang, Suvro Datta, Nick R Dee, Aimee L Desaki, Tsega Desta, Ellen Diep, Tim A Dolbeare, Matthew J Donelan, Hong-Wei Dong, Jennifer G Dougherty, Ben J Duncan, Amanda J Ebbert, Gregor Eichele, Lili K Estin, Casey Faber, Benjamin A Facer, Rick Fields, Shanna R Fischer, Tim P Fliss, Cliff Frensley, Sabrina N Gates, Katie J Glattfelder, Kevin R Halverson, Matthew R Hart, John G Hohmann, Maureen P Howell, Darren P Jeung, Rebecca A Johnson, Patrick T Karr, Reena Kawal, Jolene M Kidney, Rachel H Knapik, Chihchau L Kuan, James H Lake, Annabel R Laramee, Kirk D Larsen, Christopher Lau, Tracy A Lemon, Agnes J Liang, Ying Liu, Lon T Luong, Jesse Michaels, Judith J Morgan, Rebecca J Morgan, Marty T Mortrud, Nerick F Mosqueda, Lydia L Ng, Randy Ng, Geralyn J Orta, Caroline C Overly, Tu H Pak, Sheana E Parry, Sayan D Pathak, Owen C Pearson, Ralph B Puchalski, Zackery L Riley, Hannah R Rockett, Stephen A Rowland, Joshua J Royall, Marcos J Ruiz, Nadia R Sarno, Katherine Schaffnit, Nadiya V Shapovalova, Taz Sivisay, Clifford R Slaughterbeck, Simon C Smith, Kimberly A Smith, Bryan I Smith, Andy J Sodt, Nick N Stewart, Kenda-Ruth Stumpf, Susan M Sunkin, Madhavi Sutram, Angelene Tam, Carey D Teemer, Christina Thaller, Carol L Thompson, Lee R Varnam, Axel Visel, Ray M Whitlock, Paul E Wohnoutka, Crissa K Wolkey, Victoria Y Wong, Matthew Wood, Murat B Yaylaoglu, Rob C Young, Brian L Youngstrom, Xu Feng Yuan, Bin Zhang, Theresa A Zwingman, and Allan R Jones. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, January 2007.

[98] Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, Ke Liu, Xuetong Zhao, Jun Lin, Linzhao Cheng, Falai Chen, Tian Xue, and Kun Qu. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods*, 19(6):662–670, June 2022.

[99] Guoqiang Li, Yaping Liu, Yanxiao Zhang, Naoki Kubo, Miao Yu, Rongxin Fang, Manolis Kellis, and Bing Ren. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods*, 16(10):991–993, October 2019.

[100] Qiufu Li and Linlin Shen. Neuron segmentation using 3D wavelet integrated encoder-decoder network. *Bioinformatics*, 38(3):809–817, January 2022.

[101] Zhao-Qing Li, Tong-Xin Li, Miao Tian, Ze-Sheng Ren, Chen-Yi Yuan, Rui-Kun Yang,

Su-Juan Shi, Hui Li, and Zhen-Zhen Kou. Glial cells and neurologic autoimmune disorders. *Front. Cell. Neurosci.*, 16:1028653, October 2022.

[102] Ethan S Lippmann, Samira M Azarin, Jennifer E Kay, Randy A Nessler, Hannah K Wilson, Abraham Al-Ahmad, Sean P Palecek, and Eric V Shusta. Derivation of blood-brain barrier endothelial cells from human pluripotent stem cells. *Nat. Biotechnol.*, 30(8):783–791, August 2012.

[103] Bimin Liu, Yuxiao Li, Miao Ren, and Xiangning Li. Targeted approaches to delineate neuronal morphology during early development. *Front. Cell. Neurosci.*, 17:1259360, October 2023.

[104] Hanqing Liu, Qiurui Zeng, Jingtian Zhou, Anna Bartlett, Bang-An Wang, Peter Berube, Wei Tian, Mia Kenworthy, Jordan Altshul, Joseph R Nery, Huaming Chen, Rosa G Castanon, Songpeng Zu, Yang Eric Li, Jacinta Lucero, Julia K Osteen, Antonio Pinto-Duarte, Jasper Lee, Jon Rink, Silvia Cho, Nora Emerson, Michael Nunn, Carolyn O'Connor, Zhanghao Wu, Ion Stoica, Zizhen Yao, Kimberly A Smith, Bosiljka Tasic, Chongyuan Luo, Jesse R Dixon, Hongkui Zeng, Bing Ren, M Margarita Behrens, and Joseph R Ecker. Single-cell DNA methylome and 3D multi-omic atlas of the adult mouse brain. *Nature*, 624(7991):366–377, December 2023.

[105] J Liu, Y Huang, R Singh, J P Vert, and W S Noble. Jointly embedding multiple single-cell omics measurements. *BioRxiv*, 2019.

[106] Jialin Liu, Chao Gao, Joshua Sodicoff, Velina Kozareva, Evan Z Macosko, and Joshua D Welch. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat. Protoc.*, 15(11):3632–3662, November 2020.

[107] Yufeng Liu, Gaoyu Wang, Giorgio A Ascoli, Jiangning Zhou, and Lijuan Liu. Neuron tracing from light microscopy images: automation, deep learning and bench testing. *Bioinformatics*, 38(24):5329–5339, December 2022.

[108] Eleonora Lomi, Kate J Jeffery, and Anna S Mitchell. Convergence of location, direction, and theta in the rat anteroventral thalamic nucleus. *iScience*, 26(7):106993, July 2023.

[109] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, December 2018.

[110] Joseph Loscalzo and Diane E Handy. Epigenetic modifications: basic mechanisms and role in cardiovascular disease (2013 grover conference series). *Pulm. Circ.*, 4(2):169–174, June 2014.

[111] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nat. Methods*, 16(8):715–721, August 2019.

[112] Tiffany M Love. Oxytocin, motivation and the role of dopamine. *Pharmacol. Biochem. Behav.*, 119:49–60, April 2014.

[113] Tian Lu, Cheen Euong Ang, and Xiaowei Zhuang. Spatially resolved epigenomic profiling of single cells in complex tissues. *Cell*, 185(23):4448–4464.e17, November 2022.

[114] Malte D Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and Fabian J Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods*, 19(1):41–50, January 2022.

[115] Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D Buenrostro. Chromatin potential identified by shared Single-Cell profiling of RNA and chromatin. *Cell*, 183(4):1103–1116.e20, November 2020.

[116] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015.

[117] Jamie L Marshall, Teia Noel, Qingbo S Wang, Haiqi Chen, Evan Murray, Ayshwarya Subramanian, Katherine A Vernon, Silvana Bazua-Valenti, Katie Liguori, Keith Keller, Robert R Stickels, Breanna McBean, Rowan M Heneghan, Astrid Weins, Evan Z Macosko, Fei Chen, and Anna Greka. High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *iScience*, 25(4):104097, April 2022.

[118] Samuel A M Martin, Carola E Dehler, and Elżbieta Król. Transcriptomic responses in the fish intestine. *Dev. Comp. Immunol.*, 64:103–117, November 2016.

[119] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nat. Methods*, 18(1):9–14, January 2021.

[120] Erik Meijering. Neuron tracing in perspective. *Cytometry A*, 77(7):693–704, July 2010.

[121] Páll Melsted, A Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Joseph Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.*, 39(7):813–818, July 2021.

[122] Anthony A Mercadante and Prasanna Tadi. *Neuroanatomy, Gray Matter*. StatPearls Publishing, July 2023.

[123] Laura E Mickelsen, William F Flynn, Kristen Springer, Lydia Wilson, Eric J Beltrami, Mohan Bolisetty, Paul Robson, and Alexander C Jackson. Cellular taxonomy and spatial organization of the murine ventral posterior hypothalamus. *Elife*, 9, October 2020.

[124] Jeffrey R Moffitt, Dhananjay Bambah-Mukku, Stephen W Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D Perez, Nimrod D Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416), November 2018.

[125] Jeffrey R Moffitt, Emma Lundberg, and Holger Heyn. The emerging landscape of spatial profiling technologies. *Nat. Rev. Genet.*, 23(12):741–759, December 2022.

[126] Charles Molnar and Jane Gair. 16.1 neurons and glial cells. In *Concepts of Biology - 1st Canadian Edition*. BCcampus, May 2015.

[127] Khaled Moustafa and Joanna M Cross. Genetic approaches to study plant responses to environmental stresses: An overview. *Biology*, 5(2), May 2016.

[128] Eran A Mukamel and John Ngai. Perspectives on defining cell types in the brain. *Curr. Opin. Neurobiol.*, 56:61–68, June 2019.

[129] Hong Ni, Zhao Feng, Yue Guan, Xueyan Jia, Wu Chen, Tao Jiang, Qiuyuan Zhong, Jing Yuan, Miao Ren, Xiangning Li, Hui Gong, Qingming Luo, and Anan Li. DeepMapi: a fully automatic registration method for mesoscopic optical brain images using convolutional neural networks. *Neuroinformatics*, 19(2):267–284, April 2021.

[130] Emily Niemitz. The microarray revolution. *Nat. Rev. Genet.*, 8(1):S15–S15, October 2007.

[131] Akiko Nishiyama, Ryusuke Suzuki, and Xiaoqin Zhu. NG2 cells (polydendrocytes) in brain physiology and repair. *Front. Neurosci.*, 8:133, June 2014.

[132] Marcel Oberlaender. Neuronal morphology and its significance. In *The Neocortex*. The MIT Press, 2019.

[133] Seung Wook Oh, Julie A Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M Henry, Marty T Mortrud, Benjamin Ouellette, Thuc Nghi Nguyen, Staci A Sorensen, Clifford R Slaughterbeck, Wayne Wakeman, Yang Li, David Feng, Anh Ho, Eric Nicholas, Karla E Hirokawa, Phillip Bohn, Kevin M Joines, Hanchuan Peng, Michael J Hawrylycz, John W Phillips, John G Hohmann, Paul Wohnoutka, Charles R Gerfen, Christof Koch, Amy Bernard, Chinh Dang, Allan R Jones, and Hongkui Zeng. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, April 2014.

[134] Hanchuan Peng, Peng Xie, Lijuan Liu, Xiuli Kuang, Yimin Wang, Lei Qu, Hui Gong, Shengdian Jiang, Anan Li, Zongcai Ruan, Liya Ding, Zizhen Yao, Chao Chen, Mengya Chen, Tanya L Daigle, Rachel Dalley, Zhangcan Ding, Yanjun Duan, Aaron Feiner, Ping He, Chris Hill, Karla E Hirokawa, Guodong Hong, Lei Huang, Sara Kebede, Hsien-Chi Kuo, Rachael Larsen, Phil Lesnar, Longfei Li, Qi Li, Xiangning Li, Yaoyao Li, Yuanyuan Li, An Liu, Donghuan Lu, Stephanie Mok, Lydia Ng, Thuc Nghi Nguyen, Qiang Ouyang, Jintao Pan, Elise Shen, Yuanyuan Song, Susan M Sunkin, Bosiljka Tasic, Matthew B Veldman, Wayne Wakeman, Wan Wan, Peng Wang, Quanxin Wang,

Tao Wang, Yaping Wang, Feng Xiong, Wei Xiong, Wenjie Xu, Min Ye, Lulu Yin, Yang Yu, Jia Yuan, Jing Yuan, Zhixi Yun, Shaoqun Zeng, Shichen Zhang, Sujun Zhao, Zijun Zhao, Zhi Zhou, Z Josh Huang, Luke Esposito, Michael J Hawrylycz, Staci A Sorensen, X William Yang, Yefeng Zheng, Zhongze Gu, Wei Xie, Christof Koch, Qingming Luo, Julie A Harris, Yun Wang, and Hongkui Zeng. Morphological diversity of single neurons in molecularly defined cell types. *Nature*, 598(7879):174–181, October 2021.

[135] Mario A Penzo, Vincent Robert, Jason Tucciarone, Dimitri De Bundel, Minghui Wang, Linda Van Aelst, Martin Darvas, Luis F Parada, Richard D Palmiter, Miao He, Z Josh Huang, and Bo Li. The paraventricular thalamus controls a central amygdala fear circuit. *Nature*, 519(7544):455–459, March 2015.

[136] Diana C Peterson, Vamsi Reddy, and Debra A Mayes. *Neuroanatomy, Mammillary Bodies*. StatPearls Publishing, August 2023.

[137] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, 10(11):1096–1098, November 2013.

[138] Kendall Powell. Technology to watch in 2018. *Nature*, 553(7689):531–534, January 2018.

[139] Sebastian Preissl, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U Gorkin, Yanxiao Zhang, Brandon C Sos, Veena Afzal, Diane E Dickel, Samantha Kuan, Axel Visel, Len A Pennacchio, Kun Zhang, and Bing Ren. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.*, 21(3):432–439, March 2018.

[140] Shmuel Price and Daniel T Daly. *Neuroanatomy, Trigeminal Nucleus*. StatPearls Publishing, May 2023.

[141] Marceau Quatredeniers, Alice S Serafin, Alexandre Benmerah, Antonio Rausell, Sophie Saunier, and Amandine Viau. Meta-analysis of single-cell and single-nucleus transcriptomics reveals kidney cell type consensus signatures. *Sci Data*, 10(1):361, June 2023.

[142] Karthik R Ramanathan, Reed L Ressler, Jingji Jin, and Stephen Maren. Nucleus reuniens is required for encoding and retrieving precise, Hippocampal-Dependent contextual fear memories in rats. *J. Neurosci.*, 38(46):9925–9933, November 2018.

[143] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, 30(8):777–782, August 2012.

[144] Mohan S Rao, Terry R Van Vleet, Rita Ciurlionis, Wayne R Buck, Scott W Mittelstadt, Eric A G Blomme, and Michael J Liguori. Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from Short-Term rat toxicity studies. *Front. Genet.*, 9:636, 2018.

[145] Research leaders. Types of neurons. https://qbi.uq.edu.au/brain/brain-anatomy/types-neurons, November 2017. Accessed: 2023-12-13.

[146] Sylvia Richardson, George C Tseng, and Wei Sun. Statistical methods in integrative genomics. *Annu Rev Stat Appl*, 3:181–209, June 2016.

[147] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, 32(9):896–902, September 2014.

[148] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, March 2019.

[149] Yeonjae Ryu, Geun Hee Han, Eunsoo Jung, and Daehee Hwang. Integration of Single-Cell RNA-Seq datasets: A review of computational methods. *Mol. Cells*, 46(2):106–119, February 2023.

[150] Lindsey D Salay, Nao Ishiko, and Andrew D Huberman. A midline thalamic circuit determines reactions to visual threat. *Nature*, 557(7704):183–189, May 2018.

[151] Christian Sam and Bruno Bordoni. *Physiology, Acetylcholine*. StatPearls Publishing, April 2023.

[152] P B Sargent. What distinguishes axons from dendrites? neurons know more than we do. *Trends Neurosci.*, 12(6):203–205, June 1989.

[153] Arpiar Saunders, Evan Z Macosko, Alec Wysoker, Melissa Goldman, Fenna M Krienen, Heather de Rivera, Elizabeth Bien, Matthew Baum, Laura Bortolin, Shuyu Wang, Aleksandrina Goeva, James Nemesh, Nolan Kamitaki, Sara Brumbaugh, David Kulp, and Steven A McCarroll. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, 174(4):1015–1030.e16, August 2018.

[154] Goran Sedmak and Miloš Judaš. White matter interstitial neurons in the adult human brain: 3% of cortical neurons in quest for recognition. *Cells*, 10(1), January 2021.

[155] P Segura-Torres, U Wagner, E Massanes-Rotger, L Aldavert-Vera, M Marti-Nicolovius, and I Morgado-Bernal. Tuberomammillary nucleus lesion facilitates two-way active avoidance retention in rats. *Behav. Brain Res.*, 82(1):113–117, December 1996.

[156] F R Sharp, S M Sagar, K Hicks, D Lowenstein, and K Hisanaga. c-fos mRNA, fos, and fos-related antigen induction by hypertonic saline and stress. *J. Neurosci.*, 11(8):2321–2331, August 1991.

[157] Zachary M Sheffler, Vamsi Reddy, and Leela Sharath Pillarisetty. *Physiology, Neurotransmitters*. StatPearls Publishing, May 2023.

[158] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, November 2009.

[159] Gordon M G Shepherd. Corticostriatal connectivity and its role in disease. *Nat. Rev. Neurosci.*, 14(4):278–291, April 2013.

[160] H Shibata. Organization of projections of rat retrosplenial cortex to the anterior thalamic nuclei. *Eur. J. Neurosci.*, 10(10):3210–3219, October 1998.

[161] Christopher T Simons, Amanda H Klein, and Earl Carstens. Chemogenic subqualities of mouthfeel. *Chem. Senses*, 44(5):281–288, May 2019.

[162] Wolf Singer, Terrence J Sejnowski, and Pasko Rakic. The neocortex. https://mitpress.mit.edu/9780262043243/the-neocortex/, December 2021. Accessed: 2024-2-6.

[163] Gordon K Smyth and Terry Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, December 2003.

[164] Lingyun Song and Gregory E Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, 2010(2):db.prot5384, February 2010.

[165] Qianqian Song and Jing Su. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief. Bioinform.*, 22(5), September 2021.

[166] Staci A Sorensen, Nathan W Gouwens, Yun Wang, Matt Mallory, Agata Budzillo, Rachel Dalley, Brian Lee, Olga Gliko, Hsien-Chi Kuo, Xiuli Kuang, Rusty Mann, Leila Ahmadinia, Lauren Alfiler, Fahimeh Baftizadeh, Katherine Baker, Sarah Bannick, Darren Bertagnolli, Kris Bickley, Phil Bohn, Dillan Brown, Jasmine Bomben, Krissy Brouner, Chao Chen, Kai Chen, Maggie Chvilicek, Forrest Collman, Tanya Daigle, Tim Dawes, Rebecca de Frates, Nick Dee, Maxwell DePartee, Tom Egdorf, Laila El-Hifnawi, Rachel Enstrom, Luke Esposito, Colin Farrell, Rohan Gala, Andrew Glomb, Clare Gamlin, Amanda Gary, Jeff Goldy, Hong Gu, Kristen Hadley, Mike Hawrylycz, Alex Henry, Dijon Hill, Karla E Hirokawa, Zili Huang, Katelyn Johnson, Zoe Juneau, Sara Kebede, Lisa Kim, Changkyu Lee, Phil Lesnar, Anan Li, Andrew Glomb, Yaoyao Li, Elizabeth Liang, Katie Link, Michelle Maxwell, Medea McGraw, Delissa A McMillen, Alice Mukora, Lindsay Ng, Thomas Ochoa, Aaron Oldre, Daniel Park, Christina Alice Pom, Zoran Popovich, Lydia Potekhina, Ram Rajanbabu, Shea Ransford, Melissa Reding, Augustin Ruiz, David Sandman, La'akea Siverts, Kimberly A Smith, Michelle Stoecklin, Josef Sulc, Michael Tieu, Jonathan Ting, Jessica Trinh, Sara Vargas, Dave Vumbaco, Miranda Walker, Micheal Wang, Adrian Wanner, Jack Waters, Grace Williams, Julia Wilson, Wei Xiong, Ed Lein, Jim Berg, Brian Kalmbach, Shenqin Yao, Hui Gong, Qingming Luo, Lydia Ng, Uygar Sümbül, Tim Jarsky, Zizhen Yao, Bosiljka Tasic, and Hongkui Zeng. Connecting single-cell transcriptomes to projectomes in mouse visual cortex. *bioRxiv*, November 2023.

[167] Nicholas A Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, Susu Chen, Jennifer Colonell, Richard J Gardner, Bill Karsh, Fabian Kloosterman, Dimitar Kostadinov, Carolina Mora-Lopez, John O'Callaghan, Junchol Park, Jan Putzeys, Britton Sauerbrei, Rik J J van Daal, Abraham Z Vollan, Shiwei Wang, Marleen Welkenhuysen, Zhiwen Ye, Joshua T Dudman, Barundeb Dutta, Adam W Hantman, Kenneth D Harris, Albert K Lee, Edvard I Moser, John O'Keefe, Alfonso Renart, Karel Svoboda, Michael Häusser, Sebastian Haesler, Matteo Carandini, and Timothy D Harris. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539), April 2021.

[168] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, 14(9):865–868, September 2017.

[169] Jessica Storer, Robert Hubley, Jeb Rosen, Travis J Wheeler, and Arian F Smit. The dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA*, 12(1):2, January 2021.

[170] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of Single-Cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

[171] M Luisa Suárez-Solá, Francisco J González-Delgado, Mercedes Pueyo-Morlans, O Carolina Medina-Bolívar, N Carolina Hernández-Acosta, Miriam González-Gómez, and Gundela Meyer. Neurons in the white matter of the adult human neocortex. *Front. Neuroanat.*, 3:7, June 2009.

[172] Dongqing Sun, Zhaoyang Liu, Taiwen Li, Qiu Wu, and Chenfei Wang. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Res.*, 50(7):e42, April 2022.

[173] Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. A curated database reveals trends in single-cell transcriptomics. *Database*, 2020, November 2020.

[174] Andras Sziraki, Ziyu Lu, Jasper Lee, Gabor Banyai, Sonya Anderson, Abdulraouf Abdulraouf, Eli Metzner, Andrew Liao, Jason Banfelder, Alexander Epstein, Chloe Schaefer, Zihan Xu, Zehao Zhang, Li Gan, Peter T Nelson, Wei Zhou, and Junyue Cao. A global view of aging and alzheimer's pathogenesis-associated cell population dynamics and molecular signatures in human and mouse brains. *Nat. Genet.*, 55(12):2104–2116, December 2023.

[175] Adriaan J Taal, Ilke Uguz, Sabina Hillebrandt, Chang-Ki Moon, Victoria Andino-Pavlovsky, Jaebin Choi, Changmin Keum, Karl Deisseroth, Malte C Gather, and Kenneth L Shepard. Optogenetic stimulation probes with single-neuron resolution based on

organic LEDs monolithically integrated on CMOS. *Nature Electronics*, 6(9):669–679, August 2023.

[176] Norio Takata and Hajime Hirase. Cortical layer 1 and layer 2/3 astrocytes exhibit distinct calcium dynamics in vivo. *PLoS One*, 3(6):e2525, June 2008.

[177] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–382, May 2009.

[178] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, Osnat Penn, Trygve Bakken, Vilas Menon, Jeremy Miller, Olivia Fong, Karla E Hirokawa, Kanan Lathia, Christine Rimorin, Michael Tieu, Rachael Larsen, Tamara Casper, Eliza Barkan, Matthew Kroll, Sheana Parry, Nadiya V Shapovalova, Daniel Hirschstein, Julie Pendergraft, Heather A Sullivan, Tae Kyung Kim, Aaron Szafer, Nick Dee, Peter Groblewski, Ian Wickersham, Ali Cetin, Julie A Harris, Boaz P Levi, Susan M Sunkin, Linda Madisen, Tanya L Daigle, Loren Looger, Amy Bernard, John Phillips, Ed Lein, Michael Hawrylycz, Karel Svoboda, Allan R Jones, Christof Koch, and Hongkui Zeng. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, November 2018.

[179] S Teichmann. Method of the year 2019: Single-cell multimodal omics. *Nat. Methods*, 17(1):1, January 2020.

[180] Huihui Tian, Ke Xu, Liang Zou, and Ying Fang. Multimodal neural probes for combined optogenetics and electrophysiology. *iScience*, 25(1):103612, January 2022.

[181] Luyi Tian, Fei Chen, and Evan Z Macosko. The expanding vistas of spatial transcriptomics. *Nat. Biotechnol.*, 41(6):773–782, June 2023.

[182] Mihail Ivilinov Todorov, Johannes Christian Paetzold, Oliver Schoppe, Giles Tetteh, Suprosanna Shit, Velizar Efremov, Katalin Todorov-Völgyi, Marco Düring, Martin Dichgans, Marie Piraud, Bjoern Menze, and Ali Ertürk. Machine learning analysis of whole mouse brain vasculature. *Nat. Methods*, 17(4):442–449, April 2020.

[183] Tyler J Torrico and Sunil Munakomi. *Neuroanatomy, Thalamus*. StatPearls Publishing, July 2023.

[184] Maria Antonietta Tosches, Tracy M Yamawaki, Robert K Naumann, Ariel A Jacobi, Georgi Tushev, and Gilles Laurent. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science*, 360(6391):881–888, May 2018.

[185] Audrey Touzot, Nuria Ruiz-Reig, Tania Vitalis, and Michèle Studer. Molecular control of two novel migratory paths for CGE-derived interneurons in the developing mouse brain. *Development*, 143(10):1753–1765, May 2016.

[186] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, 21(1):12, January 2020.

[187] Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.*, 10(1):2975, July 2019.

[188] Daniel Udvary, Philipp Harth, Jakob H Macke, Hans-Christian Hege, Christiaan P J de Kock, Bert Sakmann, and Marcel Oberlaender. The impact of neuron morphology on cortical network architecture. *Cell Rep.*, 39(2):110677, April 2022.

[189] Seralynne D Vann. Re-evaluating the role of the mammillary bodies in memory. *Neuropsychologia*, 48(8):2316–2327, July 2010.

[190] K B Walhovd, H Johansen-Berg, and R T Káradóttir. Unraveling the secrets of white matter–bridging the gap between cellular, animal and human imaging studies. *Neuroscience*, 276:2–13, September 2014.

[191] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, 11(3):333–337, March 2014.

[192] David Wang and Aisha Farhana. *Biochemistry, RNA Structure*. StatPearls Publishing, July 2023.

[193] Guiping Wang, Jeffrey R Moffitt, and Xiaowei Zhuang. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci. Rep.*, 8(1):4847, March 2018.

[194] Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, Tim Dolbeare, Brandon Blanchard, Nick Dee, Wayne Wakeman, Karla E Hirokawa, Aaron Szafer, Susan M Sunkin, Seung Wook Oh, Amy Bernard, John W Phillips, Michael Hawrylycz, Christof Koch, Hongkui Zeng, Julie A Harris, and Lydia Ng. The allen mouse brain common coordinate framework: A 3D reference atlas. *Cell*, 181(4):936–953.e20, May 2020.

[195] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), July 2018.

[196] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009.

[197] Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-Cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, June 2019.

[198] A E Wiencken-barger and V A Casagrande. Visual system development and neural activity. In V S Ramachandran, editor, *Encyclopedia of the Human Brain*, pages 791–804. Academic Press, New York, January 2002.

[199] Yuan-Ting Wu, Hannah C Bennett, Uree Chon, Daniel J Vanselow, Qingguang Zhang, Rodrigo Muñoz-Castañeda, Keith C Cheng, Pavel Osten, Patrick J Drew, and Yongsoo Kim. Quantitative relationship between cerebrovascular network and neuronal cell types in mice. *Cell Rep.*, 39(12):110978, June 2022.

[200] Benyi Xiong, Anan Li, Yang Lou, Shangbin Chen, Ben Long, Jie Peng, Zhongqin Yang, Tonghui Xu, Xiaoquan Yang, Xiangning Li, Tao Jiang, Qingming Luo, and Hui Gong. Precise cerebral vascular atlas in stereotaxic coordinates of whole mouse brain. *Front. Neuroanat.*, 11:128, December 2017.

[201] Nan Xu, Theodore J LaGrow, Nmachi Anumba, Azalea Lee, Xiaodi Zhang, Behnaz Yousefi, Yasmine Bassil, Gloria P Clavijo, Vahid Khalilzad Sharghi, Eric Maltbie, Lisa Meyer-Baese, Maysam Nezafati, Wen-Ju Pan, and Shella Keilholz. Functional connectivity of the brain across rodents and humans. *Front. Neurosci.*, 16:816331, March 2022.

[202] Naoki Yamawaki and Gordon M G Shepherd. Synaptic circuit organization of motor corticothalamic neurons. *J. Neurosci.*, 35(5):2293–2307, February 2015.

[203] Jian Yang, Ming Hao, Xiaoyang Liu, Zhijiang Wan, Ning Zhong, and Hanchuan Peng. FMST: an automatic neuron tracing method based on fast marching and minimum spanning tree. *Neuroinformatics*, 17(2):185–196, April 2019.

[204] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, A Sina Booeshaghi, Ricky S Adkins, Andrew I Aldridge, Seth A Ament, Antonio Pinto-Duarte, Anna Bartlett, M Margarita Behrens, Koen Van den Berge, Darren Bertagnolli, Tommaso Biancalani, Héctor Corrada Bravo, Tamara Casper, Carlo Colantuoni, Heather Creasy, Kirsten Crichton, Megan Crow, Nick Dee, Elizabeth L Dougherty, Wayne I Doyle, Sandrine Dudoit, Rongxin Fang, Victor Felix, Olivia Fong, Michelle Giglio, Jeff Goldy, Mike Hawrylycz, Hector Roux de Bézieux, Brian R Herb, Ronna Hertzano, Xiaomeng Hou, Qiwen Hu, Jonathan Crabtree, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Yang Eric Li, Jacinta D Lucero, Chongyuan Luo, Anup Mahurkar, Delissa McMillen, Naeem Nadaf, Joseph R Nery, Sheng-Yong Niu, Joshua Orvis, Julia K Osteen, Thanh Pham, Olivier Poirion, Sebastian Preissl, Elizabeth Purdom, Christine Rimorin, Davide Risso, Angeline C Rivkin, Kimberly Smith, Kelly Street, Josef Sulc, Thuc Nghi Nguyen, Michael Tieu, Amy Torkelson, Herman Tung, Eeshit Dhaval Vaishnav, Valentine Svensson, Charles R Vanderburg, Vasilis Ntranos, Cindy van Velthoven, Xinxin Wang, Owen R White, Z Josh Huang, Peter V Kharchenko, Lior Pachter, John Ngai, Aviv Regev, Bosiljka Tasic, Joshua D Welch, Jesse Gillis, Evan Z Macosko, Bing Ren, Joseph R Ecker, Hongkui Zeng, and Eran A Mukamel. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. March 2020.

[205] Zizhen Yao, Cindy T J van Velthoven, Michael Kunst, Meng Zhang, Delissa McMillen, Changkyu Lee, Won Jung, Jeff Goldy, Aliya Abdelhak, Matthew Aitken, Katherine Baker, Pamela Baker, Eliza Barkan, Darren Bertagnolli, Ashwin Bhandiwad, Cameron Bielstein, Prajal Bishwakarma, Jazmin Campos, Daniel Carey, Tamara Casper, Anish Bhaswanth Chakka, Rushil Chakrabarty, Sakshi Chavan, Min Chen, Michael Clark, Jennie Close, Kirsten Crichton, Scott Daniel, Peter DiValentin, Tim Dolbeare, Lauren Ellingwood, Elysha Fiabane, Timothy Fliss, James Gee, James Gerstenberger, Alexandra Glandon, Jessica Gloe, Joshua Gould, James Gray, Nathan Guilford, Junitta Guzman, Daniel Hirschstein, Windy Ho, Marcus Hooper, Mike Huang, Madie Hupp, Kelly Jin, Matthew Kroll, Kanan Lathia, Arielle Leon, Su Li, Brian Long, Zach Madigan, Jessica Malloy, Jocelin Malone, Zoe Maltzer, Naomi Martin, Rachel McCue, Ryan McGinty, Nicholas Mei, Jose Melchor, Emma Meyerdierks, Tyler Mollenkopf, Skyler Moonsman, Thuc Nghi Nguyen, Sven Otto, Trangthanh Pham, Christine Rimorin, Augustin Ruiz, Raymond Sanchez, Lane Sawyer, Nadiya Shapovalova, Noah Shepard, Cliff Slaughterbeck, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Nasmil Valera Cuevas, Shane Vance, Katherine Wadhwani, Katelyn Ward, Boaz Levi, Colin Farrell, Rob Young, Brian Staats, Ming-Qiang Michael Wang, Carol L Thompson, Shoaib Mufti, Chelsea M Pagan, Lauren Kruse, Nick Dee, Susan M Sunkin, Luke Esposito, Michael J Hawrylycz, Jack Waters, Lydia Ng, Kimberly Smith, Bosiljka Tasic, Xiaowei Zhuang, and Hongkui Zeng. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature*, 624(7991):317–332, December 2023.

[206] Zizhen Yao, Cindy T J van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, Song-Lin Ding, Olivia Fong, Emma Garren, Alexandra Glandon, Nathan W Gouwens, James Gray, Lucas T Graybuck, Michael J Hawrylycz, Daniel Hirschstein, Matthew Kroll, Kanan Lathia, Changkyu Lee, Boaz Levi, Delissa McMillen, Stephanie Mok, Thanh Pham, Qingzhong Ren, Christine Rimorin, Nadiya Shapovalova, Josef Sulc, Susan M Sunkin, Michael Tieu, Amy Torkelson, Herman Tung, Katelyn Ward, Nick Dee, Kimberly A Smith, Bosiljka Tasic, and Hongkui Zeng. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241.e26, June 2021.

[207] Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, July 2022.

[208] Hongkui Zeng and Joshua R Sanes. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.*, 18(9):530–546, September 2017.

[209] Meng Zhang, Xingjie Pan, Won Jung, Aaron R Halpern, Stephen W Eichhorn, Zhiyun Lei, Limor Cohen, Kimberly A Smith, Bosiljka Tasic, Zizhen Yao, Hongkui Zeng, and Xiaowei Zhuang. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature*, 624(7991):343–354, December 2023.

[210] Yingkun Zhang, Xinrui Lin, Zhixian Yao, Di Sun, Xin Lin, Xiaoyu Wang, Chaoyong

Yang, and Jia Song. Deconvolution algorithms for inference of the cell-type composition of the spatial transcriptome. *Comput. Struct. Biotechnol. J.*, 21:176–184, 2023.

[211] Yun-Feng Zhang, Luigim Vargas Cifuentes, Katherine N Wright, Janardhan P Bhattarai, Julia Mohrhardt, David Fleck, Emma Janke, Chunjie Jiang, Suna L Cranfill, Nitsan Goldstein, Mary Schreck, Andrew H Moberly, Yiqun Yu, Benjamin R Arenkiel, J Nicholas Betley, Wenqin Luo, Johannes Stegmaier, Daniel W Wesson, Marc Spehr, Marc V Fuccillo, and Minghong Ma. Ventral striatal islands of calleja neurons control grooming in mice. *Nat. Neurosci.*, 24(12):1699–1710, December 2021.

[212] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*, 2(3):lqaa078, September 2020.

[213] Jixing Zhong, Gen Tang, Jiacheng Zhu, Weiying Wu, Ge Li, Xiumei Lin, Langchao Liang, Chaochao Chai, Yuying Zeng, Feiyue Wang, Lihua Luo, Jiankang Li, Fang Chen, Zhen Huang, Xiuqing Zhang, Yu Zhang, Hongde Liu, Xin Qiu, Shengping Tang, and Dongsheng Chen. Single-cell brain atlas of parkinson's disease mouse model. *J. Genet. Genomics*, 48(4):277–288, April 2021.

[214] Fangxu Zhou, Bohao Chen, Xi Chen, and Hua Han. Neuronal morphological Model-Driven image registration for serial electron microscopy sections. *Front. Hum. Neurosci.*, 16:846599, May 2022.

[215] Hang Zhou, Tingting Cao, Tian Liu, Shijie Liu, Lu Chen, Yijun Chen, Qing Huang, Wei Ye, Shaoqun Zeng, and Tingwei Quan. Super-resolution segmentation network for reconstruction of packed neurites. *Neuroinformatics*, 20(4):1155–1167, October 2022.

[216] Jingtian Zhou, Zhuzhu Zhang, May Wu, Hanqing Liu, Yan Pang, Anna Bartlett, Zihao Peng, Wubin Ding, Angeline Rivkin, Will N Lagos, Elora Williams, Cheng-Ta Lee, Paula Assakura Miyazaki, Andrew Aldridge, Qiurui Zeng, J L Angelo Salinda, Naomi Claffey, Michelle Liem, Conor Fitzpatrick, Lara Boggeman, Zizhen Yao, Kimberly A Smith, Bosiljka Tasic, Jordan Altshul, Mia A Kenworthy, Cynthia Valadon, Joseph R Nery, Rosa G Castanon, Neelakshi S Patne, Minh Vu, Mohammad Rashid, Matthew Jacobs, Tony Ito, Julia Osteen, Nora Emerson, Jasper Lee, Silvia Cho, Jon Rink, Hsiang-Hsuan Huang, António Pinto-Duartec, Bertha Dominguez, Jared B Smith, Carolyn O'Connor, Hongkui Zeng, Shengbo Chen, Kuo-Fen Lee, Eran A Mukamel, Xin Jin, M Margarita Behrens, Joseph R Ecker, and Edward M Callaway. Brain-wide correspondence of neuronal epigenomics and distant projections. *Nature*, 624(7991):355–365, December 2023.

[217] Liang Zou, Huihui Tian, Shouliang Guan, Jianfei Ding, Lei Gao, Jinfen Wang, and Ying Fang. Self-assembled multifunctional neural probes for precise integration of optogenetics and electrophysiology. *Nat. Commun.*, 12(1):1–9, October 2021.

[218] Györgyi Zséli, Barbara Vida, Anett Szilvásy-Szabó, Mónika Tóth, Ronald M Lechan, and Csaba Fekete. Neuronal connections of the central amygdalar nucleus with

refeeding-activated brain areas in rats. *Brain Struct. Funct.*, 223(1):391–414, January 2018.

[219] Songpeng Zu, Yang Eric Li, Kangli Wang, Ethan J Armand, Sainath Mamde, Maria Luisa Amaral, Yuelai Wang, Andre Chu, Yang Xie, Michael Miller, Jie Xu, Zhaoning Wang, Kai Zhang, Bojing Jia, Xiaomeng Hou, Lin Lin, Qian Yang, Seoyeon Lee, Bin Li, Samantha Kuan, Hanqing Liu, Jingtian Zhou, Antonio Pinto-Duarte, Jacinta Lucero, Julia Osteen, Michael Nunn, Kimberly A Smith, Bosiljka Tasic, Zizhen Yao, Hongkui Zeng, Zihan Wang, Jingbo Shang, M Margarita Behrens, Joseph R Ecker, Allen Wang, Sebastian Preissl, and Bing Ren. Single-cell analysis of chromatin accessibility in the adult mouse brain. *Nature*, 624(7991):378–389, December 2023.