Enabling Scalable Neurocartography: Images to Graphs for Discovery

by

William Roberts Gray Roncal

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

December, 2016

Abstract

In recent years, advances in technology have enabled researchers to ask new questions predicated on the collection and analysis of big datasets that were previously too large to study. More specifically, many fundamental questions in neuroscience require studying brain tissue at a large scale to discover emergent properties of neural computation, consciousness, and etiologies of brain disorders. A major challenge is to construct larger, more detailed maps (e.g., structural wiring diagrams) of the brain, known as *connectomes*.

Although raw data exist, obstacles remain in both algorithm development and scalable image analysis to enable access to the knowledge within these data volumes. This dissertation develops, combines and tests state-of-the-art algorithms to estimate graphs and glean other knowledge across six orders of magnitude, from millimeter-scale magnetic resonance imaging to nanometer-scale electron microscopy.

This work enables scientific discovery across the community and contributes to the tools and services offered by *NeuroData* and the *Open Connectome Project*. Contributions include creating, optimizing and evaluating the first known fully-automated brain graphs in electron microscopy data and magnetic resonance imaging data; pioneering approaches

ABSTRACT

to generate knowledge from X-Ray tomography imaging; and identifying and solving a variety of image analysis challenges associated with building graphs suitable for discovery. These methods were applied across diverse datasets to answer questions at scales not previously explored.

Primary Reader: Gregory D. Hager

Secondary Readers: Randal Burns and Joshua T. Vogelstein

Acknowledgments

Completing the Ph.D. journey over the past six years has required navigating a tough road filled with many challenges, but also with good friends and strong advocates. I would like to acknowledge the following people who have made this path possible and a lot more fun.

Thanks to my APL family, who were infinitely flexible in making a crazy Ph.D. schedule work, supported me financially, and always advocated for my success. A special thank you to Cyndi Utterback, Hilary Hershey and Bart Paulhamus, who supported me through personal and academic challenges and helped me to grow. Also thanks to Nathan Drenkow for his computer vision expertise and problem-solving; Mike Pekala for his deep learning wizardry; and Dean Kleissas, my brain mapping partner-in-crime who was always willing to listen to me talk and discovered many amazing things about neuroscience with me.

I want to thank the professors who have mentored me and taught me to do research over the years: Jacob Vogelstein, for convincing me that brains were cool, teaching me how to look at the bigger picture, and introducing me to the campus community; Dzung Pham and Pilou Bazin for first giving me research opportunities; Jerry Prince for teaching me

about MRI research and pipelines; Carey Priebe for teaching me about errorful graphs and sharing much wisdom; Austin Reiter and Misha Kazhdan for teaching me about computer vision and graphics and being great advocates; Mark Chevillet for helping to shape my ideas and turn them into reality; Mounya Elhilali who always had encouraging words to share; and Marsha Ottem, my high school chemistry teacher who initially cemented my passion for scientific discovery.

My sincere gratitude to my committee members and mentors, including Joshua Vogelstein, the highest degree vertex I have met, who provided support, challenged me to push harder, learn more and graduate quickly; Randal Burns, who taught me to be a computer scientist and has been an advocate and source of wisdom and expertise; and Greg Hager, who taught me to approach unknown problems and helped me to become a better researcher, thinker, and writer.

Thanks to my JHU and NeuroData family: Priya Manavalan for helping make my algorithms actually work; Ayushi Sinha for her optimism and good ideas; Kunal Lillaney for providing constant technical support and endless opinions, some of which I thankfully listened to; Jordan Matelsky for his quick puns, lightning-fast coding, and thoughtful science; Disa Mhembere, Vince Lyzinski, Cenchen Shen, Shangsi Wang, and Da Zheng who knew what to do with all those graphs I made; Alex Eusman for always being willing to help and try new things; Kwame Kutten for his thoughtful analysis and putting everything in the right coordinate frame; John Bogovic for his mentoring and helping me learn how to actually be a grad student; Manolis Tsakiris for his encouragement and tutoring;

ACKNOWLEDGMENTS

Eric Bridgeford for being an awesome pipeliner; Eva Dyer for always being positive and teaching me about how to make big science possible; Anish Simhal for his mutual love of finding synapses and support; Jesse Pastolic for always being willing to help; Tyler Tomita for making amazing classifiers; Alex Baden for his help with all things viz and processing; Colin Lea, for sharing his computer vision expertise and strategy about spines; and Greg Kiar, a fellow engineer, co-teacher, and connectome creator, who routinely celebrates and commiserates with me, often during the same occasion.

Thanks also to my friends and colleagues in the broader connectomics community who have taught me so much and helped to put my work in context while having fun, especially Bobby Kasthuri who supported me from the beginning and always had a crazy, uplifting story to share; Verena Kaynig who taught me all about finding synapses and introduced me to many exciting research problems; Stuart Berg who helped me think broadly and builds amazing tools for discovery; and Steve Plaza who is always up for finding ways to make better networks and for a great conversation. Speaking of conversations...everyone talks about making brain graphs all the time, right? This is normal? Thanks to everyone for a constant stream of lively discussions over a free happy hour, coffee, or walk around campus.

To my non-brain mapping friends and extended family who have been patient and supportive through this journey – I really appreciate your support and understanding, and I promise to have time to visit (electronically or in person) very soon. Thanks for keeping me grounded and sharing a place to stay (whether in a house or a tent), an afternoon with

your family, a supportive phone call, or a good meal. A special thanks to John Stapleton and Scott Moeller for their wisdom about completing the Ph.D. journey.

Thanks to my family for their unwavering support, and for always loving me just the way I am: my dad for always listening to me share my news, giving me encouragement and teaching me that science is cool; my mom, who has been a fierce defender of all things me, and who has edited everything I've ever written (including this dissertation); Philip, for always willing to jump in a car (or on a bike) to help out or hang out; Katherine for helping to save the world and loving me and challenging me at the same time; Maria for lots of love, infinite Peruvian cooking, and watching out for me; Carlos for welcoming me into the family and sharing lots of good food; Miguel for being the best bro-lo and helping me to keep life in perspective; and Mapita for an amazing friendship and always laughing at my jokes, even when no one else would.

Thanks to our 156 College Prep students over the past eight years, who taught me the value of perseverance and reminded me to just keep swimming.

Finally, thanks to Karla, my best friend, without whom none of this would have been possible. Your inspiration, love, and perspective have kept me going, and your creativity, patience, and support have helped me to be a better person and a better researcher.

Dedication

To my parents, Deane and John, who empowered me and taught me how to dream, and

∞

To my wife, Karla, whose encouragement and inspiration makes our dreams a reality –
cada día un poco mas . . .

Contents

List of Tables

LIST OF TABLES

List of Figures

Chapter 1

Introduction

Thesis Statement: Enabling large scale, quantitative knowledge about brain structure requires fundamentally new analysis approaches to extract information at different imaging resolutions.

1.1 Problem Statement

Recent technical progress allows neuro-experimentalists to collect ever more detailed and informative anatomical and physiological imaging data from brains of all sizes and species. These datasets span experimentally accessible spatiotemporal scales, ranging from nanometer to meter, and millisecond to monthly sampling rates. In classical neuroscientific experimental paradigms, it was feasible for neuroscientists to draw their results on paper. In contrast, many modern neuroscientific experimental paradigms break the classic

analog data analysis workflow. In particular, these large, digital datasets create significant challenges for the brain mapping community at every step of the data analysis pipeline, including storing, exploring, analyzing, and modeling the data.

Neuroscience is now entering the age of big data as laboratories from around the world begin to acquire information that exceeds their storage and computational capabilities. Data can come from a dizzying variety of experimental paradigms, ranging from serial electron microscopy to calcium imaging to multimodal magnetic resonance imaging. For many of these paradigms, data are massive three-dimensional (3D) image stacks. However, the existing computational solutions for big image datasets are often designed for many small images (rather than a single large volume), and therefore do not meet the requirements for these data. Moreover, neuroscience is rife with higher dimensional (>3D) data, as certain modalities have many different channels (e.g., array tomography datasets might have dozens of channels), and include functional data (which might have many thousands of time steps).

All of these imaging datasets contain troves of unexploited information, which are waiting to be extracted by the community, and which will reveal the underlying principles of normal, abnormal and exceptional mental function.

These advances require a paradigm shift in neuroscience from the analysis of single slides to large data volumes. Large volumes offer the raw information necessary to develop and test hypotheses at the scale of a cortical circuit. However this analysis requires new approaches that leverage automation, rather than annotations from expert neuroanatomists.

In this work, we will focus on *neurocartography*, the process of extracting knowledge about the structure of the brain [1]. It is possible to learn this information at many different scales, from lower-resolution magnetic resonance imaging that highlights major pathways between brain regions to high-resolution electron microscopy that illuminates each neuron and its chemical synaptic connections. This work highlights solutions to common challenges associated with data processing (e.g., end-to-end workflows, metrics), as well as to unique challenges specific to each imaging modality.

1.2 Graphs

Brain maps are commonly represented by *graphs*, which are a type of mathematical object. More explicitly, graphs can be represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where the vertices (V), edges (E) and attributes (A) vary depending on the imaging modality and application. These graphs can also be represented as an adjacency matrix, where vertices are represented as rows and columns of the matrix and edges at the intersection point between vertex pairs. The ability to estimate a connectome (i.e., a description of functional or structural connectivity in the brain of an individual), promises advances in many areas, from personalized medicine, to learning and education, and even to intelligence analysis [2, 3].

1.3 Scope

This dissertation summarizes our contributions in the areas of image analysis, neuroscience, and engineering. This work has concentrated on developing and deploying new paradigms for discovering knowledge at a scale not traditionally explored in neuroscience, enabling a new approach for analysis. We approach neurocartography from this perspective, developing new methodologies and processing strategies.

We do not focus on the details of imaging acquisition methodologies or preprocessing steps. We also do not repeat the excellent overviews and prior work in connectomics, preferring instead to highlight the work most relevant to ours. Instead, we point the reader to overviews of the field of connectomics [4,5,6], imaging acquisition [7,8,9], and contemporary challenges and methods [10, 11, 12, 13]. The work presented here has been influential in making decisions related to storage system architectures and frameworks for data management and retrieval, but those implementations are not the primary focus of this work.

The brain is composed of many individual parts, including neurons, glial cells, and blood vessels. In this work, we are principally concerned with estimating connectivity in the brain between *neuronal* cells, which connect at junctions known as *synapses*. Although a full understanding of our brains will require genomic, functional, and structural analyses (among others), we will focus on fundamental questions surrounding structural connectivity at different scales.

1.4 Background

In this work, we discuss novel approaches to estimate anatomical maps from large datasets across six orders of magnitude, ranging from magnetic resonance imaging (millimeter resolution) to electron microscopy (nanometer resolution). At each resolution we are able to estimate different aspects of structural brain connectivity, such as connectivity between regions, locations of blood vessels, cell bodies, and high-fidelity neuron-synapse brain graphs. Estimating connectomes is a key component of neurocartography focused on reconstructing brain circuits at different resolutions [6]. The available information for analysis is driven by the spatial resolution achieved through the combination of imaging sensors and processing methods.

The following sections provide a brief overview of various methods used to create brain maps at different scales. We provide context for the work presented in this dissertation and explain tradeoffs between data set size, resolution, processing methods, and available knowledge.

1.5 Macroscale

First, we will consider brain mapping at the macroscale, specifically diffusion Magnetic Resonance Imaging (dMRI) methods of inferring brain connectivity. In this modality, the available voxel size is approximately $1\,mm^3$, and so most of the finer details explored at lower resolution are unavailable. Indeed, a single voxel in MRI data contains approximately

50,000 neurons.

Nevertheless, many important details of brain connectivity can be learned from this data, despite the low signal-to-noise ratio. Critically, this data can be acquired *in-vivo*, and so we can ask scientific questions in living humans with covariates that directly address questions like learning and disease. In these brain graphs, regions of the brain are labeled with an atlas, generally based on structural or functional demarcations. The regions are graph vertices; edges are represented by fiber streamlines, which are estimated by observing the anisotropic diffusion of water in the brain (which correlates with the location of major axonal bundles transmitting information in the brain). An example of the raw data is shown in Figure 1.1.

1.5.1 Challenges

When considering approaches to extract knowledge from image volumes, it is important to consider the overall scientific or analysis objective. Many existing methods focus on a single part of the conceptual pipeline to translate raw images to knowledge (e.g., data acquisition, preprocessing, analysis, inference) and therefore do not always measure the quality of an individual step (e.g., algorithm, component) with the overall end-goal in mind. In MRI-based connectome estimation, many methods exist for various subfunctions of the processing pipeline, but there is no end-to-end system to automatically estimate or assess graphs. Moreover, because MRI-based connectomes are only able to provide a coarse estimate of connectivity, no clear assessment approach or metric exists to determine

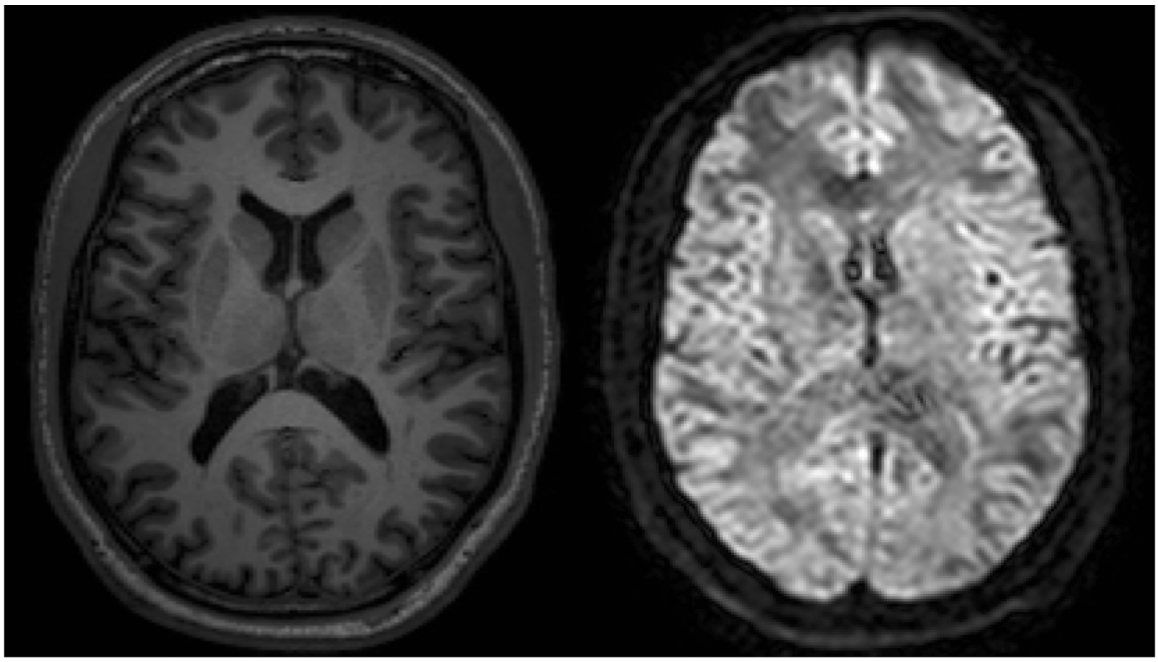

Figure 1.1: *An example slice of MRI Data.* (Left) MPRAGE anatomical data. (Right) Diffusion Tensor Imaging data. Each voxel is $\sim 1mm^3$; a putative cortical column is represented by 1 byte on disk.

whether putative brain graphs are accurate. Finally, using the graphs in a classification setting to predict cognitive or other covariates is still largely unexplored.

1.5.2 Contributions

We designed and developed the first known end-to-end automated pipeline dedicated to MR Connectome estimation [14]. Subsequent tools improved the scability and reliability of this work [15] and ultimately resulted in an open-source Python package called *ndmg* [16]. We developed various metrics for graph assessment, centered on the reliability of test-

retest data, which assesses the quality of the pipeline by comparing the reproducibility of brain graphs estimated from different scans of the same person. The pipeline was used to automatically create the largest known database of human brain graphs in the world, fueling several diverse research efforts [17, 18, 19, 20]. Finally, we developed analysis tools to look for patterns across subjects and completed proof-of-concept classification using a person's biological sex as a covariate.

1.6 Microscale

At this scale we will consider microscale approaches to brain mapping, typically including methods such as array tomography, CLARITY, X-ray microtomography, Optical Coherence Tomography, and Brainbow where each pixel (voxel) is between $100nm$ and $5\mu m$. This resolution allows for the observation of many coarse neuroanatomical features (e.g., cell bodies, blood vessels) and enables scientists to rapidly map large regions of brain tissue [21]. We focus on X-Ray microtomography, a new modality that allows for very rapid sample acquisition without requiring sectioning or subsequent alignment steps.

In X-Ray microtomography, neuron somata are typical vertices, and scientists can study brain properties such as region statistics and clustering of cells. Edges (connections) between cells are difficult to determine with certainty, but graphs (or distance matrices) can be inferred through proximity and sometimes by examining large or distinctive processes such as myelinated axons and apical dendrites. Cell types and gross cellular morphology

attributes can often be determined in this data. A small example slice of X-ray microtomography tissue is shown in Figure 1.2 [8].

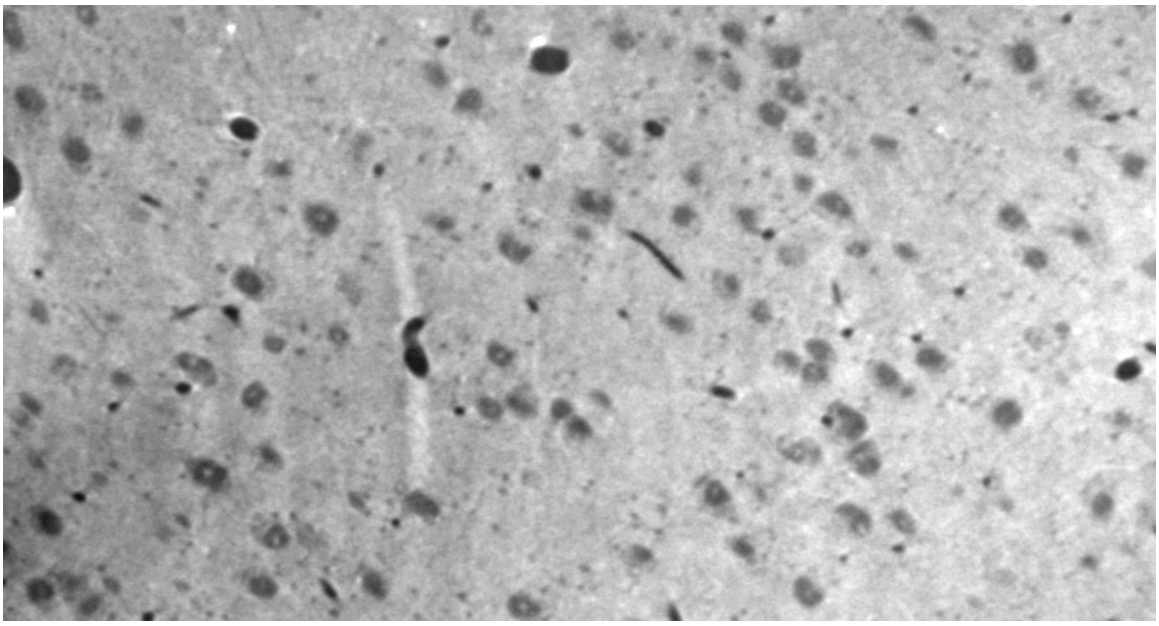

Figure 1.2: *An example slice of X-Ray Microtomography tissue.* Each voxel is $\sim$1 μm^3; a putative cortical column is represented by $\sim$1 gigabyte on disk.

1.6.1 Challenges

Current approaches to estimate large-scale maps of the brain often rely on microscale imaging methods. However, these methods are slow and often require extensive preprocessing and alignment to be useful. X-Ray Microtomography promises to greatly speed up the acquisition process, but no methods have been implemented to detect cells and vasculature in cortical data. Due to the limited spatial resolution, it can be challenging to segment individual objects and uniquely separate these objects from their surroundings.

In addition to producing high-quality results, it is important to develop tools and storage methods that can be standardized to promote reproducibility and extensibility. Providing high quality datasets and tools that allow others to leverage a scientific result is historically difficult for large neuroscience efforts, and may greatly impede progress, requiring multiple labs to do similar studies rather than utilizing a previous dataset.

1.6.2 Contributions

We present a new, scalable approach for storing and processing X-Ray Microtomography data, and pioneer scalable image analysis algorithms for this new type of data. We achieve excellent cell detection and vessel segmentation performance and demonstrate the applicability of our algorithms at large scale. Our approaches have been packaged into a toolbox for use in other applications, including future X-Ray imaging data collections. Finally, we leveraged our initial results to compute basic statistics on the data which agree with published literature, providing additional validation of our approach.

1.7 Nanoscale

A major focus of this dissertation is nanoscale connectomics, in which serial section electron microscopy (EM) is used to image small blocks of neuronal tissue. EM data produces image voxels with a resolution of about $3 - 4nm$ in the imaging plane and $30 - 50nm$ in the third dimension due to tissue sectioning limitations. These data are typically

acquired with Scanning Electron Microscopy (SEM) [22] or Transmission Electron Microscopy (TEM) [23] which offer significant advantages in imaging speed; however other approaches may offer higher resolution (e.g., FIBSEM).

Electron microscopy connectomics is focused on building "ground-truth" networks to search for local motifs and patterns, to examine local connectivity properties (e.g., Peter's rule) [22], and to test neuroscience hypotheses at a biofidelic level. This experimental modality is relatively novel, yet important because of its unique ability to reveal nanoscale anatomical structure [7].

In this setting, nanometer resolution enables the identification of individual neuronal cells as graph vertices and their synaptic connections as graph edges. Because we can observe sub-cellular architecture such as mitochondria and neurotransmitter-containing vesicles, we are able to assign rich attributes such as putative synaptic strength, cell type, direction, and path length (although assigning these attributes is not the focus of this work).

Researchers have mapped the nervous system for *C. Elegans* [24], a small nematode with just 302 neurons. More recent advances have explored properties of the visual cortex in *D. melanogaster* [25] and *M. musculus* [23, 26] at increasing scales; we expect to gain increasing knowledge about the circuitry of the brain as imaging and analysis methods are applied to larger volumes.

In this modality, one of the major challenges is scale; a trained human is capable of tracing the connections through images of an entire human brain given enough time, but actually completing this task is intractable. Even storing large brain volumes is challenging;

as mentioned above, just a cortical column (cubic millimeter) of brain data exceeds a petabyte. A small example slice of electron microscopy tissue is shown in Figure 1.3, representing approximately $24 \times 12 \times 0.03$ microns [22].

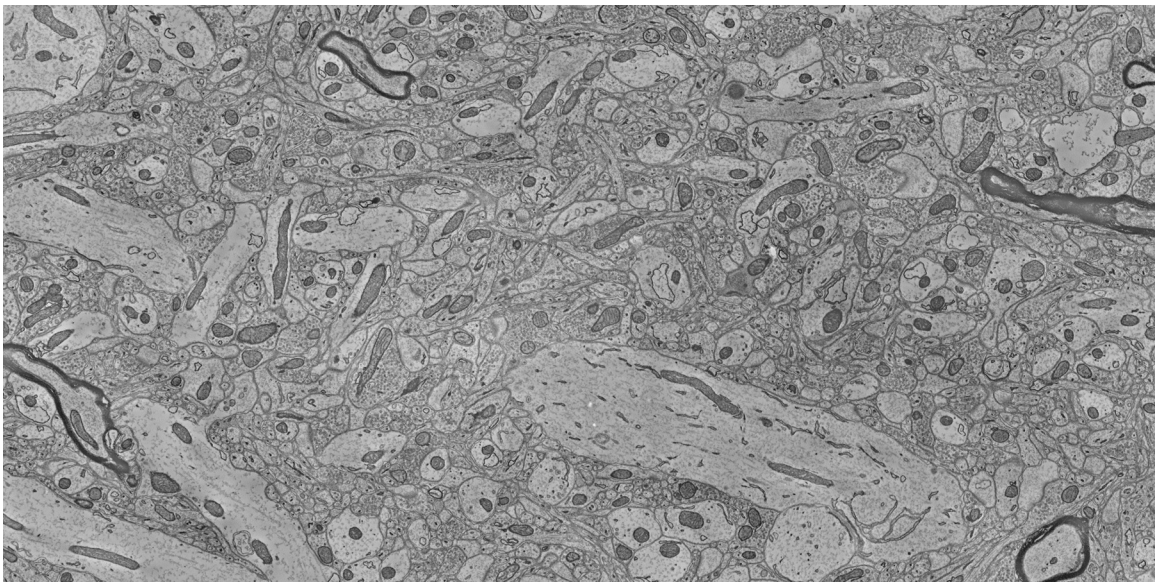

Figure 1.3: *An example slice of electron microscopy tissue.* Each voxel is $\sim 3 \times 3 \times 30\, nm^3$; a putative cortical column is represented by $\sim$ 2petabytes on disk.

1.7.1 Challenges

Using serial section electron microscopy, a human brain is estimated to be approximately 1 zetabyte (10^{21} voxels) on disk. Even a single cortical column (as proposed by Vernon Mountcastle) is approximately 2 petabytes (10^{15} voxels) at nanoscale resolution. As large volumes begin to be produced on a regular basis, new techniques will be required to store, process and analyze the samples in order to extract useful information. Such

datasets can grow by terabytes (TB) a day; multiple such datasets already exist with ~ 100 TB [26, 27]. Indeed, the collection of the first petascale neuroscience datasets are planned for the next three years [28]. Moreover, because serial EM data is single-channel (one color) and single-timestep (anatomical), addressing large-scale analysis on these datasets will be a crucial first step prior to addressing big data analysis on multi-spectral, multi-temporal, and multi-scale data.

Conventional methods are often designed to be run on a single image or small volume of data. Conventional approaches to large data processing often refer to the processing of many smaller one-dimensional (e.g., time) or two-dimensional (e.g., images). In contrast to this approach, large data neuroscience commonly deals with high dimensional data (4D volumes in MRI and potentially dozens of channels in modalities such as array tomography). Even in the simplest scenario, the demands are for automated, robust analysis of single very large volumes instead of many smaller volumes. As described above, this scalability challenge will grow in complexity and importance over the next few years.

Existing methods for creating EM graphs have high-error rates or require significant manual intervention. Algorithms exist to address key components of a conceptual pipeline, but currently operate in isolation. No approaches currently exist to automatically generate graphs or to measure overall graph performance.

1.7.2 Contributions

We present the first fully-automated approach to estimate brain graphs from electron microscopy (nanoscale) data as well as a novel metric for assessing connectivity. This more directly allows end-to-end assessment and allows us to identify and optimize components that most contribute to errors. At the heart of this work are novel image analysis solutions. Based on our initial quantification of graph errors, we identified several major problems that were integral to assessing graph connectivity. We built a state-of-the-art nanoscale synapse detector that includes a biologically-inspired, scalable random forest detector and a high-performing deep-learning method. We created the first known method to specifically link spines and shafts, which resulted in great improvements in graph error in small volumes.

We developed *NeuroData*, a platform for enabling large scale neuroscience for everyone. This ecosystem for storing, exploring, analyzing, and modeling data at large scale democratizes reproducible science and is easily accessible, even for those who may not have an extensive computer science background.

To assist in processing and retrieving data at scale, we developed a data standard called RAMON (Reusable Annotation Markup for Open Neuroscience), and created scalable analysis tools to extract knowledge from big neuroscience volumes. The code and results were released in a manner consistent with open science and can be straightforwardly extended and applied to new problems. *ndparse* provides many interfaces, algorithms and assessment tools in a Python package. We also exhibited our approach to scalable reproducible science by translating community-provided datasets to a standardized, accessible system.

1.8 Joint and Previously Published Work

Text and ideas from the following (published and draft) joint manuscripts have been incorporated in this dissertation as explained below. Papers are listed where the bulk of the text appears, but contributions may appear throughout this integrated work. These publications require a team-driven approach to science; my contributions to each paper are noted where appropriate. A full CV is provided at the end of this thesis.

- **Chapter 1:** unpublished *NeuroData Paper* [29] and *neurodata.io* contributions
- **Chapter 2:** *ndio SFN abstract:* supervised work [30]; *ndparse SFN abstract* [31].
- **Chapter 3:** *MR Connectomics:* Led research effort for *MRCAP* (c) IEEE 2010 [14] and *MIGRAINE* (c) IEEE 2013 [15] pipelines. Supervised and supported the development of the successor *ndmg* pipeline [16] with Greg Kiar; some of the work also appears in his Master's thesis [32].
- **Chapter 4:** *XBRAIN:* co-designed and co-deployed all methods for analyzing image volumes in a novel imaging modality [8].
- **Chapter 5:** *I2G:* Lead for graph error metrics; ran final experiments and led writing. Co-developed infrastructure and image analysis algorithms and pipelines [33]; *VESICLE:* Led research effort; developed VESICLE-RF [34]; *Santiago* [35].
- **Chapter 6:** *NeuroData Paper:* Led analysis and case study sections of paper and co-wrote manuscript draft. Provided input into overall ecosystem [29].
- **Chapter 7:** *Sample to Knowledge:* Co-developed idea and led manuscript [36].

1.9 Dissertation Outline

To extract maps of the brain, several major challenges need to be addressed, including developing processing pipelines and infrastructure, image analysis, and mechanisms for sharing science. The work presented here will help the community focus their efforts on the overall problem of scalable graph estimation. Our storage, processing, and assessment tools allow for modular improvements to be deployed while measuring the overall downstream impact of the change. As data volumes continue to increase in size, our image analysis methods and biologically inspired approaches provide algorithms for reliable, automated processing across many different modalities at large scale.

- **Chapter 1** provides an overview of this thesis and summarizes major contributions and context for the problems addressed in this dissertation.

- **Chapter 2** continues with a description of the challenges and solutions for enabling big data neurocartography, We use this chapter to introduce graph estimation and analysis tools that enable this work.

- **Chapter 3** builds on the approaches in earlier chapters to extend the ideas of big data processing to new domains, demonstrating generalizability of these methods and enabling novel scientific results. We particularly focus on new approaches to extract information from Magnetic Resonance Imaging (MRI) data and methods to assess graphs.

- **Chapter 4** presents methods for analyzing microscale data through a new data modality (X-Ray Microtomography) and develops robust methods to segment blood vessels and detect neurons.

- **Chapter 5** explores methods for generating knowledge in an integrated, end-to-end approach. We highlight our research to estimate the first known fully-automated electron microscopy graphs. We explore ways to quantify the resulting graph performance through a novel graph-focused metric, and use that result to build an optimal pipeline. We highlight synapse detection as a key machine vision problem and develop state-of-the-art approaches that trade-off scalability and performance. We conclude by identifying and characterizing the spine-shaft linking problem which is a major driver of graph error when assessing connectivity.

- **Chapter 6** We highlight the *NeuroData* ecosystem, which provides storage, exploration, analysis, and modeling tools for reproducible science. We explore an approach for extensible neurocartography; we apply our synapse detection techniques at scale and provide tools to test a scientific hypothesis of spatial uniformity.

- **Chapter 7** concludes with future work and a discussion of contributions and lessons learned, including a proposed paradigm for combining the tools that were developed into a multimodal, hierarchical approach.

Chapter 2

Background: Graph Estimation and Assessment

2.1 Overview

This chapter provides an introduction to the challenges, design decisions, and tools used throughout this dissertation. We focus on the process of estimating graphs from large image volumes and assessing the results. To making images-to-graphs a reality, we needed to overcome challenges in unifying **data standards**; improving the **accessibility** of large data volumes; creating **reproducible pipelines**; and developing **graph assessment** metrics.

In part to solve these challenges, we developed an ecosystem for storing, exploring, analyzing, and modeling diverse, large-scale neuroscience datasets, called *NeuroData*. This chapter summarizes the challenges and solutions for large-scale image analysis. An

integrated demonstration of the overall ecosystem is shown through two case studies in Chapter 6.

DATA STANDARDS

An acknowledged challenge in the connectomics field is annotation representation and its impact on software and institution-level interoperability [10,37]. As the field grows and data volumes increase, the sharing of data through remote and programmatic interfaces and the application of community developed algorithms and software will become common. However, research groups often currently re-engineer proprietary or lab-specific solutions for storing data that are not easily explained or implemented by outside users. This is especially true for nanoscale data; the macroscale community has made significant progress toward data standarization and laboratory information management systems [38,39,40].

ACCESSIBILITY

Conventionally, datasets were stored as tiles on disk, which were difficult to efficiently access for common processing tasks. Several research groups (e.g., DVID, Google) now have RESTful services to retrieve and store neuroscience data. However, taking full advantage of these ecosystems is still challenging, because each system has its own format and organization. Conventionally, neurocartography APIs are either non-existent (requiring users to develop their own solutions to interface with RESTful calls) or only provide interfaces to raw image and label data. This leads to stovepiped solutions that are difficult to

understand by other research groups and do not provide an easy way to combine solutions across the community. Many such algorithms exist, but progress is stymied because the community has lacked a way to combine these methods and evaluate the overall end-to-end impact of parameter changes.

REPRODUCIBLE PIPELINES

Much of the work in connectomics today has been developed to solve a particular problem that is important for graph estimation (e.g., neuron segmentation, synapse detection), but exists in isolation and is not currently integrated into an overall workflow where performance impacts on the downstream graph can be easily measured. Moreover, only limited work exists in translating community pipelining tools to the large data volumes of interest in contemporary neuroscience.

GRAPH METRICS

The graph theory community is eager to analyze neuroscience networks, but no clear approach exists to characterize the quality of the graphs produced by manual or automated methods, especially in the absence of ground truth. It is difficult to produce useful information about brain networks without being able to characterize performance or feedback information from downstream users.

We designed and implemented the first fully-automated pipelines for MR and EM-based connectomes, which we explore in detail in later chapters. This chapter provides

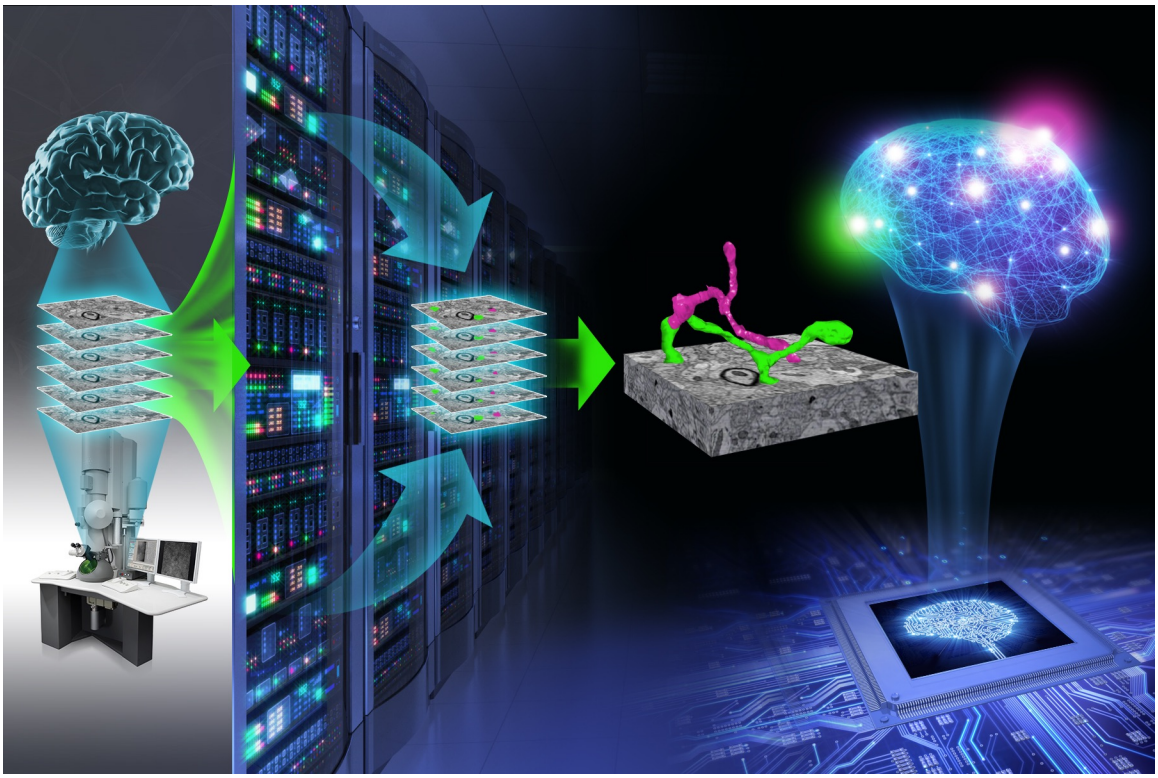

Figure 2.1: *Overview of the Images to Graphs challenge and putative pipeline.* This process begins with a brain and ends with the creation of novel, biofidelic algorithms. Image created by JHU/APL.

background and the key design decisions needed to achieve these goals. An illustration of the overall workflow is shown in Figure 2.1. These ideas have been tested across many different modalities and operating paradigms; the principles developed applied across a variety of use cases and are applicable to other users of large spatial data.

2.2 RAMON: Reusable Annotation Markup for Open Neuroscience

When developing large-scale image analysis methods, often it is desirable to incorporate data and algortihms from the broader research community. Furthermore, it is important to be able to store the results of this analysis in a standardized, accessible format for others to use. This is especially important as datasets scale in size, and repeating the underlying analysis may be computationally prohibitive. Answering this challenge requires scene parsing, rather than simply segmentation; the rich semantic annotations are critical to inferring graph structure and understanding the function and structure of neuronal circuits. We developed a standard for annotation metadata, as summarized in Table 2.1, which we call the Reusable Annotation Markup for Open Neuroscience (RAMON).

We developed RAMON to define a minimum set of annotation types and associated metadata that capture important biological information and represent the relationships between annotations that are critical for connectome generation and neuroscience exploration. Annotation metadata is trivially extensible through custom, user-defined key-value pairs. Because every group may have slightly different needs, this was not designed as a formal ontology; rather it facilitates the development of software and tools by providing a flexible, yet reliable, standard for representing annotation data. As an illustration, our synapse annotation type has metadata fields (e.g., weight, type, confidence, associated neural segments), and is extensible with arbitrary, searchable key-value pairs. By directly

writing (or converting) to this format, users can interact with dataset from many different research groups and modalities using the same data format. Moreover, scalable image analysis pipelines can be heterogenous and can use RAMON datatypes as interface and storage points to allow for more flexible processing and 'checkpoints' to return to as the workflows are optimized. *ndstore* implements the RAMON data standard and provides RESTful endpoints to access objects stored in *NeuroData*.

2.3 *ndio:* Neuroscience Data Input and Output

Our *ndio* API provides an easy-to-use, user-facing package that abstracts many of common tasks required for large-scale image analysis. One key design decision was to create classes for all RAMON annotation types (while still providing access to raw image and label data). This allows users to easily adapt existing algorithms that solve important pieces of the overall graph estimation pipeline by meeting flexible input and output endpoints. Our data-standard was developed with input from several of the major research groups involved in connectomics. Because we provide semantic labels and metadata for each annotation object, descriptive information is available to facilitate the reuse of this information, regardless of how it was generated (e.g., image analysis algorithms internal or external to a research lab, manual methods). Prior to this work, users were required to develop new tools to read in each dataset. This API, in conjunction with NeuroData, removes the barrier to access by directly providing access to over 100TB of diverse data

Table 1. An overview of the current RAMON object types. Each object is used to provide labels and attributes to objects identified in the neural tissue, facilitating interoperability and efficient data storage, retrieval, and query processing.

Type	**Description**
SYNAPSE	Junction between two NEURONs is used to connect SEGMENTs when building a GRAPH
ORGANELLE	Represents internal cell structures (e.g, mitochondria, vesicles)
SEGMENT	A labeled region of a neuron; typically a contiguous voxel set
NEURON	Container for assembling SEGMENTs
NODE	Sparse annotation format for tracing processes or objects.
SKELETON	An (organized) collection of NODEs, which are often used to represent a NEURON or arbor.
ROI	An attributed region of interest, often used for atlases and other collections of labels.
VOLUME	Used to store pixel label information; inherited by many other types
GENERIC	Extensible, used to specify arbitrary, user-defined information for a voxel set

using the same functions, allowing users to focus their resources on algorithm development and analysis tools and immediately translate their solutions to an interoperable

environment.

In addition to storing label data and metadata, the RAMON classes provide additional functionality such as tracking the annotation's global database location and visualizing a region of interest. Again, this abstracts many of the difficulties of working with large data, and allows researchers to apply a solution that works on a local volume to one of arbitrary size. The toolbox automatically handles compression, chunking, and batching of annotation data to optimize throughput and simplify software development. Finally, we have built (and are extending) tools to enable users to convert from one data storage format to another, so that the data storage challenges are separated from the large scale image analysis research. An illustration of *ndio* operations from an end-user perspective is shown in Figure 2.2.

Our API was implemented as a simple and easily-approachable software package. *ndio* is extensible and also offers easy-to-use wrappers to query many types of image and annotation data using *NeuroData* services. *ndio* provides a common vocabulary for big data neuroscience and allows researchers to focus on scientific discovery, abstracting many of the problems and common use cases associated with large data volumes. *ndio* is a successor to our earlier Application Programming Interface called Connectome Annotation for Joint Analysis of Large data (CAJAL), an open source MATLAB toolkit.

ndio is stable and is in active use by several teams across a variety of modalities, including electron microscopy, MRI, Array Tomography, and CLARITY. It facilitates access both between users and *NeuroData* services, and also between functions within *NeuroData*,

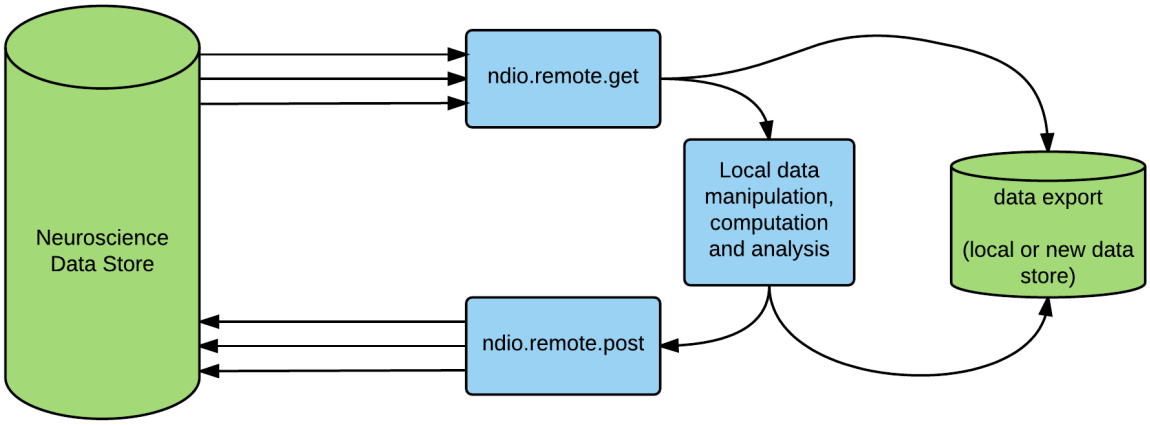

Figure 2.2: *A high-level view of ndio being used for scientific discovery.* Large data requests (gets and puts) are divided into smaller subvolumes in a process that is transparent to the end user. Data may be exported to common formats (e.g., numpy, hdf5, nifti) at any point using the *ndio* tools.

such as between *ndstore* and *ndviz*. We demonstrate use cases of these tools later in this dissertation, especially in the extensible neurocartography case study. It is publicly developed under a permissive, open-source license and can be installed from PyPI using a single command line call. We are able to trivially run powerful, client-side queries (e.g., identify orphan processes in a cuboid) and retrieve arbitrarily-sized cutouts in an efficient manner to support scalable image analysis.

2.4 Analyze

Once datasets have been ingested into a standard format using *ndstore* and explored using *ndviz*, we now seek to extract knowledge from the datasets using standard image processing and machine vision tools. This part of the *NeuroData* ecosystem, which encom-

passes preprocessing, manual labeling, algorithm development, and large scale computation is collectively referred to as *analyze*.

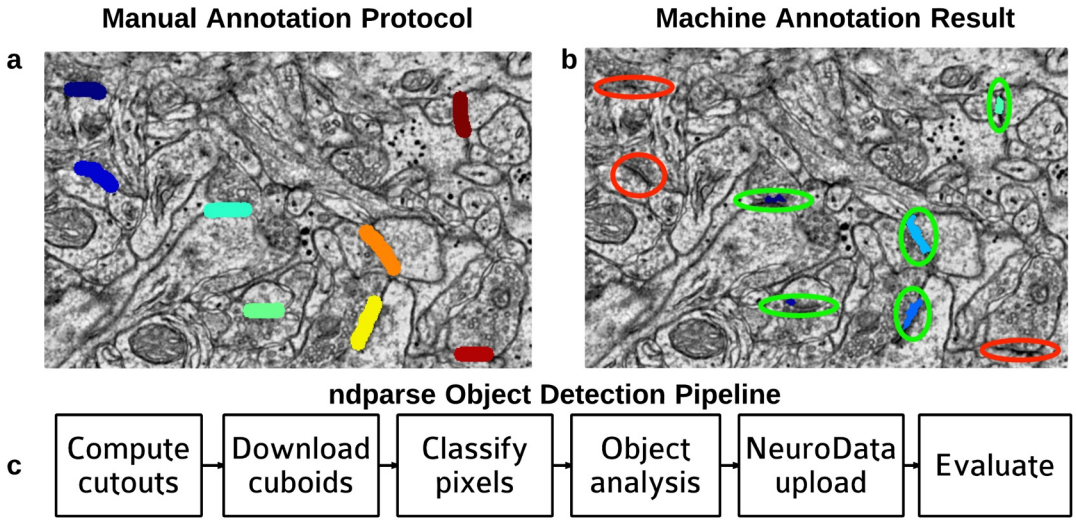

Figure 2.3: *NeuroData provides fully automatic terascale tools for processing.* (a) Example of synapses created with mana annotation tool. (b) Example of computer vision synapses created with *ndparse* . (c) Reference *deploy* object detection workflow, used to find synapses. Together these tools provide an integrated analysis environment for flexible, scientific discovery.

Given that the data has been sufficiently preprocessed (e.g., mosaiced, aligned, color corrected), we now proceed to parse the image volume "scene," assigning semantic labels (e.g., segment, synapse, neuron, mitochondria) to voxels. Parsing the scene can be divided into three main steps: (i) manually labeling a subset of training data, (ii) training and evaluating various machine vision algorithms, and (iii) deploying a chosen algorithm (and parameter set pair) at scale.

The number of data challenges and algorithms related to connectomics is exploding,

and users must choose a processing pipeline without an easy way to explore alternatives and integrate new methods. However, many pipelines follow a predictable pattern: they begin with a collection of images; align and stitch into a volume; format for compliance with a database or filesystem; process using computer vision and machine learning; and upload the result. Our *ndparse* software toolkit provides common interfaces and best practices for many of the use cases neuroimagers face when trying to extract knowledge from their data. We have written thin-layers around popular tools and developed new protocols and algorithms to facilitate the generation of results in a standard, modular format that supports many common tasks. Object detection is a canonical problem in computer vision and image analysis domains, and this framework can be easily adapted to new tasks, thus avoiding challenges such as data storage, computation, block processing, and multi-scale semantic understanding. More specifically, our Manual Annotation protocol, *mana*, downloads image data from *ndstore* in the appropriate format, leverages *ITK-SNAP* [41] to label voxels, and then uploads the resulting objects with semantic metadata. Given these training labels, *maca* uses machine vision algorithms such as *ilastik* [42] to perform pixel-level classification, and then standard morphological operations to translate high-probability clusters into discrete objects. Once an algorithm has been evaluated and a parameter set chosen, the result is deployed on a workstation or cluster environment for scalable processing. We have successfully deployed our tools to analyze large-scale image volumes by leveraging the *ndstore* architecture to parallelize at a data-block level. Our code runs on multi-core architectures using schedulers and meta-schedulers such as *qsub*, *slurm*,

and *LONI Pipeline*, and are continuing to add options popular with the user community.

These tools will work on small, stand-alone datasets for proof-of-concept testing, but are designed for large-scale analysis paradigms. More specifically, *ndparse* leverages the RESTful endpoints provided by *ndstore* and implemented by *ndio* to design workflows that are compatible with our ecosystem and storage system to efficiently read blocks of data and compute results (Figure 2.3). Additionally, we support several location queries, including the object located at a single point, all objects in a given region, and the bounding box (i.e., spatial extent) of an object.

ndparse is flexible and extensible to new algorithms. Our interfaces facilitate reproducible and interoperable science and enable rapid prototyping and optimization; throughout this work we show that research that previously required many lines of custom code can be run with only a few lines of descriptive Python, black-boxing the underlying algorithms and simplifying the computational expertise required for scalable scientific discovery.

2.4.1 Processing Framework and Pipelines

Image analysis of large neuroscience data requires unique and well-designed infrastructure; our approach is designed to eventually process petabytes of data. Specifically, scalable computer vision requires storage and retrieval of images, annotation standards for semantic labels, and a distributed processing framework to process data and perform inference across blocks that are too large to fit in RAM on a commodity workstation.

We leverage *NeuroData* infrastructure, the *ndio* API, our RAMON data standard and

reusable components to enable rapid algorithm development while simplifying the challenge of running at scale (Figure 2.5). Our tools are built on a distributed processing framework which leverages the LONI Pipeline [43] as an example workflow management tool, and interfaces with the data and annotation services provided by *NeuroData*. Our framework enables researchers to focus on developing novel image processing techniques without considering data management and processing challenges. We are able to efficiently incorporate new methods by extracting only core algorithm code (often a small fraction of the original code base). We reuse our data management framework, eliminating the need for researchers to rewrite solutions for file handling and bookkeeping. This capability enables image processing researchers to build state-of-the-art algorithms while minimizing the need to address cluster integration and scalability details.

In the appendix, we describe the compute resources used for much of the research described in this dissertation. Our front-end analysis infrastructure consists largely of commodity hardware and therefore our tools can be easily deployed in new environments. As our computing resources have matured, various cluster configurations and architectures have been tested, including "front-end" clusters at the JHU Applied Physics Laboratory and MARCC [44], and backend clusters that we maintain internally and on DataScope [45] resources.

2.4.2 Distributed Block Processing

One major open challenge (e.g., in a file-based image system) is scaling algorithms to massive volumes. We can quickly build interfaces to algorithms written by different research groups, and in different languages, to assemble a cohesive pipeline. These algorithms have well-defined interfaces and can also be repackaged for use in a different meta-scheduler environment. When running at scale, we typically divide large image volumes into smaller cuboids (accounting for overlapped and inscribed regions) which meet our computational constraints, and then process each block in parallel. We have explored a variety of merging strategies to combine results across cuboid boundaries, since the spatial database backend enables many approaches that would be much more computationally challenging with image stacks. These include overlapping cubes, merging overlapping regions along boundaries, and checking adjacent regions for writes before uploading potentially duplicate or conflicting annotations. For many usecases our favored implementation serially processes large block seams, eliminating the need for transitive closure operations. As volumes scale, a hierarchical approach may be desirable to increase throughput (Figure 2.4).

2.4.3 Processing Strategies

To date, success in connectomics has come from a combination of manual and automated processing [23,24,25,46]. As imaging advances allow for the acquisition of ever larger data

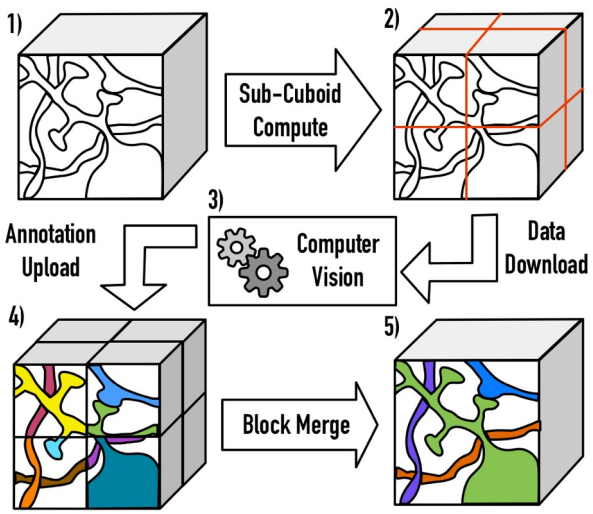

Figure 2.4: *An illustration of the distributed processing paradigm.* (1) Raw image data is (2) divided into cuboids based on user-specified parameters, with the necessary padding to perform computation. (3) After processing, the annotations are inscribed and (4) uploaded to *OCP*. (5) Finally, processes are merged across block seams, using a similarity metric of the user's choice.

volumes, the reconstruction process becomes an expensive, enormously time-consuming bottleneck [47]. This challenge becomes even more daunting if one considers the potential variability in a single organism, and that a full understanding of neuronal wiring diagrams likely requires the analysis of multiple organisms. For these reasons, we continue to advocate for a **fully-automated** strategy, with opportunities for semi-automated intervention as required. Recent research in graph theory suggests that even errorful graphs may still allow for the recovery of important neural motifs or primitives [48].

Although ultimately, fully-automated approaches will be required for scalability, **semi-automated** approaches have been used very successfully [25,37,49] in a variety of contexts

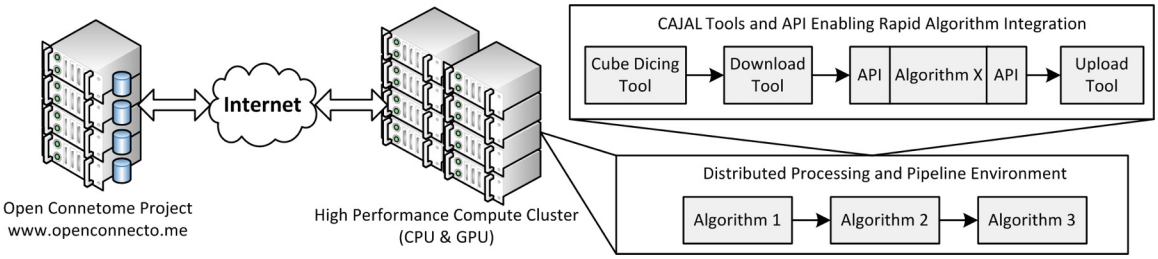

Figure 2.5: *An overall view of the processing framework, illustrating a distributed workflow paradigm.* Data and annotation stores leverage the *OCP*, and interface with a high performance compute cluster. A variety of tools are available to facilitate a distributed processing environment.

and provide a way to extract knowledge of a sufficient quality to test scientific hypotheses while automation improves. We employ an example of this approach in our nanoscale connectomics work [35].

There is still a need for **manual annotation** when creating gold-standard data. Our ecosystem easily accomodates these semantic labels as well, as we demonstrate in our neurocartography case study [22].

2.4.4 Neuron Grammars

We observe that although the detailed wiring diagram of the brain is unknown, the high-level tree structure of individual neurons obeys a predictable pattern. This pattern is analogous to a tree having a trunk, branches, and leaves in a consistent arrangement. Furthermore, although our datasets are large, the vocabulary of parts is constrained to only a few nouns, and local transitions between nouns can be considered independently of the surroundings, which can be described by a context-free grammar. This grammar is only

directly observable at nanometer resolution (e.g., serial section electron microscopy). Other modalities such as MRI and light microscopy can infer much about the brain's connectivity at a coarser scale (e.g., cell density, region connectivity). This grammar provides insight into building context-aware algorithms (e.g., *Santiago*), and eventually may provide a method for efficient error checking. This is in contrast to existing methods for automated segmentation are frequently completed without considering higher order structure [50].

At the highest level, neurons have the following parts: *cell body (C)*, *axon (A)*, and *dendrite*. As described above, connections between neurons are especially important, and so we decompose the dendrite into two parts: *dendritic spine (S)* and *dendritic shaft (D)*. We also define the additional symbols: *axonal bouton (B)*, and *synapse terminal (Y)* (Figure 2.6). Many other nouns are present in the neural tissue (e.g., glia, blood vessels, organelles), but these are not part of a basic graph and so are not included in the production rules. Because our initial experiments are of small volumes, comprising only pieces of individual neurons, all productions may also terminate at *volume edges (E)*.

We express the basic grammar productions for a single neuron; these cellular units are built up to form a graph (with synapses as the connection points between cellular units). Non-terminal symbols are capitalized and terminal symbols in lower case. Our grammar emphasizes topology rather than morphology, and is expressed as follows:

- Neuron $\rightarrow$ soma Neurites
- Neurites $\rightarrow$ Neurite | Neurite Neurites

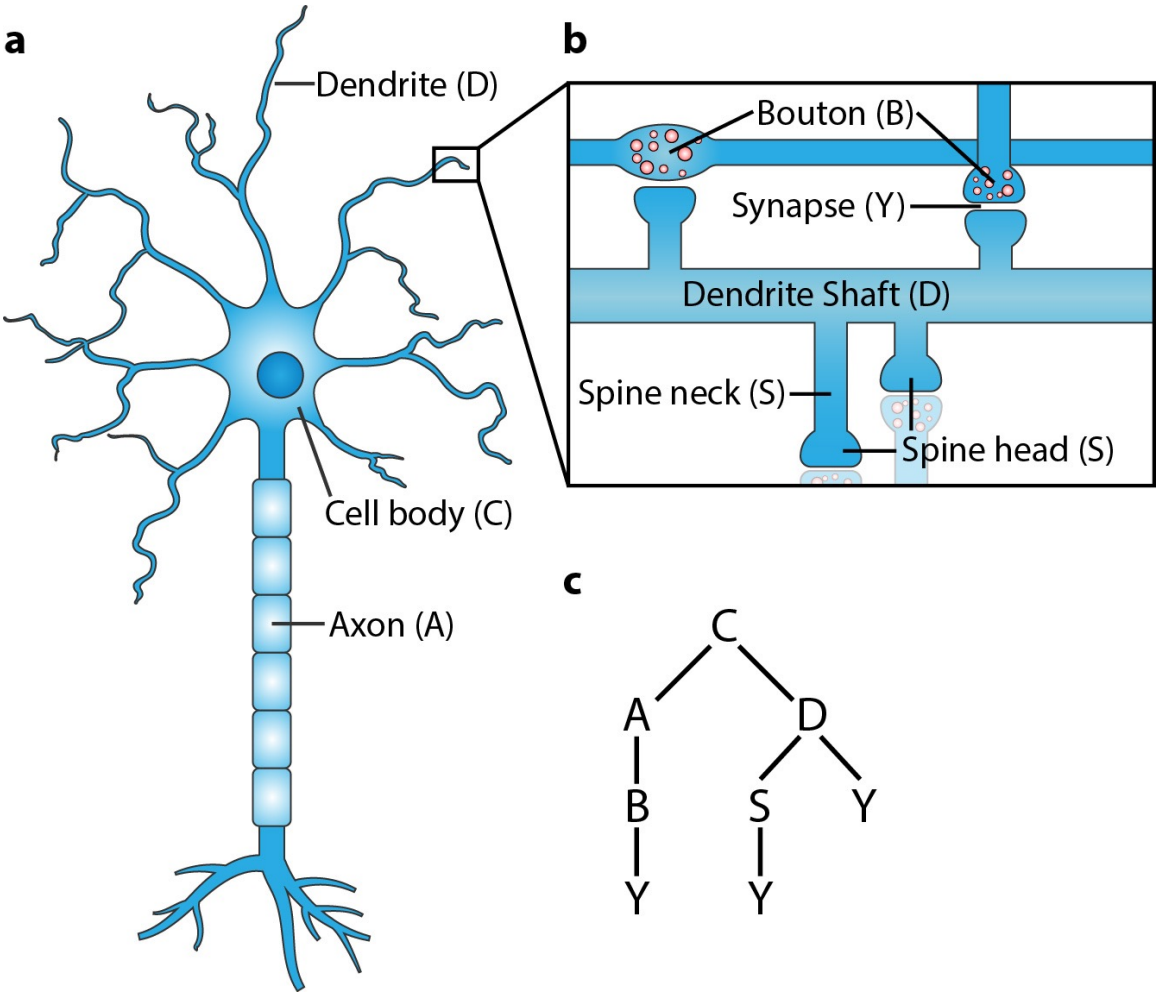

Figure 2.6: *Overview of neuron morphology.* (a) Labeled parts of a neuron, (b) with an inset showing our particular problem of interest. (c) A sample high-level parse tree capturing this information is shown for reference.

- Neurite $\rightarrow$ Dendrite | Axon
- Dendrite $\rightarrow$ shaft Spines
- Axon $\rightarrow$ axon Boutons
- Boutons $\rightarrow$ Bouton | Boutons

- Bouton $\rightarrow$ bouton synapse | bouton
- Spines $\rightarrow$ Spine | Spines
- Spine $\rightarrow$ spine synapse | spine

Typically, each synaptic terminal (Y) will be shared by a second neuron, and each spine and bouton will be associated with at least one synapse. In the vast majority of cases, we expect an axo-dendritic synapse motif, although other configurations are possible (e.g., axo-axonal, dendro-dendritic connections). We do not observe these other patterns in our training data, but our grammar could be extended if needed. The basic unit constructed by our grammar is a single neuron; these building blocks are combined (interfacing at synaptic junctions) to form brain graphs.

2.5 Metrics

One of the key contributions of this work is to develop and implement metrics for graph assessment in a variety of contexts. We find that these ideas translated well across scales and graph representations.

2.5.1 Precision-Recall

To characterize performance, we begin with a fundamental idea used regularly in detection problems in machine learning and computer vision. We compute true positives, false positives, and false negative counts, and compute precision and recall scores. Because

many representations of real brain graphs are sparse (i.e., mostly zero-valued entries), metrics that rely on true negative counts (e.g., accuracy) can produce misleadingly good scores. Precision-recall measures are robust to this type of error, and can then be combined into a single measure called an f_{beta} score. When $\beta = 1$, f_1 represents the harmonic mean of the precision and recall; β values less than 1 are biased toward precision, and those greater than 1 are biased toward recall. These metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \qquad (2.1) \qquad\qquad Recall = \frac{TP}{TP + FN} \qquad (2.2)$$

$$f_\beta = \frac{(1 + \beta^2) \times \text{true positive}}{(1 + \beta^2) \times \text{true positive} + \beta^2 \times \text{false negative} + \text{false positive}} \qquad (2.3)$$

We use this metric to characterize object detection results such as cell detection in X-ray Microtomography and synapses in electron microscopy.

2.5.2 Importance of Graph Error

A variety of error measures have been proposed for connectomics (e.g., warping error, adjusted Rand index, variation of information [50]), but are limited by their focus on an individual subtask of the entire images-to-graphs pipeline. As we will demonstrate in our electron microscopy graph estimation work (Chapter 5), the optimal results for a subtask may not translate to optimal results for the overall pipeline, and so it is important to measure

and optimize graph accuracy directly.

For example, as shown in Figure 2.7, even small neuron segmentation errors that affect graph connections are potentially very significant in terms of the resulting graph, while large errors not affecting topology may be much less significant. These small, significant errors occur frequently on narrow, fragmented spine necks, breaking connectivity between the synapse on the spine head and the parent dendritic shaft [51].

Although attributes like information direction or synaptic weight are useful for downstream analysis, the basic connectivity question we address here is perhaps the most fundamental. The metrics posed by IARPA's MICrONS program [28] and the current MICCAI CREMI challenge [52] illustrate how the community has begun to focus on these questions.

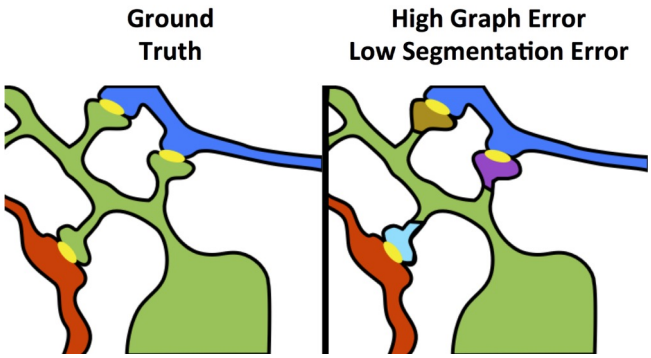

Figure 2.7: *An illustration of the spines to shafts problem.* Yellow objects represent synaptic junctions; other colors are different neurons. (Left) shows true connectivity; (Right) the effect of fragmenting neurons at the dendritic spine necks, which produces a very small change in segmentation error, but a dramatic impact to graph error.

2.5.3 Graph Matching

One of the first challenges we encountered when assessing graph performance was comparing estimated graphs to ground truth – specifically in finding a method to efficiently align graphs which is computationally challenging [53]. This challenge has been the subject of many research programs over past decades. Rather than solve this problem, we found novel ways to align our graphs; in MR data, we translate our brains to a standard reference frame and use a common atlas to find correspondences between graphs. In electron microscopy data we use line graphs to match network reconstructions.

2.5.4 Frobenius Norm

We first propose to measure graph error by simply computing the Frobenius norm between the true and test graphs, which can be applied to either MR based graphs or EM based line graphs (Equation 2.4). This metric is attractive in its simplicity, but has a few major disadvantages. This measure is unbounded, and the error will tend to increase with graph size; it is potentially misleading because it rewards the disproportionately large number of true negative edges in sparse graphs. We use this in our MR connectome work to successfully characterize small (relatively dense) graphs, but because of these disadvantages pursue other approaches for electron microscopy.

$$G_{err} = ||\mathcal{L}\{G^*_{true}\} - \mathcal{L}\{G^*_{estimated}\}||_F \tag{2.4}$$

This norm is defined [54] as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \tag{2.5}$$

2.5.5 Line Graphs

To assess graph error in electron microscopy data, we first form the *line graph*, which is the *dual* of the traditional graph, and represents connections (i.e., paths) between terminals. In this formulation, the synapses become the graph nodes and the graph edges are constructed from the neurons (Figure 2.8). A non-zero edge l_{ij} in the line graph represents a path between *synapse i* and *synapse j*. This is analogous to considering a brain graph as a communication network, where synapses are nodes (terminals) and neuronal fragments represent the paths (i.e., connections) between them [33].

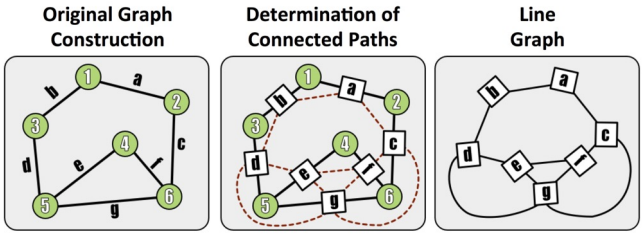

Figure 2.8: *Line Graph Construction.* Here we demonstrate the methods used to construct a line (edge-based) graph from a conventional node based network, by finding paths between edges in the original graph.

2.5.6 Graph f_1

To compute this metric, we first construct *line graphs* for both the estimated $\mathcal{L}\{G_{estimated}\}$ and true $\mathcal{L}\{G_{true}\}$ neuronal graphs, resulting in square, undirected, binary upper triangular matrices. To directly compare the graphs, we augment both matrices so that every node (i.e., synapse) in both graphs has a direct correspondence. We first find common synapses in the detected and true volumes by spatially overlapping annotations. We then augment the graphs with the synapses absent in either graph to create a superset containing all true and detected synapses, leading to true and test graphs of equal size (and corresponding nodes and edges). This graph correspondence is much easier to determine in the line graph (since synapses are small, compact objects) than in the traditional graph formulation (which often contains many neuron fragments and ambiguities introduced by split and merge errors).

In this paradigm, true positive edges (TP) occur in both the true and test graphs; false positive edges (FP) are present in the test graph but not in the true graph, and false negative edges (FN) are true edges that are missed in the test graph. Similar to an object detection evaluation, precision, recall, and the f_1 score are computed for the edges in the test graph (Equation 2.6).

Our metric is interpretible because it converts graph error to a detection problem with false positive and false negative errors. True connections are the non-zero common entries; furthermore, each incorrect entry represents a false positive (spurious connection) or false negative (missed connection). A connection between two synapses in a line graph is equivalent to those synapses being coincident on a neuron. This metric has scalability advan-

tages over voxel-based metrics, because it can be easily computed on large volumes and can be used to characterize common errors. This measure is on [0,1], and is robust to graph sparseness (i.e., true negatives do not impact the metric), and prevents a class of misleading results (e.g., an empty graph) as described earlier.

For our application, we optimize algorithm selection based on graph f_1 score. Researchers may choose a different optimization goal depending on their application (e.g., maximizing recall with acceptably high precision). A variety of metrics are computed (TP, FP, FN, precision, recall, f_1, Frobenius norm) for each graph and are available online. We extend this work to a semi-automated framework in *Santiago* [35], which is presented later in this work (Section 5.3).

$$f_1 graph = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.6}$$

2.5.7 Graph Reliability

The utility of the output brain graphs is a function of its scientific meaning. Because some of our data currently lack ground truth for the estimates, or other gold standards, investigators are left to assess the quality of estimates using only the data itself. This is a known-limitation of MR connectome estimation, but ground truth data is expensive to acquire in other modalities as well, and so these ideas can be extended to other domains.

In the absence of ground truth information (e.g., MR Connectomes), we use graph reliability to assess the efficacy of the pipeline results. The literature on reliability focuses

primarily on parametric reliability of scalar functions of the graphs. Other work (implicitly) assumes that the graphs themselves are reliable, and then asks questions about particular features of the graphs. Because the data collection and graph generation processes are so noisy, we desire to assess the reliability of their composition, and given a fixed pipeline, we can compare reliability of two different scanners or scanning sequences. Alternatively, given a fixed scanner and sequence, we can compare the reliability of two different pipelines.

Test-Retest datasets consist of multiple subjects, each of whom have been scanned multiple times. Most previous work on reliability has assessed parametric reliability of features of the data (e.g., it is standard to compare the between and within variances of, say, the number of edges [55]). However, these approaches are limited because they make parametric assumptions, and test scalar functions of the graphs.

As an example, we explain how the test-retest paradigm works in the context of our MR connectome pipeline. Let $\xi\colon \Omega \times \mathcal{T} \to \Xi$ be a brain-valued random variable (i.e., $\omega \in \Omega$ denotes a particular person, $t \in \Omega$ denotes a particular time, and $\xi_t(\omega) \equiv \xi(\omega, t)$ denotes person ω's *actual* brain at time t). Let $\psi\colon \Xi \to \mathcal{X}$ be a particular MRI scanner and scanner sequence, so $x_t(\omega) = \psi(\xi_t(\omega)) \in \mathcal{X}$ is the output from the scanner, and the input to the pipeline. The MR pipeline converts x to graphs, $\phi\colon \mathcal{X} \to \mathcal{G}$. Therefore, the scanning and pipeline together form the composition $f = \phi \circ \psi\colon \Omega \times \mathcal{T} \to \mathcal{G}$, and we can test the reliability of either ϕ or ψ over time for a particular subject.

Let $\delta\colon \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ be a graph-value metric. For vertex-aligned graphs, we adopt the

Frobenius norm of their adjacency matrices, $\delta(G, G') = \|A - A'\|_F$, due to its simplicity and its theoretical properties (in particular, under that metric, a k-nearest neighbor classifier is universally consistent for graphs [56]). Given n subjects, each with observations at two different times, we obtain $2n$ x's as input to the pipeline, and we compute $\binom{2n}{2}$ distances (because distances are symmetric). For each scan $i \in [2n] = \{1, \ldots, 2n\}$, we rank all remaining $2n - 1$ scans using δ. Let i and i' denote the first and second scan of subject i, respectively, and let $r_i \in [2n - 1]$ denote the rank of i' relative to i. We define reliability of f as $R(f) = (2n - \sum_{i \in [2n]} r_i)/(2n - 1) \in (0, 1)$, so $R = 1$ is maximally reliable and $R = 0$ is minimally reliable. Note that this notion of reliability is not limited to assessing f's or graphs, and is broadly applicable. Moreover, it is nonparametric and robust to outliers, and makes no distributional assumptions.

2.5.8 Alternative Strategies

Although our methods provide novel insights into characterizing graph quality, alternative strategies may also be useful, depending on the context. One example is to include automated or semi-automated approaches that rely on biological priors or constraints to examine errors (e.g., neuron grammars). Other ideas include extracting graph invariants to identify properties that may be useful for a particular exploitation task; however, optimizing on these summary measures may not provide an accurate characterization of the graph for other purposes, and interpretibility may be difficult. Finally, classification offers a method to test our graphs for 'signal,' by determining if populations of graphs are separable for a

particular covariate. This is a powerful technique that includes downstream processing at the expense of understanding the underlying graph structure differences (for many common algorithms).

2.6 Summary of Chapter Contributions

Our pipelines are designed around the following principles: *end-to-end optimization*, because integrated workflows allow for researchers to directly study and improve the knowledge required for their scientific discovery; *scalable processing* that interfaces with the *NeuroData* ecosystem and produces robust answers in a reasonable amount of time; and *repeatable solutions* that interface with standardized tools and data standards so that methods can be reused and extended by the community. We have developed a variety of complementary graph error metrics to assess the products of these pipelines, which gives us the information needed to create a feedback loop and improve our estimates of networks in the brain.

Chapter 3

Macroscale Graph Estimation

3.1 Overview

Currently, connectomes can be estimated in humans at $\sim 1\,mm^3$ scale using a combination of diffusion-weighted magnetic resonance imaging, functional magnetic resonance imaging, and structural magnetic resonance imaging scans. This work summarizes a novel, scalable implementation of open-source algorithms to rapidly estimate magnetic resonance connectomes, using both anatomical regions of interest (ROIs) and voxel-size vertices. To assess the performance of our pipeline, we develop a novel nonparametric non-Euclidean reliability metric (Section 2.5.7). Here we provide an overview of the methods used, demonstrate our implementation, and discuss possible extensions. A robust analysis of these brain graphs is now feasible due to recent efforts to collect large amounts of multimodal magnetic resonance imaging data [57, 58].

Prior to our work, many tools existed to process MR data; however, these were not available as an integrated graph-estimation pipeline, and it was challenging to combine the algorithms for scalable processing. Structural connectome estimation typically required an expert to manually combine these tools, run the tools in serial on a workstation, and then qualitatively assess the results. These results were typically optimized for a single dataset acquired under a particular set of scanner and subject parameters and generalizing these methods (or comparing across datasets) was very challenging.

We have built three major versions of our pipeline to robustly estimate MR brain graphs, including the first known large-scale automated approach, *MRCAP* [14]. This was subsequently improved and released as the *MIGRAINE* pipeline [15]. Following that release, we made significant improvements as joint work and released the *ndmg* pipeline [16]. We present their earlier foundational work followed by the latest pipeline and results; *ndmg* is recommended as the starting point for analysis or extension of our results. Since our first pipeline, others have published in this space, especially the Connectome Mapping Toolkit [59] and Pandas [60]. Our emphasis is on a reliable pipeline for graphs generated across populations of subjects from multiple datasets, while still allowing for flexibility to meet the requirements of individual researchers.

3.2 Introduction

An ideal connectomics methodology would enable scalable computing of graphs and functionals that yield reliable estimates. Moreover, such a tool would be open source, and would make the data it processes open source and user-friendly. Building such a tool, however, is challenging. The data for each subject consists of $\sim$1 gigabyte (GB) of multimodal MRI data. Converting from raw data to graphs and functions requires daisy-chaining many subroutines, each of which implements a different transformation of the data. *MRCAP* is the first pipeline that we developed for graph inference [14]. However, *MRCAP* has robustness and scalability limitations; it requires about 10 hours per subject to generate a final output, and has scheduler constraints. Moreover, *MRCAP* only generates small graphs, with 70 vertices, rather than voxel-wise graphs and functions. Other pipelines (e.g., [60]) have similar problems.

This section presents our MRI Graph Reliability Analysis and INference for ConnEctomics (MIGRAINE) pipeline, including methodology and experimental results. In addition to satisfying the desiderata (scalability and reliability) mentioned above, our pipeline, and much of our data, are provided in accordance with open science. The tested data originate from a wide assortment of institutions and projects, demonstrating pipeline robustness. We demonstrate the improved reliability of our pipeline over the previous state-of-the-art method, in addition to its improved scalability. Moreover, we utilize a notion of reliability related to the mean reciprocal rank often used in the information retrieval community. Finally, a useful metric for assessing pipeline and graph reliability is constructed and

demonstrated.

3.3 Raw Data

Table 3.1 provides summary statistics for the datasets processed via MIGRAINE. For each dataset, we collected both diffusion and structural MRI (MPRAGE). The two test-retest datasets (KKI–42 and NKI-24) were used to assess reliability; the other datasets were processed because they contained interesting phenotypic information (covariates) that can be utilized in future studies. The pipeline accepts diffusion weighted (dMRI) and structural (sMRI) images, the associated metadata, and user-specified parameters as inputs in a variety of formats (e.g., XML, PAR/REC, NIFTI, DICOM).

Table 3.1: *Various datasets successfully processed via MIGRAINE.* Key for covariates: S=standard (sex, age, handedness), C=cognitive, B=behavioral, L=language, D=diagnostic (e.g., bipolar). (c)2013 IEEE.

name	# subjects	covariates	ref
KKI–42	21	S	[61]
NKI–24	12	S	[62]
MRN–111	111	S, C	N/A
MRN–1313	1313	S,C,D	N/A
CASL–36	36	S,C,B,L	N/A

3.4 MIGRAINE Pipeline

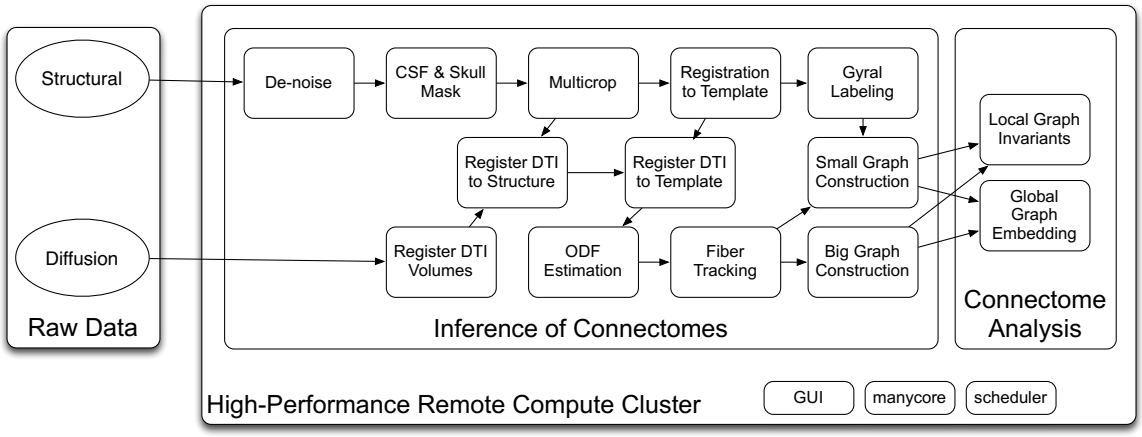

Figure 3.1: *MIGRAINE Pipeline Overview.* This figure illustrates each of the major components of the MIGRAINE pipeline; each block corresponds to a step in the integrated pipeline. (c) 2013 IEEE.

3.5 Graph Generation

MRCAP was implemented using the Java Image Science Toolkit (JIST) [63], consisting of 22 Java modules; for *MIGRAINE*, this was wrapped and integrated within the LONI pipeline framework [64]. Processing time is significantly reduced by swapping out some modules with improved functions and reducing I/O and communication. *MIGRAINE* is flexible and can be modified using existing neuroimaging modules already incorporated in LONI (e.g., [65]), or custom code that is command-line executable. Our graph generation workflow is comprised of structural, diffusion, and connectivity layouts, each of

which consists of a collection of modules. The modules themselves are assembled from a variety of algorithms, authors, and methods; we integrated these tools into one automated processing flow.

Our graph generation routines (summarized in Figure 3.1) consist of several key steps. Small graphs (e.g., 70 vertices and up to $\binom{70}{2} = 2415$ edges) are computed as detailed in [14], except for additional steps needed to register the structural and diffusion data for each subject to a common template (e.g., the MNI atlas [66]). The voxels in the template brain volume are labeled in accordance with the Desikan atlas regions [67]. We estimate tensors and perform deterministic tractography (FACT [68]) to estimate fibers in the brain. (These functions can easily be swapped with more sophisticated, but time consuming, options such as ODF estimation and probabilistic tractography.) Finally, an estimate of connectivity between each pair of regions is recorded (e.g., the number of times each region pair is connected by a fiber). Much of the connectome literature uses *region*-wise rather than *voxel*-wise graphs [69]. These results therefore enable comparison with previous analysis, allow for the assessment of reliability between pipelines, and support existing classification methods.

3.5.1 Structural Processing

Structural image processing begins with the SPECTRE algorithm [70], which removes the skull and non-CNS (central nervous system) tissue using a joint registration and tissue classification technique. The tissue classification is performed using FANTASM, a robust

fuzzy C-means intensity classification algorithm [71]. This allows for the identification of high-intensity skin and adipose tissue, and low-intensity bone matter, all of which can be subsequently eliminated. This result is smoothed and is used as an input for dMRI co-registration in the connectivity layout.

In the second step of the structural processing layout, the brain is divided into a set of 70 regions defined by the Desikan gyral label atlas [67]. Parcellation is achieved by registering one or more template brains to the subject brain using VABRA, a vectorized form of the Adaptive Bases registration Algorithm (ABA) [72]. This algorithm performs nonrigid intensity-based registration using normalized mutual information as a cost function, and models the deformation as a linear combination of radial basis functions. The results from the different template registrations are subsequently combined using STAPLE [73].

3.5.2 Diffusion Processing

Using the dMRI information, the *diffusion tensor* is estimated for each voxel using a log-linear minimum mean squared error measure [74]. A diffusion tensor is a local model of the diffusion process, which is influenced by tissue microstructure, particularly axonal projections. These tensors enable the computation of fractional anisotropy (FA), a scalar value derived from the tensor that roughly describes the relative intensity of diffusion along a given direction, and can indicate the *coherency* of axonal bundles summarized by the tensor.

From the computed tensors, streamlines are derived with the FACT algorithm [68], a

fast, deterministic algorithm for reconstructing fibers. FACT is a classical method that has been shown to recover many important fiber tracts [75], and is widely applied in the neuroscience community despite its inability to resolve crossing fibers. If desired, probabilistic tractography or other algorithms may be used instead [76,77,78]. Each computed fiber tract represents the estimated location of a large group of axons, which are signaling pathways (i.e., connections) between brain regions.

3.5.3 Connectivity Processing

At the beginning of the connectivity layout, the dMRI image data is preprocessed and co-registered to the structural output data using VABRA. Each fiber streamline traverses a (potentially large) number of voxels. We postulate that axonal fibers exist and connect any pair of voxels that a streamline traverses; therefore the two regions containing the same fiber streamline are assumed to be connected (Figure 3.2). To obtain an estimate of connection strength, various strategies may be used (e.g., mean FA value along fiber streamline), but for simplicity we use the raw count of fibers connecting each pair of regions.

Because we divide each brain into 70 regions, our MR connectomes are theoretically characterized by $\binom{70}{2} = 2415$ values. Because we cannot assign a polarity to a streamline, the connections are undirected, implying that the 70 x 70 connectivity matrix is symmetric. Furthermore, we do not compute connections within a region, implying that the matrix is hollow (i.e., the diagonal is empty). Therefore, the final dimensionality of the output is 2,415, representing the connection strength between each of the 2,415 pairs of cortical

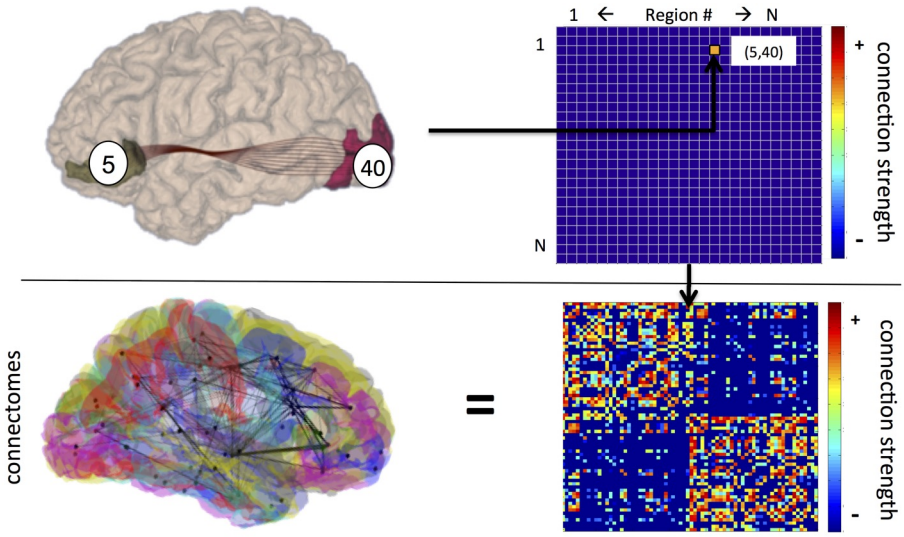

Figure 3.2: *Graph estimation workflow.* This figure illustrates the process of constructing a graph from a parcellated brain and a set of fiber streamlines. For each pair of regions, *i* and *j*, the number of fiber streamlines that are incident to both regions are counted. This value is recorded in the graph as the edge count between those regions (i.e., G_{ij}). All region pairs are evaluated to construct the map of connectivity for the subject of interest.

regions. Example MR connectomes are shown in Figure 3.4. Region pairs exhibiting no connectivity are assigned a value of zero, shown as dark blue in the figure. An alternative three-dimensional view of selected cortical connections is shown in Figure 3.2.

3.5.4 Big Graph Estimation

To generate big graphs, we utilize the Magnetic Resonance One-Click Pipeline (MROCP) code base as detailed in [79]. Initially a mask (e.g., ROIs) is applied to the fiber streamlines created during small graph estimation. Each surviving voxel becomes a vertex in

the sparse, column-compressed graph. Here edges represent a single fiber connecting a pair of vertices within the bounds defined by the mask. Finally, we iterate over each fiber streamline, recording an edge between every two vertices that can be reached (i.e., that are connected) by a single fiber.

All of these graphs are aligned by construction; for each MR scan we obtain a big graph with $\sim 10^7$ *aligned* vertices and $\sim 10^{10}$ edges. Because we are conservative with the mask, many of these voxels are noise. We therefore reduce these graphs to their largest connected component, which essentially keeps all white matter voxels, consisting of $\sim 10^5$ vertices and $\sim 10^8$ edges.

Computing analytics (i.e., multivariate glocal invariants [79]) on big graphs is a challenging endeavor due to the computational intensity associated with processing graphs with $\sim 10^8$ edges. Equivalent computational tasks are generally designated to specialized hardware like GPUs, graph processing engines like GraphLab [80], or distributed solutions like MapReduce. We utilize MROCP to efficiently compute several multivariate graph analytics, including Latent Position (LP-k), Number of Local 3-Cliques (NL-3), Clustering Coefficient (CC), Scan Statistic-1 (SS-1), Degree and Edge count (see [79] for details).

3.6 MR Graph Results

We successfully processed subjects from a variety of datasets (both existing and new), totaling over 1500 subjects from multiple centers and acquisition paradigms using a rapid,

extensible, automated framework. In addition to producing small graphs, we have demonstrated additional processing capability through the estimation of big graphs and analytics. The pipeline is scalable and has internal validation and packaging scripts to enable efficient analysis. The resulting graphs and analytics are currently being used to develop classifiers, to provide new insight into the way brains are wired, and to determine which aspects of the network are informative in predicting cognitive properties. A univariate measure of total fiber count per subject is shown in Figure 3.3.

The *MIGRAINE* pipeline offered significant improvements in scalability and processing time relative to *MRCAP* as demonstrated on a small compute cluster (248 cores, 1TB total RAM). On average, the MIGRAINE pipeline takes $\sim$3 hours/subject to compute small graphs (i.e., the output from *MRCAP*, which took $\sim$10 hours/subject on this cluster), an additional 5 hours/subject to produce big graphs, and 3.5 hours/subject for graph invariants, for a total of 11.5 hours/subject. Much of this improvement is obtained by utilizing a common registration template, allowing for anatomical labels to be computed only once and then reused. Multi-core capabilities only contribute marginally for a single subject (in both pipelines) because the most intensive computations occur serially. However, there are significant efficiencies in scheduling when evaluating a large number of subjects, with the number of cores being the limiting factor.

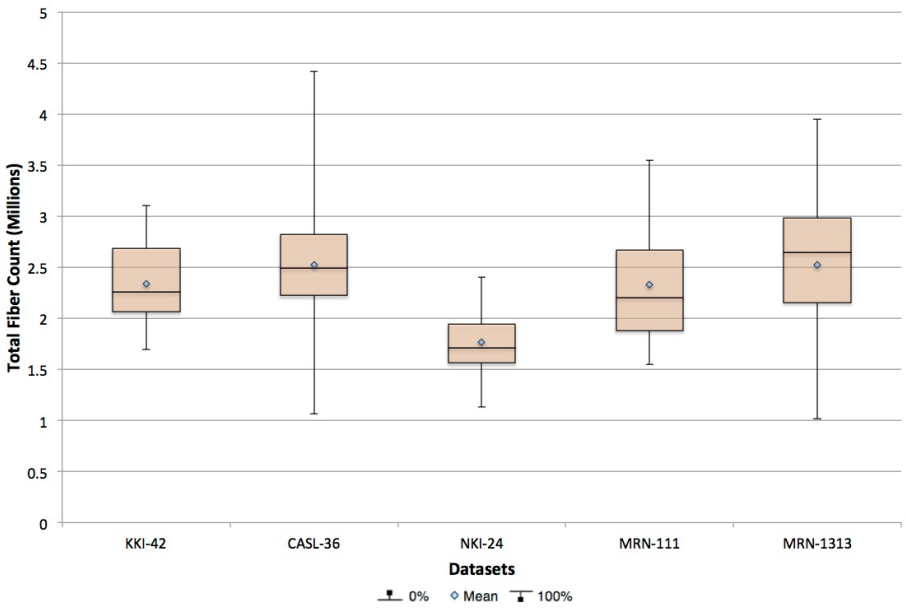

Figure 3.3: *Box plots for each data set.* This figure shows the total fiber count for each subject in each dataset. (c) 2013 IEEE.

3.6.1 Reliability

A variety of tools have been developed to analyze intermediate pipeline products (e.g., matrix comparison tools, analysis of fiber counts). *MIGRAINE* leverages several algorithmic improvements versus *MRCAP*, including changes to data preprocessing and the registration to a common registration space. The results differed by 13% from the *MRCAP* baseline and produced better subject separability (Table 3.2).

To validate that our graphs convey a repeatable signal, we used the KKI–42 Test-retest Data [61] to analyze graph estimation reliability, as described in Section 2.5.7. We demonstrated that the *MIGRAINE* pipeline produced a stable connectivity measurement across multiple scans of the same subject.

Table 3.2: *Validation showing improved discrimination.* This table shows *MIGRAINE* performance relative to *MRCAP* using the KKI–42 dataset (c) 2013 IEEE.

Pipeline	Intra-Sub Mean Diff	Inter-Sub Mean Diff	Closest Inter-Sub	# Matches
MRCAP	26032	51584	38451	40/42
MIGRAINE	20378	56126	42663	42/42

For all 42 graphs, the most closely-related graph (as computed with the Frobenius norm on the adjacency matrix) belonged to the same person, scanned at a different time. A visualization of the graphs for six test-retest pairs are shown in Figure 3.4, and the results of all individual subject comparisons are shown in Figure 3.5.

3.7 *ndmg* pipeline

Following the success of the *MRCAP* and *MIGRAINE* pipelines, we jointly developed a new, more robust pipeline as an open-source package called *ndmg* [16] that had significant improvements in scalability and robustness as well as new functionality (Figure 3.6). Although algorithmically the pipeline is very similar to the steps outlined above to generate graphs, we transitioned to *dipy* [81] and *fsl* [65] tools, as well as our own graph estimation and quality control measures. This new pipeline is written as an open source python

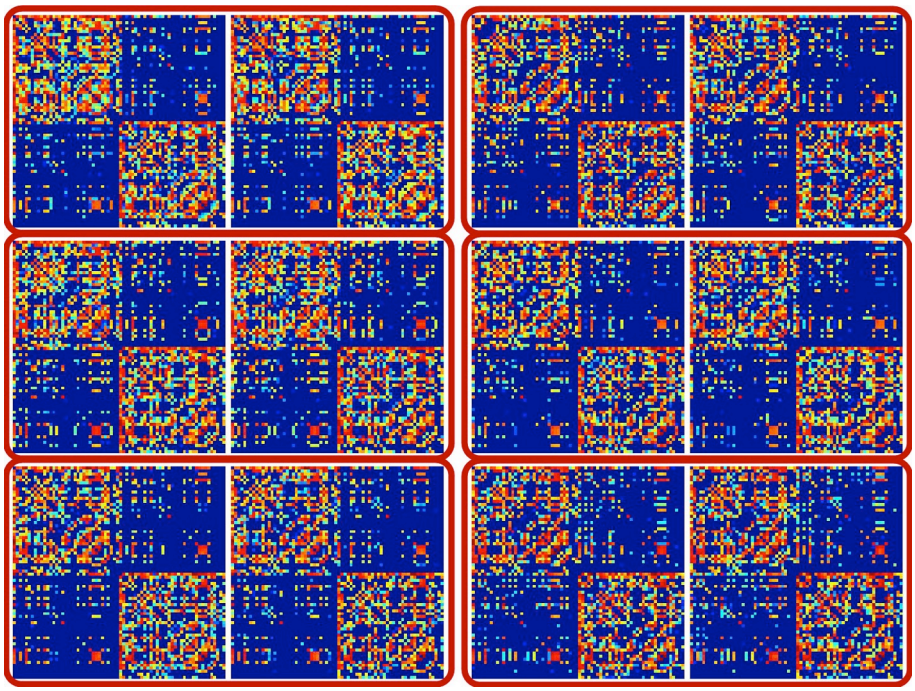

Figure 3.4: *Six example Test-Retest graphs.* Top (L-R): Male, 25 years old (M25), Female, 26 years old (F26), Middle: M25, F30, Bottom: M38, F61. (c) 2013 IEEE.

package that has significant improvements in robustness and scalability. Generating a graph using *ndmg* now typically takes less than an hour on a commodity machine and runs on a variety of platforms, including a mid-range laptop.

3.7.1 Current Software Package

Our current pipeline is modular and other algorithms can be added or exchanged to meet the demands of a particular research question. Our work is focused on producing a one-click robust pipeline that is reliable across all of the available open datasets (Figure 3.7).

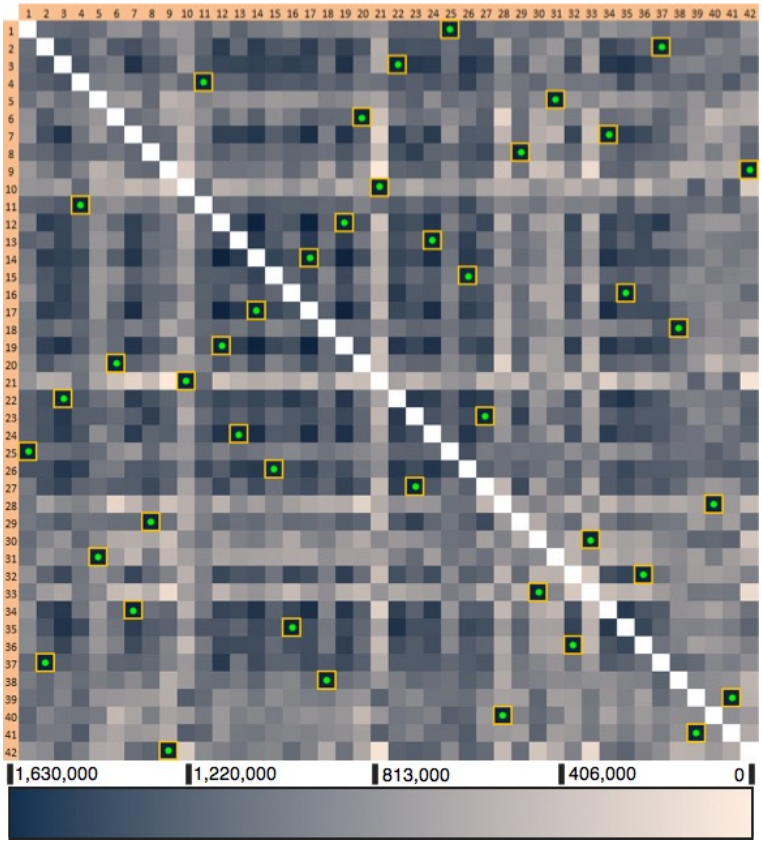

Figure 3.5: *KKI Test-Retest Results.* Yellow boxes: Highest similarity, Green dots: True pairs, White: Self-comparison. (c) 2013 IEEE.

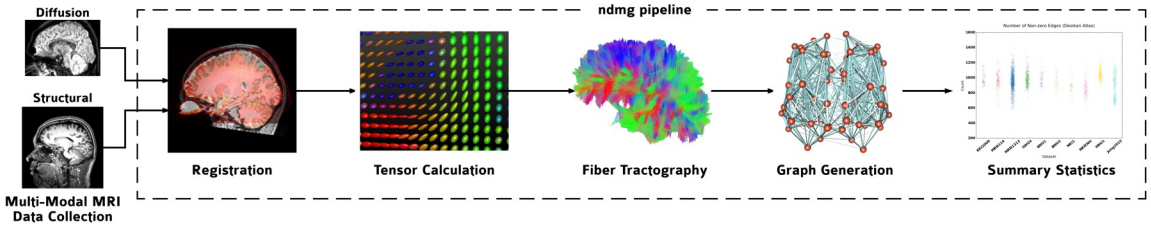

Figure 3.6: *Overall depiction of ndmg pipeline.* This pipeline begins with sMRI and dMRI data and produces brain graphs using various parcellations.

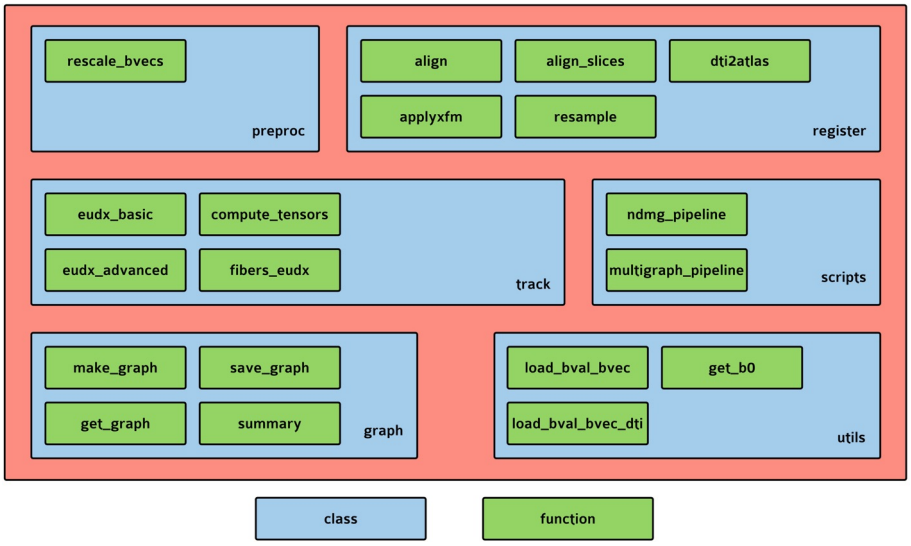

Figure 3.7: *ndmg package architecture.* This diagram includes classes and functions to process data at different stages of the pipeline.

3.7.2 Data Derivatives

We leveraged the increased capacity of this new pipeline to generate brain graphs for more than twenty different atlas parcellations and thousands of scans from ten different datasets, including many of the datasets originally processed with *MIGRAINE* and a total of five Test-Retest datasets. (In this iteration of the pipeline we did not process the Human Connectome Project data because the properties of that data are incongruous with other publicly available datasets and our focus is on building generalizable tools that work across studies using more common methods. Future work will extend our pipeline to accomodate this dataset because of its unusually high quality and emerging datasets using similar protocols.) We believe that our tens of thousands of graphs (across thousands of brain scans) reprsents the largest, most diverse collection of brain graphs available.

3.8 MR Graph Analysis

3.8.1 Reliability

We extended our notion of reliability to incorporate *Mean Normalized Rank* [82], which is defined as: $MNR = p(||a_{ij} - a_{ij'}|| \leq ||a_{ij} - a_{i'j'}||)$, where *a* represents an adjacency matrix (i.e., brain graph), indexed by *i*, representing the subject and *j* representing the scan [83]. This reliability score $\in [0, 1]$, where 1 indicates that the retest pair(s) are all in the best position (closest match). We validated the *ndmg* pipeline by running this metric against two Test-Retest datasets and found a value of 1.0 (perfect reliability) for the KKI2009 dataset and 0.984 for SWU (after rejecting outliers) as visualized in Figure 3.8.

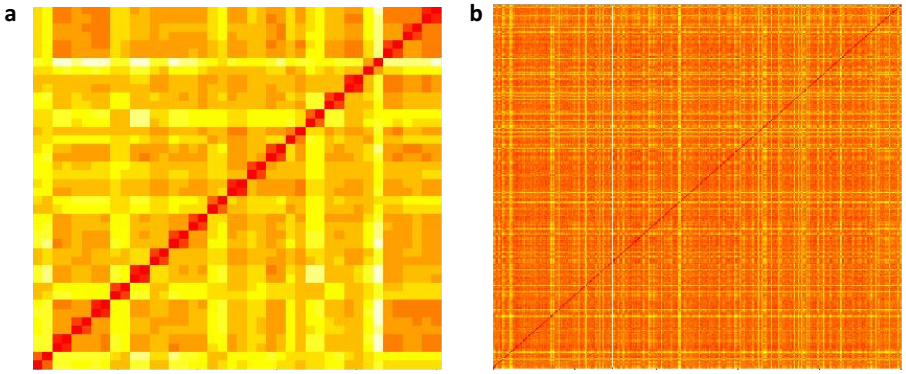

Figure 3.8: *ndmg reliability visualization.* This figure shows the high reliability of the (a) KKI2009 dataset and the (b) SWU4 dataset; locations where the brighest red values appear clustered in pairs on the diagonal represent a correct pairing.

3.8.2 Mean Connectome Estimation

Once these pipelines have processed thousands of subjects at various resolutions, we can begin to address many important research questions related to connectivity disorders and cognitive processing. As an illustrative example, we can create a consensus mean connectome, which may be generated simply by computing the sample mean of the vertex-aligned graphs across a dataset. We compute this graph for the KKI2009 dataset, and also compute an average of the binarized (threshold: 0) graphs, highlighting the probability of each edge existing (Figure 3.9). This result provides insight into the major connectivity pathways of the brain and can be used to assess individual differences, compute graph statistics, and as an input to assess pipeline quality.

3.8.3 Connectome Classification

While this work has been primarily focused on graph estimation and assessment we have also used these graphs for classification. One notable proof-of-concept result is the sex-classification result [84] which develops a novel classifier based on the raw edge differences between sub-populations of graphs. The classifier finds the *signal subgraph* consisting of nodes (and corresponding edges) that are most discriminative between these two groups and then classifies each subject based on this information. This result is interpretable because it directly uses the underlying graph structure (rather than graph invariants), and achieved state-of-the-art (16% misclassification rate) leave-one-out (LOO)

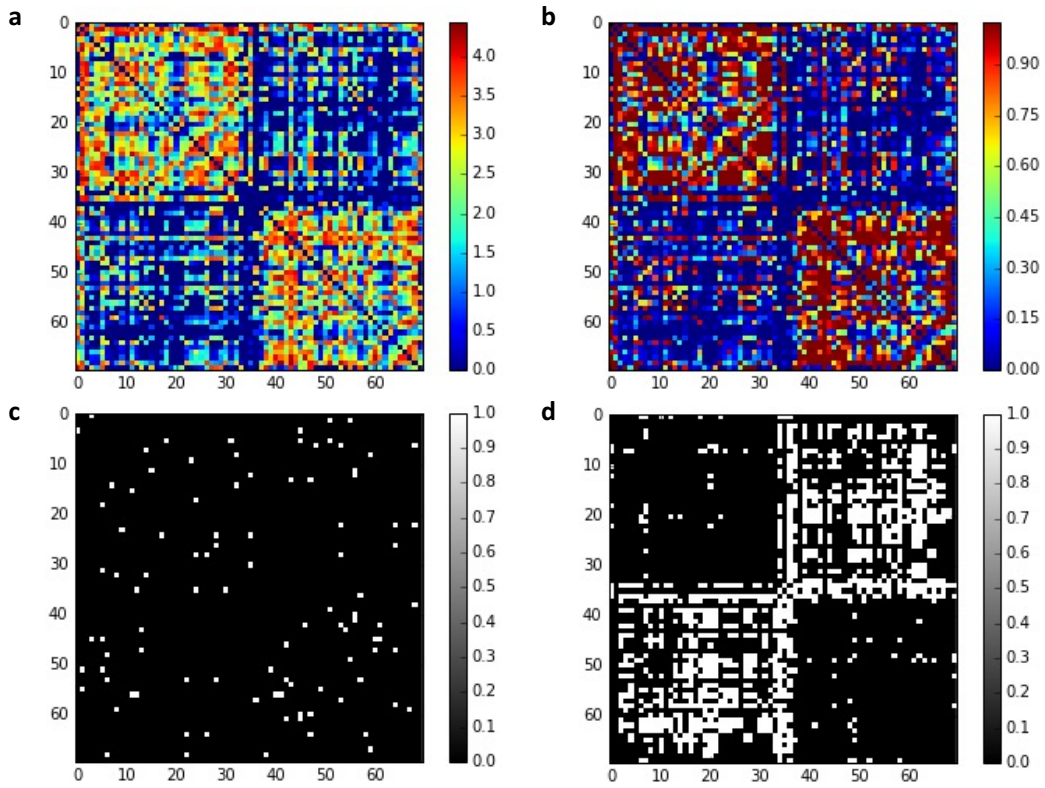

Figure 3.9: *ndmg mean connectome assessment.* This figure shows the mean connectome visualizations for the KKI2009 dataset with the Desikan parcellation, including: (a) the mean (log) sample connectome; (b) the probabilistic (binarized) graph; (c) the intersection of all graphs (white: present in all graphs); (d) the union of all graphs (white: edge not present in any graphs).

performance on a small dataset.

3.9 Summary of Chapter Contributions

In this chapter we apply our tools and scalable methodology to many relatively small data volumes to produce the first automated, scalable approach for reliable brain graphs at a macroscale. We currently host the largest, most diverse set of MR-based brain graphs in the world. Future work with the pipeline will focus on understanding and mitigating batch effects across datasets, understanding the neuro-fidelity of the produced graphs, and investigating downstream applications using the graphs to better understand disease and injury and potentially intervene to mitigate adverse effects. Future chapters will explore similar concepts in a paradigm focused on extracting knowledge from single, large-scale data volumes.

Chapter 4

Assessment of X-Ray Microtomography data for Open Neuroscience

4.1 Overview

In contrast to MR imaging, which allows for the *in-vivo* assessment of putative large-scale connections in the white-matter of the brain, X-ray microtomography provides a new method to resolve the 3D microstructure of the brain at cellular resolution. Similar methods start by thinly slicing and staining the brain, and then imaging each individual section with visible light photons or electrons. In contrast, X-rays can be used to image thick samples, providing a rapid approach for producing large 3D brain maps without sectioning.

Here we develop and deploy new methods for automated cell detection and blood vessel segmentation, along with subsequent statistical analysis of the resulting brain structures.

This work provides the first known algorithms for cell-detection and vessel-segmentation in this imaging modality and provides a highly efficient method for assessing the microstructure of the brain at a fraction of the time and computational cost of other approaches. Our results demonstrate a method to rapidly quantify cells and blood vessel in large brain volumes, complementing other brain mapping and connectomics efforts.

4.2 Introduction

Large-scale brain maps that provide a glimpse into the cellular and vascular architecture of the brain are essential for understanding neuroanatomy, and its relation to function and disease [85]. Unfortunately, acquiring high resolution brain maps is still difficult and time intensive [86]. Conventional light and electron microscopy (EM) methods require sectioning tissue into thin slices (μm scale), imaging each slice individually, and then stitching the images back together to get a 3D brain map. For example, stitching BigBrain—a 3D reconstruction of a full human brain at $20\ \mu\mathrm{m}$ isotropic resolution—required approximately $1,000$ hours to complete [87]. Electron scatter occurs at even smaller depths than visible light and as a consequence, EM requires even thinner slices ($\sim$30nm). Therefore, it takes approximately three months to image a cubic millimeter of brain tissue at 20 nm resolution [88], requiring approximately two petabytes on disk. Methods for quickly imaging the brain's microstructure are critical for understanding and comparing the structure and function of many brains.

Tissue clearing approaches such as CLARITY [89] and expansion microscopy [90] address some of the challenges associated with large samples. However, unlike EM, these techniques produce sparse reconstructions which reveal only subsets of neurons in the volume. In addition, tissue clearing requires the removal of scattering membranes in tissue samples, which renders them incompatible with subsequent serial section electron microscopy to identify individual neuronal connections. As a consequence, interrogation of the sample is primarily limited to the mesoscale and it is challenging to re-investigate the same tissue at higher resolution. Therefore, new approaches capable of producing large-scale complete mesoscale reconstructions of the brain are required.

X-ray microtomography (μCT) provides a unique and largely untapped opportunity for brain mapping. Theoretically, X-rays can penetrate through centimeter-scale brain volumes with micron resolution, without the need for sectioning. Recent studies have demonstrated the utility of benchtop μCT systems for neuroscience [91, 92]. However, using benchtop systems for large-scale brain mapping efforts is difficult due to the long exposure times needed to collect even a single image; imaging a cubic mm brain sample at $1\mu\mathrm{m}$ resolution would take at least 13 hours on state-of-the-art scanners [93]. Fortunately, synchrotron-based μCT offers far higher photon flux and thus provides an avenue for the rapid acquisition (two orders of magnitude speedup) of large brain volumes [94, 95, 96, 97]. However, μCT has not yet been adapted to meet the demands of large-scale brain mapping efforts.

Here we introduce a pipeline for quantifying mesoscale neuroanatomy with μCT. We demonstrate that samples fixed with aldehydes, stained with osmium, and embedded in

plastic can be imaged with high-energy synchrotron radiation. The resulting image datasets provide sufficient isotropic resolution ($1\ \mu\mathrm{m}^3$) and contrast to resolve the 3D structure of neuronal and glial cell bodies, vasculature, and segments of large apical dendrites and myelinated axons. We can subsequently section the same samples and image them with an electron microscope; the result shows excellent preservation of the ultrastructure and straightforward correspondence (leading to easy co-registration) between X-ray and EM datasets. These results confirm that μCT can be used to produce imaging data with sufficient resolution to compute mesoscale brain maps containing information about the cyto- and myelo-architecture of cortex. We developed a suite of open-source tools, XBRAIN (X-ray Brain Reconstruction, Analytics and Inference for Neuroanatomy), for cell detection, blood vessel segmentation, and statistical analyses of X-ray image volumes. μCT in combination with image parsing techniques offers an effective path from brain specimens to mesoscale brain maps.

We developed methods to image, segment, and analyze the neuroanatomical structure of brain volumes to quantify neuroanatomy with μCT.

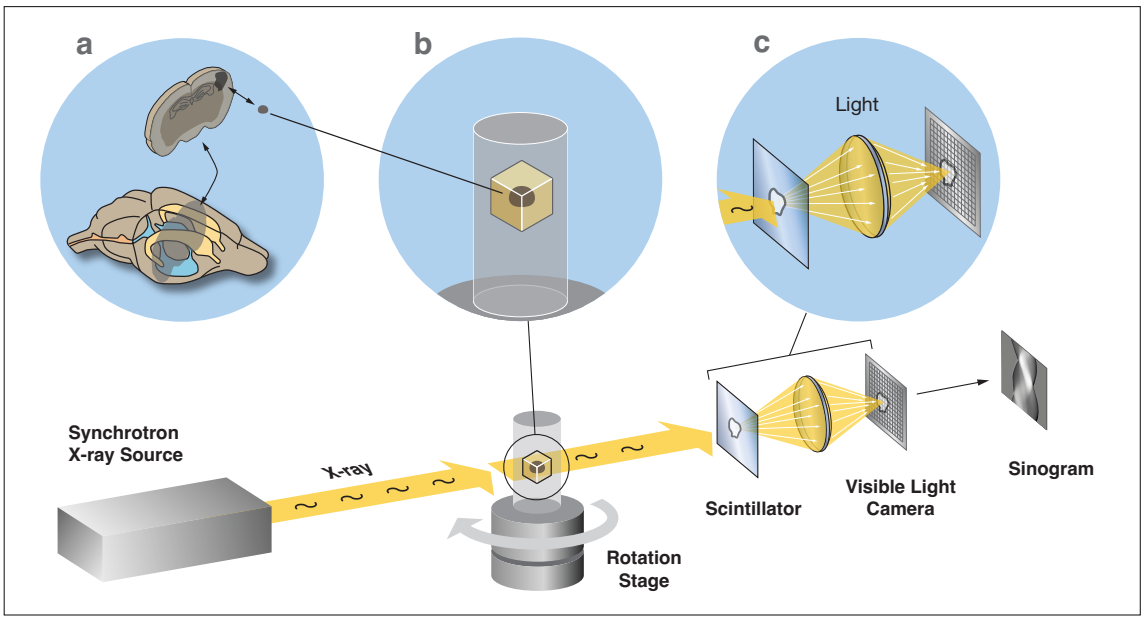

Figure 4.1: *Synchrotron X-ray imaging of millimeter-sized brain volumes.* A schematic of our sample preparation and imaging setup are displayed along the bottom: from left to right, we show the synchrotron X-ray source interacting with a embedded sample of brain tissue as it is rotated to collect multi-angle projections. To collect projection data, X-rays are passed through a scintillator crystal which converts X-rays into visible light photons, and then focused onto a light camera sensor. Finally, we obtain a sinogram from the sample by collecting data from a row of sensor pixels. Above, we show a more detailed depiction of the (a) sample preparation, (b) sample mounted in the instrument, and (c) conversion and focusing of X-rays to light photons.

4.3 Image Acquisition Methods

4.3.1 X-ray tomography on a millimeter-scale brain sample

Using the 2-BM synchrotron beamline at the Advanced Photon Source (APS) [98], tomography data was obtained from cubic mm volumes of brain tissue (Figure 7.1). We

used techniques compatible with large volume EM [22, 99]. Mice were anesthetized and transcardially perfused with aldehydes (2 percent PFA and 2 percent Glutaraldhyde), stained with heavy metals (osmium, uranium, and lead), dehydrated, and embedded in plastic (EPON). The main dataset analyzed in this paper is taken from mouse somatosensory cortex (S1). After calibrating the instrument, collecting the main dataset studied in this paper took approximately six minutes and requires no volume alignment or registration process post-acquisition. We are able to obtain data around 130 times faster than with laboratory sources, and with higher image quality.

4.3.2 X-rays reveals diverse neural structures

X-ray images allow for resolving the putative location and morphology of cell bodies, blood vessels, and segments of large neurites (Figure 7.2b). The voxels inside cells are estimated to be 4.56 ± 1.13 dB (mean$\pm$std) brighter on average than their immediate surroundings. At this contrast level and resolution, it is possible to discern the location and size of cells in the sample. Blood vessels are also visible in the sample and provide even stronger contrast than cell bodies, making them much easier to track. This signal strength suggests that it should be possible to segment the tissue into cell bodies and blood vessels, which we validate with our automated techniques.

After collecting μCT data, ultra-thin sectioning and electron microscopic imaging was performed on the same sample. EM confirmed the identity of the cell bodies, myelinated axons, and blood vessels, corresponding to those annotated in the μCT dataset (Figure 7.2c),

suggesting that details seen in the X-ray dataset are bona fide and not spurious results of our imaging and processing pipelines. In addition, no changes were observed in the microtome sectioning properties of the epon-embedded brain tissue, nor were any obvious signs of irradiation-induced structural damage seen in the scanning electron micrographs obtained from these sections. Structures like synapses and mitochondria are still clearly evident (Figure 7.2c). These results confirm that μCT and EM can be coupled to produce multi-resolution brain maps. Since the labeling approaches are species independent (i.e., they do not depend on transgenic strategies), the approaches can be applied to human (and other) brain biopsies in future work.

4.3.3 Volume of the analyzed sample

The image volume that we analyze is $1400 \times 2480 \times 1547$ voxels, which corresponds to a volume of size $910 \times 1612 \times 1005$ microns (1.474 cubic millimeters). As the sample is rotated within the field of view (sample plane), we compute that the number of unmasked voxels represents a volume of approximately 0.41 cubic mm.

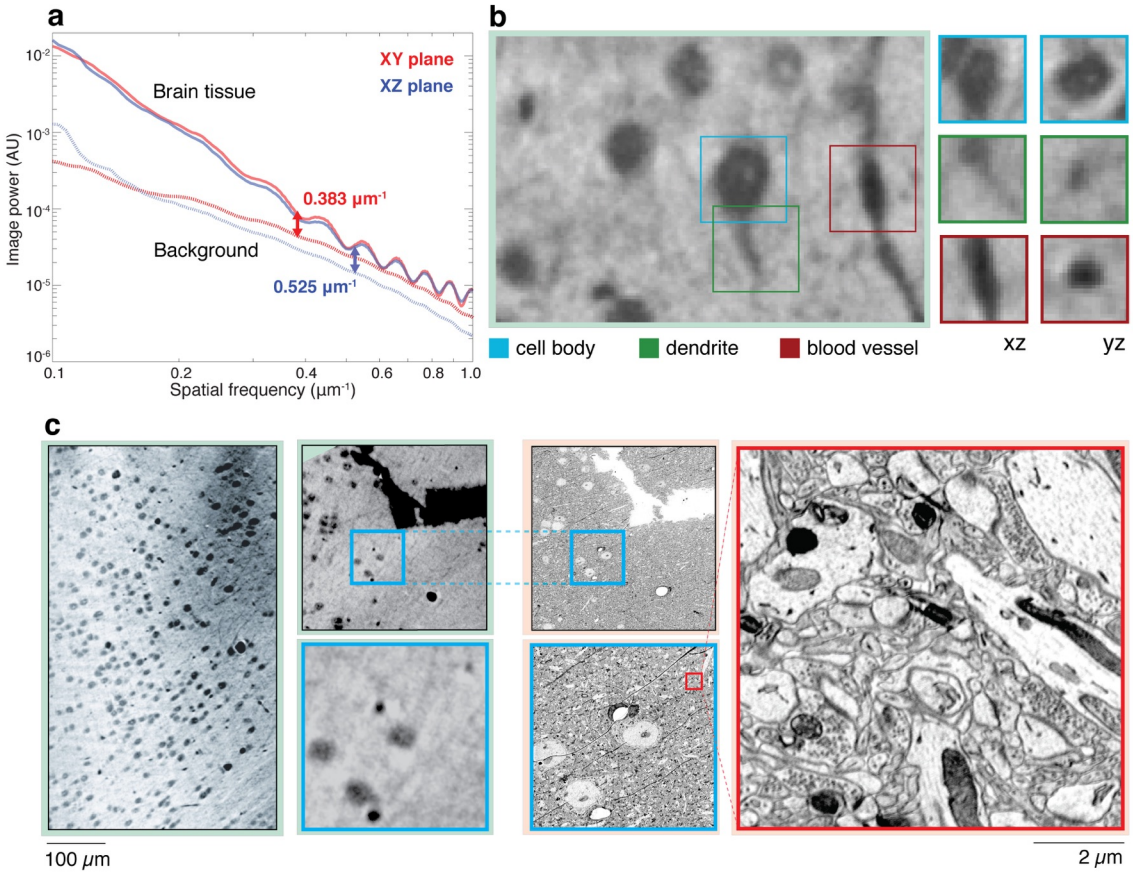

Figure 4.2: *Synchrotron X-ray imaging provides micron resolution of brain volumes.* (a) From our reconstructed volumes, we compared signal (SPS) and noise (NPS) regions to assess the feasibility of the detection task. (b) We show multi-view projections of X-ray image volumes, where the 3D structure of cells, vessels, and dendrites is visible. (c) We show μCT and EM images of the same sample, collected at three different pixel sizes ($0.65\ \mu\mathrm{m}$, $100\ \mathrm{nm}$, $3\ \mathrm{nm}$). Using landmarks observed in the μCT scan, we located the same configuration of cells in the EM dataset (outlined in blue) and observe that the EM ultrastructure is well preserved after μCT (outlined in red).

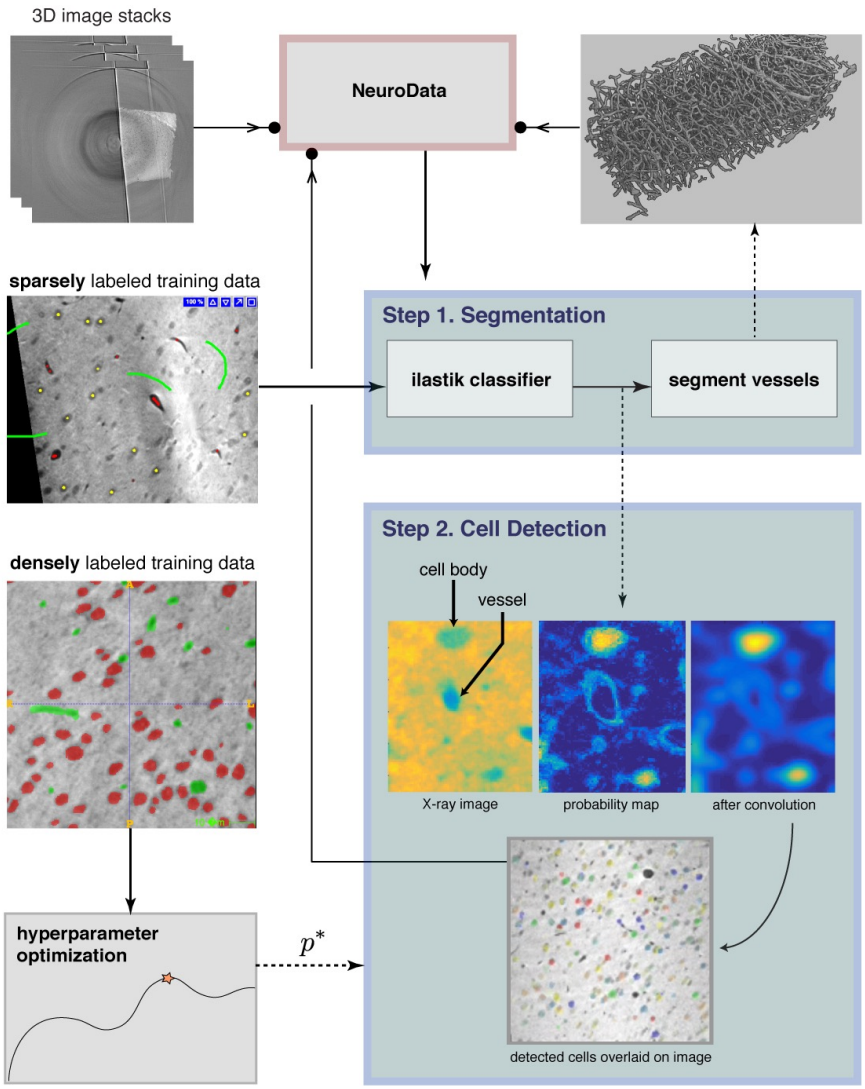

Figure 4.3: *Image processing and image analysis pipeline for segmentation and cell detection.* Sparsely labeled training data is integrated into our segmentation module (Step 1) to train a Random Forest classifier using *ilastik*. Densely annotated training data is used to perform hyperparameter optimization for the cell detection algorithm (Step 2). The final detected cells are displayed at the bottom of Step 2, with detected cells overlaid on top of the original X-ray image. Solid arrows indicate inputs into a module, outputs are indicated by dashed arrows, and outputs that are stored in *NeuroData* are indicated with a filled circle terminal.

4.4 Image Analysis Methods

4.4.1 Automated image analysis methods

The datasets afforded by X-ray tomography are too large to be realistically analyzed by humans. Therefore, we developed X-BRAIN, which provides automatic 3D segmentation algorithms to extract cells and vessels from the image volumes (Figure 4.3). More specifically we provide image processing and image analysis methods for preprocessing and artifact removal, segmentation, estimating the location and size of cells, and vessel segmentation. Methods are also provided for large-scale analyses, to compute relevant statistics on the reconstructed maps of the cells and vessels. X-BRAIN is implemented in Matlab and Python and both code and data are openly available through the project website, providing a community resource for the automated segmentation and quantification of mesoscale brain anatomy.

Our main image processing and image analysis pipeline (Step 1-2 in Figure 4.3) consists of methods for segmenting blood vessels and detecting the location and size of cells in the volume. In the initial step of our workflow, a Random Forest classifier is trained to predict the probability that each brain voxel belongs to each of the three classes: cell body, blood vessel, and background (other). *ilastik* is used to sparsely annotate data and build the classifier using intensity, edge, and gradient features computed on the image volume [42]. This classification procedure returns three probability maps $\mathcal{P} = \{P_c, P_v, P_{bg}\}$, which collectively provide the probability tuple $p(x, y, z) = \{P_c(x, y, z), P_v(x, y, z), P_{bg}(x, y, z)\}$

that each voxel, whose position is denoted by (x, y, z), is a cell, vessel, or lies in the background (output of *ilastik* in Step 1 of Figure 4.3). This classification procedure provides an easy and intuitive way to estimate which voxels correspond to cell bodies and blood vessels.

The simplest way to convert a probability map to a (binary) segmentation, is to threshold the probabilities and group the resulting structures that pass this test into connected components. In the case of vessel segmentation, we can employ this procedure with minimal tweaks. To segment vessels in the sample, we threshold the vessel probability map and then apply simple morphological filtering operations to clean and smooth the resulting binary data. Visual inspection and subsequent quantification of precision and recall of vessel segmentation suggests a high-degree of accuracy through this simple post-processing of the *ilastik* outputs.

GREEDY SPHERE FINDING APPROACH FOR CELL DETECTION

While *ilastik* provides a good starting point for identifying cell body locations, individual cells and vessels are often hard to distinguish by simply thresholding the probability map. Applying the same thresholding procedure used for vessel segmentation, to the segmentation of cells, is difficult because neurons and blood vessels are often densely packed. In this case, simple thresholding-based approaches tend to group clusters of cells and vessels together (Figure 4.4). We developed an algorithm for cell detection (Step 2 in Figure 4.3), which produces estimates of the centroids and radii of detected cells. Our method leverages

prior knowledge of the approximate size and spherical shape of cells to select sphere-like objects from the pre-filtered probabilities to resolve situations where neurons and blood vessels appear in close proximity. To separate these components into their constituent parts (cells and vessels), we developed a greedy approach which is similar in spirit to matching pursuit algorithms for sparse signal recovery [100]. The main idea behind our approach for cell finding is to iteratively refine our estimate of the cell position and then "remove" this cell from the data. We do this by first creating a spherical template with a diameter roughly equal to that of the cell; the exact choice of parameter was learned through a hyperparameter search. We apply a 3D-FFT to convolve the spherical template with the cell probability map produced by *ilastik*. This produces a "sphere map" which gives us high responses in regions that are likely to contain cell bodies. At each step of our algorithm, we select the global maxima of the sphere map to be the centroid of the next detected cell. After finding this cell, we then zero out the probability map in this region so that we cannot select a candidate cell in this same location again, and repeat this matching procedure until convergence. We define convergence as the point at which the correlation between the probability map and our template drops below a user-specified threshold or reaches the maximum number of iterations.

CELL SIZE ESTIMATION

After finding the centroids of all detected cells, we can then efficiently estimate their sizes. To do this, we center a small spherical template at the detected center of each cell

and estimate the cell size by varying the template size. When the template can no longer be inscribed within the cell body, we observe a sharp decay in the correlation. We compute the correlation between the probability map while increasing the diameter of the spherical template, find the maximum decrease in correlation, and select this diameter as our estimate of the cell size. This operation has low complexity and can be performed on the entire (cubic mm) dataset on a single workstation. Once we have detected cells, estimating the diameter of the cell body is a simple one-dimensional fitting problem.

HYPERPARAMETER SEARCHES

We developed a tool to run hyperparameter searches over our methods to maximize performance on the ground truth volume V1. After exploring the parameter space, we ran a grid search over the most critical parameters (cell size, dilation, and threshold cutoff) to find a stable, optimal point. We select the parameters (cell size: 18, dilation: 8, threshold 0.47) that maximized f_1, the harmonic mean between precision and recall. Because voxels on the edge of volumes have inherent ambiguity for both human and machine annotators, we choose to disregard objects at the edge (of both detected and truth volumes) when computing precision and recall scores throughout this manuscript to ensure the most representative result.

NON-PARAMETRIC DENSITY ESTIMATION.

To compute the density of detected cells within a volume, a k-nearest neighbor (kNN) density estimation algorithm [101,102] was applied, which estimates the density using only distances between the samples (cells) and their k^{th} nearest neighbor. More concretely, the distance between a centroid vector $\mathbf{x} \in \mathbb{R}^3$ and a matrix $\mathbf{A}$ is defined as

$$\rho_k(\mathbf{x}, \mathbf{A}) = \|\mathbf{x} - \mathbf{a}_k\|_2^2,$$

where $\mathbf{a}_k$ is the k^{th} nearest neighbor to $\mathbf{x}$ contained in the columns of $\mathbf{A}$. The value of the empirical distribution p at $\mathbf{v} = (x, y, z)$ was then estimated using the following consistent estimator [101]:

$$p(\mathbf{v}) \propto \frac{k}{N\rho_k(\mathbf{v}, \mathbf{V})},$$

and $\mathbf{V}$ contains the centroids of the rest of the detected cells in the sample. We computed this quantity over a 3D grid, where the volume of each bin in the sample grid is $\text{Vol} = (8.44\mu\text{m})^3$. We selected this bin size to ensure that detected cells will lie in roughly a single grid point. This choice was further confirmed by visually inspecting the resulting density estimates. After computing the density for each 3D bin in our selected grid, we normalized to obtain a proper probability density function. Finally, we computed an estimate of the number of cells per cubic millimeter as, $p_d(\mathbf{v}) = (p(\mathbf{v})N/\text{Vol}) \times 10^9$. The intuition behind this approach is that in regions where we have higher density of samples, the quantity $\rho_k(\mathbf{v}_i, \mathbf{V})$ will be very small, and therefore the probability of generating a sample at this location is large. In practice, we set $k = \sqrt{N}$ which guarantees that the estimates of p will

asymptotically converge to the exact point estimates of the distribution since ρ_k converges to 0 as $N \to +\infty$ [101].

DETAILS OF EXPERIMENTS ON LARGE-SCALE DATASETS.

After validating and benchmarking our algorithms, processing was scaled to the entire dataset of interest (x voxels: 610-2010, y: 1-2480, z: 390-2014, resolution: 0), using the LONI processing environment [43]. *NeuroData* was used to obtain and store data: image data was requested for each computed block and the results were written to a spatially co-registered annotation channel [103]. Each block was retrieved an processed in an embarassingly parallel manner with sufficient padding to provide edge context; the results were uploaded to a *NeuroData* annotation project. We have also implemented an alternative merging strategy to account for cells near boundaries. Briefly, we eliminated putative detections touching an edge or that overlap an object already present in the database to further reduce edge effects.

4.4.2 Evaluation metrics

To compute human-to-human agreement and evaluate the performance of our methods, we developed tools to compare segmentations at both the pixel and object level. Detected pixels/objects that do not appear in the manual segmentation are counted as false positives, and manually identified pixels/objects not found by the automatic segmentation algorithm result in false negatives (misses). When evaluating the performance of our methods for

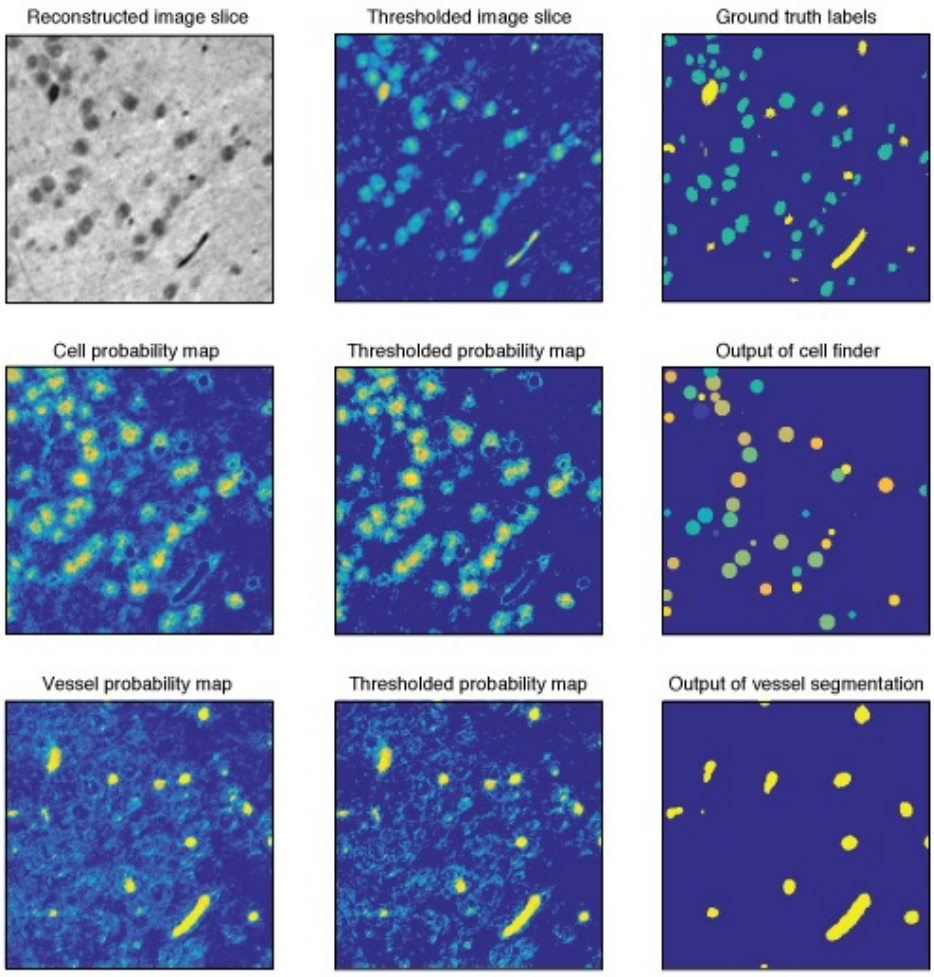

Figure 4.4: *Results of X-BRAIN pipeline for vessel segmentation and cell detection.* In the top row, (left) a reconstructed image slice in false color, (middle) mean thresholded slice, and (right) ground truth labels for both cells (green) and vessels (yellow). In the second row, (left) the cell probability map we obtained after training a Random Forest classifier on the data with *ilastik*, (middle) the mean thresholded probability map, and (right) the output of our greedy sphere finder approach which operates on the cell probability map to obtain an estimate of the centroid and diameter of cells. In the third row along the bottom (left) the vessel probability map, (middle) the thresholded map, and (right) the output of our segmentation algorithm.

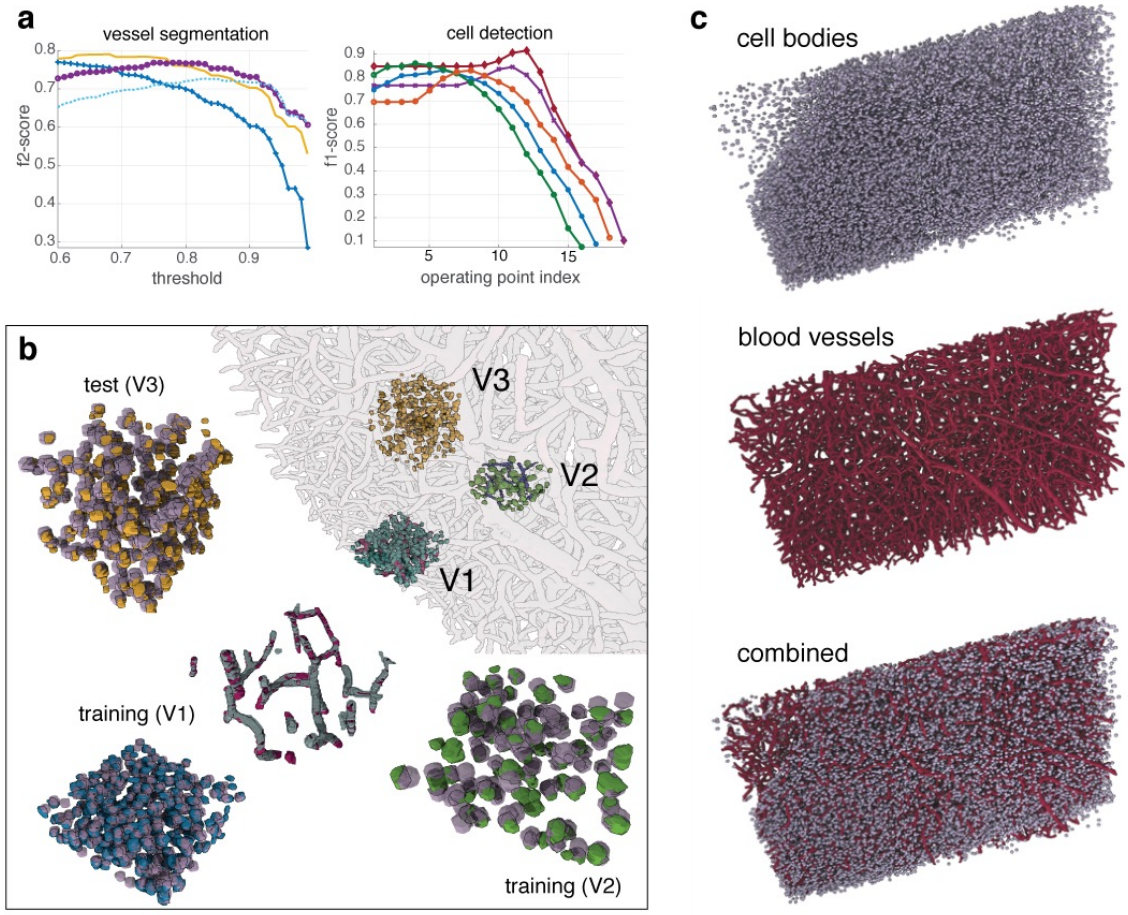

Figure 4.5: *Automatic methods for segmentation and cell detection reveal dense mesoscale brain maps.* (a) f_2 score performance of our vessel segmentation; each curve represents a varying vessel segmentation threshold (left) and f_1 score for cell detection (right) as we increase the stopping criterion (x-axis) in our greedy cell finder algorithm. In (b), the results of our cell detection and vessel segmentation algorithms are visualized for training (V1, V2) and test (V3) volumes, both inside the entire volume (right) and individually (left). We overlay the results of X-BRAIN on the three volumes, based upon the best operating point selected in (a). In (c), we show renderings of the output of our cell detection and vessel segmentation algorithms on the entire cubic mm sample.

detecting cells (object-level errors), we compute matches between two sets of centroids by identifying cell pairs in different segmentations that are nearest neighbors. If the matching centroids are within a fixed distance ($10\ \mu\mathrm{m}$) from one another, we label them a match and remove both cells from the dataset to avoid duplicate assignments. The matching process iterates until all possible matches are found, and precision and recall metrics are computed. For cell detection, we compute the f_1 score as it places equal weight on precision and recall. However, in the case of the pixel-level segmentation of vessels, we observe that optimizing the f_2 score produces more accurate results (confirmed by visual inspection).

4.4.3 Manual labeling and human-to-human agreement

In order to obtain ground truth datasets to quantify the performance of our algorithms and to assess human-to-human agreement, we used a total of four trained annotators (A0, A1, A2, A3) and five novices to label different sub-volumes (V0, V1, V2, V3) of our image dataset using ITK-Snap [41]. Two of the trained experts (A0, A1) and the five novices labeled cells and vessels in V1, a $195 \times 195 \times 65$ micron cube of data ($300 \times 300 \times 100$ voxels). Annotator A0 was instructed to produce a *saturated reconstruction*, where all cells and vessels were fully labeled. A1 produced a saturated segmentation of a sub-volume of V1, which we denote as V0.

A good dataset, at minimum, should allow human annotators to clearly see the structures of interest and in turn, reliably annotate them. We thus measured human annotator ability in finding and labeling cell bodies and blood vessels in multi-view projections

(orthogonal 2D projection planes) of the 3D image data.

To then compute human-to-human agreement across annotators, we computed the voxel-wise precision and recall between V0-A0 (ground truth) and V0-A1, which we computed to be $(p, r) = (0.93, 0.58)$ for cell bodies and $(p, r) = (0.99, 0.29)$ for vessels. While precision is high in both cases, the recall is much lower. This is due to the fact that A1 produces an underestimate of A0's labels; we tested this by dilating A1's labels until we maximized the f_1 score between both annotations. In this case, we obtained a precision $(p, r) = (0.84, 0.76)$ for cell bodies and $(p, r) = (0.85, 0.73)$ for vessels.

We then computed the agreement between these annotators in detecting cell centroids. We first cleaned each segmentation to ensure all cells are disconnected from one another. We then applied a connected component algorithm and found the center of mass of each component to estimate the centroid of each cell. We matched centroids across the two annotations and computed object-level precision and recall. When ignoring cells along the boundaries of the volume, there are no cells identified by A1 that are not identified by A0 and only one cell identified by A0 that was not identified by A1. While precise manual segmentation of the boundaries of cell bodies and vessels is challenging, human-to-human agreement is nearly perfect when identifying cell centers ($(p, r) = (1, 0.989)$). We conclude that the data is of sufficient quality to segment cell bodies and vessels with automated methods.

In addition to computing human agreement, we also acquired additional volumes for testing our algorithms. For these purposes, we had another expert annotator (A2) densely

label the cells and partially label the vessels in a training volume (V2). Annotator A1 then edited all cells in this volume, which we denoted as V2-A12. Finally, to test our methods, we had an external party randomly select a sub-volume to be used as a hold-out test volume at a location unknown to the authors of this manuscript. Annotator A0 and A3 then iteratively refined a common estimate of the cell centroids in volume (V3); this annotation is referred to as V3-A03. V3 is used as a hold-out test set to evaluate the accuracy of our cell detection method.

To quantify the time required to label the centroids of cell bodies, we recruited five subjects with no previous experience to label the centers of cell bodies in 3D. Each subject was instructed to label as many cells as possible in thirty minutes. The average number of cells that these subjects labeled was 51.2 and the median was 62. These results suggests that a novice can label the centroids of around 100 cells in one hour. In practice, we find that it takes experts around 5 hours to reliably label all cell centers in a $(100\ \mu\mathrm{m})^3$ volume. From estimates of the cell density in mouse cortex, we expect around 120,000 cells per cubic millimeter; therefore, to manually annotate all cells in a cubic mm would require a projected 1200 person-hours or 50 days working 24 hours per day.

4.4.4 Data Accessibility and Reproducibility

We uploaded the raw and masked images into *NeuroData*. Additionally, we stored the annotations and segmentations resulting from our analysis in the *NeuroData* spatial database to facilitate rapid access, dissemination, and analysis of the data. This framework

allows for researchers to freely download arbitrary volumes of raw data, manual labels, or automated annotations for algorithm development or analysis. Users may also query the metadata of detected cells within a volume, which enables rapid knowledge extraction from the X-ray datasets and statistical analyses at scale.

4.5 Results

4.5.1 Computing the signal-to-background (SNR)

To estimate the intrinsic difficulty of separating cells from their background, we calculated the ratio of the intensity between cells and their exteriors. To do this, we sampled 10 cells every 25 slices ($15.6\ \mu\mathrm{m}$) in each of the three manually annotated volumes (V1, V2, V3) using ITK Snap. We placed a small circular marker within the cell's boundary and a marker outside of the cell in a location where the cell's boundary is clearly resolved. This generated 30 samples in both V1 and V2 and 89 samples in V3, of the brightness inside (signal) and outside (noise). We then computed the signal-to-noise ratio (SNR) for the i^{th} cell as follows:

$$SNR = 20 \log_{10} \left(\frac{s_i}{n_i} \right),$$

where s_i (signal) and n_i (noise) contains the mean value of the labeled pixels within and outside of the i^{th} labeled cell, respectively. The mean and standard deviation of the SNR (dB) across each subvolume is: V1 = $(4.73, 0.69)$, V2 = $(4.59, 1.49)$, V3 = $(4.49, 1.13)$.

Thus, we observe the largest variance in SNR in V2 and the lowest average SNR in V3. The training volume V1 appears to have the highest mean and lowest variance out of the three volumes. Our estimates of SNR appear to be predictive of the difficulty of the segmentation task, and are correlated with the accuracy of our segmentation results on the different volumes.

4.5.2 Detection Performance

To optimize each stage of our segmentation pipeline, we performed an exhaustive grid search to find the set of hyperparameters (i.e., threshold parameters for cell/vessel detection, the size of spherical template, and the stopping criterion for the cell finder) that maximize a combination of the precision and recall (f-score) between our algorithm's output and manually annotated data from volume V1 (Figure 4.5a-b). After tuning our cell detection algorithm to find the best set of hyperparameters, we were able to obtain a precision and recall of $(p, r) = (0.86, 0.84)$ on the same volume. Our initial results on this training volume and visual inspection of large-scale runs (Figure 4.5c) suggest that our methods provide reliable maps of the cells and vessels in the sample.

The image data varies across space, due to various details of the imaging and reconstruction pipeline. Therefore, it is important to test that our segmentation algorithm works reliably across regions previously unseen during classifier training. We labeled and tested our cell detection algorithm on two additional test cubes V2 and V3 (Figure 4.5b) that are spatially disjoint from V1 and each other. V2 served an initial test set, as we added some

sparse training data from this volume to train our *ilastik* classifier. V3 served as a held-out test set, as the location of this cube was unknown before tuning and running the algorithm on the entire dataset. After obtaining ground truth labels, we ran X-BRAIN on V2 and V3, using the set of parameters selected by optimizing our method on V1. The precision and recall is given by $(p, r) = (0.83, 0.76)$ and $(p, r) = (0.94, 0.78)$, for V2 and V3 respectively. These results suggest that X-BRAIN generalizes well across different regions of the sample, and is robust to fluctuations in brightness and contrast.

The variation in training and test volume performance can be partially explained by fluctuations in the brightness, introduced during tomographic image reconstruction. To understand the connection between the fluctuations in contrast and difficulty of the cell detection problem, we computed the SNR across multiple cells within each of the labeled volumes. The mean and standard deviation of the signal-to-noise (SNR) between cells and their background in all three volumes was V1 = $(4.73, 0.69)$, V2 =$(4.59, 1.49)$, and V3 = $(4.49, 1.17)$. As expected, the precision and recall (for cell detection) seem to be correlated with the variance of the SNR in the volume (providing a measure fluctuations in contrast). In particular, we obtained the lowest precision and recall for V2, and indeed, this volume exhibited the highest variance in the contrast between cells and their background. Even in light of these brightness fluctuations, our sensitivity analysis (Figure 4.5a) and results (Figure 4.5b) on training and test volumes suggest that X-BRAIN generalizes well across different regions of the volume. Furthermore, visual inspection of our large-scale results (Figure 4.5c), reveals a good correspondence between cells and vessels that are visible in

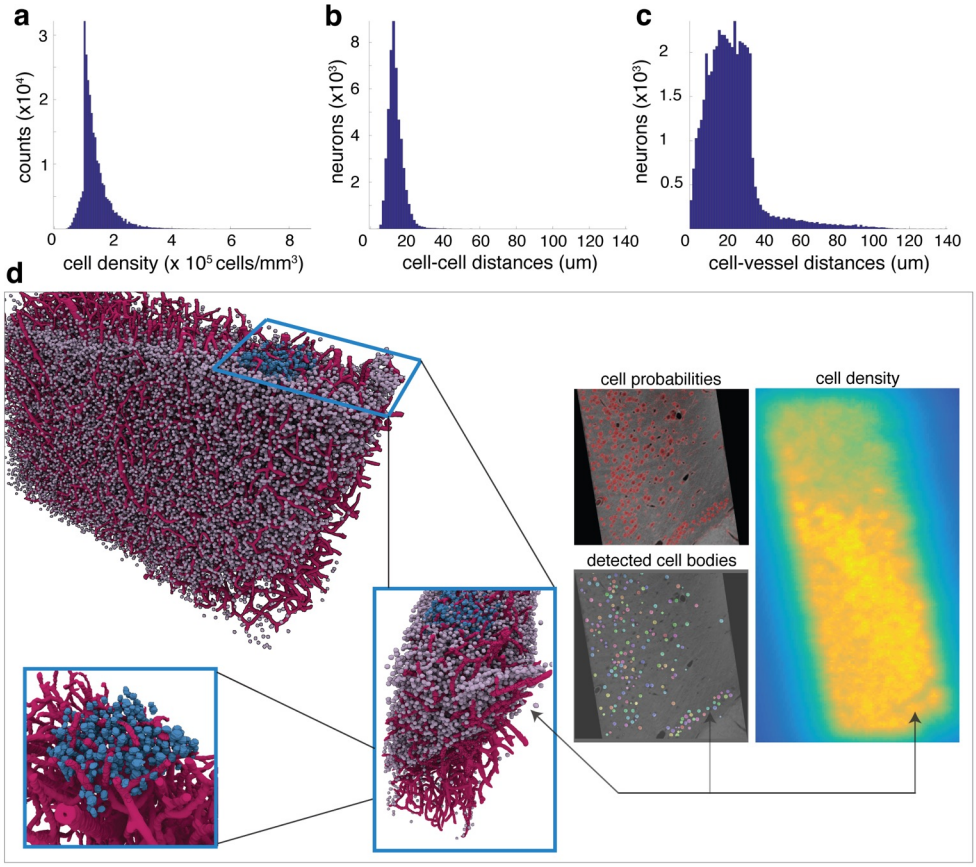

Figure 4.6: *Spatial statistics of X-ray volumes reveal layering and spatially-diverse distribution of cell bodies.* We display histograms of: (a) the estimates of the cell density over the extent of the entire sample of mouse cortex, (b) distances between the center of each cell and its nearest neighbor (cell-to-cell distances), and (c) distances between the center of each cell and the closest vessel voxel (cell-to-vessel distances). In (d), we visualize the data and confirm neuroantomical structure. We show a 3D rendering of the detected cells and vessels in the entire sample, with a manually labeled cube (V1) highlighted in blue. To confirm the 3D structure seen in these visualizations, on the right, we confirm the same 3D structure in the cell probability maps (red indicating high probability), detected cell maps (each detected cell displayed in a different color), and density estimates. This result provides further confirmation that the 3D structure of the sample is preserved in our density estimate.

slices and those detected algorithmically. These results suggest that our methods are robust and can be applied at large scale.

4.5.3 Scalable processing

We applied our pipeline to segment vessels and detect cells in a cubic mm sample ($2560 \times 2560 \times 1624$ voxels) of excised brain tissue collected from mouse somatosensory cortex (Figure 4.5). To apply X-BRAIN to large datasets, we created an analytics workflow that uses (but does not require) the LONI Pipeline environment [43] to automatically distribute jobs across a cluster environment. Our workflow is parallelized by dividing our large dataset into small data blocks which can be processed independently, based upon a user-specified graphical (xml-based) description of the dependencies between various algorithms. Running our analytics pipeline on a cubic mm sample took approximately six hours on a small 48-core cluster. As a result, we detected $48{,}689$ cells over the extent of the analyzed sample ($\sim$0.42mm^3).

4.5.4 Quantifying cellular and vascular information

CELL DENSITY

To compute the spatially-varying density of cells, we applied a robust non-parametric approach for density estimation. Adopting a non-parametric approach enables us to obtain an accurate estimate of the distribution without making any restrictive assumption on its

form. In particular, we rely on the popular k-nearest neighbors (kNN) density estimation algorithm [101,102], which estimates a distribution using only distances between the samples (cells) and their k^{th} nearest neighbor. When applied to the entire volume, we calculated an average density of 1.3×10^5 cells per mm^3 (Figure 4.6a). These results are comparable to other studies that estimate an average of 1.2-2.5×10^5 cells per mm^3 in mouse cortex [104], both in terms of our average, and the spread in the distribution. These density estimates provide important information about the spatially-varying distribution of cells within the sample.

VASCULATURE DENSITY

The location of cell bodies, relative to one another, and relative to the vasculature, is important for studying diseases that afflict the brain [104]. We developed automated tools to compute distances between detected cell centers (cell-to-cell distances, Figure 4.6b) and distances between each cell and the closest segmented vessel (cell-to-vessel distances, Figure 4.6c). Cell-to-vessel distances are spread between 10-40 μm, with very few cells exceeding this distance (34.3 ± 533.4 μm). In contrast, the cell-to-cell distances appear to be much more concentrated, with a strong peak at 12.7μm and with much smaller variance (21.3 ± 43.1 μm). The distribution of distances between cells and vessels (Figure 4.6b) aligns with previous results [104, 105] and confirmed the accuracy of our approach for large-scale analysis. We further estimated that the fractional volume of vessels in the sample was 1.85%. This estimate is in agreement with previous studies [104, 105, 106],

which estimate the fractional density of vessels in the cortex to range from $0.97 - 3.64\%$. These results further confirm that our methods can be used to compute information about the relationship between cells and vasculature in the brain. The information described in this section has been summarized in Table 4.1.

VISUALIZATION

To complement our analysis tools, we developed methods to produce and visualize mesoscale maps, with the cellular density and vasculature as their output. These methods are integrated into the NeuroData framework; after running a sample through our pipeline, users can download different descriptions of the neuroanatomy, either alone, or combined with the image data to help reveal relevant structures in the images. Using these multiple modes of visualization (Figure 4.6d), we identified a 3D structure with extremely high cell density, clustered at the bottom of the sample (Layer 6). We confirmed this structure in both 3D visualizations (left), in X-ray micrographs, the cell probability maps, and in our estimate of the cell density (right). All of these representations provide information and descriptions of the data that can be used to further visualize and quantify its neuroanatomy. The combination of dense reconstructions of cells and blood vessels provide a unique approach for studying the joint distribution of brain cytoarchitecture and vasculature. A snapshot of the XBRAIN results is shown in Figure 4.7 in the *NeuroDataViz* environment.

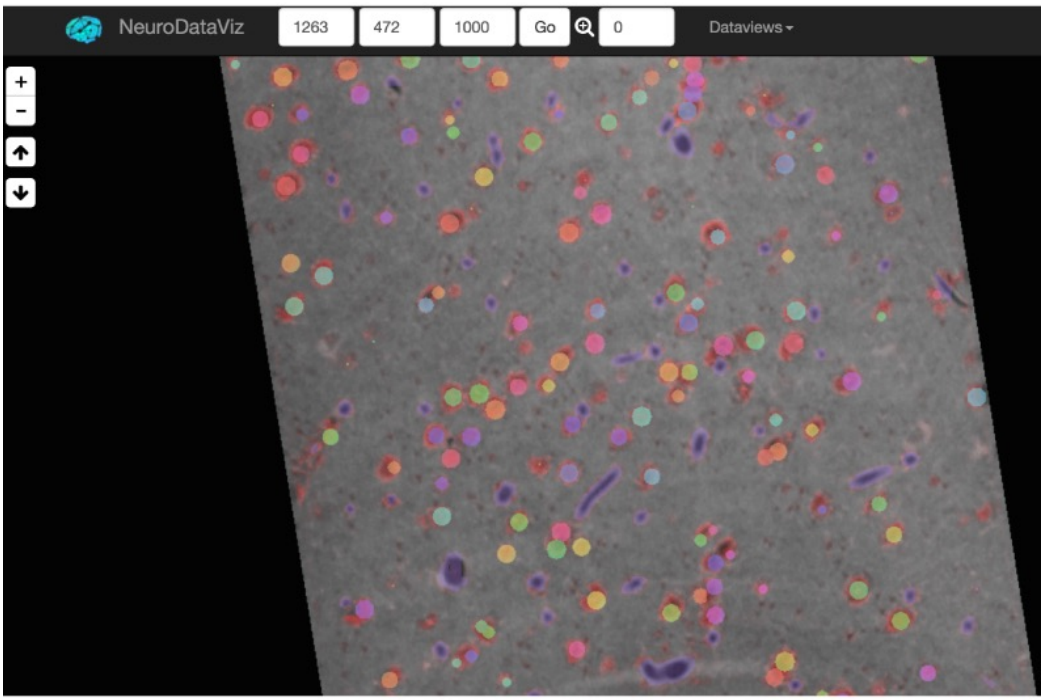

Figure 4.7: *Visualization of cell and vessel segmentation performance.* The results of X-BRAIN processing on our data sample as visualized through NeuroData's visualization service (ndviz) and users can easily traverse through the volume using NeuroData's web-based GUI. The cell probabilities (translucent red) and final cell detections (opaque multi-color, where each color represents a unique ID for a cell), and the vessel segmentation (translucent purple), are all overlaid on the corresponding X-ray image.

Annotation	# Cells	Cell area	Volume ($\%$ of mm^3)	Density ($10^5/\mathrm{mm}^3$)
V0-A0	97	$(2136, 2060)$	0.06	1.63
V0-A1	96	$(1489, 1499)$	0.06	1.28
V0-XBRAIN	94	$(1983, 2123, 51)$	0.06	1.57
V1-A0	321	$(1997, 2035)$	2.5	1.28
V1-XBRAIN	302	$(1983, 1963, 56)$	2.5	1.21
V2-XBRAIN	112	$(1918, 1963, 62)$	0.06	1.87
V3-A03	281	N/A	0.2	1.41
V3-XBRAIN	240	$(1419, 1385, 42)$	0.2	1.20
Vtot-XBRAIN	$48,689$	$(1454, 1385, 60)$	42	1.02

Table 4.1: *Statistics of manually labeled volumes, cell counts, and sizes for different volumes and annotators.* In the first column, we display the name of the volume (V0, V1, V2, and V3) and annotator to identify each manual (A0, A1, A2, A3) or automated (XBRAIN) annotation. In the second column and third columns, we report the number of detected cells and the mean/median size of annotated cell bodies (number of labeled voxels). The training datasets include V0 (a subset of V1), V1, and V2. Volume V3 is held-out test set which whose location was unknown during training and tuning the parameters of the algorithm.

4.6 Discussion

We have shown that μCT can be used to rapidly quantify mesoscale neuroanatomy in a millimeter scale sample without sectioning. Our results demonstrate how osmium-stained

and plastic-embedded brains, in conjunction with a synchrotron X-ray source, produce sufficient contrast and resolution to automatically detect blood vessels and cell bodies. Our approach to automated anatomy is uniquely poised to provide detailed, large, mesoscale maps of the brain.

The high precision and recall of our algorithms suggest that our segmentation and cell detection methods can be used to reliably and quickly survey data volumes and identify cells and vessels in the sample. We can use these methods to build more systematic studies of regions of interest with EM, once the large-scale structure is identified using μCT. Information about where cells and vessels lie can be used as a prior in segmentation algorithms (in EM) and also to improve subsequent registration and alignment. As our pipeline for X-ray image analysis has been integrated in *NeuroData*, we can readily combine existing EM analysis pipelines with our methods to analyze the same dataset with μCT and EM. These results can be combined to create a multi-modal brain map that contains information about the cytoarchitectural and cerebrovascular properties of a sample, in addition to the fine-scale information about the processes and synapses afforded by EM.

Knowledge about the macro-scale organization of the brain, such as Brodmann maps [107], have been based primarily on human anatomists working with thin, sparsely-labeled slices of brains. However, with developments in large-scale connectomics with EM [33,49] and the techniques we present here for μCT, far larger and more comprehensive datasets become possible, but are too large for manual analysis. The capabilities of synchotron source X-ray microscopy, combined with staining approaches for entire brain preparation

[92], offers the possibility of imaging entire brains at the mesoscale. With these capabilities, it should become possible to obtain brain maps in a new, data-driven fashion, enabling the massive-scale quantification of a broad set of effects related to disease, development, and learning in the brain.

4.7 Summary of Chapter Contributions

In this chapter we explore mesoscale neurocartography and demonstrate the ability to analyze millimeter scale data at sub-micron resolution using a method that does not require tissue sectioning and post-imaging alignment. This approach is significantly faster than conventional methods and still allows for the automatic reconstruction of neuronal cells and vasculature. Using this framework, we create a coarse map that can be refined through subsequent imaging. We leverage the *NeuroData* framework and demonstrate our framework for open science on a new modality to enable large-scale discovery.

Chapter 5

Nanoscale Images to Graphs

Reconstructing a map of neuronal connectivity is a critical challenge in contemporary neuroscience. Recent advances in high-throughput serial section electron microscopy (EM) have produced massive 3D image volumes of nanoscale brain tissue for the first time. The resolution of EM allows for individual neurons and their synaptic connections to be directly observed. Because recovering neuronal networks by manually tracing each neuronal process at this scale is unmanageable, researchers are developing automated image processing modules. To date, state-of-the-art algorithms focus only on the solution to a particular task (e.g., neuron segmentation or synapse identification).

We present the first fully-automated images-to-graphs pipeline (i.e., a pipeline that begins with an imaged volume of neural tissue and produces a brain graph without any human interaction). To evaluate overall performance and select the best parameters and methods, we also develop a metric to assess the quality of the output graphs. We develop

state-of-the art methods for detecting synapses, and also evaluate a set of neuron segmentation algorithms and parameters, searching possible operating points to identify the best available brain graph for our assessment metric. Finally, we deploy a reference end-to-end version of the pipeline on a large, publicly available data set. This provides a baseline result and framework for community analysis and future algorithm development and testing. All code and data derivatives have been made publicly available toward eventually unlocking new biofidelic computational primitives and understanding of neuropathologies.

In existing approaches, one of the key, commonly-occurring error modes is dendritic shaft-spine fragmentation. We posit that directly addressing this problem of *connection identification* may provide critical insight into estimating more accurate brain graphs. To this end, we develop a network-centric approach motivated by biological priors and image grammars. We build a image analysis pipeline to reconnect fragmented spines to their parent dendrites using both fully-automated and semi-automated approaches. Our experiments show we can learn valid connections despite uncertain segmentation paths. We curate the first known reference dataset for analyzing the performance of various spine-shaft algorithms and demonstrate promising results that recover many previously lost connections. Our automated approach improves the local subgraph score by a factor of 4, and the full graph score by 60 percent. These data, results, and evaluation tools are all available to the broader scientific community. This reframing of the connectomics problem illustrates a semantic, biologically-inspired solution to remedy a major problem with neuron tracking.

5.1 VESICLE

Synapses, which are a key communication structure in the brain, are particularly difficult to detect due to their small size and limited contrast. Prior work in automated synapse detection has relied upon time-intensive, error-prone biological preparations (e.g., isotropic slicing, post-staining) in order to simplify the problem.

Here we present VESICLE (Volumetric Evaluation of Synaptic Interfaces using Computer Vision at Large Scale), the first known approach designed for mammalian synapse detection in anisotropic, non-poststained data. Our methods explicitly leverage biological context, and the results exceed existing synapse detection methods in terms of accuracy and scalability. We provide two different approaches - a deep learning classifier (VESICLE-CNN) and a lightweight Random Forest approach (VESICLE-RF), to offer alternatives in the performance-scalability space. Addressing this synapse detection challenge enables the analysis of high-throughput imaging that is soon expected to produce petabytes of data, and provides tools for more rapid estimation of brain-graphs. Finally, to facilitate community efforts, we develop tools for large-scale object detection, and demonstrate this framework to find $\sim$50,000 synapses in 60,000 μm^3 (220 GB on disk) of electron microscopy data.

5.1.1 Overview

Mammalian brains contain billions to trillions of interconnections (i.e., synapses). To date, the full reconstruction of the neuronal connections of an organism, a "connectome,"

has only been completed for nematodes with hundreds of neurons and thousands of synapses [24, 108]. It is generally accepted [5, 109] that such wiring diagrams are useful for understanding brain function and contributing to medical advances. For example, many psychiatric illnesses, including autism and schizophrenia, are thought to be "connectopathies," where inappropriate wiring mediates pathological behavior [110]. Reliably and automatically identifying synaptic connections (i.e., brain graph edges) is an essential component in understanding brain networks.

Although the community has made great progress towards automatically and comprehensively tracking all neuron fragments through dense electron microscopy data [111,112], current state-of-the-art methods for finding synaptic contacts are still insufficient, especially for large-scale automated circuit reconstruction.

In order to detect synapses in electron microscopy data, neuroscientists typically choose to image at $\sim$5 nm per voxel in plane, with a slice thickness of $\sim$5-70 nm. Capturing complete neurons therefore requires processing terabytes to petabytes of imaged tissue. The largest datasets currently available (and of sufficient size to begin estimating graphs) are acquired using scanning electron microscopy (SEM) or transmission electron microscopy (TEM) due to their high throughput capability [23, 113]. These methodologies scale well, but provide a challenging environment for object detection. The slices are thick relative to in-plane resolution (i.e., anisotropic), due to methodological limitations, and often do not have optimal staining to visually enhance synaptic contacts. The detection algorithms proposed in this paper are specifically implemented to address these challenges. We train a

random forest classifier (VESICLE-RF), which leverages biological context by restricting detections to voxels that have a high probability of being membrane. This classifier also relies on the identification of neurotransmitter-containing vesicles, which are present near chemical synapses in mammalian brains. We also present a deep learning classifier (VESICLE-CNN) to find synapses, which improves performance at the expense of additional computational complexity.

Both of the VESICLE classifiers provide state-of-the-art performance, and users may choose either method, depending on their environment (e.g., the importance of performance vs. run time, computational resources, availability of human proofreaders). Our classifiers provide new opportunities to assess neuronal connectivity and can be extended to other datasets and environments.

5.1.2 Previous Work

Previous methods for synapse detection have taken several approaches, including both manual and automated algorithms. Two recent approaches, Kreshuk2011 [114] and Becker-2013 [115] address the synapse detection problem in post-stained, isotropic, focused ion beam scanning electron microscopy (FIBSEM) data. Kreshuk2011 uses a Random Forest voxel-based classifier and texture-based features to identify pronounced post-stained post-synaptic densities. This approach is insufficient for our application because of the anisotropy and much lower contrast of our synaptic regions (Figure 5.1), as well as the computational expense. Becker2013 also uses a voxel-based classification approach and features similar

to Kreshuk2011; however, Becker2013 extends the approach by considering biological context from surrounding pre- and post-synaptic regions at various sizes and locations, based on the synapse pose. This technique relies on full 3D contextual information and greatly reduces false positives compared to Kreshuk2011.

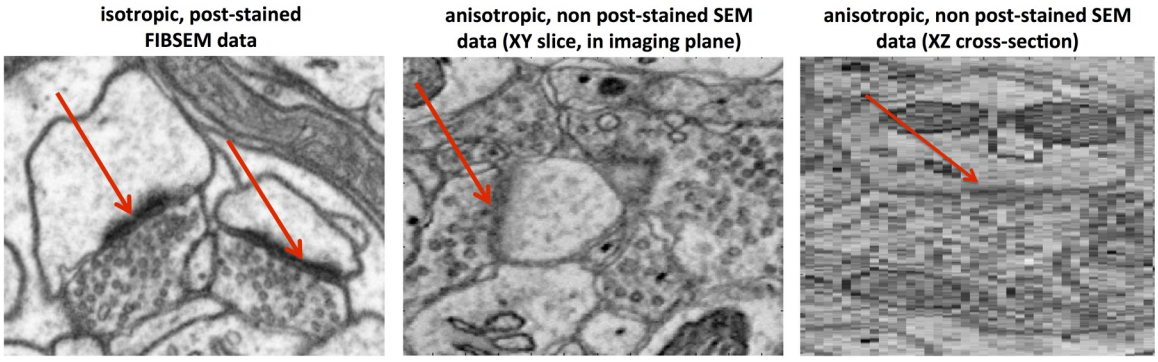

Figure 5.1: *Examples highlighting the synapse finding challenge.* (Left) Previous work on synapse detection has focused on isotropic post-stained data, which shows crisp membranes and dark fuzzy post synaptic densities (arrows) from all orientations. (Middle, Right) The alternative imaging technique of non post-stained, anisotropic data promises higher throughput, lack of staining artifacts, reduction in lost slices, and less demanding data storage requirements - all critically important for high-throughput connectomics. (Right) The XZ plane of a synapse in anisotropic data is shown, illustrating the effect of lower resolution. We address this more challenging environment, in which membranes appear fuzzier and are harder to distinguish from synaptic contacts. Data courtesy of Graham Knott (left) and Jeff Lichtman (middle, right).

Our result was directly compared to the Becker2013 method [115] (which was found to be superior to Kreshuk2011 [114]). Other work on synapse detection exists but was not used as a comparison method in this manuscript; some methods rely on post-stained

data [116], post-stained data and accurate cell segmentation [117], or post-staining and tailoring for *Drosophila* synapses, which have very different appearances [37].

5.1.3 Methods

BIOLOGICAL CONTEXT

Synapses occur along cell membrane boundaries, between (at least) two neuronal processes. Although synapses occur in many different configurations, the majority of connections annotated in this dataset are axo-dendritic connections. The pre-synaptic axonal side is known as a bouton, characterized by a bulbous end filled with small, spherical vesicles. The synaptic interface is often characterized by a roughly ellipsoidal collection of dark, fuzzy voxels. In VESICLE-RF, we attempt to directly capture these features.

Prior to feature extraction, we leverage membranes (found using the deep learning approach [118]) which greatly reduces the computational burden and provides a more targeted learning environment for the classifier (Figure 5.2). The membrane-finding step is computationally intensive (requiring about 3 weeks on 27 Titan GPU cards); however current approaches to neuron detection require membrane probabilities (e.g., [50, 51]) and so this leverages previously computed information.

We also identify clusters of vesicles by finding maximal responses to a matched filter extracted from real data followed by clustering to suppress false positives. This detector acts as a putative bouton feature to localize regions containing synapses. The vesicles are

also of biological interest (e.g., for synapse strength estimation). Vesicle detection is very lightweight and contributes negligibly to total run time (requiring only 3 hours for the entire 60,000 μm^3 evaluation volume).

RANDOM FOREST CONTEXT AWARE CLASSIFIER (VESICLE-RF)

We opted for a random forest classifier [119] due to its robust finite sample performance in relatively high-dimensional and nonlinear settings. Furthermore, recent research suggests that this approach significantly outperforms other methods on a variety of tasks [120, 121]. The output of the random forest is a scalar probability for each pixel, which we threshold and post-process to obtain a class label.

We began with a large set of potential feature descriptors (e.g., Haralick features, Gabor wavelets, structure tensors), and evaluated their performance based on a combination of Random Forest importance on training and validation data, computational efficiency, and ability to capture biologically significant characteristics. We pruned this feature set to ten features, retaining state-of-the-art performance in a computationally lightweight package. For efficiency reasons, we computed features in a two-pass approach. We first computed several data transforms on the two-dimensional EM slice data (to better account for the image anisotropy). We then created our features by convolving box kernels of different bandwidths with the results from the previous step. This allowed information to be summarized at different scales, as explained further in Table 5.1. Finally, we computed a feature capturing the minimum distance to a neurotransmitter-containing vesicle.

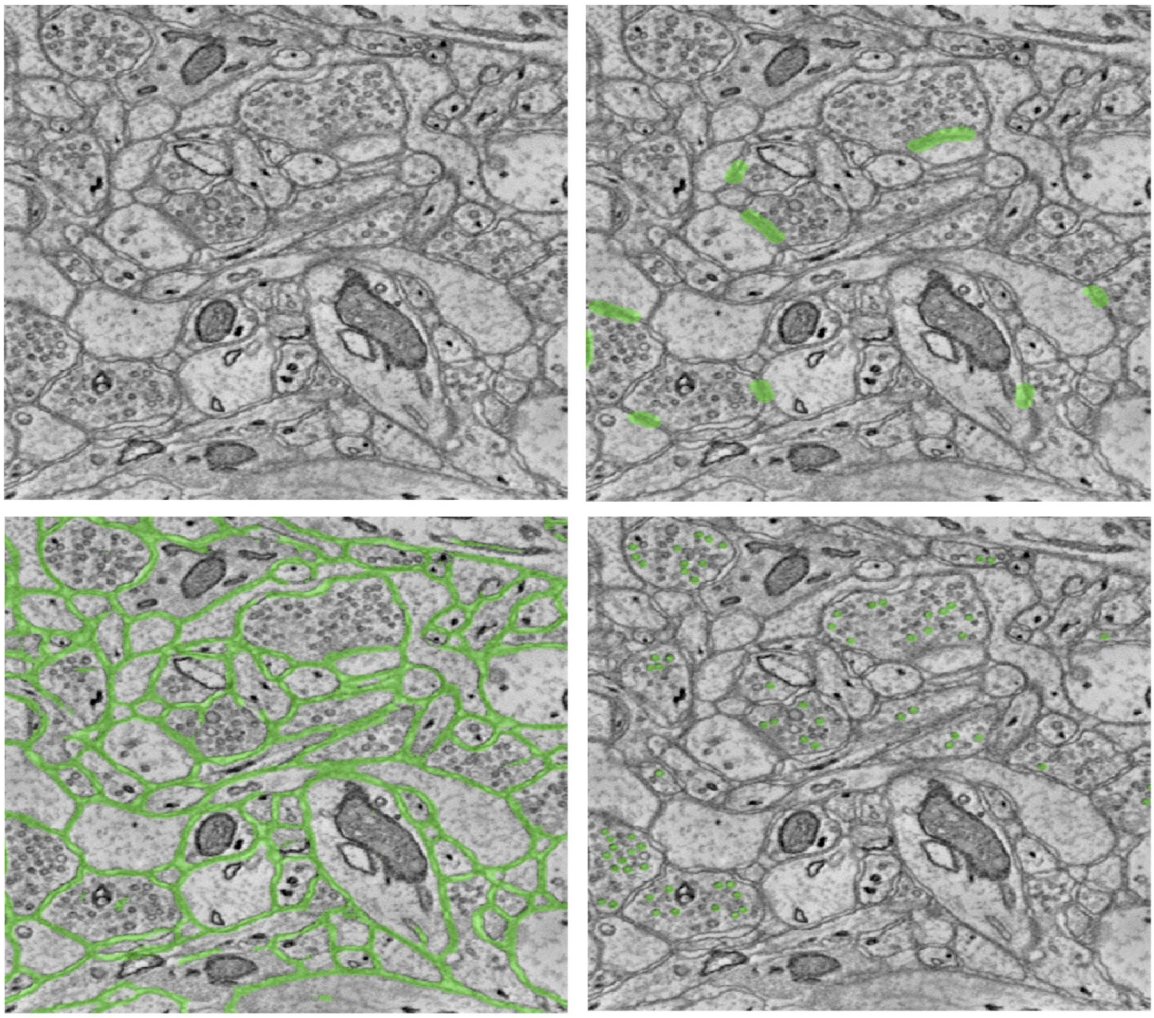

Figure 5.2: *Biologically inspired features.* (Upper left) A single cross-section of EM data is shown. (Upper right) The detection task is to identify synapses shown in green. (Lower left) These synapses are known to exist at the interface of two neurons; these boundaries can be approximated by previously computed membranes, allowing us to restrict the evaluation regions to the green pixels. (Lower right) Clusters of vesicles are a good indicator of an axonal bouton, suggesting that one or more synaptic sites is likely nearby. Vesicles found by our automated detection step are highlighted in green.

Table 5.1: *Description of features used in VESICLE-RF.* Data transforms are summarized using different kernel bandwidths: $\theta_0 : [5, 5, 1]$, $\theta_1 : [15, 15, 3]$, $\theta_2 = [25, 25, 5]$, $\theta_3 = [101, 101, 5]$, $\theta_4 =$ minimum vesicle distance.

Data Transform	**Box Kernel**
Intensity	θ_0, θ_1
Local Binary Pattern	θ_0
Image Gradient Magnitude	θ_1, θ_2
Vesicles	$\theta_2, \theta_3, \theta_4$
Structure Tensor	θ_1, θ_2

We train our classifier using 200,000 samples (balanced synapse, non-synapse classes). Putative synapse candidates are fused into 3D objects by thresholding and size filtering as described in Section 5.1.4. This method is scalable, requiring only a small amount of computational time and resources to train and test.

DEEP LEARNING CLASSIFIER - VESICLE-CNN

Deep convolutional neural networks (CNNs) have recently provided state-of-the-art performance across a wide range of image and video recognition problems. These successes include a number of medical imaging applications; a small sample includes mitosis detection [122], organ segmentation [123] and membrane detection in EM data [118]. The recent success in membrane detection is particularly compelling given the common imaging modality

and visual similarity between membranes and synapses. To test the hypothesis that CNNs may also provide an effective means of classifying synapses, we adopt and re-implement the pixel-level classification approach of [118] suitably adapted for our application.

As in section 5.1.3, each pixel in the EM cube is presumed to be either a synapse pixel or a non-synapse pixel. The features used for classification consist of a 65×65 tile centered on the pixel location of interest. For classification we use a CNN with three convolutional layers and two fully connected layers, roughly corresponding to the CNN designated "N3" [118]. This CNN is implemented using the Caffe deep learning framework [124]; the full architecture specification (e.g., types of nonlinearites and specific layer parameters) is encoded in the Caffe configuration files which are provided as part of our open source code.

During training we balance the synapse (target) and non-synapse (clutter) examples evenly; since synapse pixels are relatively sparse, this involves substantially subsampling the majority class. To focus the training on examples that are presumably the most challenging, we threshold the membrane probabilities described above, and use the result as a bandpass filter. Negative examples are drawn randomly from the set of non-synapse membrane pixels. We also add synthetic data augmentation by rotating the tiles in each mini-batch by a random angle (the insight behind this step is that synapses may be oriented in any direction). Our test paradigm does not rely on membrane probabilities; once trained, the deep learning classifier requires only EM data as input. The neurotransmitter-containing vesicles used in VESICLE-RF are not used in VESICLE-CNN for training or test.

5.1.4 Results

Our VESICLE classifiers were trained and evaluated for both performance and scalability, exceeding existing state-of-the-art performance. VESICLE was evaluated on an anisotropic ($3 \times 3 \times 30\ nm$ resolution) color-corrected [125] dataset of non-poststained mouse somatosensory cortex [113]. This is the largest known dataset of its kind. Prior to all processing, we downsampled the data to $6 \times 6 \times 30\ nm$ resolution. The training and test volumes were extracted from this larger EM volume. For training, each method used a $1024 \times 1024 \times 100\ \mu m^3$ region of data (denoted AC4). For testing, a non-contiguous, equally-sized cube from the same dataset was evaluated (denoted AC3). For the deep learning algorithm, a different size pad region was used due to training methodologies. A padded border was used on the test region for all algorithms to ensure that all labeled synapses in the volume were available for evaluation.

Gold standard labels for synapses were provided by expert neurobiologist annotators. The training labels were assumed to be correct (our classification result was evaluated in an open-loop process).

PERFORMANCE EVALUATION

We assess our performance by evaluating the precision-recall of synaptic objects. Pixel error, while potentially useful for characterizing synaptic weight and morphology, is a less urgent goal for connectomics, which must first identify the connections between neurons before ascribing attributes. A focus on pixel accuracy also can obscure the actual task of

connection detection.

A quantitative comparision of our performance relative to existing work is presented in Figure 5.3. Of particular significance is our performance at high recall operating points. For many connectomics applications, it is essential to ensure that the majority of connections are captured (i.e., low false negative rate); false positives can be remediated through a variety of approaches (e.g., biological plausiblity based on incident neurons, manual proofreading).

To construct precision-recall curves from VESICLE-RF and VESICLE-CNN pixel-level classification results, we developed a procedure to sweep over probability score thresholds (0.5-1.0) and create initial objects through a connected component analysis. We generated additional operating points by varying biologically motivated size (2D: 0-200 minimum, 2500-10000 maximum; 3D: 100-2000 minimum) and slice persistence (1-5 slices) requirements. For Becker2013, we ran the statically linked package provided by the author on our data volumes. We then followed their suggested method (similar to the VESICLE approach) to create synaptic objects from raw pixel probabilities by thresholding probabilities (0.0-1.0), running a connected component algorithm and rejecting all objects comprised of fewer than 1000 voxels [115].

When computing object metrics, we computed true positives, false positives, and false negatives by examining overlapping areas between truth labels and detected objects. We added the additional constraint of allowing each detection to count for only one truth detection, to disallow large synapse detections that cover many true synapses and provide

little intuition into connectomics questions. A version of the classifier was trained without the vesicle features to provide insight into the importance of biological context in this problem domain.

A qualitative visualization of our VESICLE-RF performance is shown in Figure 5.4.

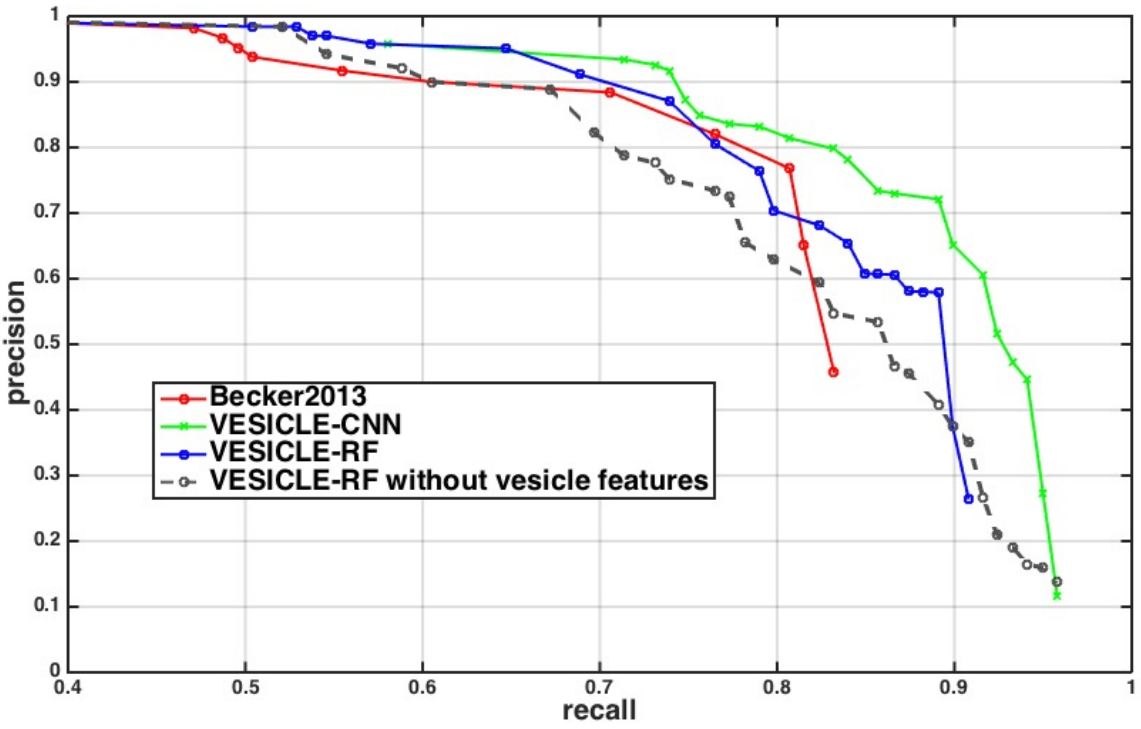

Figure 5.3: *VESICLE-RF and VESICLE-CNN significantly outperform prior state-of-the art, particularly at high recall rates.* The relatively abrupt endpoint of the Becker2013 method occurs because beyond this point, thresholded probabilities are grouped into large detected regions rather than individual synapses, which are disallowed.

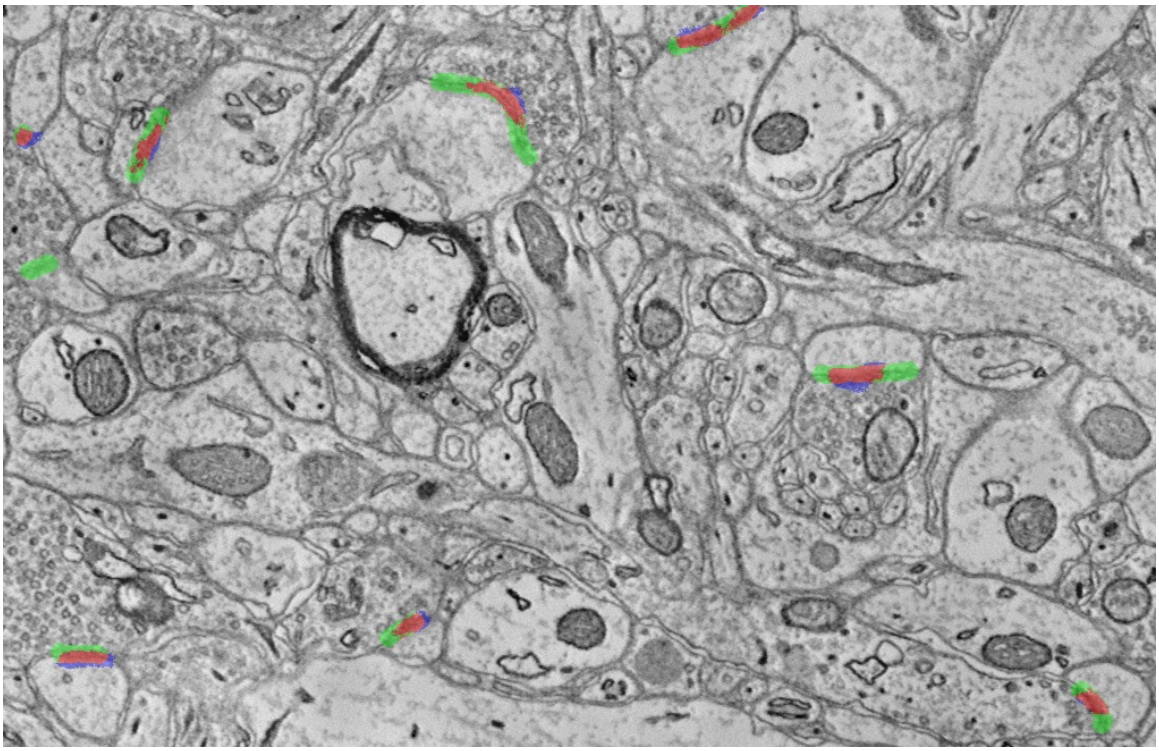

Figure 5.4: *Example VESICLE result.* Gold standard labels are shown in green, and VESICLE-RF detections are shown in blue. Red pixels represent True Positives (TP). Objects that are only green are False Negatives (FN) and objects that are only blue are False Positives (FP). Object detection results are analyzed in 3D, so single slices may be misleading.

SCALABILITY ANALYSIS

VESICLE-RF enables large-scale processing because of its light computational footprint and ability to be easily parallelized in a High Performance Computing (HPC) CPU environment. Relative to Becker2013 [115], this approach is dramatically less computationally intensive for training (8 GB RAM, 10 minutes v. 20 GB RAM, 11 hours). During evaluation our approach is approximately twice as fast (10 minutes versus 20 minutes, using

unoptimized Matlab code), and has one-eighth the maximum computational load with half the maximum RAM requirement. VESICLE-CNN required 56 hours to train and 39 hours to evaluate the test cuboid on a single GPU.

To demonstrate the scalability of our approach and our distributed processing framework, we applied our VESICLE-RF classifier to the largest available non-poststained, anisotropic dataset [113]. The inscribed cuboid is ~ 220 GB on disk; we downsample by a factor of two in the X and Y dimensions prior to processing (60,000 μm^3, 56GB on disk after downsampling). In Figure 5.5, we show a visualization of the 50,335 synapses found in this analysis. We chose the VESICLE-RF method here to emphasize the advantages of scalable classifiers; this method is ~ 200 times faster than VESICLE-CNN (evaluated as a single job on the same data cube). When deploying our classifier at scale, we increase our pad size to be larger than a synaptic cleft, and discard border detections; this allows us to avoid boundary merge issues. We also optionally allow the detection threshold to vary based on pixel probabilities to improve robustness.

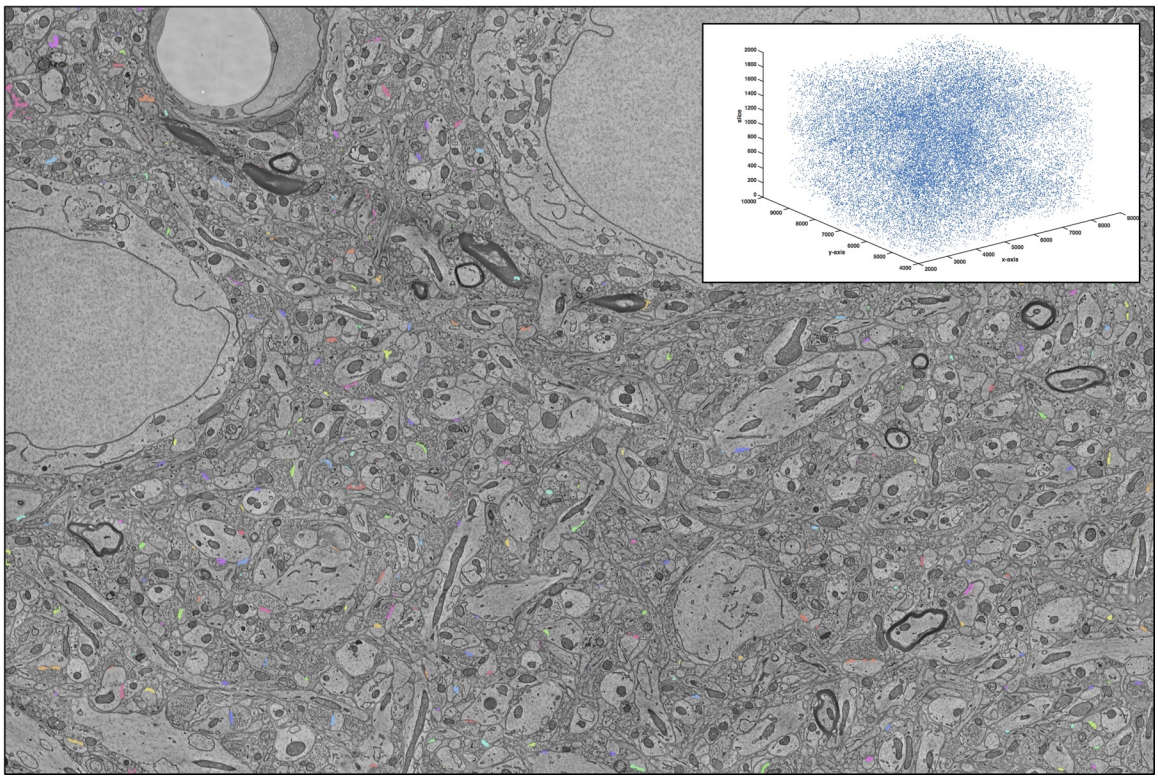

Figure 5.5: *Visualization of large-scale synapse detection results.* We found a total of 50,000 putative synapses in our volume. An XY slice showing detected synapses is shown, and a point cloud of the synapse centroids are also visualized (inset). A full resolution version of this image is available via RESTful query. Each synapse is represented by a different color label.

5.2 Images to Graphs

Brain tissue volumes imaged using electron microscopy today already contain many thousands of cells that can be resolved at the scale of a single synapse. The amount of information is daunting: in just $1\,mm^3$ of brain tissue, we expect petabytes of data containing 10^5 neurons and 10^9 synapses [126]. While this region is very small in physical volume compared to an entire brain, it is roughly the scale of a cortical column, a hypothetical

fundamental organizing structure in the cortex [127].

Our goal is to transform large 3D electron microscopy volumes of neural tissue into a detailed connectivity map, called a connectome. This approach will directly estimate brain graphs at an unprecedented level of detail. Each neuron is represented in the graph as a node, and each synapse is represented as an edge connecting these nodes. Manual human annotation, while currently the most accurate method of reconstruction, is unrealistic as volumes scale. A recent study estimated that manual annotation of a cortical column requires hundreds of thousands of person-years [49].

Therefore, an automated method to run algorithms at scale is needed to realize the promise of large-scale brain maps. We developed a novel ecosystem of algorithms, software tools and web services to enable the efficient execution of large-scale computer vision algorithms to accomplish this task.

We also introduce a fully automated images-to-graphs pipeline and an assessment metric for the resulting graphs. This metric allows us to directly assess the connectivity properties of the graph, rather than relying on intermediate measures (e.g., synapse precision-recall or segmentation pixel error). We run a grid search over a collection of parameters (i.e., both individual modules and their settings) using our pipeline to determine the best available result for analysis and interpretation. Once this optimal operating point was determined, we estimate the brain graph for a volume of neural tissue in our scalable framework. We believe that assessing graph error directly, as a system level evaluation (rather than component assessment) provides an improvement in the state-of-the-art research.

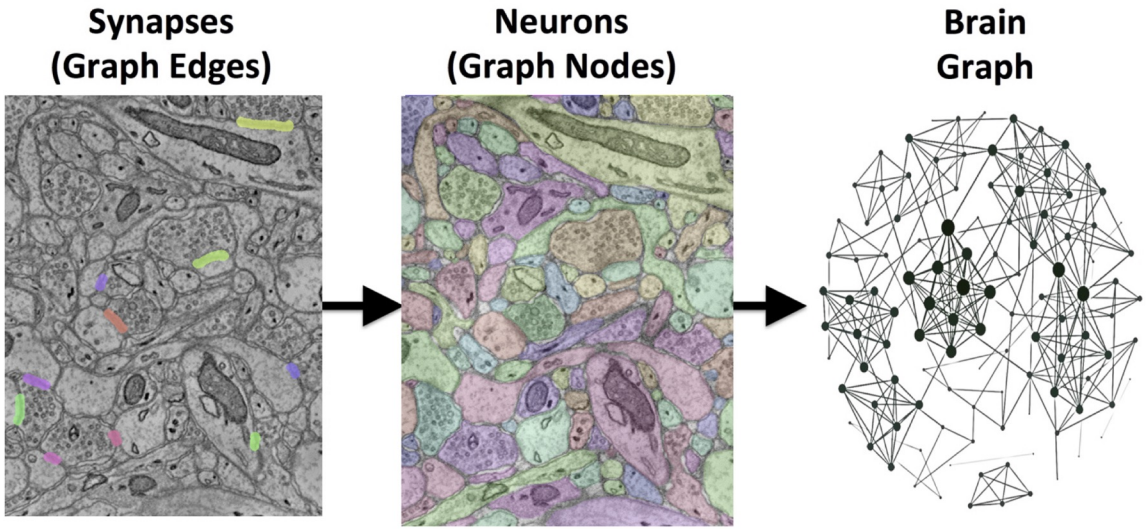

Figure 5.6: *An illustration of the images-to-graphs process.* (Left) Detected synapses are superimposed on raw EM data; (Middle) these are overlaid and combined with multicolored neuron segments to (Right) estimate a graph. Nodes are represented by neurons and edges by synapses. The data shown here are a subset from a small, hand-labeled region of brain tissue. The graph (right) therefore represents a gold-standard brain network from this region of tissue using a standard graph layout (not spatial position.)

5.2.1 Previous Work

Previous research has produced methods that advance the field of connectomics in important ways, but none have provided an end-to-end, automated, scalable approach. Several manual or semi-automated approaches have been used to construct brain circuits [23, 25, 117]. Other groups have produced automated algorithms [42, 50, 51] that solve important pieces of the overall puzzle (e.g., neuron segmentation, synapse detection). These modules have generally been evaluated on small subvolumes without considering the overall

graph result; additional work is needed to improve both algorithm accuracy and scalability for large graph inference.

In building our images-to-graphs pipeline, we leverage previous work whenever available. To detect cell membranes, we reimplement the ISBI 2012 challenge-winning approach [118], which frames membrane detection as a per-pixel classification problem and obtains state-of-the-art results using a small ensemble of Deep Neural Networks (DNN). We segment the neuronal structures by incorporating Rhoana [51], an open-source algorithm, which selects and fuses candidate 2D contours into 3D objects using conditional random fields (CRFs) and Integer Linear Programming (ILP). We also integrate Gala [50], an agglomerative approach that combines super pixels into neurons. Together these two methods represent the two major approaches currently used in neuron segmentation; other methods can be readily adapted to this framework if desired. Scalable synapse detection (i.e., the edges in our graph) is accomplished using the lightweight, scalable synapse detector highlighted in Section 5.1.

Finally, *NeuroData* [103] provides a high-performance spatial database optimized for processing large neuroimaging volumes. These tools facilitate scalable processing and provide significant advances over a flat-file architecture in terms of data storage and access.

5.2.2 Pipeline

Following the approach described in Chapter 2, we built a framework for connectomics processing that was agnostic to the underlying algorithms and provided reusable modules

for common steps such as volume dicing, image data download, annotation upload, and annotation merging. By leveraging RAMON, Open Connectome Project, our API, and the LONI Pipeline workflow manager [43], we built a system capable of rapidly integrating, running, and evaluating connecomics algorithms on a single workstation or on a high-performance compute cluster. To evaluate the graphs produced by the pipeline, we used the f_1 score of the line graph, which compares the detected graph edges to the true (expected) edges (Equation 2.5).

PIPELINE TRAINING

To prepare for algorithm evaluation and testing, we trained several algorithms used in the pipeline. For these tasks, we selected a data region separate from our evaluation region. Our primary training region was a $1024 \times 1024 \times 100$ voxel region (known to the community as AC4). Gold standard annotations for both neurons and synapses exist for this volume, based on expert neuroanatomist tracings. Our training tasks included: selecting a template for our vesicle detection module; training our deep-learning membrane classifier on 2D patches; building a random forest classifier for our synapse detection module; and training a Gala agglomerative classifier.

PIPELINE EVALUATION

To evaluate the optimal setting for generating graphs from volumes of brain images, we constructed a fully automated pipeline to conduct a hyper-parameter search of different

algorithms and their parameters and evaluate them based on community-suggested measures of synapse error, segmentation error, and our novel graph error metric (Figure 5.7). Other metrics can be straightforwardly added if desired. For evaluation, we used a separate, previously unseen region ($1000 \times 1000 \times 100$ voxels), known to the community as (part of) AC3. Gold standard annotations for both neurons and synapses exist for this volume, based on expert neuroanatomist tracings.

PIPELINE DEPLOYMENT

In this pipeline, we process a large volume for connectomics analysis (Figure 5.9). Based on the classifiers created in the training workflows and the operating points found in the evaluation pipeline above, we select an operating point and deploy our end-to-end images-to-graphs pipeline as a reference implementation over a large volume (the entire inscribed dataset); an example slice is shown in Figure 5.8.

5.2.3 Algorithms

Our approach transforms an image volume of cortical tissue into a wiring diagram of the brain. To assemble this pipeline, we begin with membrane detection [118], and then assemble these putative two-dimensional neuron segments into three-dimensional neuron segments using Rhoana [51], Gala [50], or a watershed-based approach. These are the nodes in our graph, and are evaluated using the Adjusted Rand Index and neuroanatomist curated ground-truth, following community convention. For edge detection, we leverage

the state-of-the-art method, VESICLE, as described earlier in this chapter.

Synapse and neuron association is completed by finding the neuron labels (i.e., graph nodes) that overlap most frequently with the labeled voxels from each synapse object (i.e., graph edge). This association is recorded via bidirectional linkages in the RAMON objects' metadata fields. Metadata assigned to each object can be traversed server side to construct a graph [103], or the graph can be built client side at small scales. Output graphs are converted via a web-interface to a community compatible format of choice using MROCP [79], such as GraphML.

5.2.4 Data

Our experiments utilized a large publicly available volume of mouse somatosensory (S1) cortex, imaged using scanning electron microscopy at $3 \times 3 \times 30\ nm$ per voxel (8-bit data) [128], aligned and ingested into the Open Connectome Project infrastructure. All images were color-corrected [125] and downsampled by a factor of two in the imaging plane. The entire raw data volume is approximately 660GB. The inscribed cube for our deployment workflow is $6000 \times 5000 \times 1850$ voxels (56 GB), or roughly 60,000 um^3.

5.2.5 Results

The images-to-graphs pipeline allows us to address the question of graph quality and begin to optimize results. We take a systems view of the connectomics problem and

evaluate a set of hyper-parameters (i.e., algorithm selection as well as parameters within algorithms) to determine the best operating point. In principle, parameters across all modules could be explored; however, we limit our experiment to variations in neuron segmentation and synapse detection methods for simplicity.

EXPERIMENTS

We initially performed a parameter sweep to determine the best operating point for our chosen metric, and then applied those parameter settings in a deployed setting.

EVALUATION

We used our pipeline to examine the interaction and settings of the segmentation algorithm and the synapse detector that achieve the optimal graph f_1 score. Our evaluation varied neuron segmentation parameters (e.g., membrane strength, thresholds, number of segmentation hypotheses). Our synapse operating points were chosen by sweeping over size and probability thresholds. Combinations of these parameters were tested, and the results are displayed as a matrix in Figure 5.10. We examined 1856 possible graphs, requiring approximately 8,000 cluster jobs and over 3TB of data manipulation. The entire evaluation workflow took approximately 13 hours.

After synapses and neurons were combined to construct a graph, we evaluated the line graph error. A permutation test was run to compute the null distribution of this test statistic. Specifically, we calculated the graph error by uniformly sampling a random graph with

the same line graph density as the observed graph for B=10,000 samples. The p-value was then the value of the resulting cumulative distribution function, evaluated at the test-statistic value of the observed graph. We chose a significance threshold of less than 0.001; non-significant operating points are shown in gray in Figure 5.10. The results are shown in sorted synapse and segmentation error order. Each cell in the matrix represents a single graph, and the optimal result is circled in the table.

The maximum f_1 graph score was achieved with a segmentation corresponding to an ARI score much worse than optimal. It is clear that constructing the best graph (according to our chosen metric) is more complicated than simply choosing the point with the best synapse f_1 score and lowest segmentation adjusted rand index error. Figure 5.11 further demonstrates the non-linear relationship between graph error and intermediate measures. By considering the overall problem context, we could select and tune the available algorithms to determine the best result (i.e., operating point) for a given task, such as motif finding. The optimal graph was computed using the Gala segmentation algorithm with an agglomeration threshold of 0.8; the synapse detection probabilities were thresholded at 0.95, and a connected component analysis was used to form the final synapse objects. Objects with a size greater than 5000 pixels in 2D or less than 1000 voxels in 3D were removed to reduce erroneous detections. The optimal f_1 score was 0.16, indicating that significant improvement was needed.

DEPLOYMENT

The deployment workflow provides a capability demonstration and produced 12234 neurons with non-zero degree and 11489 synaptic connections in a volume of $\approx 60,000$ cubic microns. Total runtime on 100 cores was about 39 hours, dominated by the block-merging step, which is currently performed on each seam serially. Membrane computation currently takes an additional 3 weeks on our small GPU cluster; this process can be run in parallel and recent advances suggest methods to dramatically speed up this step [129].

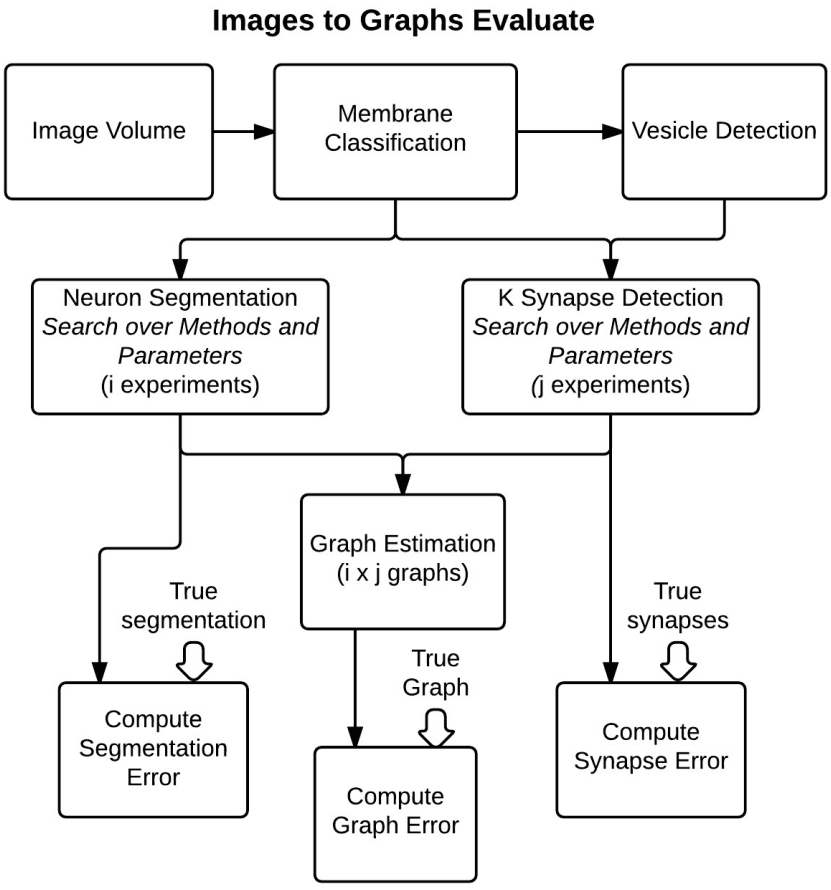

Figure 5.7: *An overall view of the Images-to-Graphs Evaluation Pipeline, beginning with image data and ending with graph creation.* Graphs are estimated and evaluated for each combination of i segmentation experiments and j synapse detection experiments.

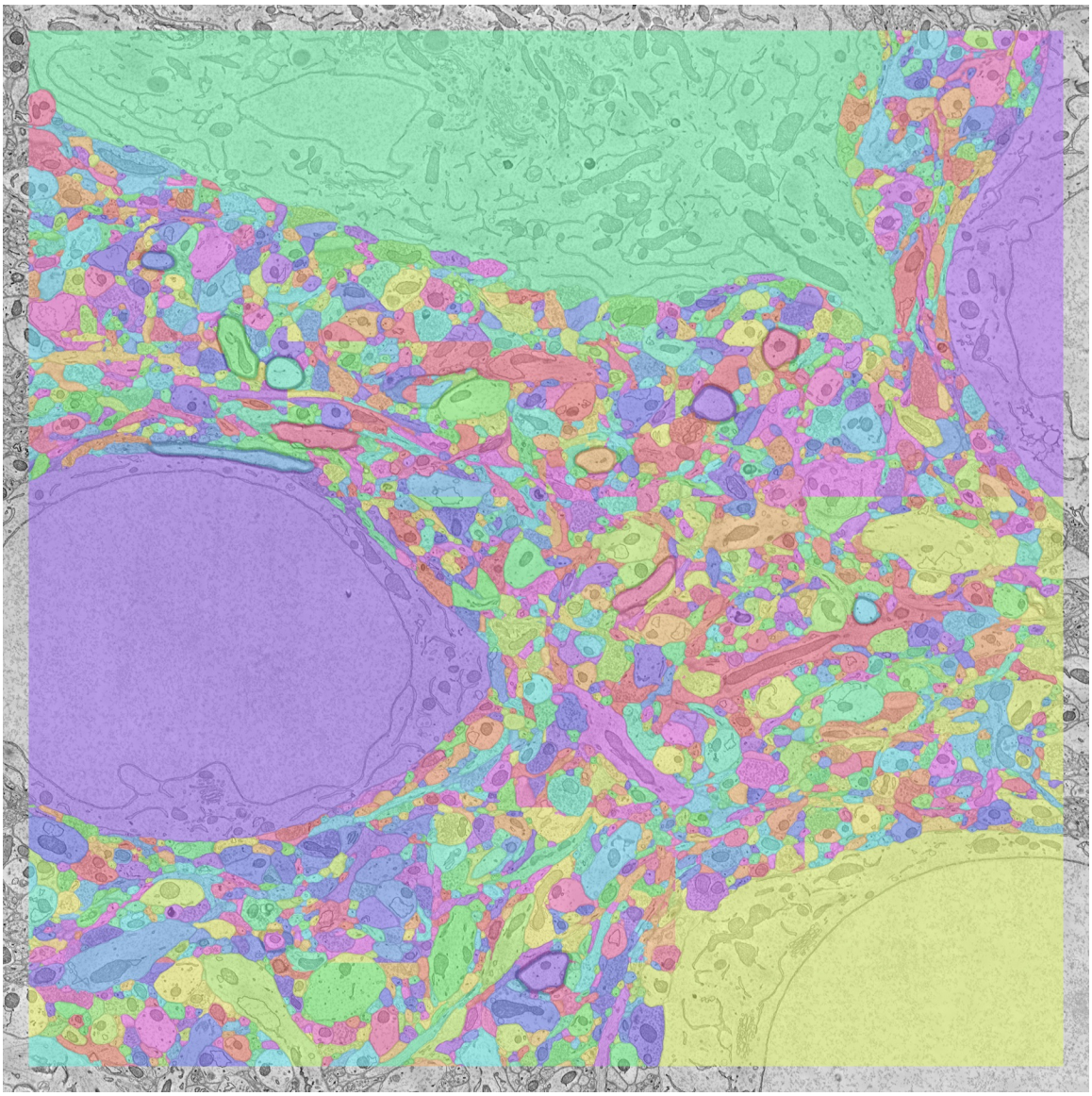

Figure 5.8: *An example XY slice of the entire inscribed cuboid from the Kasthuri2015 reference dataset.* This result illustrates a large-scale end-to-end segmentation and merge result that can be used to construct graphs, after selecting an operating point using the methods shown in this chapter.

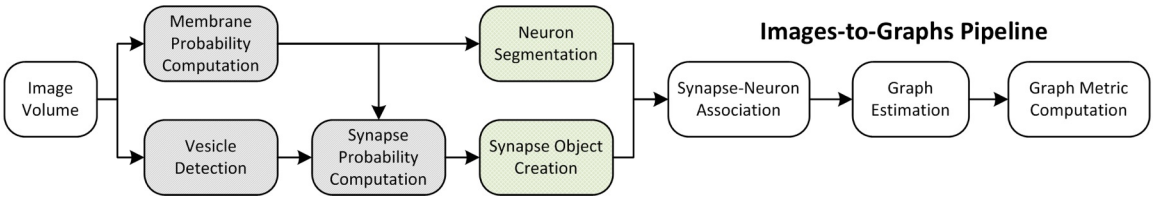

Figure 5.9: *An overall view of the Images-to-Graphs Deploy Pipeline, beginning with image data and ending with graph creation.* Modules in white are executed each time, modules that are gray (darkly-shaded) are executed once and not varied in our analysis, and modules lightly-shaded represent our parameter space.

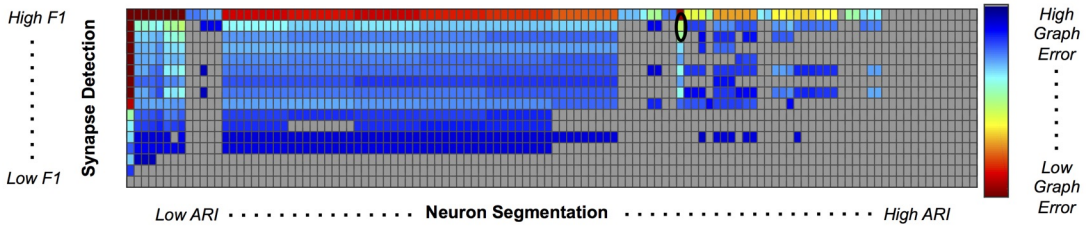

Figure 5.10: *Experimental Graph Based Error.* 1856 graphs were created by combining 13 synapse detector operating points (rows) with 100 neuron segmentation operating points (columns). The rows are ordered by synapse f_1 score, and the columns by segmentation adjusted Rand index. The first row and column represent truth, and the upper left corner of the matrix has an error of 0. Cell color reflects graph error (clipped to show dynamic range), with a dark red indicating lowest error and dark blue indicating highest error. Values shaded in gray are not significant; the selected operating point (max f_1 graph score is circled in black.

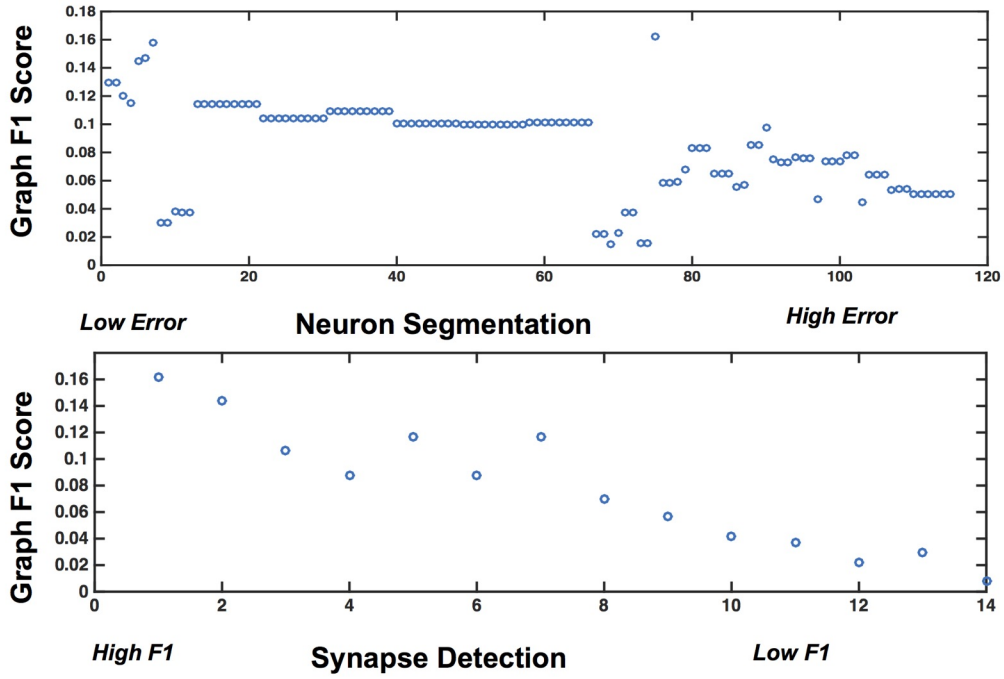

Figure 5.11: *Error Variability.* Plots demonstrating the variability of graph error with segmentation error (top) and synapse error (bottom), for the rows and columns associated with the best operating point.

5.3 SANTIAGO

5.3.1 Introduction

In adult vertebrate brains, each spine is typically associated with an excitatory synaptic connection, making their detection and association a critical task for building brain graphs [130, 131]. Many current algorithms in connectomics are designed and evaluated using surrogate metrics (e.g., voxel-level segmentation of neurites) rather than global graph measures. Therefore, even the best available segmentation, coupled with ground-truth synapses, may produce a poor estimate of connectivity. Current results are quite accurate at capturing large process segmentation, but one large contributor to network degradation in vertebrate brains is spine neck fragmentation. This is caused by small cross-sectional areas, densely packed structures, limited contrast, and poor overlap due to anisotropy. Biologically, spines are small projections from the dendritic shafts of neuronal cells. Spines occur predominantly in vertebrates and are both prolific (i.e., a single human brain likely contains many billions of these structures) and difficult to track in existing imaging methods due to their small length (a few microns) and volume (1 femtoliter) [131]. The cross-sectional area of spine necks are typically $\sim$0.2 microns [131], corresponding to only a few pixels across a single imaging plane at the resolution typically used in serial section electron microscopy. Spines were discovered by Santiago Ramon y Cajal [132], in the late 19th century and it is hypothesized that understanding their function will unlock many of the secrets of neuronal computation [131].

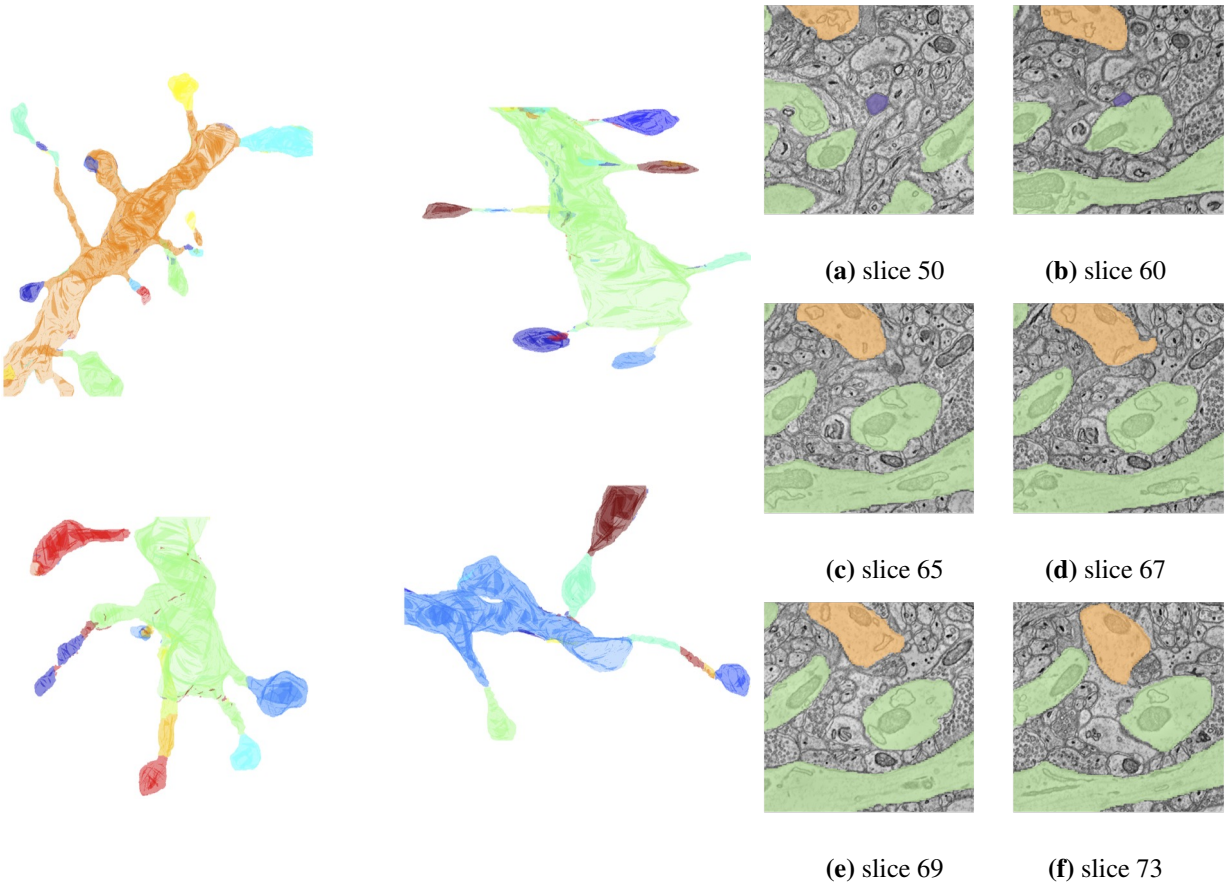

Figure 5.12: *Examples of the spine fragmentation problem.* (Left) Images illustrate typical split errors made in reconstructing spines by superimposing the automated segmentation labels on the ground-truth for individual neurites. If reconstructed correctly, each object should be only a single color. These illustrations actually understate the problem, as they do not show merge errors for labels that extend beyond the ground-truth mask. (Right) A typical spine merging problem is illustrated with the spine is shown in blue but incompletely linked; the true parent is in orange and other potential parent shafts are shown in green.

As described in Section 2.5.5, the network graph can be represented as a line graph in which synapses may be thought of as vertex terminals, with edges (i.e. neuron fragments) between them as pathways [33]. Although more complicated information is useful for downstream analysis (e.g., attributes like direction or weight), this basic connectivity question is perhaps the most fundamental of the unanswered connectomics questions. Indeed, when segmentation algorithms fail to connect these processes, the graph contains many disconnected nodes and an inaccurate picture of connectivity. An illustration showing the challenges inherent in reconstructing spine-shaft linkages is shown in Figure 5.12.

In this section we introduce *Santiago (Spine Association for Neuron Topology Improvement and Graph Optimization)*. We believe that ours is the first work to introduce an algorithm for solving the spine-shaft linking problem in serial section electron microscopy data. However, several other groups have noted the difficulties associated with reconstructing dendritic spines and have developed semi-automated workflows to correct errors, including these spine fragments [11, 51, 117]. Of particular significance is research to assess and prioritize error proofreading based on connectivity impact [37]. Other work has extensively studied dendritic spines (e.g., [131, 133]), providing a rich set of priors and information when reconstructing neuronal circuits. Other methods suggest related ideas, albeit from a different perspective or targeting a different setting [134, 135, 136].

In this work, we carefully explore prior EM segmentation results and develop an algorithm to reattach the fragmented spines, thus reconnecting many of the synapses that were previously graph isolates. We leverage our understanding of local image grammars to develop

a classifier that determines the best merge strategy for each spine. We specifically focus on the spine-shaft problem, while acknowledging that related challenges such as synapse association, long range axonal projections, semantic typing, and merges across cuboid boundaries will need to be addressed when developing a comprehensive automated solution. We believe that this is the first algorithm to explicitly address the spine problem in the context of nanoscale connectomics, and we provide the datasets and methods from this work as a testbed for future researchers.

5.3.2 Methods

In vertebrate brains, anisotropic neuroimaging methods (e.g., electron microscopy) have the most difficulty resolving the finest processes [117]. When considering basic questions of connectivity, the spine necks are among the most important, yet the most difficult to trace. Tracing large axons and dendrites may be possible at low-resolution, however, resolving spine necks requires sub-10 nm resolution.

In this chapter we examine (grammar) production rules (Section 2.4.4) which are easy to exploit and the target of this algorithm, focused on the relationship between dendritic spines and their parent shafts.

- Dendrite $\rightarrow$ shaft Spines
- Spines $\rightarrow$ Spine | Spines
- Spine $\rightarrow$ spine synapse

Although this paper targets the spine-dendrite production rule, we hope that this demon-

stration encourages the incorporation of other higher-level (biological) inference rules toward better circuit reconstructions.

METRICS

We focus on two metrics leveraging the f_{beta} score [137]. In this paper we fix $\beta = 1$, although we explored other values for estimating graph properties in internal experiments. We first consider raw precision-recall scores of the spine-shaft association problem; this simply captures whether putative links are correct in an automated setting. Next, we put these scores into a Top-K ranking setting, where we identify shafts that are likely partners for each spine. This latter approach has applications for speeding up semi-automated proof-reading workflows by allowing proofreaders to quickly choose from amongst a few choices rather than manually segmenting paths in an unconstrained environment. This is an active area of research and promises to greatly impact circuit quality while improvements are made in fully automated algorithms [11, 37]. When reattaching spines, we also compute the f_1 graph error as described in Section 5.6.

PIPELINE

We leverage the ideas developed above to guide our image processing and classification features to predict the best candidate shaft for each spine. We first assess the characteristics of this problem and use them to develop a solution to improve the resulting network topology in an automated or semi-automated setting. A block diagram of our approach is

shown in Figure 5.13.

DATA PREPROCESSING

Our baseline method begins with known, semantically labeled shafts, spines, and synapses. Estimating these labels is a problem that has been carefully studied, for neuron segmentation [50,51,112,138], synapse detection [34,115,116], and semantically labeling objects [111, 139]. This work instead focuses on the linking problem that results after these algorithms have been run.

More specifically, we begin with ground-truth synapses and shafts, and use Gala [111] to segment the best spine candidate using agglomerative segmentation. Gala is an often-used, high-performing [33] technique that allows us to automatically generate a realistic estimate of the spine volume (reserving the spine truth information only for semantic labeling). Other segmentation methods may be used as inputs to *Santiago* as the computed features are independent of any Gala specific metadata.

SPANNING TREES

Spines have a known distribution of distances between their head and the shaft [131], which we exploit by looking for all shaft partners within a defined radius of each orphan spine. As illustrated by the biological structure outlined above in Figure 6.1, each neuron has a tree structure, with each spine connected to exactly one parent. There should be no orphan spines when neglecting boundary effects. Therefore, we construct a spanning forest,

consisting of a set of minimum spanning trees (one rooted at each spine). To minimize computational complexity, we treat each spine orphan as an independent subproblem; the preserved spine-shaft link is the maximum probability edge in the graph. Future versions of *Santiago* could be extended to consider more complex interdependencies (e.g., periodic spine anchor locations, non-uniform spine distribution across dendrites).

We extract features by observing link distance, direction, and path cost. More specifically, we compute the following quantities: minimum distance between spine and shaft; minimum distance between synapse and shaft; shaft size in window; minimum distance from end of spine to shaft; minimum distance from linearly propagated spine path; minimal path cost from spine to shaft (currently computed using membrane probabilities [118]); and branching angle between spine and shaft. These features are robust to a variety of settings and noise, and provide an excellent estimation of the correct spine link. For each feature, except spine-shaft branching angle, we use both a raw score and relative ranking score to improve classifier robustness. A version of *Santiago* that uses only geometric label relationships could also be deployed to reduce data dependencies and speed processing.

CLASSIFICATION AND ASSIGNMENT

To determine the weight of each edge in a spanning tree (i.e., probability of a spine-shaft link), we use a random forest classifier composed of these features. We follow a cross-validation approach, with all spines in the same dendritic parent group considered together (i.e., either in training or test) to minimize overfitting. For each fold, we reserve

one of these groups for testing and use the remaining groups to train the classifier.

Each link receives a probability score when applying the random forest classifier, and we compute precision-recall on these links along with f_{beta} scores of 0.5, 1, and 2. To compute a best overall graph-f_1 score, we construct a spanning forest using a hard classification to predict the best candidate shaft for each spine. When constructing a spanning tree, the associated edge weight is inversely-related to these link probabilities (i.e., a high probability edge has a low link cost).

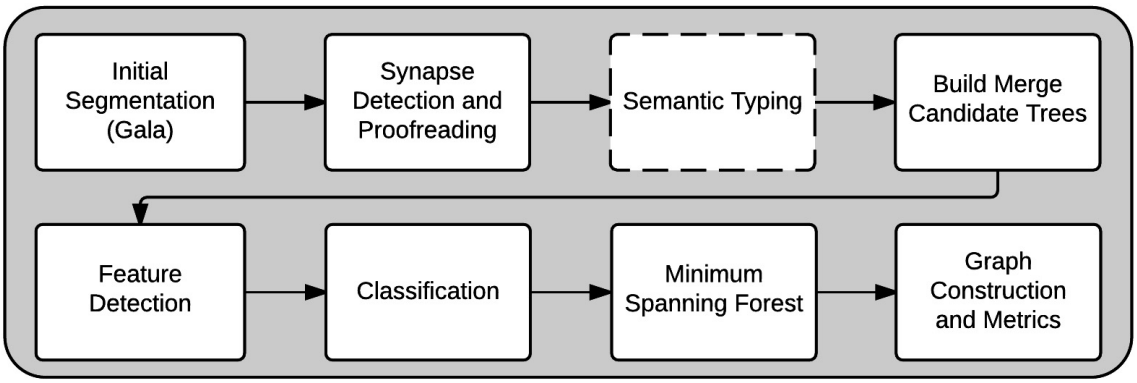

Figure 5.13: *A block diagram of our proposed approach for identifying fragmented spines.* We begin with a Gala segmentation and end with a graph. Semantic typing is shown with a dashed line because this step is outside the scope of this work.

5.3.3 Results

We first develop data sets appropriate for characterizing and optimizing spine association. We also carefully assess the impact of spines on overall connectivity, and demonstrate our algorithms on real data.

DATA

Gold standard annotations for connectomes (especially in cortex) are still limited and challenging to leverage for automated analysis. In this work we develop the first baseline dataset specifically designed to evaluate the spine problem, derived from a saturated, manual tracing in somatosensory cortex [22].

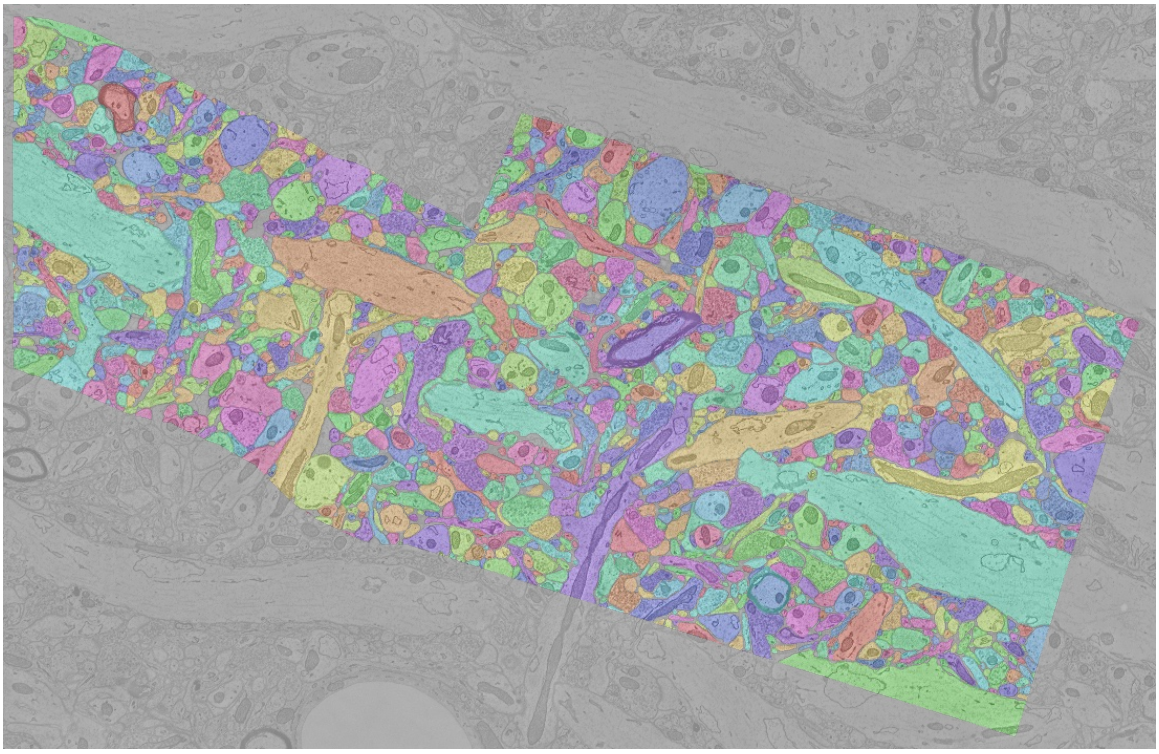

Figure 5.14: *Experimental Data.* A single slice of the primary segmentation (gold standard) dataset used in this experiment is shown above. Each color corresponds to a unique object (e.g, dendrite, glia, spine, axon). Synapses are annotated in a different, spatially co-registered channel.

3-CYLINDER DATASET:

This dataset contains several thousand neurite fragments, 1700 synapses, and over 1000 spines. We curate this information to produce a dataset suitable for training and assessment. To avoid conflating the spine assignment problem with earlier segmentation challenges, we work with data centered around a synapse with a biologically motivated cube size of about $2\mu m$ in each direction (corresponding to $700 \times 700 \times 140$ voxel cuboids [131]. Due to boundary conditions of the cylinder, some parts of these cutouts have no shafts and other shafts are cut-off; however, the resulting candidate merge trees provide a realistic, challenging scenario. We partially mitigate these edge effects and spurious labels by admitting only objects explicitly labeled as spines by the original authors, and restrict shafts to large objects of at least a cubic micron. Only those spines having a corresponding ground-truth shaft parent in this restricted set are analyzed in this work. These represent a significant fraction of the connections in the full data volume, but eliminate other connections, which should be analyzed in future work. This preprocessing procedure results in 531 spines and 38 target shafts for analysis. The data used in this analysis can be explored online in a *NeuroData ndviz* project.

IMAGES-TO-GRAPHS DATASET:

Additionally a small region of this dataset containing reconstructions from part of the 'AC3 region' were used in a recent analysis to assess graph-f_1 error [33]. We use this data in our simulations to assess the impact of fragmented synaptic connections to partially justify

the importance of this work. Because this volume is small (spanning $1024 \times 1024 \times 100$ pixels), some edge effects in labeling are also present in these data, as processes transiting the edges of the volume can lead to association information unavailable to an automated algorithm.

SIMULATION RESULTS

As others have noted [33, 37, 51], spines are a major issue in graph connectivity, but to date, the impact has not been quantified in the context of electron microscopy graph estimation. Here we explore a quantitative assessment through simulation.

3-CYLINDER DATASET

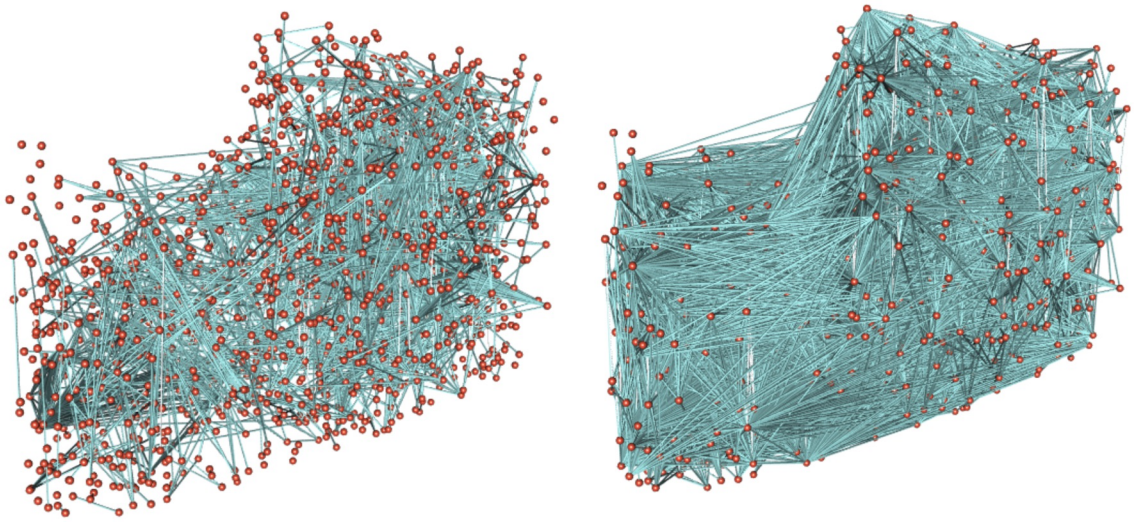

Figure 5.15: *Spine Importance.* (Left) Graph with disconnected spines and (Right) Gold standard graph. These illustrations emphasize the large impact of spines on the overall graph connectivity.

We begin with the true connectivity matrix where all spines are correctly associated

with their parents in the 3-cylinder dataset. We investigate the impact of spine-shaft linking on overall connectivity (i.e., the line graph), by detaching spines. This process creates a separate segmentation label for the spine, leading to a synapse that is effectively disconnected on the dendritic side. Because high-degree nodes impact the graph disproportionately, we repeat our simulations at different levels of spine fragmentation, quantifying error, average degree and their variances. In the true line graph, the total number of edges is 31,980 (average degree: 18.8). In the graph with all spines disconnected, the total number of edges is 2,718 (average degree: 1.6); a vizualization highlighting these differences can be seen in Figure 5.15.

For the 3-cylinder dataset, we examined this spine fragmentation and conclude that spine association is necessary, but not sufficient, to ensure an accurate connectome. Most of the connections are carried on spines and are lost when these spines are fragmented. We show this quantitatively in Figure 5.16.

We further explored this idea by running Gala on a region surrounding each spine and found that nearly all (86%) of the spine-shaft linkages were missed. A successful match was scored whenever the most common segmentation label for the spine truth and shaft truth were identical; this assessment disregards overmerging failures. Our cutout region was a $700 \times 700 \times 140$ voxel window, corresponding to a cutout of $4 \times 4 \times 4\,\mu m$, centered about the synapse of interest. In our anisotropic data, this was sufficient to capture nearly all (97%) of the shaft partners for the spines of interest.

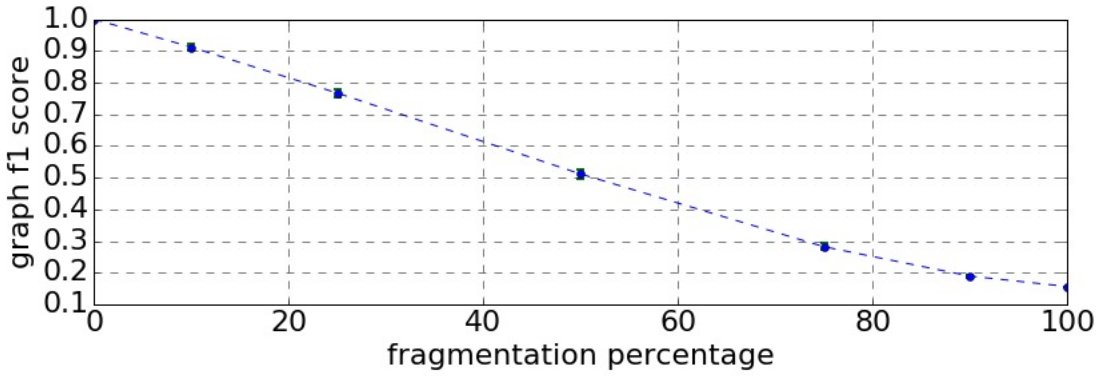

Figure 5.16: *Spine Impact on Graph Error.* Graph error as a function of spine fragmentation (0-100%) showing the f_1 graph error. This firmly establishes the importance of spines on connectivity, especially at a local scale. 1,000 iterations were performed with different spines removed each time.

IMAGES-TO-GRAPHS DATASET

To further understand the spine problem, we conduct an additional simulation using the images-to-graphs dataset. We identify all synapse orphans as putative spines in a Gala segmentation (threshold: 0.5, tuned to reduce the possibility of overmerging). Because no semantic labels are available, we select all orphans incident to a synapse. We then synthetically merge those objects to their parent "shaft," which we select as the largest object in the dataset with the same truth label, based on overlap with corresponding ground-truth labels.

The resulting score improves the baseline f_1-graph score of 0.31 to 0.64. This again emphasizes the importance of spines, as we can double the graph-f_1 score by identifying and linking these orphans without altering the other segmentations. This result suggests

that existing algorithms are accurate at reconstructing large processes. By focusing on these small, disproportionately important objects that violate a known biological constraint (i.e., connectedness), we can address many of the deficiencies of conventional algorithms with a hierarchical approach.

COMPUTER VISION RESULTS

Table 5.2: *Table showing computer vision results on each dataset.* Baseline score prior to this algorithm is zero matches, as we operate only on spines that are missed by Gala.

3-Cylinder Kasthuri Dataset [22]	**Scores**
Top-1 (match)	203 / 455 = 0.45
Top-2	299 / 455 = 0.66
Top-3	352 / 455 = 0.77
Top-5	405 / 455 = 0.89
Top-10	436 / 455 = 0.96
Maximum f_1	0.47
Median rank (when available)	2
Mean rank (when available)	2.41
Spines with truth in window	440 / 445 = 0.97
Average shafts / spanning tree	9.5

We apply our spines-shafts pipeline to identify candidate shafts for each orphan spine.

In this work, we leverage previously computed membranes [33] and compute Gala segmentations on a small region surrounding each orphan synapse (as discussed above). Running Gala takes about 2 hours and 20GB of RAM on a single core for this sized volume. NeuroProof [140], a successor to Gala, offers faster computation and additional options, but was not specifically evaluated for this work. Computing features and classification took approximately 15 minutes per spine; the bulk of this time was spent computing path-finding features.

In all of the results reported below, we only show orphan spines (excluding the 76 Gala spines that successfully matched). Therefore, the baseline for prior state-of-the-art is that all of these spines are disconnected from their parents (0/455). We note that a few (15/455) spines do not have shafts present due to the window size chosen; these are not excluded and are treated as errors when reporting algorithm performance.

In Table 5.2 we provide a detailed reporting of results showing conventional f_1-detection metrics for both automated processing of edges (e.g., maximum f_1) and Top-K performance, for use in semi-automated proofreading approaches. We also report our mean and median ranks when the true parent shaft is present, as well as the number of available shafts available on average for each linking scenario.

In Table 5.3 we report the results of our algorithm on graph-f_1 error. We show two columns of values: the *Santiago* subgraph contains only the connections and their immediate partners used in our test dataset (i.e., the axon and dendrite fragment and corresponding parent neuron information); and the 3-cylinder graph contains all connections and demon-

strates that fixing errors in the local subgraph translates to overall graph quality improvement. The spanning forest results shows our best overall automated performance, while the Top-K results show the simulated impact of humans successfully (perfectly) proofreading the Top-K results. In a real-world proofreading setting, a user could be presented with options and would either identify the true partner or return no match. As K increases, the operator workload (and potential performance) will increase correspondingly.

Table 5.3: *f_1 graph scores on Santiago subgraph and full 3-cylinder graphs.* The table below shows a baseline for performance prior to *Santiago* and after running. The post-run numbers include an assessment using fully-automated and semi-automated approaches.

***Santiago* subgraph**	**f_1 graph score**	**3-cylinder graph**	**f_1 graph score**
no spines	0.035	no spines	0.205
gala spines only	0.089	gala spines only	0.246
spanning forest (auto)	0.404	spanning forest (auto)	0.398
top-1 (proofread)	0.518	top-1 (proofread)	0.462
top-2 (proofread)	0.711	top-2 (proofread)	0.573
top-3 (proofread)	0.816	top-3 (proofread)	0.638
top-5 (proofread)	0.910	top-5 (proofread)	0.701
all (proofread)	0.983	all (proofread)	0.752

5.4 Discussion

5.4.1 VESICLE

We have presented two algorithms for synapse detection in non-poststained, anisotropic EM data, and have shown that both perform better than state-of-the-art methods. The Random Forest approach offers a scalable solution with an approach inspired by expert human annotators, while the deep learning result achieves the best overall performance.

We built and demonstrated a reusable, scalable pipeline using the Open Connectome Project services and used it to find putative synapses on large cubes of mammalian EM data. We presented the largest result known by orders of magnitude (in both volume processed and synaptic detections). In the future, we plan to refine our estimates of synapse morphology and position by implementing a region-growing algorithm and incorporating additional contextual information. We also plan to utilize supervoxel methods in both of our approaches, and consider VESICLE-RF texture features inspired by CNN results.

For the VESICLE-CNN approach, future work includes exploring alternative CNN architectures (such as the very deep networks considered in [141]), enhancing the tile-based input features (e.g., to include 3D context) and improving the computational complexity (through the use of sampling techniques guided by our biological priors and/or computational techniques (e.g., [142]).

5.4.2 Images to Graphs

We have demonstrated the first framework for estimating brain graphs at scale using automated methods. We recast this problem as a graph estimation task and consider algorithm selection and parameter tuning to optimize this objective by leveraging a novel graph error metric. Our work provides a scaffolding for researchers to develop and evaluate methods for the overall objective of interest.

We evaluated our pipeline across a set of parameters and modules, leveraging a combination of published methods and novel algorithms. Additional insights may be gained at larger scales and through additional optimization. Although our error metric currently considers only binary, unweighted graphs, there are opportunities to extend this to apply to attributed graphs, as well as to weight the metric by error types (e.g., the number of false positives or false negatives).

Automated results do not need to be perfect to draw statistical conclusions, and errorful graphs may be used as the basis for inference and exploitation of "big neuroscience" challenges [48]. Bias in errors is another important factor in constructing exploitable graphs which has not been fully explored. With the ability to efficiently compare combinations of different algorithms and operating points, we can begin to answer the question of graph quality and how to optimize the results. Having the ability to examine the process from an end-to-end systems approach will enable rapid improvement in graph quality. The infrastructure demonstrated in this work provides a community test-bed for further exploration and at-scale computation. Although this chapter focuses exclusively on electron micro-

graphs, our framework is extensible to many other modalities, including Array Tomography [143], CLARITY [144], and two-photon calcium imaging data.

5.4.3 *Santiago*

We reframe the connectomics problem to focus explicitly on connectivity, and measure progress using a convenient, descriptive metric. We illustrate a semantic, biologically inspired solution to partially remedy one of the major problems of neuron reconstruction (i.e., linking spines to dendritic shafts). By inferring paths that may not be clear at a voxel-level, we are able to recover connections that would have otherwise been lost. The field of connectomics is still in its infancy; this work provides an early example of the untapped potential for combining well-studied biological phenomena to computer vision approaches. Although we demonstrate this idea in the context of electron microscopy and a particular segmentation algorithm, the underlying principles are very general and potentially could be leveraged in other settings including light microscopy (where spine necks are difficult or impossible to resolve due to resolution constraints). The tree structure and biological priors provide a scaffolding that constrains the reconstruction puzzle, and may greatly facilitate both estimation and error-checking as models improve.

We emphasize the importance of connection, rather than segmentation, and propose a new solution that allows for many spines to be recovered that are missed using conventional approaches. Our algorithm is agnostic to the segmentation "preprocessing" method and will likely improve as coarse segmentations improve. Future work will apply these

techniques to different data sets and problem settings, and will also explore fully-automated approaches (e.g., semantic typing) and integration into a complete pipeline. Our preprocessing segmentation algorithm, Gala, missed most of the spines, requiring *Santiago* to do extensive reassembly. As segmentation algorithms improve or incorporate these biological priors directly, the post-processing required may be substantially reduced, improving graph fidelity.

We produced a database of spines and shafts which enable future algorithm development and testing. Our methods and approach are scalable and fit into a broader effort that seeks to transform images into graphs. Our code and data are publicly available in accordance with reproducible science.

5.5 Summary of Chapter Contributions

In this chapter we provide state-of-the-art machine vision solutions to a challenging synapse detection problem and provide two methods, one that is easily scalable to large volumes, and a second with superior performance. We use this method in the creation of the first, fully-automated graph estimation pipeline for nanoscale connectomics, and develop a novel error metric to assess performance. The resulting performance is poor, partly due to the fragmentation of spines from their parent dendritic shafts. We propose automated and semi-automated methods to repair these links and improve graph estimation using a novel, biologically motivated approach.

Chapter 6

NeuroData: Enabling Big Data Neuroscience for Everyone

6.1 Overview

NeuroData has been developed to lower the barrier to entry into big data neuroscience. We have designed and built a comprehensive ecosystem for enabling terascale neuroscience. This includes our flagship project, the *NeuroData Connectome Project* (previously called the *Open Connectome Project*). Our infrastructure enables anyone in the world with internet access to visualize, download, analyze, upload, and interact with a large number of public datasets. Moreover, we provide an ecosystem using the *NeuroData* infrastructure to generate results that are fundamentally reproducible and extensible.

We demonstrate the utility of these tools via two serial electron microscopy case studies. First, we show an example of extensible neurocartography by reproducing many of the quantitative results from a recent landmark EM paper [22] that included a saturated annotation of a region of cortex. Next, we build tools to test a novel hypothesis about the distribution of synapse locations in cortex on a larger, complementary dataset. *NeuroData* democratizes the scientific process, enabling anyone, regardless of background, computational resources, or expertise, to study neuroscience at scale. Although this work has contributed significant leadership to ideas and implementations of the overall ecosystem, this chapter will focus primarily on the integration of image analysis, pipelines, and data models while still providing a broader context. In this chapter we highlight our overall ecosystem, including applications of the analysis tools detailed earlier in this work.

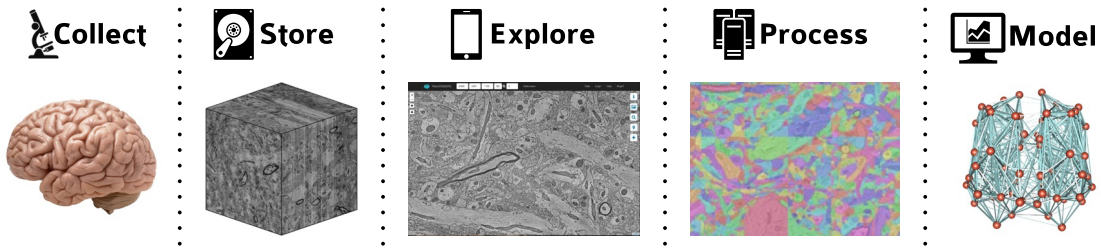

Figure 6.1: *An overview of the steps in the scientific discovery process supported by NeuroData.* This process begins with data collection, and includes steps to store, explore, process, and model the data, ultimately resulting in new knowledge.

6.2 Introduction

In the 21st century, big data (i.e., data too large to fit on a workstation) is transforming industry, governments, and science. The growth of data acquisition has greatly outpaced the growth of data analysis, rendering current computational and statistical tools insufficient for extracting meaning from large datasets in many domains.

As explored below, existing ecosystems for data storage, exploration, analysis, and modeling are ill-equipped to meet the challenges that arise with conducting science on big neuroscience data. For example, imagine that we have collected 20 TB of serial EM data from part of a cortical column, and we have the very fundamental neuroanatomical question to answer: "Is the 3D distribution of synapses uniform throughout the volume?" We divide the data analysis requirements into four components, each of which encompasses unique challenges:

- Store: Many labs do not have many terabytes of storage, and even if they did, each scientist would need to dynamically choose where to put data, how to organize it, and what format to use. This leads quickly to siloed data, as others cannot easily access or operate on the data without learning the organization and specific data storage formats.
- Explore: Simply visualizing data that are larger than RAM is non-trivial. Current local solutions typically involve sub-sampling or downsampling the data, and loading just a subset of the data into RAM [145], a compute and storage intensive process. Web visualization solutions offload the additional computation and storage to a remote

server [146], but must then integrate multiple data types from multiple sources to be useful for analysis.

- Analyze: The raw data is typically not collected in a form that is amenable to parsing into semantic objects. A reference preprocessing workflow might include stitching images, adjusting chromatic aberrations, and aligning data to reference coordinate frames for comparison. Current solutions often require all of the data to be loaded into RAM locally, or require the high-overhead sampling techniques described above. Manually identifying all of the $\sim$10,000,000 synapses, at a rate of about 1 synapse per second, would take over a year (40 hrs/week, 50 weeks/year), assuming that the infrastructure existed to trivially record these detections. Existing computational methods do not readily integrate with data volumes of this scale without external tool development. Additionally, many different tools have been developed for neuroscience applications, such as *ilastik* [42] and *Thunder* [147], and it is difficult for users to choose, integrate, and apply these methods to their experimental paradigm. Finally, as datasets grow in size and complexity, no good method exists to store, share, and reproduce this analysis.
- Model: Modeling the statistical regularities (and irregularities) from the resulting data derivatives requires statistics incorporating space, shape, and graphs. Existing packages for such analyses often have limited capabilities, are expensive, or do not scale sufficiently.

If a single group did manage to overcome all of those barriers, then they could answer this

question and publish the results. But science is fundamentally a collective endeavor, where each result builds upon previous research. With small data, the full analysis can easily be reproducible using existing technology by depositing the digital data into a data repository (e.g., FigShare, Dropbox, Google Drive) and the code into a code repository (e.g., GitHub). The above paradigm is not yet standard practice in neuroscience, but it could be. With big data, however, the above suggestions for reproducibility would not work, for a variety of reasons:

- Most data repositories will not accept multi-terabyte datasets.
- Even if they did, there is not a standard protocol for how other researchers would access the data. For a given result, researchers would have to develop tools and support resources to either request, process, and combine subsets of data, or develop an infrastructure to support downloading and processing the entire large dataset.
- Code to analyze terascale data is not yet "turn-key" (e..g, not as simple as running a Matlab script). Running the code often means having installed a number of libraries, some of which may only be compatible with certain operating systems and environments. Running, optimizing and debugging other researchers' code often takes weeks per tool, and many different tools may be required for a single scientific challenge.
- The cost and complexity of storing and analyzing big data locally is much more than doing so on dedicated community infrastructure with economies of scale and amortization of costs across different resources.

Thus, in practice, the effort of simply reproducing a previous result might require many

months or years of work, often in fields that are distant from neuroscientists' core competencies and research interests. Moreover, science proceeds not merely by reproducing previous results, but by confirming and extending them. This includes simple modifications such as checking the robustness of previous analyses by modifying some of the assumptions, tests, or the data on which the results rely. To the extent that one can make relatively minor changes to a previous analysis, while re-using as much infrastructure as possible, the barrier to confirming and extending previous results to novel findings can be significantly lowered.

We have built the *NeuroData* ecosystem to address the challenges described above, therefore enabling *reproducible* and *extensible* neuroscience regardless of scale. In particular, we have made a large number of big neuroscience datasets openly accessible. By storing all of the datasets, regardless of scale and modality, in a common framework, a single infrastructure is sufficient to process all datasets. This includes each of the above four steps (store, explore, analyze, and model). Figure 6.2 provides a summary schematic illustrating all of the tools we have developed.

When processing very large datasets, a paradigm shift is required to access data efficiently. Workflows to store, explore, analyze, and model data are necessarily distributed, since they do not fit in RAM and often not on a single hard drive. Moreover, many current experiments with big data rely on many individually small, self-contained 1D or 2D signal representations. The 3D+ neuroscience data instead often consists of a single large sample.

For *storing* this data, we build a datastore containing hierarchical, compressed cuboids

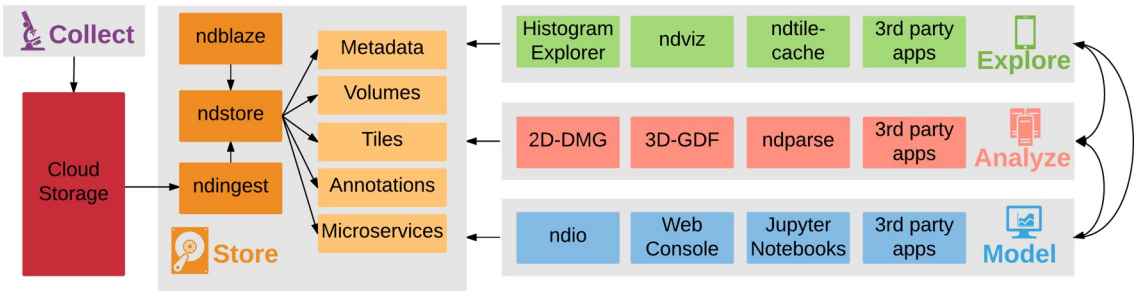

Figure 6.2: *The NeuroData Ecosystem is made up of four collections of services, each designed to run in a different environment.* **Store** runs on big data clusters, and allows users to download data, run queries remotely, and upload new data (including metadata). **Explore** runs on servers close to the client (e.g., using a Content Delivery Network), and allows efficient data exploration regardless of geographical location. **Analyze** runs on high performance compute clusters (close to the data), and enables distributed machine annotation. Finally, **Model** runs locally, on an end user's laptop or workstation, which allows programmatic interaction with the data (e.g., using Python). Our ecosystem leverages the scalability of modern compute systems to allow users to store, explore, analyze, and model data all over the world.

of image and annotation data, and high-performance tools to facilitate rapid reading and writing of arbitrary blocks of data. This datastore is optimized for machine vision algorithms and is unconstrained by traditional 2D file formats (e.g., PNG, TIFF). For *exploring* neuroscience data, we provide a web-interface that allows for rapid, universal image and annotation exploration and basic computation from within a browser, enabling anyone to understand large-scale neuroscience from any internet-connected platform. For *analyzing* neuroscience data, we provide scalable tools for translating data from the microscope to a format suitable

for machine vision tools. We also provide wrappers, algorithms and workflows that interface with RESTful services provided by our datastore. Because the data are too large to analyze in a single block, this strategy allows users to prototype on small volumes and then scale that solution to many small volumes, and collect the results into a coherent scientific result. Finally, in *modeling*, we demonstrate the previous steps of store, explore and analyze on two case studies to obtain results that are not possible via conventional methods.

6.3 Datasets

The image data stored in *NeuroData* comprise approximately 80 teravoxels of (uncompressed) public image data and approximately 150 teravoxels of (uncompressed) image data inclusive of private data across approximately 100 datasets. These datasets span spatial scales of experimental neuroscience, ranging from nanometer scale with electron microscopy to millimeter scale with MRI. We also store time-series data [147], and a calcium imaging time-series dataset. By storing disparate datasets in the same format, anyone can access and analyze many different datasets with the same functionality and syntax. Because of this, *NeuroData* is one of the largest and most diverse public neuroscience data repositories in the world. To complement the image databases, we have manually and machine-generated annotation datasets leveraging the RAMON data standard described previously. These data provide semantic objects suitable for analysis and extensibility. In addition to the public data, we also host a large number of private image and

annotation datasets, including a variety of additional modalities. Our repository includes >300 array tomography datasets associated with the *NeuroData* Synaptome Project; approximately 20 different whole brain CLARITY datasets (including several multispectral datasets); and several expansion microscopy and X-ray microtomography datasets.

6.4 Reproducible Science

The capability we have enabled for the greater neuroscience community is more important than the particular questions we have addressed in this work. Anyone, professional or citizen scientists alike, can use our services to learn about the brain. Power users can modify our scripts to answer different questions, and developers can add additional capabilities for the community. As the community increasingly moves towards collecting massive datasets, we hope these tools can be useful to address a wide range of scientific questions, and for other disciplines that are collecting 3D+ data.

Crucially, all of these tools are interoperable, extensible, and open source with a very permissive license (Apache 2.0), meaning that anyone can use, copy, or modify them for free. The brain graphs and other derivatives are provided to the public for download and analysis. Once data are collected, our basic organization allows anyone to use these tools to (i) store, (ii) explore, (iii) analyze and (iv) model properties of the data. Our website (`http://neurodata.io`) provides full documentation and tutorials; because everything is developed open source, as we continue to scale and add features, these improve-

ments will be available nearly instantaneously.

6.5 Extensible Neurocartography

Neuroscience data is growing rapidly in size and complexity, leading to new results that push the frontiers of scientific discovery. However, the various data formats and the sheer size of these datasets make reproducibility and extensibility difficult. Historically, scientific data are stored in a lab-specific format that becomes especially challenging to parse as we enter the age of big data. Furthermore, often operations are common across analyses and are reimplemented in slightly different ways, impeding the pace of progress.

NeuroData provides an answer to many of these challenges that are especially important for the very large datasets starting to emerge. We demonstrate our framework by examining the results from a recent landmark paper [22]. The authors heroically manually annotated a subvolume of somatosensory cortex, and provided three things:

- a paragraphical "parts list" of all the anatomical objects in a small volume;
- an Excel file listing all 1,700 synapses, as well as over 20 properties of each (e.g., post-synaptic density centroid location and size); and
- a VAST remote volume file containing manual labels, as well as additional information (e.g., object type, and parent-child hierarchy).

We converted this information into the RAMON annotation data standard using RESTful calls, and uploaded the data to *ndstore* and *RamonDB*. One immediate implication is that

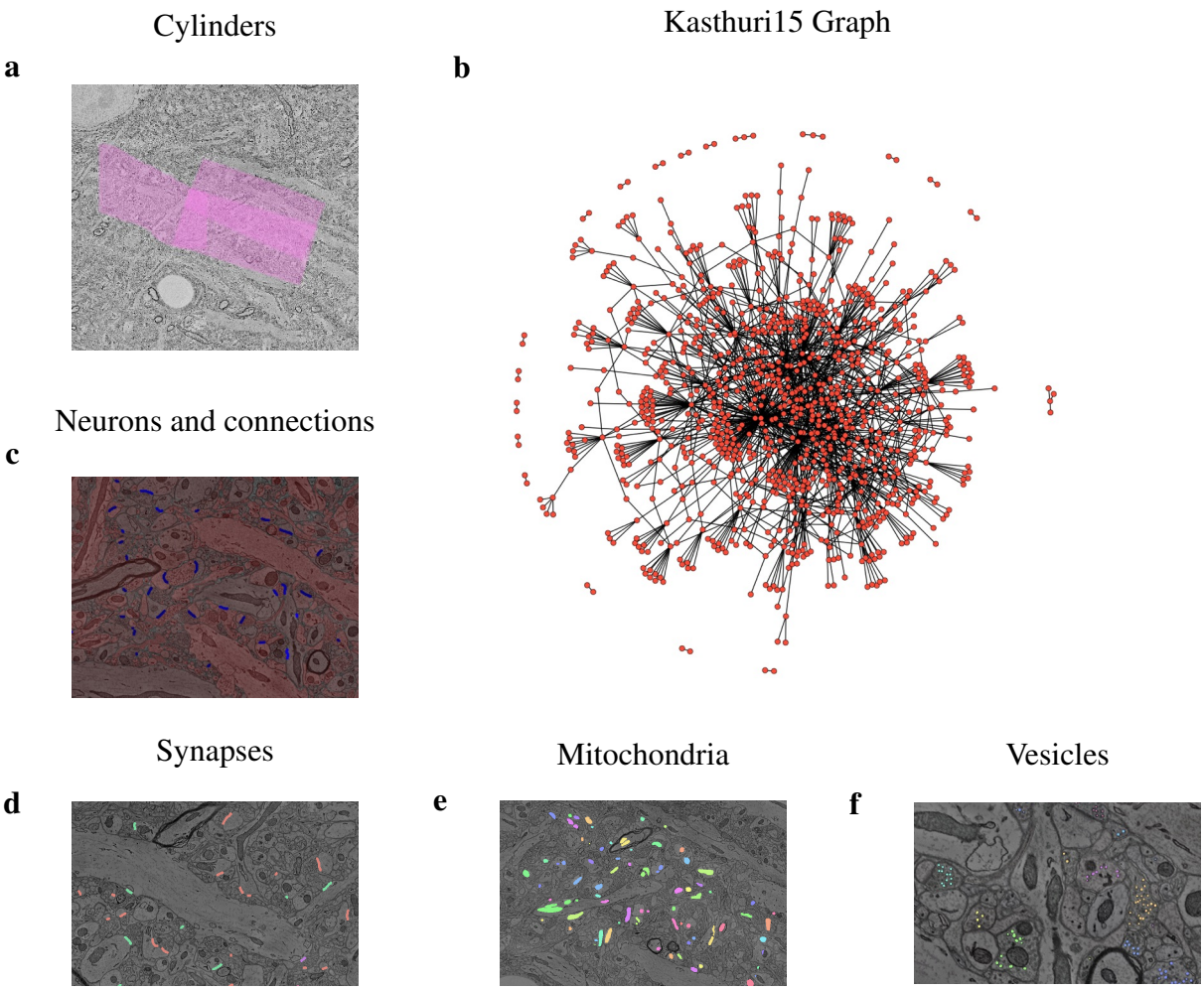

Figure 6.3: *An illustration showing several of the key automatically-produced neurocartography claims.* a. visualization of the three-cylinder mask volume; b. synapses and their neurite partners; c. synapses overlaid on EM data; d. mitochondria overlaid on EM data; and e. vesicles overlaid on EM data. A graph showing the connections between neurons is shown at the upper right.

we could check for (and improve) consistency between the provided excel spreadsheet and the VAST data. Indeed, we discovered that for 2% of the synapses in the spreadsheet, the location provided did not match any synapses in the actual VAST export. Similarly,

the VAST export contained synapse fragments (many were small painting artifacts) that we discarded because they were not indicated as a connection in the spreadsheet. For manually-segmented neurons, we applied the volume mask for the "3-cylinder region" to ensure that our analyses were based only on the voxels that had been most carefully annotated and proofread. While this may seem to be a banality, checking the data consistency is the first step for any data analysis, and often the first "result" are idiosyncrasies that can be corrected. Of the inconsistencies, we ignored some that we could not determine how to correct, and corrected the rest to proceed.

We proceeded to assess many quantitative claims from the manuscript (e.g, counting particular kinds of objects, computing volumes). Within VAST, there is no way to compute these quantities, and the spreadsheet does not contain enough information to extract or reproduce all of these results (because it only includes metadata and does not include bounding boxes or volumetric data). However, using RamonDB (object metadata and labeled voxels) one can automatically and reproducibly compute all of these properties. Table 6.1 reflects the major quantitative claims reproduced in Jupyter notebooks using *ndio* and other *NeuroData* tools.

We show cutouts with annotation overlays for each of the major annotation channels (Figure 6.3a,c-f) used in this analysis, which can be retrieved using our RESTful API or viewed in *ndviz*. Additionally, we vizualize the graph automatically created from the three-cylinder data (Figure 6.3b).

claim #	**description**	**results**
00	get data	ndio can be used to retrieve datasets of interest
01	Data Stats	total dataset volume: 2,859,750 μm^3 number of 2d contours: 1,117,335 labeled voxels: 243,813,763; 42131 μm^3 (1.47%) segment count: 3,945 in cylinders (RAMON); 6,655 total
01	mask data	mask data: 90,152,653 voxels; 1,558 μm^3 segments labeled in mask: 82,232,896; 1,421 μm^3 extracellular space: 8.8% cylinder 1: 747 μm^3, cylinder 2: 629 μm^3, cylinder 3: 686 μm^3 = 1,700 μm^3 total
02	RAMON neurons	1907 objects
02	Dendrites	segments: 1,807 parent dendrites: 306 excitatory: 290 (95%) percentage smooth: 5% Total number of spines in the three cylinders is: 1,295
02	Axons	segments: 1,766; parent axons: 1,423 excitatory: 1,310 (92%) inhibitory 99 (7%)
02	Other Objects	Other objects in the three cylinders is: 21 The total number of myelinated axons in the three cylinders is: 8 The total number of astrocytes in the three cylinders is: 10 The total number of oligodendrocytes in the three cylinders is: 333 The total number of glia in the three cylinders is: 343
02	Area stats	Total voxel area of axons in volume: 14,676,088 Total voxel area of dendrites in volume: 18,343,562 Total voxel area of neurites in volume: 33,019,650 Total voxel area of glia in volume: 2,780,602 Total voxel area of masked region: 35,953,525 Percentage of voxels that are neurites in cellular volume: 0.92 Percentage of voxels that are glia in cellular volume: 0.08 Percentage of voxels that are extracellular space: 0.0043
02	Axon-Dendrite comparisons	The voxel ratio of axons to dendrites is: 0.80
02	Miscellaneous	Number of orphans is: 25 Total number of spines in red cylinder volume: 753 Total number of axon neurites in red cylinder volume: 923
03	Synapses	density of synapses: 1.09 / umˆ3 fraction of volume that is psd: 0.009
04	Mitochondria	mitochondria in cylinder 1 is: 650
05	spines	spines for apical ”red” dendrite are: 139
06	vesicles	vesicles manually annotated in cylinder 1 is: 161,368
07	connectivity	Synapse-based connectivity matrix can be automatically retrieved Touch-based connectivity matrix can be automatically retrieved
08	kasthuri spreadsheet	A simplified version of the paper spreadsheet can be automatically reproduced by our infrastructure

Table 6.1: *Enumeration of Kasthuri2015 dataset claims. NeuroData* provides the infrastructure to retrieve data according to the RAMON data standard, and to reproducibly generate statistics and visualizations of the scientific claims. This data is available for future discovery and analysis.

6.6 Synapse Spatial Distribution

The spatial distribution of synapses is a fundamental property of neural tissue, and one that has been studied in smaller volumes in the past. Large electron microscopy datasets now offer the possibility of investigating these questions at much larger scales, but require new tools and approaches for processing, analysis and assessment. *NeuroData* provides these tools and showcases a scalable, reproducible approach in this case study. The goal of this work is to enable similar research by other investigators, regardless of resources or prior experience in big data neuroscience.

Two recent studies have examined the spatial distribution of synapses in rat cortex [148, 149] in smaller volumes. Using a few thousand synapses, these studies fail to reject the hypothesis that synapses are distributed according to a random sequential adsorption model. This implied that their synapses were distributed almost randomly, with the only constraint being that they could not overlap. Their experimental design included multiple small volumes from different animals, which is often a wise experimental design choice, to mitigate batch effects [150].

In contrast to the above small volume studies, we used a single large volume of mouse visual cortex [23]. However, extracting the scientific information in large datasets requires adapting analysis paradigms and algorithms to operate on (one) large volume rather than many small volumes. This requires considering tradeoffs in terms of algorithm complexity and robustness, and incorporating strategies for dealing with block boundaries and distributed computing. *NeuroData* leverages tools such as *ndstore*, *ndio* and *ndparse* to deploy a

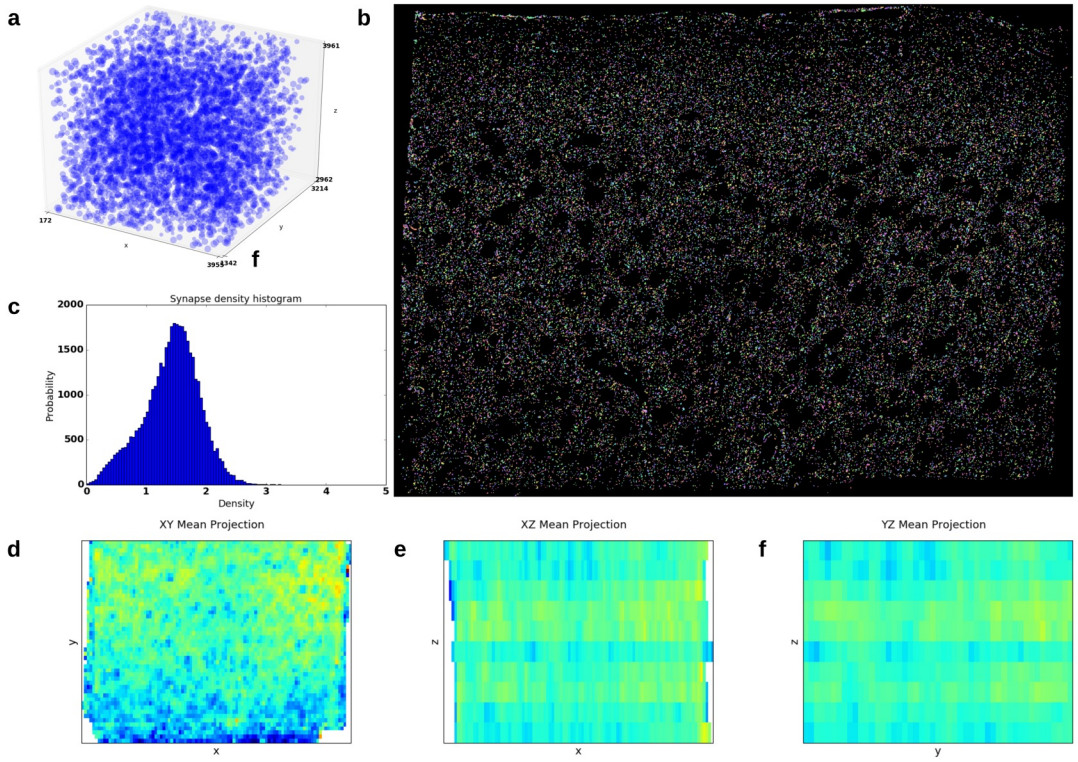

Figure 6.4: *Case Study 2 - Spatial Synapse Detection.* a. A sampling of synaptic densities are shown in 3D space, with marker size proportional to density. b. A low-resolution image of the putative synapses found on a single slice of EM tissue; note the voids corresponding to soma locations. c. A histogram showing synapse window frequency, grouped by density. A uniform distribution should have a single peak with very small sidelobes. d, e, f. The 3D mean projections along the XY, XZ, and YZ axes, respectively, better depict the non-uniform distribution of synapses found by our classifier.

state-of-the-art synapse detection algorithm [34] at scale, to vizualize and extract putative synapses, and to analyze the resulting densities.

Briefly, we began with a small volume of EM data that was locally aligned and ground

truthed (derived from [116]), and divided it into non-overlapping training and validation sets of $\sim 50-100\mu m^3$, each containing theimately 100 synapses. We quantified our performance on the small validation dataset, and chose a relatively balanced detection operating point of 0.69 precision, and 0.62 recall for our automated analysis. We used this data to train, evaluate and deploy a VESICLE-RF classifier [34] across the entire volume ($\sim$20 TB at native resolution, downsampled to $\sim$5 TB at $8 \times 8 \times 45$ nm resolution), using a variety of distributed computing tools.

We found a total of 11.7 million putative synapses, leading to a mean density estimate of 1.35 synapses/ μm^3, which is commensurate with the expected density of synapses per volume reported in the literature. This value is somewhat higher than than the density that we report for the Kasthuri data; however, local volumes in the Bock dataset have putative densities across a large range (Figure 6.4**c**). This further highlights the importance of larger volume analysis for density exploration. An image projection showing the putative synapse detections is shown to emphasize the scale of our detection result (Figure 6.4b). We investigated the result using our scalable visualization tools and identified biases introduced by our detector (not unexpected when scaling a classifier to large volumes). We reduced their effect by focusing on cortical layer 2/3, and created a mask using *ilastik* to remove regions where our detector did not perform well (e.g., pia, vessels, soma). Next, we estimated synapse densities across $\sim$ 100,000 $5 \times 5 \times 5\mu m$ non-overlapping blocks, counting only unmasked volumes; a sampling of these windows is shown (Figure 6.4a), as well as mean projections (Figure 6.4d-f).

Our complementary experimental design enabled us to consider the spatial distribution across a much larger volume and to carefully visualize and analyze the results. This prevented false results for cases in which parts of the volume locally exhibited a uniform synaptic distribution while other areas did not. Indeed, by looking at the computed histogram (Figure 6.4c) and heatmap density plots (Figure 6.4d-f), it is immediately apparent that putative synapses are *not* distributed uniformly in space and that there is a clear cluster pattern. Each of the slices in the lower inset represent a 2D slice from a 125 cubic micron volume, approaching the size of the volumes used in the previous publications. It would be very difficult to detect the anisotropic densities we observe in these smaller volumes given the nature of the non-uniformity.

Using *ndio*, we generated additional slice samples to check for bias. In most regions of the data, the classifier qualitatively performed similarly; two areas with inaccurate results were regions containing artifacts (e.g., tissue folds), and slices with degraded contrast. To partially mitigate this effect, we removed values greater than two standard deviations from the mean when testing for uniformity. Finally, to check for non-uniformity, we ran a Chi-Squared test, with the null hypothesis equal to synapses exhibiting a uniform density. We initally rejected the null with a p-value ~ 0. However, upon closer inspection, we found that many of the lower density regions contained unmasked cell bodies that affect this statistical test, and additional efforts to qualitatively validate the result by separating blocks into high and low density regions were unsuccessful when performed by a non-expert. Therefore, this result should be considered preliminary, and assessment of the underlying detector bias

by an expert neuroanatomist is needed to confirm or reject this result. That is, we provided the framework to retrieve and assess targeted volumes of tissue, but scientific validation is still ongoing.

This large-scale analysis represents a new frontier in neuroscience that is only possible using big data neuroscience techniques; indeed this is the largest known assessment of synaptic density to date. This volume is much smaller than the cubic millimeter or whole brain (in light microscopy) datasets beginning to be generated, and these tools can be used for diverse, larger analyses as volumes become available.

6.7 Summary of Chapter Contributions

We have developed all of the tools to enable big data neuroscience for anyone with internet access. This required developing tools for (i) storing, (ii) exploring, (iii) analyzing, and (iv) modeling big neuroscience data of several types, including images (raw data) and shapes (annotated data). This collection of tools has enabled us to standardize big neuroscience experiments from many different modalities, including electron microscopy, array tomography, CLARITY, Expansion Microscopy, X-ray Microscopy, Optophysiology (Calcium Imaging), and Magnetic Resonance Imaging.

Because we started with EM data, we focused our scientific discovery use cases on two reference nanoscale datasets, Bock et al. (2011) [23] and Kasthuri et al. (2015) [22]. In the original Kasthuri manuscript, the authors included a set of image data, VAST files,

Excel files, and thousands of lines of MATLAB code to obtain answers to fundamental neuroanatomy questions at the nanoscale. We were able to build Jupyter notebooks that approximately reproduced a large fraction of their results. More importantly, utilizing our services, anyone in the world can now run those *exact* same notebooks to obtain the *exact* same answers.

In the second case study, we used a significantly larger dataset ($\sim$20 TB) from another recent landmark EM paper [23]. From that data, we characterize the three-dimensional distribution of $\sim$10 million synapses in mammalian cortex, and build tools to test a hypothesis about uniformity at a scale much larger than previously explored.

We are in the process of scaling up the number of datasets, the range of experimental modalities, and the web-services we enable. All of our code and data are available online, in accordance with best practices of open science, allowing others to replicate and extend both our scientific results as well as our computational infrastructure.

Chapter 7

Conclusion

Based on the large-scale image analysis experiments conducted in this work, we propose future experiments to enhance integrated and multimodal discovery.

7.1 Integrated, End-to-End Discovery

Imaging methods used in modern neuroscience experiments are quickly producing large amounts of data capable of providing new knowledge about neuroanatomy and function. A great deal of information in these datasets is relatively unexplored and untapped. One of the bottlenecks in knowledge extraction is that often there is no feedback loop between the knowledge produced (e.g., graph, density estimate, or other statistic) and the earlier stages of the pipeline, such as acquiring the data. We advocate for the development of sample-to-knowledge discovery pipelines that can be used to optimize acquisition and processing

steps toward a targeted objective. We therefore propose that optimization takes place not just within each processing stage, but also between adjacent and non-adjacent steps of the pipeline.

Much of scientific exploration involves three main stages to translate raw data to knowledge suitable for making scientific discoveries: acquisition, processing, and analysis. These stages, which begin with sample collection and result in mathematical analysis (knowledge), are typically performed independently. These stages are typically optimized in a feed-forward manner, without the ability to revise previous steps. This is problematic because it is important to consider the best set of parameters in a global context, where the question of interest might well lead to a solution that is not obvious at a particular stage of the pipeline. Many challenges and potential improvements have been identified (e.g., [10, 37]), and we believe that significant advancements may be made by combining these ideas with an integrated approach.

7.2 Multimodal Discovery

We have shown the ability to extract knowledge from different modalities. These methods were generally conducted on different imaging volumes, but because MRI is non-destructive and the preparation for X-ray microtomography is compatible with nanoscale imaging, these tools could be combined to extract brain maps at all three resolutions following a hierarchical approach. Targeted high-resolution studies could be used to augment the scaffold gained from the faster, cheaper, lower-resolution approach.

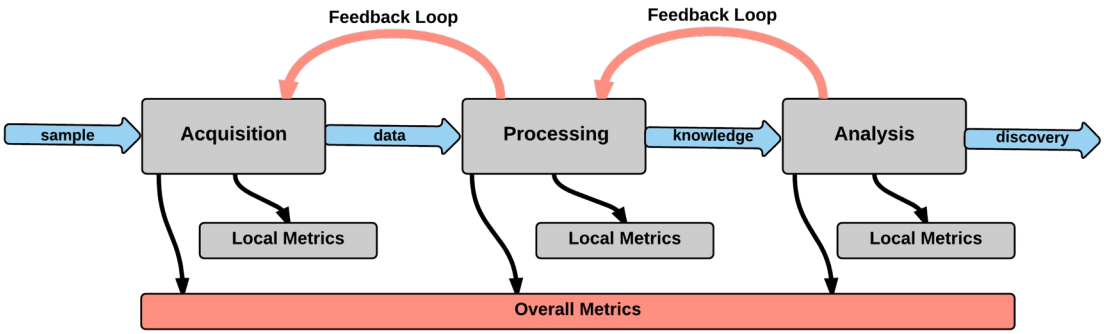

Figure 7.1: *Anatomy of an experiment.* We illustrate our augmented processing sample-to-knowledge workflow for a scientific experiment. A traditional workflow consists of a feed-forward chain of stages (gray), which represent major (often disparate) building blocks. The products of these stages (blue arrows) represent the current interface points. Our augmented pipeline adds feedback loops between stages and interfaces to an overall knowledge metric which may lead to improved performance.

We briefly outline an experimental paradigm (Figure 2) leveraging advances in multi-modal brain mapping approaches (e.g., electron and X-Ray microscopy). Our goal is to better estimate synaptic density and investigate community structure through strategic sampling. We detail how large mesoscale maps of brains can complement higher resolution EM maps and may potentially bridge the scale gap.

7.3 Future Work

Since the field of connectomics is young, many challenges remain to reconstruct biofidelic graphs. The challenges highlighted below are some of the key drivers for improvement,

and researchers are already beginning to make progress in many of these areas.

7.3.1 Metrics

Understanding the impact that a change to the structure of a graph has on a particular function on the graph is still challenging. Indeed, many functions that we wish to compute on a graph (e.g., clustering coefficient, degree distribution), do not produce smooth mappings when we perturb our estimate of the graph structure. When we change our estimate of a graph, it is possible to produce dramatic changes in the sufficient statistic of interest. Therefore it is especially critical to consider the inference impacts on graph changes when designing robust and efficient metrics.

7.3.2 Scalable Tools

To our knowledge, there are few examples of fully integrated pipelines in neuroimaging, although much research has focused on building well-engineered stages that optimize a chosen local metric. We assert that these stages can be readily combined to produce a true sample-to-knowledge pipeline, leading to improved efficiencies and performance. We believe that automation is necessary to effectively address data at this scale, and that additional research is needed to scale from terabytes to petabytes and beyond.

7.3.3 Error Checking

Errorful results may still allow for successful analysis and inference, especially if those errors can be quantified. Meaningfully assessing and correcting errors in large-scale neuronal maps (e.g., through an analysis of expected biological priors) is an exciting, but challenging avenue for future research.

7.3.4 Community Outreach

Building common tools will enable diverse research groups from across the community to contribute, and will more easily enable reproducible and extensible science. Moreover, many of the challenges dealing with very large image volumes are common across groups and solutions can be readily shared with agreed upon standards.

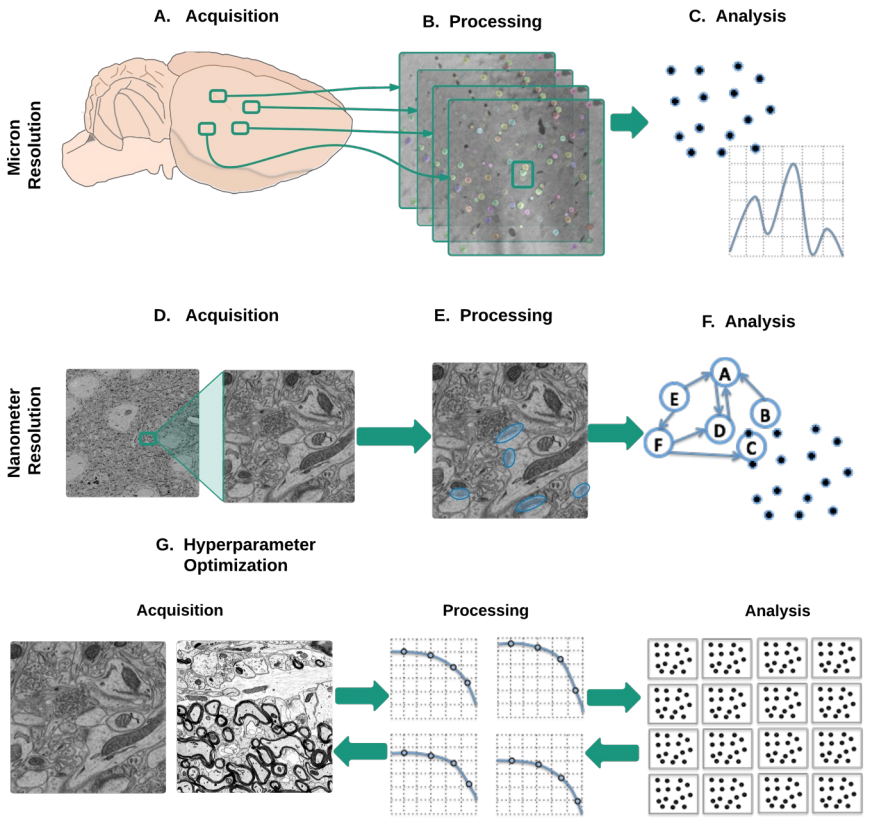

Figure 7.2: *Multimodal Synapse Motifs.* This figure illustrates the proposed experimental paradigm: (A) Initially a mesoscale image of the brain is reconstructed using X-ray microtomography. (B) This volume is then used by image analysis algorithms to estimate of cell body location and size. (C) This knowledge can be represented as a map of the cell body locations in space, along with relevant attributes such as confidence and size. (D) High resolution electron microscopy imaging occurs for selected blocks; X-ray imaging is non-destructive, and so it is possible to re-image interesting locations in the same sample. (E) Next, we locate all synapses using automated approaches, which leads to (F) knowledge about relative position and densities for each block in support of scientific discovery. (G) At each stage opportunities exist for local and global optimization of knowledge.

7.4 Summary

In this dissertation we outline methods to automatically estimate and assess maps of the brain at multiple resolutions in challenging operating environments. We develop several metrics that are particularly informative for measuring graph error, and help to focus research on end-to-end approaches designed to optimize the knowledge of interest (e.g., graphs). We use those results to identify remaining challenges and drive the development of new image analysis algorithms. Our image analysis, metrics, and graph reconstruction algorithms yield results for open neuroscience and connectomics at large-scale. We built a scalable platform for scientific discovery that can be used as a starting point to build larger, more accurate brain graphs, and to extract knowledge for inference tasks in health care and biofidelic computing.

Chapter 8

Appendix

8.1 Websites and Links

Links to projects, code, and documentation can be found on my github site at: `github.com/wrgr/links/blob/master/dissertation.md`, and the following websites:

- NeuroData: Enabling Big Data Neuroscience (`neurodata.io`)
- JHU/APL Intelligent Systems Center Github page (`www.github.com/iscoe/`)
- My github site (`www.github.com/wrgr`)

8.2 Front-End Processing Infrastructure

Our infrastructure continues to evolve; the following specifications are provided as an illustration of the compute and storage resources used for our experiments. Data and results were stored using *NeuroData*. Our CPU compute cluster for images-to-graphs evaluation and deployment had a peak usage of 100 cores and 1TB RAM. The (larger) overall cluster configuration consisted of AMD Opteron 6348 cores (2.8GHz, 12-core/processor, 4 processors/node) and 256GB RAM per node. For membrane detection, we used a small GPU cluster containing 27 GeForce GTX Titan cards with 6GB RAM. We leveraged Son of Grid Engine (SGE) for task scheduling and the LONI Pipeline for workflow management [43]. Because we parallelized at a data block level, each task is embarassingly parallel, and so we used traditional scheduling methods.

8.3 Back-End *NeuroData* Services

On the backend, *OCP* uses a load-balancing webserver (2x Intel Xeon X5650, 2.67GHz, 12 core/processor and 48 GB of RAM). This webserver distributes jobs across three data servers running a distributed database (each with 2x Intel Xeon X5690, 3.47GHz, 12 core/processor and 48GB of RAM). Additionally, 100TB of network mounted disk-space is available for storage [103]. An overall schematic of our infrastructure is shown in Figure 2.5.

Bibliography

[1] N. Kasthuri and J. W. Lichtman, "Neurocartography," *Neuropsychopharmacology*, vol. 35, no. 1, pp. 342–343, 01 2010.

[2] O. Sporns, "Networks of the Brain," *Learning*, no. August, p. 375, 2010.

[3] J. W. Lichtman and J. R. Sanes, "Ome sweet ome: what can the genome tell us about the connectome?" *Current opinion in neurobiology*, vol. 18, no. 3, Jun. 2008.

[4] L. W. Swanson and J. W. Lichtman, "From cajal to connectome and beyond," *Annual Review of Neuroscience*, vol. 39, no. 1, pp. 197–216, 2016/09/03 2016.

[5] H. S. Seung, *Connectome: How the Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt, 2012.

[6] O. Sporns, G. Tononi, and R. Kötter, "The human connectome: a structural description of the human brain," *PLoS Computational Biology*, vol. 1, no. 4, p. e42, 2005.

[7] W. Denk and H. Horstmann, "Serial block-face scanning electron microscopy to

reconstruct three-dimensional tissue nanostructure." *PLoS biology*, vol. 2, no. 11, p. e329, Nov. 2004.

[8] E. L. Dyer, W. Gray Roncal *et al.*, "Quantifying mesoscale neuroanatomy using x-ray microtomography," *In Review*, "2016".

[9] D. G. Nishimura, *Principles of Magnetic Resonance Imaging*. Stanford University, 2010.

[10] J. W. Lichtman, H. Pfister, and N. Shavit, "The big data challenges of connectomics," *Nature Neuroscience*, vol. 17, no. 11, 2014.

[11] M. Helmstaedter, "Cellular-resolution connectomics: challenges of dense neural circuit reconstruction." *Nature methods*, vol. 10, no. 6, pp. 501–7, Jun. 2013.

[12] S. Mori, *Introduction to Diffusion Tensor Imaging*. Elsevier B.V., 2007.

[13] H. Johansen-Berg and T. E. Behrens, Eds., *Diffusion MRI: From Quantitative Measurement to In vivo Neuroanatomy*. Elsevier B.V., 2009.

[14] W. Gray Roncal, J. A. Bogovic *et al.*, "Magnetic resonance connectome automated pipeline: An overview," *IEEE Pulse*, vol. 3, no. 2, pp. 42–48, Mar (c) 2010 IEEE.

[15] W. Gray Roncal, Z. H. Koterba *et al.*, "Migraine: Mri graph reliability analysis and inference for connectomics," Dec (c) 2013 IEEE.

[16] G. Kiar, W. Gray Roncal *et al.*, "ndmg: Neurodata's mri graphs pipeline," https://doi.org/10.5281/zenodo.60206, 2016.

BIBLIOGRAPHY

[17] D. L. Sussman, D. Mhembere *et al.*, "Massive diffusion mri graph structure preserves spatial information," *Organization for Human Brain Mapping*, 2013.

[18] K. J. Gorgolewski, A.-A. Fidel *et al.*, "Bids apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods," *BioArxiv*, 2016.

[19] R. Tang, M. Ketcha *et al.*, "Law of large graphs," *arXiv preprint*, 2016.

[20] G. Kiar, K. J. Gorgolewski *et al.*, "Science in the cloud (sic): A use case in mri connectomics," *arXiv preprint*, 2016.

[21] E. S. Lein, M. J. Hawrylycz *et al.*, "Genome-wide atlas of gene expression in the adult mouse brain," *Nature*, vol. 445, no. 7124, pp. 168–176, 01 2007.

[22] N. Kasthuri, K. J. Hayworth *et al.*, "Saturated Reconstruction of a Volume of Neocortex," *Cell*, vol. 162, no. 3, pp. 648–661, 2015.

[23] D. D. Bock, W.-C. A. Lee *et al.*, "Network anatomy and in vivo physiology of visual cortical neurons," *Nature*, vol. 471, no. 7337, pp. 177–182, 2011.

[24] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The Structure of the Nervous System of the Nematode Caenorhabditis elegans," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 314, no. 1165, pp. 1–340, Nov. 1986.

[25] S. Takemura, A. Bharioke *et al.*, "A visual motion detection circuit suggested by Drosophila connectomics," *Nature*, vol. 500, no. 7461, pp. 175–181, Aug. 2013.

BIBLIOGRAPHY

[26] W.-C. A. Lee, V. Bonin *et al.*, "Anatomy and function of an excitatory network in the visual cortex," *Nature*, vol. 532, no. 7599, pp. 370–374, 2016.

[27] J. L. Morgan, D. R. Berger *et al.*, "Article The Fuzzy Logic of Network Connectivity in Mouse Visual Thalamus," *Cell*, vol. 165, no. 1, pp. 192–206, 2016.

[28] "MICrONS: Machine Intelligence from Cortical Networks," iarpa.gov/index.php/research-programs/microns, accessed: 2016-05-19.

[29] "NeuroData: Enabling Big Data Neuroscience for Everyone," *In Preparation*, 2016.

[30] J. Matelsky, S. Berg *et al.*, "ndio: Neuroscience Discovery Input and Output," in *Society for Neuroscience Abstract*, San Diego, 2016.

[31] W. Gray Roncal, M. Pekala *et al.*, "ndparse: Tools and Interfaces for Scalable Neuroscience Discovery," in *Society for Neuroscience Abstract*, 2016.

[32] G. Kiar, "Gremlin: Graph estimation from mr images leading to inference in neuroscience," Ph.D. dissertation, Master's Thesis, Johns Hopkins University, 2016.

[33] W. Gray Roncal, D. M. Kleissas *et al.*, "An Automated Images-to-Graphs Framework for High Resolution Connectomics," *Frontiers in neuroinformatics*, pp. 1–13.

[34] W. Gray Roncal, M. Pekala *et al.*, "VESICLE: Volumetric Evaluation of Synaptic Inferfaces using Computer vision at Large Scale," in *26th British Machine Vision Conference (BMVC)*, 2015, pp. 1–9.

[35] W. Gray Roncal, C. Lea, A. Baruah, and G. D. Hager, "Santiago: Spine association for neuron topology improvement," *arXiv preprint*, 2016.

[36] W. Gray Roncal, E. L. Dyer, G. Doga, and K. Kording, "From sample to knowledge: Towards an integrated approach for neuroscience discovery," *arXiv preprint*.

[37] S. Plaza, "Focused proofreading - efficiently extracting connectomes from segmented em images," *arXiv preprint*, 2014.

[38] T. Sherif, P. Rioux *et al.*, "Cbrain: a web-based, distributed computing platform for collaborative neuroimaging research," *Frontiers in Neuroinformatics*, vol. 8, p. 54, 2014.

[39] S. Das, T. Glatard *et al.*, "The {MNI} data-sharing and processing ecosystem," *NeuroImage*, vol. 124, Part B, pp. 1188 – 1195, 2016, sharing the wealth: Brain Imaging Repositories in 2015.

[40] D. Landis, W. Courtney *et al.*, "{COINS} data exchange: An open platform for compiling, curating, and disseminating neuroimaging data," *NeuroImage*, vol. 124, Part B, pp. 1084 – 1088, 2016, sharing the wealth: Brain Imaging Repositories in 2015.

[41] P. A. Yushkevich, J. Piven *et al.*, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[42] C. Sommer, C. Straehle, U. Koethe, and F. Hamprecht, "ilastik: Interactive Learning and Segmentation Toolkit," *8th IEEE International Symposium on Biomedical Imaging (ISBI 2011)*, 2011.

[43] D. E. Rex, J. Q. Ma, and A. W. Toga, "The LONI Pipeline Processing Environment," *NeuroImage*, vol. 19, no. 3, pp. 1033–1048, Jul. 2003.

[44] "MARCC: Maryland Advanced Research Computing Center," http://www.marcc.jhu.edu, accessed: 2016-09-05.

[45] "Johns Hopkins University Data-Scope," http://idies.jhu.edu/resources/datascope/, accessed: 2016-09-05.

[46] J. Lu, J. C. Tapia, O. L. White, and J. W. Lichtman, "The Interscutularis Muscle Connectome," *PLoS Biology*, vol. 7, no. 2, 2009.

[47] A. H. Marblestone *et al.*, "Conneconomics: The Economics of Large-Scale Neural Connectomics," Dec. 2013.

[48] C. E. Priebe, D. L. Sussman, M. Tang, and J. T. Vogelstein, "Statistical inference on errorfully observed graphs," *Journal of Computational and Graphical Statistics*, vol. 24, no. 4, pp. 930–953, 2015.

[49] M. Helmstaedter, K. L. Briggman, and W. Denk, "High-accuracy neurite reconstruction for high-throughput neuroanatomy," *Nature Neuroscience*, vol. 14, no. 8, pp. 1081–1088, 2011.

[50] J. Nunez-Iglesias, R. Kennedy *et al.*, "Machine learning of hierarchical clustering to segment 2D and 3D images." *PloS one*, vol. 8, no. 8, p. e71715, Jan. 2013.

[51] V. Kaynig-Fittkau, A. Vazquez-Reina *et al.*, "Large-scale automatic reconstruction of neuronal processes from electron microscopy images," *IEEE transactions on medical imaging*, vol. 1, no. 1, pp. 1–14, Mar. 2013.

[52] "Miccai challenge on circuit reconstruction from electron microscopy images," http://www.cremi.org.

[53] V. Lyzinski, D. L. Sussman *et al.*, "Seeded graph matching for large stochastic block model graphs," *arXiv preprint*, p. 1310.1297, Oct 2013.

[54] G. Golub and C. Van-Loan, *Matrix Computations*, 1996, vol. 10, no. 8.

[55] U. Braun, M. M. Plichta *et al.*, "Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures." *Neuroimage*, vol. 59, no. 2, pp. 1404–1412, 2012.

[56] J. T. Vogelstein, R. J. Vogelstein, and C. E. Priebe, "Are mental properties supervenient on brain properties?" *Nature Scientific Reports*, p. 11, 2011.

[57] D. C. Van Essen, K. Ugurbil *et al.*, "The Human Connectome Project: a data acquisition perspective." *NeuroImage*, vol. 62, no. 4, pp. 2222–31, Oct. 2012.

[58] M. Mennes, B. B. Biswal, F. X. Castellanos, and M. P. Milham, "Making data sharing work: The FCP/INDI experience." *NeuroImage*, Oct. 2012.

[59] A. Daducci, S. Gerhard *et al.*, "The connectome mapper: an open-source processing pipeline to map connectomes with MRI." *PloS one*, vol. 7, no. 12, p. e48121, Jan. 2012.

[60] Z. Cui, S. Zhong *et al.*, "PANDA: a pipeline toolbox for analyzing brain diffusion images." *Frontiers in human neuroscience*, vol. 7, no. February, p. 42, Jan. 2013.

[61] B. A. Landman, A. J. Huang *et al.*, "Multi-parametric neuroimaging reproducibility: a 3-T resource study." *NeuroImage*, vol. 54, no. 4, pp. 2854–66, Mar. 2011.

[62] K. B. Nooner, S. Colcombe *et al.*, "The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry," *Frontiers in Neuroscience*, vol. 6, no. 152, 2012.

[63] B. C. Lucas, J. a. Bogovic *et al.*, "The Java Image Science Toolkit (JIST) for rapid prototyping and publishing of neuroimaging software." *Neuroinformatics*, vol. 8, no. 1, pp. 5–17, Mar. 2010.

[64] I. D. Dinov, J. D. Van Horn *et al.*, "Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline." *Frontiers in neuroinformatics*, vol. 3, p. 22, Jan. 2009.

[65] M. Jenkinson, C. F. Beckmann *et al.*, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[66] G. Grabner, A. L. Janke *et al.*, "Symmetric Atlasing and Model Based Segmentation:

An Application to the Hippocampus in Older Adults," in *Lecture Notes in Computer Science*, 2006, vol. 4191, pp. 58–66.

[67] R. S. Desikan, F. Ségonne *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest." *NeuroImage*, vol. 31, no. 3, pp. 968–80, Jul. 2006.

[68] S. Mori, B. Crain, V. Chacko, and P. Van Zijl, "Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging," *Annals of neurology*, vol. 45, no. 2, pp. 265–269, Feb. 1999.

[69] R. C. Craddock, S. Jbabdi *et al.*, "Imaging human connectomes at the macroscale." *Nature methods*, vol. 10, no. 6, pp. 524–39, Jun. 2013.

[70] A. Carass, M. B. Wheeler *et al.*, "A joint registration and segmentation approach to skull stripping," in *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, April 2007, pp. 656–659.

[71] D. L. Pham, "Spatial models for fuzzy clustering," *Computer Vision and Image Understanding*, vol. 84, no. 2, pp. 285–297, 2001.

[72] G. K. Rohde, A. Aldroubi, and B. M. Dawant, "The adaptive bases algorithm for intensity-based nonrigid image registration," *IEEE Transactions on Medical Imaging*, vol. 22, no. 11, pp. 1470–1479, Nov 2003.

[73] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance

level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, July 2004.

[74] D. Le Bihan, J.-F. Mangin *et al.*, "Diffusion tensor imaging: Concepts and applications," *Journal of Magnetic Resonance Imaging*, vol. 13, no. 4, pp. 534–546, 2001.

[75] S. Wakana, A. Caprihan *et al.*, "Reproducibility of quantitative tractography methods applied to cerebral white matter," *NeuroImage*, vol. 36, no. 3, pp. 630 – 644, 2007.

[76] T. Behrens, M. Woolrich *et al.*, "Characterization and propagation of uncertainty in diffusion-weighted mr imaging," *Magnetic Resonance in Medicine*, vol. 50, no. 5, pp. 1077–1088, 2003.

[77] G. J. Parker, H. A. Haroon, and C. A. Wheeler-Kingshott, "A framework for a streamline-based probabilistic index of connectivity (pico) using a structural interpretation of mri diffusion measurements," *Journal of Magnetic Resonance Imaging*, vol. 18, no. 2, pp. 242–254, 2003.

[78] D. K. Jones, "Tractography gone wild: Probabilistic fibre tracking using the wild bootstrap with diffusion tensor mri," *IEEE Transactions on Medical Imaging*, vol. 27, no. 9, pp. 1268–1274, Sept 2008.

[79] D. Mhembere, W. Gray Roncal *et al.*, "Computing Scalable Multivariate Glocal

Invariants of Large (Brain-) Graphs," in *Global Conference on Signal and Information Processing*, 2013.

[80] J. E. Gonzalez, D. Bickson, and C. Guestrin, "PowerGraph : Distributed Graph-Parallel Computation on Natural Graphs," pp. 17–30, 2012.

[81] E. Garyfallidis, M. Brett *et al.*, "Dipy, a library for the analysis of diffusion mri data," *Frontiers in Neuroinformatics*, vol. 8, p. 8, 2014.

[82] S. Wang, Z. Yang *et al.*, "Optimal decisions for discovery science via maximizing discriminability: Applications in neuroimaging," *In Preparation*, 2016.

[83] S. Wang *et al.*, "Optimal design for discovery science: Applications in neuroimaging," *Organization for Human Brain Mapping*, 2015.

[84] J. T. Vogelstein, W. Gray Roncal, R. J. Vogelstein, and C. E. Priebe, "Graph classification using signal-subgraphs: applications in statistical connectomics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1539–1551, Jul 2013.

[85] J. W. Lichtman and W. Denk, "The big and the small: challenges of imaging the brain's circuits." *Science (New York, N.Y.)*, vol. 334, no. 6056, pp. 618–23, Nov. 2011.

[86] M. N. Economo, N. G. Clack *et al.*, "A platform for brain-wide imaging and reconstruction of individual neurons," *eLife*, vol. 5, p. e10566, 2016.

BIBLIOGRAPHY

[87] K. Amunts, C. Lepage *et al.*, "Bigbrain: an ultrahigh-resolution 3D human brain model," *Science*, vol. 340, no. 6139, pp. 1472–1475, 2013.

[88] A. Eberle, S. Mikula *et al.*, "High-resolution, high-throughput imaging with a multibeam scanning electron microscope," *Journal of microscopy*, vol. 259, no. 2, pp. 114–120, 2015.

[89] K. Chung and K. Deisseroth, "Clarity for mapping the nervous system," *Nature methods*, vol. 10, no. 6, pp. 508–513, 2013.

[90] F. Chen, P. W. Tillberg, and E. S. Boyden, "Expansion microscopy," *Science*, vol. 347, no. 6221, pp. 543–548, 2015.

[91] E. A. Bushong, D. D. Johnson *et al.*, "X-ray microscopy as an approach to increasing accuracy and efficiency of serial block-face imaging for correlated light and electron microscopy of biological specimens," *Microscopy and Microanalysis*, vol. 21, no. 01, pp. 231–238, 2015.

[92] S. Mikula and W. Denk, "High-resolution whole-brain staining for electron microscopic circuit reconstruction," *Nature Methods*, vol. 12, no. 6, pp. 541–546, 2015.

[93] A. Arillo, E. Penalver *et al.*, "Long-proboscid brachyceran flies in Cretaceous amber (Diptera: Stratiomyomorpha: Zhangsolvidae)," *Systematic Entomology*, vol. 40, no. 1, pp. 242–267, Jan. 2015.

[94] S. E. Hieber, C. Bikis *et al.*, "Tomographic brain imaging with nucleolar detail and automatic cell counting," *Scientific Reports*, vol. 6, pp. 32 156 EP –, 09 2016.

[95] R. Mizutani, A. Takeuchi *et al.*, "Microtomographic analysis of neuronal circuits of human brain," *Cerebral Cortex*, 2009.

[96] R. Mizutani, A. Takeuchi *et al.*, "Unveiling 3d biological structures by x-ray microtomography," *Microscopy: Science, Technology, Applications and Education., Formatex Research Center, Badajoz*, pp. 379–386, 2010.

[97] R. Mizutani and Y. Suzuki, "X-ray microtomography in biology," *Micron*, vol. 43, no. 2, pp. 104–115, 2012.

[98] F. De Carlo, D. Gürsoy *et al.*, "Scientific data exchange: a schema for hdf5-based storage of raw and analyzed data," *Journal of synchrotron radiation*, vol. 21, no. 6, pp. 1224–1230, 2014.

[99] J. C. Tapia, N. Kasthuri *et al.*, "High-contrast en bloc staining of neuronal tissue for field emission scanning electron microscopy," *Nature protocols*, vol. 7, no. 2, pp. 193–206, 2012.

[100] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[101] B. Póczos and J. G. Schneider, "On the estimation of alpha-divergences," in *Int. Conf. on Artif. Int. and Stat.*, 2011, pp. 609–617.

[102] D. O. Loftsgaarden, C. P. Quesenberry *et al.*, "A nonparametric estimate of a multivariate density function," *Ann. Math. Stat.*, vol. 36, no. 3, pp. 1049–1051, 1965.

[103] R. Burns, W. Gray Roncal *et al.*, "The Open Connectome Project Data Cluster: Scalable Analysis and Vision for High-Throughput Neuroscience," *Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM)*, Jun 2013.

[104] P. S. Tsai, J. P. Kaufhold *et al.*, "Correlations of neuronal and microvascular densities in murine cortex revealed by direct counting and colocalization of nuclei and vessels," *The Journal of Neuroscience*, vol. 29, no. 46, pp. 14 553–14 570, 2009.

[105] J. Wu, Y. He *et al.*, "3d braincv: simultaneous visualization and analysis of cells and capillaries in a whole mouse brain with one-micron voxel resolution," *Neuroimage*, vol. 87, pp. 199–208, 2014.

[106] S. Heinzer, T. Krucker *et al.*, "Hierarchical microimaging for multiscale analysis of large vascular networks," *Neuroimage*, vol. 32, no. 2, pp. 626–636, 2006.

[107] K. Zilles and K. Amunts, "Centenary of brodmann's map—conception and fate," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 139–145, 2010.

[108] D. J. Bumbarger, M. Riebesell, C. Rödelsperger, and R. J. Sommer, "System-wide Rewiring Underlies Behavioral Differences in Predatory and Bacterial-Feeding Nematodes." *Cell*, vol. 152, no. 1-2, pp. 109–19, Jan. 2013.

[109] J. L. Morgan and J. W. Lichtman, "Why not connectomics?" *Nature methods*, vol. 10, no. 6, pp. 494–500, Jun. 2013.

[110] J. Fitzsimmons, M. Kubicki, and M. E. Shenton, "Review of functional and anatomical brain connectivity findings in schizophrenia." *Current opinion in psychiatry*, vol. 26, no. 2, pp. 172–87, Mar. 2013.

[111] A. Vazquez-Reina, M. Gelbart *et al.*, "Segmentation fusion for connectomics," *2011 International Conference on Computer Vision*, pp. 177–184, nov 2011.

[112] J. Funke, B. Andres *et al.*, "Efficient automatic 3D-reconstruction of branching neurons from EM data," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2012, pp. 1004–1011.

[113] N. Kasthuri, K. Hayworth *et al.*, "The brain on tape: Imaging an Ultra-Thin Section Library (UTSL). Society for Neuroscience Abstracts," *Society for Neuroscience Abstract*, 2009.

[114] A. Kreshuk, C. N. Straehle *et al.*, "Automated detection and segmentation of synaptic contacts in nearly isotropic serial electron microscopy images." *PloS one*, vol. 6, no. 10, p. e24899, Jan. 2011.

[115] C. Becker, K. Ali, G. Knott, and P. Fua, "Learning context cues for synapse segmentation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 10, pp. 1864–1877, 2013.

[116] A. Kreshuk, U. Koethe *et al.*, "Automated detection of synapses in serial section transmission electron microscopy image stacks." *PloS one*, vol. 9, no. 2, p. e87351, Jan. 2014.

[117] Y. Mishchenko, T. Hu *et al.*, "Ultrastructural Analysis of Hippocampal Neuropil from the Connectomics Perspective," *Neuron*, vol. 67, no. 6, pp. 1009–1020, 2010.

[118] D. D. Ciresan, A. Giusti, L. M. L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[119] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[120] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," *Proceedings of the 25th International Conference on Machine Learning (2008)*, pp. 96–103, 2008.

[121] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[122] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer, 2013, pp. 411–418.

[123] H. R. Roth, A. Farag *et al.*, "Deep convolutional networks for pancreas segmentation in CT imaging," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2015, pp. 94 131G—-94 131G.

[124] Y. Jia, E. Shelhamer *et al.*, "Caffe : Convolutional Architecture for Fast Feature Embedding," *ACM Conference on Multimedia*, 2014.

[125] M. Kazhdan, K. Lillaney *et al.*, "Gradient-Domain Fusion for Color Correction in Large EM Image Stacks," *arXiv preprint*, pp. 1–7, 2015.

[126] V. Braitenberg and A. Schüz, *Anatomy of the cortex: Statistics and geometry.* Springer-Verlag Publishing, 1991.

[127] V. B. Mountcastle, "Modality and topographic properties of single neurons of cat's somatic sensory cortex." *Journal of neurophysiology*, vol. 20, no. 4, pp. 408–34, Jul. 1957.

[128] K. Hayworth, N. Kasthuri, R. Schalek, and J. Lichtman, "Automating the Collection of Ultrathin Serial Sections for Large Volume TEM Reconstructions," *Microscopy and Microanalysis*, vol. 12, no. S02, pp. 86–87, Jul. 2006.

[129] J. Masci, A. Giusti *et al.*, "A Fast Learning Algorithm for Image Segmentation with Max-Pooling Convolutional Networks," *arXiv preprint*, 2013.

[130] J. I. Arellano, R. Benavides-Piccione, J. DeFelipe, and R. Yuste, "Ultrastructure

of Dendritic Spines: Correlation Between Synaptic and Spine Morphologies," pp. 131–143, nov 2007.

[131] R. Yuste, *Dendritic Spines*. MIT Press, 2010.

[132] S. Ramon y Cajal, "Estructura de los centros nerviosos de las aves," *Rev Trim Histol Norm Pat 1*, pp. 1–10.

[133] J. N. Bourne and K. M. Harris, "Dendritic Spines Distinctive Structural Features of Dendritic Spines," *Encyclopedia of Life Sciences*, 2007.

[134] J. D. Jackson, A. Yezzi, W. Wallace, and M. F. Bear, "Segmentation of Coarse and Fine Scale Features Using Multi-scale Diffusion and Mumford-Shah," pp. 615–624, 2003.

[135] E. Türetken, F. Benmansour *et al.*, "Reconstructing Loopy Curvilinear Structures Using Integer Programming," 2013.

[136] P. Strandmark and F. Kahl, "Shortest Paths with Higher-Order Regularization," vol. 37, no. 12, pp. 2588–2600, 2015.

[137] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[138] T. Parag, A. Chakraborty, S. Plaza, and L. Scheffer, "A Context-Aware Delayed Agglomeration Framework for Electron Microscopy Segmentation," pp. 1–19, 2015.

[139] N. Krasowski, T. Beier *et al.*, "Improving 3D EM Data Segmentation by Joint Optimization Over Boundary Evidence and Biological Priors," in *ISBI*, 2015.

[140] "NeuroProof," https://github.com/janelia-flyem/NeuroProof, accessed: 2016-10-20.

[141] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recoginition," pp. 1–14, 2015.

[142] A. Giusti, D. Ciresan *et al.*, "Fast Image Scanning with Deep Max-Pooling Convolutional Neural Networks," *arXiv preprint arXiv:1302.1700*, 2013.

[143] K. D. Micheva and S. J. Smith, "Array tomography: A new tool for imaging the molecular architecture and ultrastructure of neural circuits," *Neuron*, vol. 55, no. 5, p. 824, 2016/09/02.

[144] K. Chung, J. Wallace *et al.*, "Structural and molecular interrogation of intact biological systems," *Nature*, vol. 497, no. 7449, pp. 332–337, 05 2013.

[145] A. Cardona, S. Saalfeld *et al.*, "TrakEM2 software for neural circuit reconstruction." *PloS one*, vol. 7, no. 6, p. e38011, Jan. 2012.

[146] S. Saalfeld, A. Cardona, V. Hartenstein, and P. Tomanvcak, "CATMAID: collaborative annotation toolkit for massive amounts of image data," *Bioinformatics*, vol. 25, no. 15, pp. 1984–1986, 2009.

[147] J. Freeman, N. Vladimirov *et al.*, "Mapping brain activity at scale with cluster computing," *Nature Methods*, no. July, Jul 2014.

[148] A. Merchán-Pérez, J.-R. Rodríguez *et al.*, "Three-dimensional spatial distribution of synapses in the neocortex: a dual-beam electron microscopy study," *Cerebral Cortex*, vol. 24, no. 6, pp. 1579–1588, 2014.

[149] L. Anton-Sanchez, C. Bielza *et al.*, "Three-dimensional distribution of cortical synapses: a replicated point pattern-based analysis," *Frontiers in Neuroanatomy*, vol. 8, no. August, pp. 1–15, Aug. 2014.

[150] J. T. Leek, R. B. Scharpf *et al.*, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

Johns Hopkins University
Applied Physics Laboratory
11100 Johns Hopkins Road
Mail Stop: ISC (MP2)
Laurel, MD 20723
✉ *wgr@jhu.edu*
neurodata.io

William R. Gray Roncal

Education

2010 – 2016 **Ph.D. in Computer Science**, *Johns Hopkins University*, Committee Chair: Gregory D. Hager, **Dissertation Title:** Enabling Scalable Neurocartography: Images to Graphs for Discovery. In this work I developed a scalable computer vision framework to create and assess connectomes across multiple modalities, from millimeter (magnetic resonance imaging) to nanometer (electron microscopy). This work supports efforts to democratize science and develop novel algorithms and a deeper understanding of the brain.

2004 – 2005 **M.S. in Electrical Engineering**,
University of Southern California.

1999 – 2003 **B.E. in Electrical Engineering, Math Minor**,
Vanderbilt University, Nashville, TN.

2002 **Successfully completed Study Abroad program**, *University of Leeds, England.*

2012 **Successfully completed Introduction to Connectomics course**, *Co-offered by MIT and Harvard Universities, Massachusetts, USA.*

Professional Experience

2010 – present **Researcher**, *Neurodata and the Open Connectome Project*, Johns Hopkins University.
Lead and organize a variety of research thrusts in the areas of mesoscale and nanoscale connectomics, community analytics, and computer vision. Responsible for translating raw image volumes to knowledge for subsequent statistical inference. Supported researchers from multiple institutions in achieving their scientific goals. `neurodata.io`

2007 – present **Senior Staff Engineer, Project Manager**, *Research and Exploratory Development Department*, Johns Hopkins University, Applied Physics Laboratory, Laurel, MD.
Task lead for ground truthing and evaluation effort for IARPA MICrONS, a cutting-edge program to store, map, assess, and develop algorithms from a cubic millimeter (i.e., cortical column) of mammalian cortex.

Provide technical and project leadership for several efforts to build and analyze brain graphs (ranging from millimeter to nanometer scale). Developed the first scalable, fully automated pipeline to estimate and assess electron microscopy brain graphs. Contributed system engineering, computer vision, and computational neuroscience expertise to support various scientific endeavors. Projects have included the NIH Synaptomes of Mice and Men Transformative Research Award and MICrONS, which are part of the Presidential BRAIN Initiative.

Previously led team to develop a passive sonar target tracking solution that applied across diverse projects and sponsor communities. Provided machine learning and data analysis expertise for several different target detection applications.

2009 – present **Co-Director and Co-Founder**, *College Prep Program at APL*, Laurel, MD.
Lead all-volunteer annual summer program to support and encourage under-served students who have the desire and academic potential to excel in college, but who lack the mentoring and resources necessary to succeed. 100% of program graduates are on track to earning a 4-year college degree. 500+ hour annual volunteer commitment.

2001 – 2007 **Electrical and System Engineer**, *Northrop Grumman Corporation*, Information Technology (TASC), Chantilly VA, Space Technology, Redondo Beach, CA.
Subject Matter Expert and Responsible Engineer for a group of satellite systems. Regularly interfaced with customers and provided support. Program received 100% Award Fee. Responsible for the design, execution and assessment of several system compatibility tests. (Intern 2001-02)

Skills and Memberships

Skills **Proficient with Python, MATLAB, LaTeX, MS Office, Linux, OSX, and Windows** · **Experience with big data analytics** · **Six Sigma Greenbelt, Engineer Intern (EIT)**.

Memberships **Secret Clearance** · **Society for Neuroscience, Eta Kappa Nu, Tau Beta Pi Honor Societies**.

Awards & Honors

2015 **2015 JHU/APL Author's First Paper Publication Award for Images to Graphs Manuscript**, *JHU Applied Physics Laboratory*.

2014 **Hart Prize Award for Best Research Project**, *JHU Applied Physics Laboratory*.

2014 **Volunteer of the Year Award**, *Howard County, Maryland*.

2010 – 2016 **Full-tuition PhD Fellowship**, *JHU Applied Physics Laboratory*.

2010 **Post-Master's Certificate in Electrical Engineering**, *Johns Hopkins University*.

2009 – 2013 **APL Diversity Awards**, *JHU Applied Physics Laboratory*.

2009 **Diversity Leadership Award**, *Johns Hopkins University*.

2004 – 2005 **Full-tuition Master's Degree Fellowship**, *Northrop Grumman Corporation*.

2003 ***Graduated Magna Cum Laude***, *Vanderbilt University*.

2003 **Program Award for Top EECS Student**, *Vanderbilt University*.

2003 **Dean's Award for Outstanding Service**, *Vanderbilt University*.

1999 – 2003 **Full-tuition Harold Stirling Vanderbilt Scholarship**, *Vanderbilt University*.

Teaching

2015, 2016 **Introduction to Connectomics Intersession Course**, *Johns Hopkins University*, Conceived, designed and taught intensive research-based class, culminating in a student poster session. Class evaluation score of 4.8/5 was one of highest in the Whiting School.

2011 – present **Mentor**, *Johns Hopkins Unversity*, Supported eight high school, undergraduate and graduate students' research projects.

Publications

1 W. Gray Roncal, C. Lea, A. Baruah, and G. D. Hager. SANTIAGO : Spine Association for Neuron Topology Improvement. *arXiv preprint*, 2016.

2 G. Kiar, K. J. Gorgolewski, D. Kleissas, W. Gray Roncal, B. Litt, B. Wandell, R. A. Poldrack, M. Wiener, V. R. Jacob, R. Burns, and J. T. Vogelstein. Science In the Cloud (SIC): A use case in MRI Connectomics. *arXiv preprint*, 2016.

3 E. L. Dyer, W. Gray Roncal, H. L. Fernandes, G. Doga, V. D. Andrade, R. Vescovi, K. Fezzaa, X. Xiao, J. T. Vogelstein, C. Jacobsen, and P. K. Konrad. Quantifying mesoscale neuroanatomy using X-ray micro-tomography. *In Review*, pages 1–28, 2016.

4 W. Gray Roncal, E. L. Dyer, G. Doga, and K. Kording. From sample to knowledge : Towards an integrated approach for neuroscience discovery. *arXiv preprint*, pages 1–8.

5 W. Gray Roncal, D. M. Kleissas, and J. T. Vogelstein. An Automated Images-to-Graphs Framework for High Resolution Connectomics. *Frontiers in Neuroinformatics*, 9:1–10, 2015.

6 W. Gray Roncal, M. Pekala, V. Kaynig-fittkau, D. M. Kleissas, J. T. Vogelstein, H. Pfister, R. Burns,

R. J. Vogelstein, M. A. Chevillet, and G. D. Hager. VESICLE : Volumetric Evaluation of Synaptic Interfaces using Computer vision at Large Scale. *British Machine Vision Conference*, pages 1–9, 2015.

7 N. Kasthuri, K. J. Hayworth, D. R. Berger, R. L. Schalek, J. A. Conchello, S. Knowles-Barley, D. Lee, A. Vázquez-Reina, V. Kaynig, T. R. Jones, M. Roberts, J. L. Morgan, J. C. Tapia, H. S. Seung, W. Gray Roncal, J. T. Vogelstein, R. Burns, D. L. Sussman, C. E. Priebe, H. Pfister, and J. W. Lichtman. Saturated Reconstruction of a Volume of Neocortex. *Cell*, 162(3):648–661, 2015.

8 M. Kazhdan, K. Lillaney, W. Gray Roncal, D. Bock, J. T. Vogelstein, and R. Burns. Gradient-Domain Fusion for Color Correction in Large EM Image Stacks. *arXiv*, 2015.

9 W. Gray Roncal, Z. H. Koterba, D. Mhembere, D. M. Kleissas, J. T. Vogelstein, R. Burns, A. R. Bowles, D. K. Donavos, S. Ryman, R. E. Jung, L. Wu, V. Calhoun, and R. J. Vogelstein. MIGRAINE: MRI Graph Reliability Analysis and Inference for Connectomics. *GlobalSIP*, dec 2013.

10 R. Burns, W. Gray Roncal, D. Kleissas, K. Lillaney, P. Manavalan, E. Perlman, D. R. Berger, D. D. Bock, K. Chung, L. Grosenick, N. Kasthuri, N. C. Weiler, K. Deisseroth, M. Kazhdan, J. Lichtman, R. C. Reid, S. J. Smith, A. S. Szalay, J. T. Vogelstein, and R. J. Vogelstein. The Open Connectome Project Data Cluster: Scalable Analysis and Vision for High-Throughput Neuroscience. *Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM)*, jun 2013.

11 J. T. Vogelstein, W. R. Gray Roncal, R. J. Vogelstein, and C. E. Priebe. Graph classification using signal-subgraphs: applications in statistical connectomics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1539–1551, jul 2013.

12 D. Mhembere, W. Gray Roncal, D. Sussman, C. E. Priebe, R. Jung, S. Ryman, R. J. Vogelstein, J. T. Vogelstein, and R. Burns. Computing Scalable Multivariate Glocal Invariants of Large (Brain-) Graphs. *GlobalSIP*, dec 2013.

13 W. R. Gray Roncal, J. A. Bogovic, J. T. Vogelstein, B. A. Landman, J. L. Prince, and R. J. Vogelstein. Magnetic resonance connectome automated pipeline: an overview. *IEEE Pulse*, 3(2):42–48, mar 2012.

Posters

1 J. Matelsky, S. Berg, A. Eusman, K. Lillaney, J. T. Vogelstein, G. D. Hager, and W. Gray Roncal. ndio: Neuroscience Discovery Input and Output. In *Society for Neuroscience Abstract*, San Diego, 2016.

2 W. Gray Roncal, M. Pekala, D. M. Kleissas, J. T. Vogelstein, and G. D. Hager. ndparse: Tools and Interfaces for Scalable Neuroscience Discovery. In *Society for Neuroscience Abstract*, 2016.

3 W. Gray Roncal, A. Simhal, J. Vogelstein, F. Collman, E. L. Dyer, M. Chevillet, R. Burns, G. Sapiro, and G. Hager. Scalable automated (synapse) detection using the Open Connectome Project. In *Society for Neuroscience Abstract*, 2015.

4 A. Baden, K. Lillaney, W. Gray Roncal, J. Vogelstein, and R. Burns. Web Visualization of Massive Neuroscience Datasets using the Open Connectome Project. In *Society for Neuroscience Abstract*, 2015.

5 A. Simhal, W. Gray Roncal, A. Baden, K. Lillaney, K. Kutten, M. Miller, J. Vogelstein, R. Burns, L. Ye, R. Tomer, K. Deisseroth, and G. Sapiro. Computational statistics for whole brain CLARITY analysis using the Open Connectome Project. In *Society for Neuroscience Abstract*, 2015.

6 J. Vogelstein, S. Smith, W. Gray Roncal, R. Vogelstein, R. Burns, K. Lillaney, A. Baden, G. Kiar, and P. Manavalan. Open Connectome Project and NeuroData: Enabling Data-Driven Neuroscience at Scale. In *Society for Neuroscience Abstract*, 2015.

7 G. Kiar, W. Gray Roncal, D. Mhembere, E. Bridgeford, D. Clark, R. Millham, Michael Craddock, Cameron Burns, and J. T. Vogelstein. Community Connectomics via Cloud Computing Utilizing m2g - a Reference Pipeline. In *Organization for Human Brain Mapping*, 2015.

8 W. Gray Roncal, O. M. Akmal, M. Encarnacion, T. Latchman, A. Baruah, J. T. Vogelstein, D. Dementhon, N. Kasthuri, R. Burns, C. E. Priebe, and G. D. Hager. A Semantic Framework to Guide Computer Vision in (EM) Connectomics. In *Society for Neuroscience*, 2014.

9 A. Sinha, W. Gray Roncal, N. Kasthuri, J. W. Lichtman, and R. Burns. Automatic Annotation of 3D Axoplasmic Reticula for Neuron Segmentation. In *Resting State Brain Connectivity*, page 4(9): A26, 2014.

10 D. M. Kleissas, W. Gray Roncal, P. Manavalan, K. Lillaney, A. Sinha, J. T. Vogelstein, G. D. Hager, R. Burns, M. A. Chevillet, and R. J. Vogelstein. Automated Neuronal Graph Creation Using the Open Connectome Project RAMON Data Open Connectome Project (OCP) Services. *Society for Neuroscience*, page 2, 2014.

11 W. R. Gray Roncal and Others. Towards a Fully Automatic Pipeline for Connectome Estimation from High-Resolution EM Data. In *OHBM*, 2013.

12 D. M. Kleissas, W. Gray Roncal, P. Manavalan, J. T. Vogelstein, D. D. Bock, R. Burns, R. J. Vogelstein, H. Moreno, M. Perez, and W. Reyes. Large-Scale Synapse Detection Using CAJAL3D. *Neuroinformatics*, 2013.

13 W. R. Gray Roncal and Others. Towards a Fully Automatic Pipeline for Connectome Estimation from High-Resolution EM Data. In *Cold Spring Harbor Laboratory, Neuronal Circuits*, 2012.

14 J. T. Vogelstein, W. R. Gray Roncal, R. J. Vogelstein, J. Bogovic, S. Resnick, J. Prince, and C. E. Priebe. Connectome Classification: Statistical Graph Theoretic Methods for Analysis of MR-Connectome Data. In *Organization for Human Brain Mapping*, 2011.

15 J. T. Vogelstein, W. Gray Roncal, J. G. Martin, G. C. Coppersmith, M. Dredze, J. Bogovic, J. L. Prince, S. M. Resnick, C. E. Priebe, and R. J. Vogelstein. Connectome Classification using statistical graph theory and machine learning. In *Society for Neuroscience*, 2011.

16 W. R. Roncal, Gray, J. A. Bogovic, J. T. Vogelstein, C. Ye, B. A. Landman, J. L. Prince, and R. J. Vogelstein. Magnetic resonance connectome automated pipeline and repeatability analysis. In *Society for Neuroscience*, 2011.

17 J. T. Vogelstein, J. Bogovic, A. Carass, W. R. Gray Roncal, J. L. Prince, B. Landman, D. Pham, L. Ferrucci, S. M. Resnick, C. E. Priebe, and R. J. Vogelstein. Graph-Theoretical Methods for Statistical Inference on MR Connectome Data. In *Organization for Human Brain Mapping*, 2010.

18 W. R. Gray, J. T. Vogelstein, J. Bogovic, A. Carass, J. L. Prince, B. Landman, D. Pham, L. Ferrucci, S. M. Resnick, C. E. Priebe, and R. J. Vogelstein. Graph-Theoretical Methods for Statistical Inference on MR Connectome Data. In *DARPA Neural Engineering, Science and Technology Forum*, 2010.

Invited Talks

1 Volumetric Evaluation of Synaptic Interfaces using Computer Vision at Large Scale. Hopkins Imaging Initiative, 2015.

2 BigNeuro 2015: Making sense of big neural data workshop. NIPS, 2015.

3 *Images to Graphs for Inference*, European Research Data Alliance, 2015.

4 *Images to Graphs: Techniques and Strategies for Mapping the Brain*, Oxford University, 2015.

5 *Machine Intelligence from Cortical Networks (MICRONS) IRAD*, JHUAPL Hart Prize Colloquium, 2015.

6 *Graph Inference for Connectomics*, JHUAPL Graphs Seminar Series.

7 *Predicting Human Performance from Brain Connectivity*, JHU Lattman Lecture Series, 2011.

8 *Connectome Annotation for Joint Analysis of Large 3-Dimensional Data: Research Progress*, Harvard University, 2013.

9 *Towards A Fully Automatic Pipeline for Connectome Estimation from High-Resolution EM Data*, Janelia Farm Turning Images to Knowledge Conference.

Vita

William Gray Roncal is a Project Manager in the Research and Exploratory Development Department at the Johns Hopkins University Applied Physics Laboratory (APL). In 2005, Will received a Master of Electrical Engineering from the University of Southern California. He earned his Bachelor of Electrical Engineering Degree from Vanderbilt University in 2003. He is a member of the Society for Neuroscience, Eta Kappa Nu, and Tau Beta Pi.

Will applies computer vision algorithms to solve big data challenges at the intersection of multiple disciplines. Although he has experience in diverse environments ranging from undersea to outer space, he currently works in connectomics, an emerging discipline within neuroscience that seeks to create a high-resolution map of the brain.

Will co-founded The College Prep Program at APL, a free, volunteer-led program which has supported and encouraged over 150 underserved students over the past 8 years who have the desire and academic potential to excel in college, but who lack the mentoring and resources necessary to succeed.