CARNEGIE MELLON UNIVERSITY

DIETRICH COLLEGE OF HUMANITIES AND SOCIAL SCIENCES
DISSERTATION

Submitted in Partial Fulfillment of the Requirements
For the Degree of DOCTOR OF PHILOSOPHY

Title: "Understanding the form and function of neuronal physiological diversity"

Presented by: Shreejoy J. Tripathy

Accepted by: The Center for the Neural Basis of Cognition
October 31, 2013

Thesis Committee:
Nathan Urban, Chair
Rob Kass
William Cohen
Aryn Gittis
Etienne Sibille
Anne-Marie Oswald
Gordon Shepherd, Yale University

Understanding the form and function of neuronal physiological diversity

Shreejoy J. Tripathy

September 18

Center for the Neural Basis of Cognition
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Nathan Urban, Adviser
Rob Kass
William Cohen
Aryn Gittis
Etienne Sibille
Anne-Marie Oswald
Gordon Shepherd, Yale University

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Shreejoy Tripathy is supported by a National Science Foundation graduate research fellowship and a Presidential Fellowship in the Life Sciences from the Richard King Mellon Foundation.

Keywords: neuron diversity, neuron coding, stimulus decoding, olfactory bulb, neurophysiology, text mining

Dedicated to the memory of Aaron Swartz,
shine on you crazy diamond

Abstract

For decades electrophysiologists have recorded and characterized the biophysical properties of a rich diversity of neuron types. This diversity of neuron types is critical for generating functionally important patterns of brain activity and implementing neural computations. In this thesis, I developed computational methods towards quantifying neuron diversity and applied these methods for understanding the functional implications of *within-type* neuron variability and *across-type* neuron diversity.

First, I developed a means for defining the functional role of differences among neurons of the same type. Namely, I adapted statistical neuron models, termed generalized linear models, to precisely capture how the membranes of individual olfactory bulb mitral cells transform afferent stimuli to spiking responses. I then used computational simulations to construct virtual populations of biophysically variable mitral cells to study the functional implications of within-type neuron variability. I demonstrate that an intermediate amount of intrinsic variability enhances coding of noisy afferent stimuli by groups of biophysically variable mitral cells. These results suggest that within-type neuron variability, long considered to be a disadvantageous consequence of biological imprecision, may serve a functional role in the brain.

Second, I developed a methodology for quantifying the rich electrophysiological diversity across the majority of the neuron types throughout the mammalian brain. Using semi-automated text-mining, I built a database, NeuroElectro, of neuron type specific biophysical properties extracted from the primary research literature. This data is available at `http://neuroelectro.org`, which provides a publicly accessible interface where this information can be viewed. Though the extracted physiological data is highly variable across studies, I demonstrate that knowledge of article-specific experimental conditions can significantly explain the observed variance. By applying simple analyses to the dataset, I find that there exist 5-7 major neuron super-classes which segregate on the basis of known functional roles. Moreover, by integrating the NeuroElectro dataset with brain-wide gene expression data from the Allen Brain Atlas, I show that biophysically-based neuron classes correlate highly with patterns of gene expression among voltage gated ion channels and neurotransmitters. Furthermore, this work lays the conceptual and methodological foundations for substantially enhanced data sharing in neurophysiological investigations in the future.

Acknowledgments

I have been truly blessed by the range and depth of support I have received throughout graduate school. My experiences in Pittsburgh as a graduate student have been among most formative times of my life and I'm very grateful for every moment.

First and foremost, a huge thank you to my adviser and fearless leader, Nathan Urban. Thank you for accepting me into your lab. Thanks for inspiring me to only work on problems that are worth solving. And an extra thanks for giving me the freedom to work on the things which I thought were the most interesting, even though they may have made little sense as the bud of an idea in my head. One of my biggest fears about being an adviser one day is having to live up to the example that you've set for me. Thanks for running your lab in a way that makes each day an intellectual adventure and for hiring smart creative people who love discussing ideas almost as much as actually doing the science.

Thanks to Krishnan Padmanabhan. You were among the first people in the lab to whom I turned in times of confusion. Though I regularly give you a lot of scientific grief, you taught me that the best thing that I can do as a scientist is to work on the things that I am the most passionate about. You also taught me that there's nothing wrong with putting forth absolutely ridiculous ideas, because occasionally some of those ideas will strike gold. As a computational person, I was lucky to work with such an open-minded experimentalist as yourself. Your willingness to share your hard-earned data will all who ask is truly an example to all experimentalists.

My journey into neuroinformatics would never have begun without Rick Gerkin. You have been a second adviser to me, and have probably done more to shape my research interests than anyone else. Thanks for encouraging me to go to that first neuroinformatics conference a couple years ago which set my current research path in motion. I've really enjoyed our conversations over the years, especially the ones over lunch, where we talked about how the best things we could do as programming-savvy scientists is to build tools to make other scientists more productive. I hope we continue collaborating long into the future.

Thanks to the greater Urban Lab. From my early days and the great bunch of people who trained me, including: Anne Marie Oswald, Jason Castro, Ken Hovis, Sonya Giridhar, Elie Kanal, Daniel Jimenez; to the current lab: Shawn Burton, Matt Geramita, Santosh Chandrasekaran, Pete Jones, Jing Wen, Annie Liu and Yiyi Yu. I've loved our conversations each day. And a special thanks to Shawn and Matt for helping do experiments for me to help validate NeuroElectro data. One of my favorite things is to walk by the main lab when I'm bored or stuck and a problem and chat about anything with anyone while slowly sipping a soda. And thanks to everyone for accepting someone as computational as me into the lab. I hope that the special spark and love for data that I brought continues in lab meetings into the future.

I'm very grateful to my wonderful thesis committee: Anne Marie Oswald, Aryn Gittis, Rob Kass, William Cohen, Etienne Sibille and Gordon Shepherd. I appreciate the comments and perspective that each of you brought to our discussions. I am especially grateful to Gordon, who saw my work on NeuroElectro at a very early stage and who convinced me that it was worth continuing to pursue.

Thanks to the Center for the Neural Basis of Cognition and the greater Pittsburgh neuroscience community. I'm so lucky to have landed in an academic community with the enormous breadth and depth of the CNBC. When I think about the kind of department I would love to work in some day, one of my requirements is that it be something like the CNBC where psychologists and statisticians and biologists are all conversant and jovial with one another. I see the growing interdisciplinary nature of science and in neuroscience as a feature which will only increase. I'm also very grateful to the Program in Neural Computation. Looking back, the PNC was perfect for someone like me who wanted to blaze their own unique academic path.

I've gotten to know and enjoy spending time with such a great group of friends. Thanks for being with me in good times and bad. Thanks for listening to my "next greatest idea" or my baseless claims of my grandiosity. Events which especially stand out for me include: shooting off fireworks in the park, rolling down the ski slopes at 7 springs, and countless nights in Squirrel Hill at the CAGE. My time in Pittsburgh wouldn't have been anywhere near as joyful and memorable without you all.

Lastly, thanks to my wonderful, infinitely loving family. To my Mom, I thank for teaching me the value of hard work and to never settling for second best. Thanks for teaching me math and writing at the same time that I was learning to talk. To my Dad, thank you for showing me the wonder of science and for letting me perform my crazy experiments even if they broke the computer or stopped up the sink. Thanks for buying me a subscription to Scientific American and for going to all of my tennis matches and hockey games. Thanks to both my parents for impressing on me (and continuing to impress on me) the value of being a good human being. To my brother Shreep, thanks for being a great big brother. Thanks especially for going to medical school, which got Mom and Dad off my back about being a doctor which let me follow my passion for the brain and go to graduate school.

Contents

Chapter 1

Introduction

Neuronal biophysics describes the complex process by which a neuron's membrane transforms synaptic and electrical inputs to subthreshold and spiking outputs. Through studying this input-output transformation, neurophysiologists can relate how the intrinsic biophysical properties of individual neurons shape the specific computations that each neuron performs on its inputs. Thus understanding intrinsic biophysics provides insights into each neuron's computational role within its larger neural circuit as well the neuron's potential role in producing organism-level behaviors (Koch, 1999; Izhikevich, 2010). Because of the explicit link between neuronal membrane properties and circuit function, studies of neuronal biophysics have led to tremendous insights into the detailed, mechanistic processes underlying certain neurological disorders such as epilepsy and channelopathies (Zuberi et al., 1999; Rajakulendran et al., 2007; Depienne et al., 2009).

Rather than each neuron being functionally identical, there are many different types of neurons. Furthermore, there are a number of different electrophysiological phenotypes that neurons can express which are associated with unique computational roles in the brain. For example, cortical basket neurons display a "fast-spiking" phenotype, and are implicated in mechanisms of gain control and facilitating synchrony among cortical pyramidal neurons (Moore et al., 2010). A neuron's biophysical properties arise through expression of combi-

nations of ion channels that collectively define the neuron's electrophysiological phenotype (Koch, 1999; Llinás, 1988). This functional diversity among neurons is analogous to antibody generation in the immune system, and has been hypothesized to contribute to the functional robustness of brains (Singer et al., 2010). Intrinsic biophysical diversity among neurons goes hand-in-hand with other kinds of diversity, including morphological (Parekh and Ascoli, 2013), molecular (Sugino et al., 2006), and neurotransmitter diversity (Nelson et al., 2006).

Given this rich electrophysiological diversity among neurons, a long-standing challenge is defining how exactly to choose and define the parameters for use in determining how neurons differ (Ascoli et al., 2008; Hamilton et al., 2012). For example, when recording from neurons *in vitro*, it is common to inject depolarizing current into the neuron's cell membrane while recording the neuron's membrane voltage (Connors et al., 1982). A common question to then ask is: when the neuron fires action potentials, do they occur regularly spaced in time or do they instead come in groups of bursts (Ascoli et al., 2008; Markram et al., 2004)? The neurophysiologist will then use the differences in spike patterns as a defining characteristic for use in distinguishing between neuron types.

Ideally, these characteristics that neurophysiologists use to define neurons will correspond to differences that are functionally relevant to neuronal computation. However, there are not formal community-based standards in place to decide when observations of neurons should be "split" (into neuron subtypes) or "lumped" (into a single neuron super-class) (Ascoli et al., 2008; Hamilton et al., 2012). Furthermore, a challenge with interpreting the data from electrophysiological recordings is that they are notoriously sensitive to experimental conditions, making it difficult to directly compare results across studies. Throughout this thesis, I address the following 3 questions in order to better describe and define neuronal electrophysiological diversity:

1. How do neurons transform their inputs to outputs? How should this transformation

be described?

2. How should electrophysiological differences among neurons be quantified?
3. How should researchers communicate results on the electrophysiological findings from neurons?

Furthermore, I develop these methodologies to answer the following scientific questions on the form and function of neuronal electrophysiological diversity:

1. What is the computational role of electrophysiological variability within a neuron type? Is there an optimal level of within-type neuron variability?
2. How electrophysiologically diverse are neuron types throughout the brain? Are there unknown electrophysiological similarities among neuron types previously thought to be functionally distinct?

In the remainder of this chapter, I present background and review material that is most relevant to the work in this thesis.

1.1 Background and Related Work

1.1.1 Intracellular electrophysiology and the origins of intrinsic biophysical properties

Neuronal intrinsic biophysical properties are typically obtained using electrophysiological methods to record membrane potentials from neurons contained in acute brain slices (Schwartzkroin, 1975; Connors et al., 1982; Stuart et al., 1993). Using glass electrodes with fine tips, neurophysiologists can both measure the transmembrane voltage and manipulate this voltage by injecting current (Fig. 1.1).

Rather than the neuron simply acting as a passive electrical element to an injected current (i.e. a resistor and capacitor in parallel) and just integrating the current, neurons

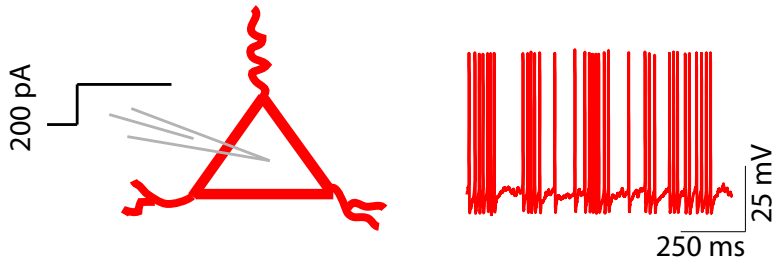

Figure 1.1: *Schematic and example data for intracellular electrophysiology.* Left: Procedure of recording from a neuron with a glass electrode and injecting current into the neuron's cell body. Right: Example voltage trace showing trains of action potentials.

usually respond to input currents with complex dynamics (Llinás, 1988). Hundreds of types of ion channels are expressed by neurons (Harmar et al., 2009). Individual ion channels are made of one or more proteins, each of which is encoded in that cell's genome. The opening and closing (gating) of these ion channels is regulated by many factors, including the transmembrane voltage (Hille, 2001; Ranjan et al., 2011). Thus the conductance of ion channels is dependent on the current state and past history of the neuron's activity. The critical role of ion channels in regulating neuronal activity was first proposed by Hodgkin and Huxley in their groundbreaking work studying the action potential of the squid giant axon (Hodgkin and Huxley, 1952) and later confirmed by Erwin Neher and Bert Sakmann while studying the gating of single ion channels in a patch of cell membrane (Neher and Sakmann, 1976).

A neuron's transfer function, which describes the relationship between a neuron's inputs and its outputs (e.g., given by the neuron's frequency-current relationship), is itself a function of the kinds of ion channels the neuron expresses in its membrane. For example, some neurons, like entorhinal cortex cells (Giocomo et al., 2007) or inferior olive neurons (Lampl and Yarom, 1997), display subthreshold resonances, allowing the neuron to selectively amplify and transmit inputs delivered at a specific frequencies. Neuronal transfer functions define the computation that the neuron's membrane performs on its inputs, and contributes to what makes different types of neurons unique in their response

properties. Because ion channels influence single neuron computation, channelopathies and other disorders that affect the efficacy of a single channel are often associated with behavioral deficits (e.g. Dravet's syndrome (Depienne et al., 2009)). Thus a large part of the study of cellular neurophysiology is to develop a better understanding of the relationship between electrophysiological phenotypes and expression of specific kinds of ion channels and corresponding currents.

One complication in this analysis arises because of the mismatch between the description of voltage-gated currents and voltage gated channels (Crasto et al., 2007; Harmar et al., 2009). Neurophysiologists initially reported the properties of neurons in terms of the currents that they expressed. For example, Hodgkin and Huxley described the behavior of the squid giant axon in terms of the fast sodium current and the delayer rectified potassium current (Hodgkin and Huxley, 1952). Later, molecular and genetic analyses have described the behavior of neurons in terms of the channel proteins (or even the channel genes) that these neurons express (Coetzee et al., 1999; Toledo-Rodriguez et al., 2004; Harmar et al., 2009; Marder and Taylor, 2011). Unfortunately the relationship between currents and channels and genes is not always simple. For example, the sodium currents in a given neuron may be mediated by a number of types of sodium channels encoded by one or more sodium channel genes (Momin and Wood, 2008). Similarly, what physiologists have called A-type potassium current likely is mediated by mixtures of channel subunit proteins encoded by genes of the Kv1.x and Kv4.x families (Carrasquillo and Nerbonne, 2013).

1.1.2 Describing neuronal electrophysiological phenotypes

Neurons can be described and subdivided based on many different characteristics, including morphology (Parekh and Ascoli, 2013), gene expression (Lein et al., 2007) and physiology (Migliore and Shepherd, 2005). In studying the electrophysiology of neurons,

neurophysiologists will typically measure a number of electrophysiological characteristics from each neuron for use in describing the neuron's electrophysiological phenotype (Woody and Gruen, 1978; Connors et al., 1982; Toledo-Rodriguez et al., 2004; Bean, 2007). These include measurement of passive electrical properties like the neuron's resting membrane potential, the voltage of the neuron that it "rests" at upon no stimulation, and its input resistance, which reflects the neuron's membrane resistivity to a current injection. Active properties that are typically measured include characteristics of the neuron's action potentials, including the action potential threshold, width, and height (Fig. 1.2). Additionally, neurophysiologists will measure regular patterns in which the neuron produces trains of action potentials. For example, does the neuron tend to fire multiple spikes as a burst or are action potentials fired at regular intervals in time (Llinás, 1988; Migliore and Shepherd, 2002; Markram et al., 2004)? Does the neuron fire spikes spontaneously in the absence of an external driving stimulus? To quantify these, neurophysiologists will compute statistics like the neuron's frequency-current (FI) curve, which relates the amount of positive current injected into the neuron to the number of evoked action potentials, or the coefficient of variation of the neuron's interspike interval distribution, which provides a metric to use in quantifying burst versus regular spiking.

A common approach to use in partitioning neurons into electrophysiologically-based classes is to measure a number of electrophysiological characteristics (such as resting membrane potentials and input resistances) across a set of neurons and then use a clustering analysis to partition the recorded neurons into different subsets (Markram et al., 2004; Migliore and Shepherd, 2005; Antal et al., 2006; Padmanabhan and Urban, 2010; Druckmann et al., 2012). While a strength of this approach is that it reflects the observed data and does not rely on a single characteristic such as whether the neuron bursts or not, there is no guarantee that the electrophysiological characteristics used to separate the neurons reflect functionally relevant features of the neurons. An extreme example of this is that

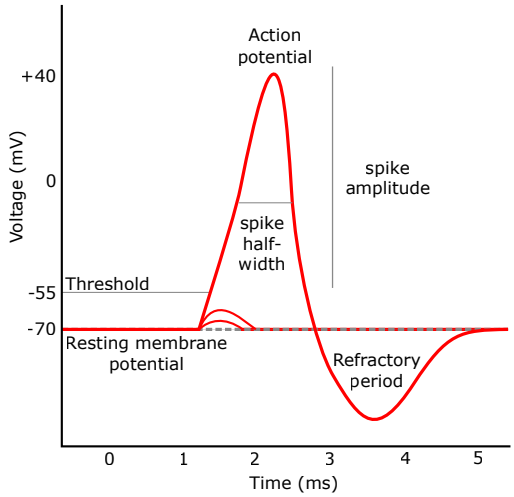

Figure 1.2: *Example electrophysiological characteristics that are computed from a neuron's action potential.* Image modified from Wikipedia user Synaptidude.

though neuron bursting is often a useful characteristic for partitioning neurons (Markram et al., 2004; Ascoli et al., 2008), a deep understanding of the computational or functional role of burst firing versus regular firing has generally been lacking (though see (Lisman, 1997; Oswald et al., 2004; Marsat and Pollack, 2006)).

1.1.3 Defining neurons using parameters of a computational model

A different approach towards defining neurons along electrophysiological dimensions is to first construct a computational model that captures certain features of the recorded neuron's activity. Then, to compare neurons, one can simply assess differences in model parameters between the neurons.

Hodgkin Huxley modeling approach

The most prolific example of this approach in cellular neuroscience is the usage of Hodgkin-Huxley-type neuron models to capture or recapitulate recorded neuron electrophysiological activity (Hodgkin and Huxley, 1952; Rall and Shepherd, 1968; De Schutter and Bower, 1994; Bhalla and Bower, 1993; Hines and Carnevale, 1997; Koch, 1999; Ermentrout, 2002;

Migliore et al., 2003; Prinz et al., 2004; Galán et al., 2006; Brette et al., 2007; Marder and Taylor, 2011). The Hodgkin-Huxley modeling approach treats neurons as the sum of passive electrical components (e.g. the neuron's membrane is treated as a capacitor) and active components, which are modeled via different ion channel conductances which each have time and voltage dependent kinetics (Fig. 1.3A, (Koch, 1999; Hille, 2001; Druckmann et al., 2011)). Often this approach involves the specification of ion channel properties in multiple or hundreds of individual electrical compartments, corresponding to segments of axon or dendrite (e.g., (Rall and Shepherd, 1968; Bhalla and Bower, 1993; De Schutter and Bower, 1994)). The goal of the modeling approach is to find a set of ion channel conductances that effectively or adequately describes the recorded neuron's behavior. This assessment of whether the neuron model adequately captures the real neuron is often done in a qualitative way in these kinds of models, where the computational modeler will ask if the model captures some particular phenotypic behavior of the recorded neuron, like a precise mechanism underlying burst firing (Bhalla and Bower, 1993; Davison et al., 2003).

One reason for the wide adoption of this approach by the community is its biological plausibility, because the modeling approach directly reflects known features of neurons and ion channels and allows for the predicted values of ion channel parameters and densities to be explicitly tested. However, given the inherent difficulty in obtaining best-fit solutions to systems of non-linear dynamical equations (Strogatz, 2000), finding best sets of model parameters (i.e. the values of specific ion channel conductances) that best match recorded neuron data can be difficult. Furthermore, it has been shown that multiple, disparate sets of model parameters (i.e different sets of ion channel conductances) can lead to the same or highly similar neuron and neuron network electrophysiological phenotypes (Prinz et al., 2004). Therefore two neuron models, which may differ greatly in their sets of ion channel conductances, may yet still possess identical or similar electrophysiological phenotypes. Thus even if the model perfectly fits the data, inferences to the ion channel properties are

difficult to make conclusively (Marder and Taylor, 2011).

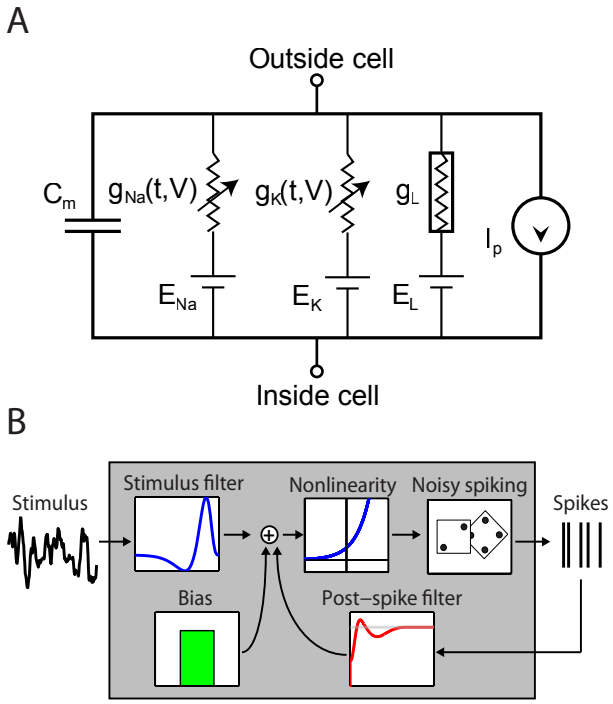

Figure 1.3: *Example single neuron models.* A) Schematic showing equivalent circuit form of Hodgkin-Huxley model. Features of neuron and its membrane are modeled explicitly. For example, ion channel conductances are modeled by time and voltage varying resistors. B) Structure of the basic single neuron generalized linear model form. Phenomenological features of the neuron, like its responses to a stimulus or its average response following an action potential are modeled directly.

Statistical/Phenomenological modeling approach

Given these difficulties with the Hodgkin-Huxley modeling methodology, an alternative is to take a statistical, phenomenological approach towards neuron modeling (Kass and Ventura, 2001; Paninski, 2004; Badel et al., 2008). Rather than modeling neurons using models with parameters which can be mapped onto known physical neuron features like the expression and kinetics of different ion channels, this approach directly models specific, phenomenological features of the neuron's activity (Fig. 1.3B). For example, given that neurons go into a refractory period following an action potential, this refractory period duration is often modeled directly in statistical neuron modeling frameworks; this is in contrast to the Hodgkin Huxley approach where the refractory period is not modeled

directly and only results as an emergent property of the interaction of multiple ion channel conductances. An advantage of this approach is that since these models are based in statistical modeling, the models are typically constructed such that there is a single, unique "best-fit" solution of model parameters which best capture a given piece of neuron recording data (Paninski, 2004). This uniqueness of model parameters given a set of recorded data is an advantage when using this modeling approach as a first step for characterizing neurons and then using the fit model parameters for the purpose of comparing different neurons on the basis of model parameters (Mensi et al., 2012).

1.1.4 Criteria for partitioning neurons into neuron types

Neurons are defined according to a number of criteria, including electrophysiological characteristics, which are the focus of this thesis. The most common criteria for characterizing neurons is by determining the location of the cell body both in the brain and within a specific cell layer. This approach was used extensively by Santiago Ramón y Cajal, the father of modern neuroscience, who extensively studied the cellular architecture of many parts of the brain through investigating stained slices of brain tissue under a microscope and drawing his observations of the neural cytoarchitecture in fine detail (Ramón y Cajal, 1995). He used neuron location as well as neuron shape to define the major neuron types throughout the brain; his neuron type-ology remains the basis for the common neuron types in use today. Moreover, modern tools allow for tracing the detailed morphology of neurons and digitizing and publicly sharing morphological traces (Parekh and Ascoli, 2013) as well as clustering neurons based on morphological similarity (Ascoli et al., 2008).

As technology has progressed, neuron types are now defined using additional criteria. For example, the development of techniques to record from neuron cell bodies using sharp and patch-clamp electrodes has allowed neurophysiologists to record electrical activity from neurons and use electrophysiological criteria like burstiness and fast-spiking phenotypes to

help define neuron classes (Woody and Gruen, 1978; Connors et al., 1982; Llinás, 1988; Migliore and Shepherd, 2002; Markram et al., 2004). Additional defining metrics include assessing whether individual neurons express particular marker proteins, such as specific calcium binding proteins like parvalbumin or calciretinin or neuropeptides like somatostatin or cholecystokinin (Kawaguchi and Kubota, 1997; Heintz, 2004; Sugino et al., 2006; Ascoli et al., 2008; Huang and Zeng, 2013).

A challenge with the current practice for defining neurons into distinct classes and types is that these classes are usually not unambiguous and non-overlapping. For example, virtually all neocortical basket cells display a "fast-spiking" electrophysiological phenotype, have a "basket-shaped" cell body, and also express the marker protein parvalbumin (Connors and Gutnick, 1990; DeFelipe, 1997; Markram et al., 2004; Moore et al., 2010). However, neocortical basket cells are an exception because it is typically rare for other defined cell types to also display clear relationships across multiple classification criteria (Wang et al., 2004; Ma et al., 2006; Ascoli et al., 2008). For example, mitral cells of the olfactory bulb, defined by Ramón y Cajal because their cell bodies are located in a clearly defined cell layer in the olfactory bulb and have a cell body shaped like a bishop's mitre, display significant variability among their electrophysiological properties which vary from regular firing to burst firing (Padmanabhan and Urban, 2010; Angelo et al., 2012). Given these findings, should neurophysiologists choose to separate this cell class into multiple sub-classes on the basis of electrophysiological evidence? Or treat these neurons as a single neuron type on the basis of morphological criteria alone?

In light of these challenges for obtaining comprehensive consensus-based definitions for each neuron type, there have been a number of working groups formed with the goal of defining the comprehensive list of neuron types in a given brain region and throughout the brain. Among the best known examples of such a group is the Petilla group for defining neocortical interneurons (Ascoli et al., 2008); the Neuron Registry Taskforce of the In-

ternational Neuroinformatics Coordinating Facility is an another such example (Hamilton et al., 2012; Larson and Martone, 2013). However, the success of the Petilla group has been mixed. On one hand, the members of the Petilla group can agree upon the basic classes formed using morphological, electrophysiological, or molecular criteria. For example, when recording from a neuron, based on its spiking responses to a series of current steps, the neuron can be placed into one of approximately ten electrophysiologically defined phenotypic classes (Ascoli et al., 2008). Contrastingly, putting these multiple criteria together into a single prescriptive guideline for when neuron observations should be "split" or "lumped" has remained challenging, however. Moreover, due to the inherent diversity and complexity of neurons, it is unclear whether a single "neuron type-ology" that bridges multiple definitional criteria will ever emerge for the brain, like it is beginning to crystallize for neuron types in the vertebrate retina (Field and Chichilnisky, 2007; Siegert et al., 2012; Helmstaedter et al., 2013). As a practical consequence of the inherent challenges in classifying and naming neurons, communicating results and findings on specific neuron types is made more difficult as scientists need to reconcile multiple neuron naming schemes.

1.1.5 Databases in cellular and systems neuroscience

There are a number of databases that contain structured information specific to neurons and their properties. For example, information on the detailed shapes of neurons (i.e., their morphology) is being compiled by NeuroMorpho (Parekh and Ascoli, 2013) which contains user-submitted neuron morphological reconstructions made using the NeuroLucida format (Glaser and Glaser, 1990). The SenseLab database, ModelDB, ((Migliore et al., 2003),`http://senselab.med.yale.edu/modeldb/default.asp`) compiles user-submitted computational models developed for simulating the electrophysiological and neurochemical properties of single neurons and networks of neurons (e.g., multi-compartment Hodgkin-Huxley type conductance based models). Similarly, other SenseLab databases, includ-

ing NeuronDB and CellPropDB ((Crasto et al., 2007), http://senselab.med.yale.edu/NeuronDB/), contain information on the ionic currents and neurotransmitters expressed by each neuron and how these are distributed with respect to neuronal morphology. Detailed information on ion channel subtypes, including voltage and temporal dynamics, genetic homology, and corresponding literature references is being compiled by Channelpedia ((Ranjan et al., 2011)), a subproject within the Blue Brain Project (Markram, 2006). A notable missing resource on neuron properties is a database that compiles information on neuron type specific electrophysiological phenotypes and basic biophysical properties.

The Allen Institute for Brain Sciences provides brain-wide gene expression atlases, where the expression of each of the genes in the mammalian genome has been systematically quantified throughout the brain at the resolution of brain regions and cell layers for a number of model organisms and across stages of neural development ((Lein et al., 2007), http://brain-map.org). Similarly, the Allen Institute also provides information on the anatomical connectivity of different brain regions. Parallel to this effort is the Brain Architecture Management System (BAMS, (Bota et al., 2005), http://brancusi1.usc.edu) in which neural connectivity information has been manually curated by domain experts from the existing research literature. The WhiteText Project takes a complementary approach to BAMS and uses biomedical Natural Language Processing (bioNLP) to "text-mine" statements on brain region connectivity from literature abstracts (French et al., 2009, 2012). Another resource which uses text-mining is the NeuroSynth Project which mines fMRI-based brain activation maps from published x, y, z coordinate data tables from neuroimaging publications (Yarkoni et al., 2011). An advantage of automated approaches for content extraction is their scale since they can be applied to arbitrary numbers of publications; however, they typically extract information with less accuracy than human experts.

In addition to these neuroscience subdomain-specific databases are meta-databases that provide linking facilities for cross-resource data integration. For example, NeuroLex ((Lar-

son and Martone, 2013),http://neurolex.org), provides a platform for community annotation of neuron types on the basis of morphological, neurochemical, or electrophysiological properties. Similarly, OpenSourceBrain (http://www.opensourcebrain.org/) is a community platform for collaborative development of computational neuron and network models that utilizes open standards such as NeuroML (Gleeson et al., 2010) to facilitate interoperability of models developed by different researchers. Given this wealth of neuroscience resources, the Neuroscience Information Framework (NIF, (Gardner et al., 2008),http://www.neuinfo.org), provides tools for semantically searching across these diverse databases through the development and incorporation of neuroscience domain-specific ontologies (Bug et al., 2008; Larson and Martone, 2009; Hamilton et al., 2012; Imam et al., 2012). For example, in NIF, the search query "mitral cell" returns a number of database records including relevant research literature from PubMed, computational models from ModelDB, and connectivity information from BAMS.

A challenge with neuroscience databases, especially those that are populated through user submitted content, is ensuring that these resources are well known in the larger community and that investigators voluntarily contribute content. This is in contrast to other fields such as genetics or molecular biology, where uploading data to a publicly accessible database such as GenBank (Benson et al., 2013) or the protein data bank (PDB,(Bernstein et al., 1977)) is viewed as necessary and required for publication and continued funding.

In Chapter 3 of this thesis, I develop a database called NeuroElectro in which I use semi-automated approaches to mine measurements on neuronal biophysics from the existing neurophysiology literature. Namely, with my semi-automated extraction procedures I make use of both automated text mining methods (as used by WhiteText and NeuroSynth (Yarkoni et al., 2011)) as well as expert curation (like BAMS (Bota et al., 2005)) which combine the benefits of scale provided by automated approaches with the accuracy provided by manual approaches.

1.1.6 Investigator-to-investigator variability in neurophysiology

A challenge with electrophysiological data is that it is notoriously difficult to directly compare results across investigators and laboratories. Obtained results are highly dependent upon the exact conditions in which each experiment was done. For example, neurophysiologists have observed that the animal species, strain, and age all influence the values of some of the most commonly measured parameters (e.g., (Zhu, 2000; Spruston and Johnston, 1992)). Furthermore, neurophysiologists have anecdotally reported that subtle investigator-specific preparation details that would generally be very difficult to fully specify in an article's methods section, such as how neurons are selected for electrophysiological recording, also influence collected electrophysiological measurements. Therefore, directly pooling and comparing results across investigators, even within the same lab, is usually not done or is done only in a qualitative rather than a quantitative way. Moreover, this question of data standardization is often the reason given for why data collected for large-scale projects such as the Blue/Human Brain Project or Allen Institute should be collected vertically within a single institute (Markram, 2006; Lein et al., 2007; Kandel et al., 2013)). Thus any effort to draw inferences from the results collected across investigators will have to account for these investigator-level sources of variability. Though there are no simple or definitive answers for how to address these issues of investigator-level variability, potential solutions include working towards standardizing electrophysiological data collection and reporting practices.

1.1.7 Thesis outline

In the remainder of this thesis, I investigate the form and function of neuron diversity using novel methodological approaches. Specifically, in Chapter 2 I describe methods and analyses for understanding the role of electrophyisological variability within a neuron type. In Chapter 3, I discuss methods towards building a database of basic electrophys-

iological properties across mammalian neuron types by extracting this information using semi-automated text-mining from the existing research literature. Chapter 4 then shows novel analyses and utilizations of this database of electrophysiological properties. Lastly, in Chapter 5 I summarize the major work described in this dissertation and discuss potential future work.

Chapter 2

Intermediate intrinsic diversity enhances neural population coding

2.1 Chapter Summary

In this chapter, I develop and apply analyses towards understanding the role of neuronal *within-type* variability among olfactory bulb mitral cells (MCs). When working with *in vitro* data collected by Krishnan Padmanabhan, a post-doc in the lab, I applied statistical generalized linear models (GLMs) towards describing how spikes evoked in MCs result from dynamic stimuli injected as current into MC cell bodies. Though MC input-output transfer functions are quite complex and vary considerably across MCs, the GLM models could sufficiently capture this richness and variability among neurons. Importantly, these models gave me the ability to easily quantify how any two MCs differ in their intrinsic biophysics by simply comparing their GLM model parameters. Using this statistical framework, I then used stimulus decoding methods to investigate the computational role of neuronal variability. I constructed many populations of MCs, which vary in their level of cell-to-cell variability, and asked how well each population can encode a shared stimulus. I found that MC populations which best encode stimuli are those with balance neuronal variability with

neuronal redundancy.

In performing this work, I learned that though neuronal biophysical properties are quite complex, a sufficiently rich and flexible statistical model, fit with an appropriate amount of data, can often sufficiently capture the underlying neural complexity.

This chapter has been published in its entirety in Tripathy et al. (2013).

2.2 Abstract

Cell-to-cell variability in molecular, genetic and physiological features is increasingly recognized as a critical feature of complex biological systems, including the brain. While such variability has potential advantages in robustness and reliability, how and why biological circuits assemble heterogeneous cells into functional groups is poorly understood. Here, we develop novel analytic approaches towards answering how neuron-level variation in intrinsic biophysical properties of olfactory bulb mitral cells (MCs) influences population coding of fluctuating stimuli. We capture the intrinsic diversity of recorded populations of neurons through a statistical approach based on generalized linear models. These models are flexible enough to predict the diverse responses of individual neurons yet provide a common reference frame for comparing one neuron to the next. We then use Bayesian stimulus decoding to ask how effectively different populations of MCs, varying in their diversity, encode a common stimulus. We show that a key advantage provided by physiological levels of intrinsic diversity is more efficient and robust encoding of stimuli by the population as a whole. However, we find that the populations that best encode stimulus features are not simply the most heterogeneous, but those that balance diversity with the benefits of neural similarity.

2.3 Introduction

Biological systems including brains must function efficiently under many constraints, including constraints on the numbers of individual neurons dedicated to a given task. Brain function therefore depends on an appropriate division of labor, with specific neurons dedicated to different functions. For example, different types of retinal ganglion cells represent visual information at different time scales (Puchalla et al., 2005), and distinct classes of cortical interneurons play diverse roles in coordinating network activity (Moore et al., 2010). Whereas attempts to understand how distinct classes of cells encode information have proven successful (Puchalla et al., 2005), the importance of within-type variability remains poorly understood (Altschuler and Wu, 2010; Marder and Taylor, 2011) though has recently become a topic of great interest (Angelo and Margrie, 2011; Padmanabhan and Urban, 2010).

Though neuron-to-neuron variability is often viewed as an epiphenomenon of biological imprecision (Altschuler and Wu, 2010; Marder and Taylor, 2011), having neurons of the same type that respond to different stimulus features may improve stimulus encoding. This variability may be leveraged to improve functions such as stimulus encoding if heterogeneous output of neurons of a single type is collectively used for population coding. Such populations of neurons could efficiently represent complex stimuli by collectively covering the relevant stimulus space (Marsat and Maler, 2010; Puchalla et al., 2005; Schneider and Woolley, 2010). Network interactions could further increase the efficiency of information transmission by decorrelating neural responses and reducing the redundancy between their outputs (Giridhar et al., 2011; Schneidman et al., 2003; Tkacik et al., 2010). In contrast, eliminating redundancy (also referred to as biological degeneracy) may make stimulus coding less robust to noise or to damage (Azouz and Gray, 1999), thus we hypothesized that an optimal coding strategy would require balancing diversity with feature similarity or overlap.

While theorists have previously explored this issue (Stocks, 2000; Tkacik et al., 2010), analysis of the function of the diversity of real populations of neurons requires overcoming methodological hurdles associated with studying cell-to-cell variability (Altschuler and Wu, 2010; Marder and Taylor, 2011). Cell-level differences (that are typically averaged away) must be captured and quantified. Once these differences have been quantified, one must compare the functional output of populations differing in their variability. In the context of neural coding these issues translate to answering the questions: What properties of neurons determine their response to stimuli? How are these properties distributed? And how do these distributions of properties influence the encoding of stimuli by populations? While previous experimental approaches have identified neuron diversity using standard receptive field analyses, these typically do not describe the full complexity of neural responses to stimuli (Pillow et al., 2008; Butts et al., 2007; Slee et al., 2005) nor do they allow the source of the response heterogeneity to be identified as either synaptic or intrinsic. Additionally, simplistic readouts of population spiking output may underestimate the richness of the underlying neural code (Narayanan et al., 2005; Marsat and Maler, 2010; Puchalla et al., 2005; Schneider and Woolley, 2010). Our approach allows the influence of intrinsic diversity to be isolated from synaptic differences and captures the full potential of these diverse populations for stimulus encoding.

Specifically, we developed measures of neuronal population diversity based on statistical generalized linear models (Kass and Ventura, 2001; Pillow et al., 2008) that accurately reproduce the responses of recorded individual olfactory bulb mitral cells (MCs). These cells have been shown to express significant biophysical variability from neuron-to-neuron (Angelo and Margrie, 2011; Padmanabhan and Urban, 2010). We then used the framework of model-based stimulus decoding (Pillow et al., 2008, 2010) to compare how populations varying in their diversity optimally encode varieties of stimuli. This approach enables us to determine whether specific advantages arise from the intrinsic diversity of these neurons,

and how MC populations balance the competing benefits of diversity and feature similarity.

2.4 Results

2.4.1 Statistical neuron models capture mitral cell response diversity.

We generated models of individual MCs from data collected during in vitro whole-cell recordings in which somatic current injection of broad-band filtered noise (Padmanabhan and Urban, 2010) evoked action potential trains (2.1; n = 44 neurons). Synaptic transmission was blocked pharmacologically, so that differences in the cells' spiking responses reflected only differences in their intrinsic firing properties (e.g. due to biophysical conductances and/or morphology). Each neuron's spiking response to input current was fit by a generalized linear model (GLM). GLMs extend stimulus-based reverse correlation or linear-nonlinear-Poisson (LNP) models (Warland et al., 1997; Slee et al., 2005) by including terms that describe how a neuron's spike probability is modulated via its previous spikes (Kass and Ventura, 2001; Pillow et al., 2008). Here each GLM had a constant (bias) term to match baseline firing, a linear stimulus filter determining the neuron's stimulus preference, and a spike history function capturing the neuron's refractory and bursting properties.

This approach captures the spiking responses of neurons without explicitly modeling the many ion channels expressed by individual cells (Angelo and Margrie, 2011; Padmanabhan and Urban, 2010; Marder and Taylor, 2011). Furthermore, GLMs modeled MC activity better than a simpler model that did not include spike history effects (LNP; 2.1 and 2.5), indicating that post-spike refractory and bursting effects substantially contribute to action potential generation in these neurons. Since the parameters of the GLM model emergent physiological features of the recorded neurons, comparing GLM parameters across neurons

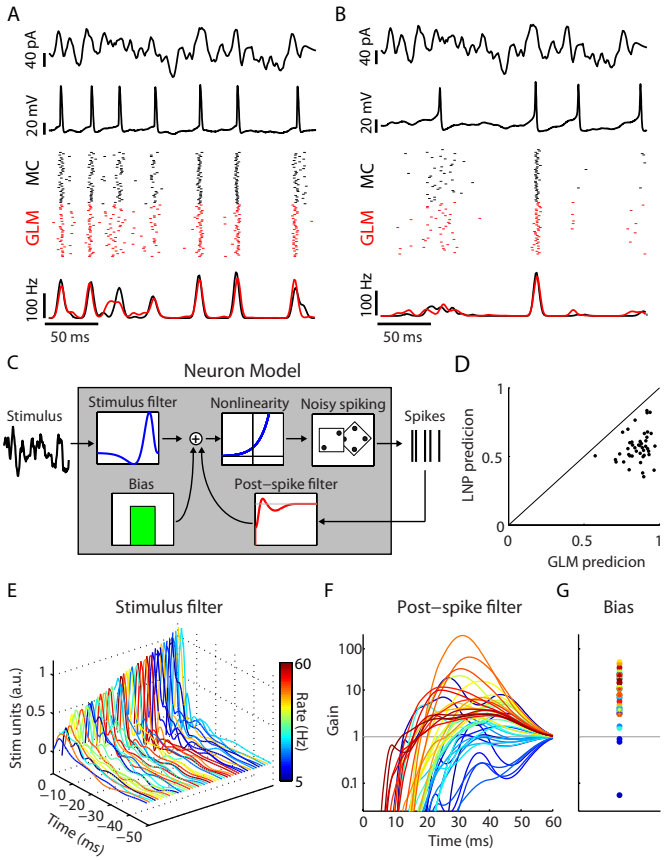

Figure 2.1: Simple models capture mitral cell stimulus-evoked responses and intrinsic diversity. (A) Mitral cell (MC) intrinsic properties are probed using filtered broadband stimuli (1st row) injected somatically to evoke changes in membrane voltage (2nd row). Spike rasters (3rd row; black) and PSTH (4th row; black) for repeated stimulus presentations (n = 40 trials) show that this MC spikes to the stimulus with temporal jitter and displays a coarse stimulus preference. Model neuron rasters (3rd row, red) and PSTH (4th row, red) show that the model accurately predicts MC activity on novel stimuli. (B) Same as A but for a different neuron. (C) Structure of the generalized linear model (GLM) neuron. Model parameters describe a temporal stimulus filter, a post-spike filter, and a constant bias term. An exponential nonlinearity defines an instantaneous spike-rate and is used to draw noisy spikes. (D) GLM models accurately predict 86 +/- 11 % (mean +/- s.e.m.) of stimulus-evoked activity across all MCs, computed as the correlation coefficient between MC and model PSTH. For all neurons, the GLM fits were better than LNP models. (E-F) Model parameters for all MCs. Each line corresponds to parameters for a unique neuron and are colored by mean firing rate. (E) Temporal stimulus filters model differential stimulus specificity of neurons. (F) Exponentiated post-spike filters, plotted as a multiplicative gain in spike probability following a spike at t= 0 ms. Values less (greater) than 1 indicate a decreased (increased) spike probability. (G) Bias terms also show considerable variation. Same y-axis as F.

illustrates the diversity among MCs. For example, the diversity reflected in post-spike (i.e. spike-history) filters potentially corresponds to a recently characterized variability in the rebound depolarization current of these neurons recorded in vivo (Angelo and Margrie, 2011). Furthermore, the interaction of each MC's GLM parameters defines how it responds to stimuli and dictates the complex stimulus features that each neuron best encodes. We note that the efficacy of the GLM approach in capturing MC responses was not specific to the precise stimulus statistics delivered to the neurons here (2.6).

2.4.2 Diversity enables efficient stimulus representation.

Because the GLM approach captures the intrinsic diversity across MCs, different model MCs generate unique spike trains when presented the same dynamic stimulus (2.2 and 2.7). We utilized this model-based approach to ask which features of these individual neurons influence the amount of information about the stimulus that each neuron captures (2.2). Quantifying the quality of stimulus representation using information theoretic methods (Pillow et al., 2008, 2010), we found that neuron information rates were strongly correlated with firing rate (r = .87), in line with previous findings (Borst and Haag, 2001). However, we note that we found examples of neurons that had identical firing rates and yet differed almost two-fold in their information rates, suggesting the importance of additional factors other than firing rate governing the amount of transmitted information. For example, neurons whose spike times were reliable across stimulus repeats and whose spikes were strongly stimulus driven (i.e. minimal contributions from bias or spike-history terms) were more informative per spike (2.2 and 2.7). We note that the large range and diversity of firing rates observed among the MCs here is concordant with those found in vivo (Shusterman et al., 2011).

We extended this information-based framework to examine how populations of recorded MCs encode a common stimulus, considering two broad possibilities. First, stimuli might

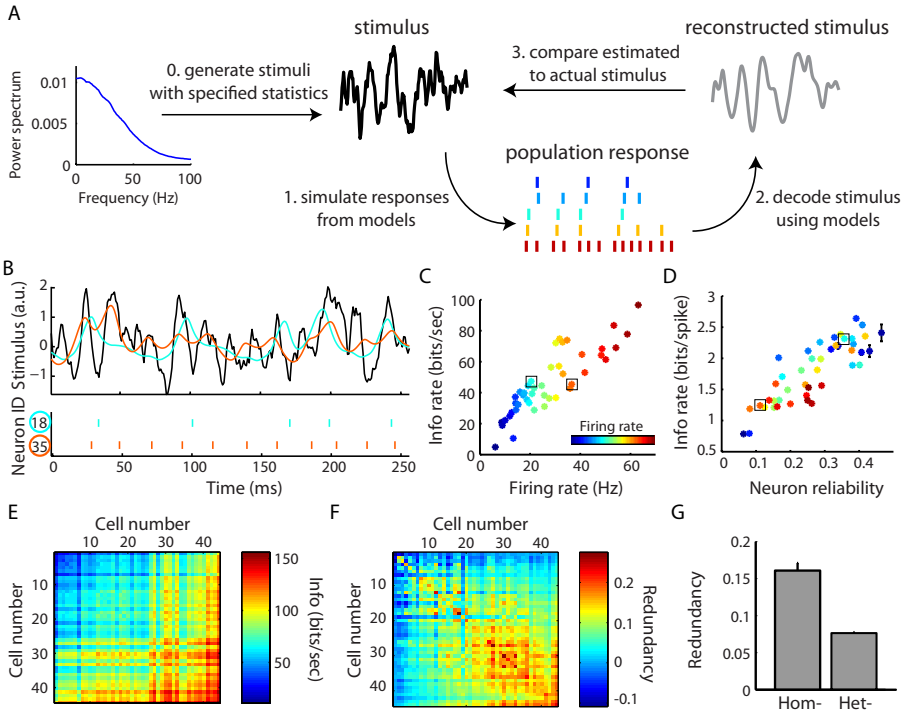

Figure 2.2: Using simulated ensemble responses to study stimulus representation in diverse neural populations. (A) Schematic of the paradigm used to study how neural populations represent stimuli. Following the generation of noisy stimuli, population spike responses were simulated using the MC models. Bayesian decoding was used to estimate the most probable stimulus given the population response and then compared to the actual stimulus. (B-D) Stimulus encoding by single neurons. Stimulus statistics and coloring of neurons same as in Fig. 1. (B) Stimulus (top; black), spike trains (bottom), and reconstructions (top; cyan, orange) for two example neurons. These neurons encode the same stimulus differently, as evidenced by their unique spike trains and stimulus reconstructions. (C) Quantifying stimulus representation using mutual information (mean +/- s.e.m., n = 44 cells) shows that a neuron's information rate is strongly correlated with its firing rate (r = .87). Boxes indicate neurons shown in B. (D) Same as C, but plotted as average information conveyed by single spikes as a function of neuron reliability. (E-G) Stimulus encoding by neuron pairs. (E) Mutual information for all neuron pairs with neurons ordered along axes by increasing firing rate. Values on (off) diagonal correspond to homogeneous (heterogeneous) pairs. (F) Same as E but plotted as the normalized informational redundancy of the neuron pair. Positive (negative) values indicate population redundancy (synergy) where zero indicates independent stimulus encoding. 90% of pairs were redundant, yet overall redundancy values were small, indicating near-independent encoding. (G) Homogeneous pairs (Hom) are significantly more redundant than heterogeneous pairs (Het; p = 2.5*10-16, Wilcoxon, n = 44 and 946 pairs, respectively).

be best encoded by groups of highly similar neurons, suggesting that averaging across the population of recorded neurons can compensate for unreliable spiking in any single neuron (Schneider and Woolley, 2010). Alternatively, stimuli might be best encoded by groups of heterogeneous neurons, suggesting that maximizing the representation of temporal features of the stimuli is important (Atick and Redlich, 1993; Tkacik et al., 2010). We specifically chose to study how diverse groups collectively represent an identical stimulus to mimic features of the olfactory bulb, where 25-50 sister MCs projecting to the same glomerulus (Padmanabhan and Urban, 2010) each receive highly correlated stimulus- and respiration-driven synaptic input (Schoppa and Westbrook, 2001; Shusterman et al., 2011; Spors et al., 2006; Friedrich, 2006; Verhagen et al., 2007).

We created populations of uncoupled virtual mitral cells by randomly selecting groups of model neurons (i.e. fit from the recorded MCs). Spiking responses in these virtual populations were then simulated using the GLM models, enabling us to probe ensemble responses to many more stimuli than could be delivered during experimental recordings. The neurons in these synthetic populations varied in the diversity of their GLM parameters, allowing us to determine how neuronal diversity influences the encoding of fluctuating stimuli. To this end, we used Bayesian model-based decoding, which optimally reconstructs the input to a population (i.e. its "perceived stimulus") given its ensemble response (Pillow et al., 2008, 2010). This approach solves the high-dimensional problem of interpreting dynamic population responses (Schneidman et al., 2003; Pillow et al., 2010) without making undue simplifications or assumptions about the nature of the neural code (Narayanan et al., 2005; Schneider and Woolley, 2010). However, we note that we could have instead focused on alternative metrics of population output instead of stimulus representation efficacy.

We first employed the analysis described above on populations consisting of pairs of simulated neurons. Homogeneous pairs, composed of two copies of the same model neuron (with identical stimulus filtering properties), encoded 73 +/- 11 % more informative about

the stimulus than a single neuron copy alone (2.8). In other words, because spiking is a stochastic process, decoding is improved by considering multiple spike trains from identical model neurons. This allows for averaging out the effect of any single neuron's noise. Next, we considered both homogeneous and heterogeneous pairs of neurons, and quantified the informational redundancy of these pairs. This method compares the information of the pair relative to the sum of each neuron's information independently (Schneidman et al., 2003), and gives an indication of the efficiency of information representation by the population. For example, do neurons together represent information redundantly (i.e. both neurons communicate identical or partially overlapping messages)? Or do they instead represent information synergistically (i.e. both neurons communicate more information together than both individually)? While we found that most homogeneous and heterogeneous populations represented information redundantly (2.2), homogeneous pairs were twice as redundant as heterogeneous pairs (16% versus 8% informational redundancy). Given that these neurons do not directly communicate, we note that the appearance of synergism among neurons pairs here is somewhat surprising and is likely due to limitations in our ability to estimate information rates among low firing rate neurons (see Section 2.6 for further explanation). Nonetheless, these results demonstrate that while pooling responses over multiple neurons even multiple copies of the same neuron is beneficial, the heterogeneity in intrinsic properties in actual mitral cells is beneficial for efficiently representing sensory information.

2.4.3 Intrinsic diversity enables populations to generalize across stimulus types.

We next investigated the effect of diversity on stimulus coding in larger neuronal populations. In 2.3 we plot actual and reconstructed stimuli for two example populations: the first, a homogeneous group composed of five copies of the highest firing rate, most

informative neuron from 2.2C; the second, a population composed of neurons with diverse parameters (Fig. 3D). Both populations encode stimuli composed of high frequencies with high fidelity (2.3A); however, the diverse population is more effective in representing lower frequency stimuli (2.3E) than the homogeneous one (2.3B,C). Thus though the diverse population has 45% fewer spikes than the homogeneous one, the diverse population better utilizes its allocation of five neurons by representing more of the relevant stimulus space with its (temporal) receptive fields.

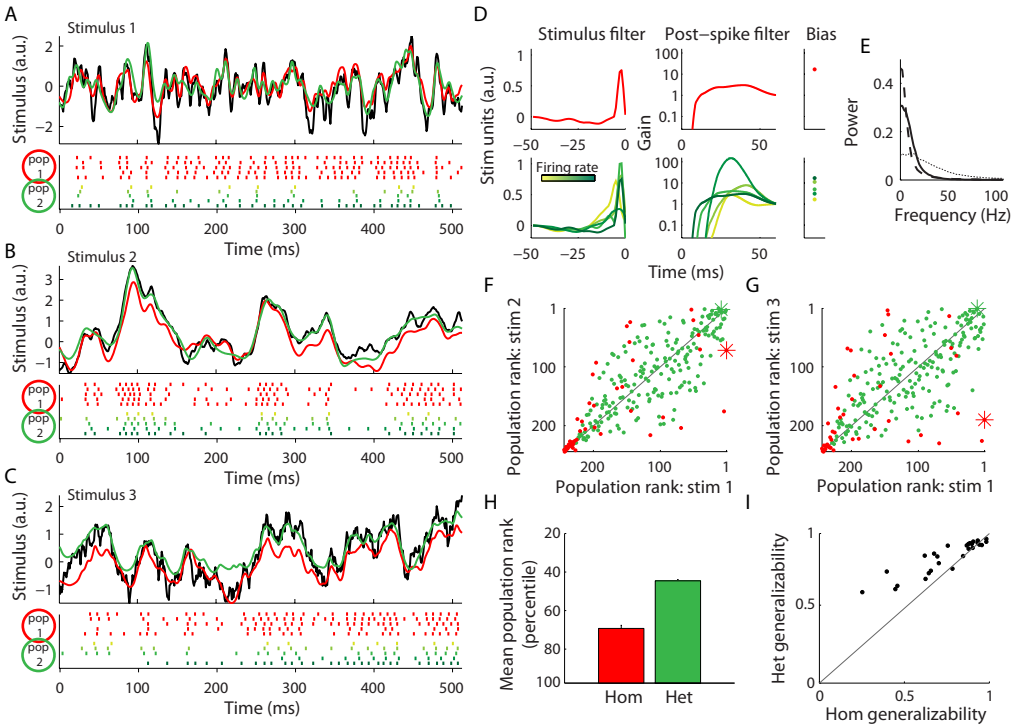

Figure 2.3: Populations composed of diverse neurons effectively encode stimuli with very different frequency spectra. (A-C) Example stimulus (top; black), rasters (bottom), and reconstructions (top) for a homogeneous population composed of 5 copies of the most informative neuron (pop 1, red) and a heterogeneous population composed of 5 neurons with diverse properties (pop 2, green) for 3 stimuli with different power spectra: Stimulus 1, gaussian white noise (GWN) convolved with an alpha function with $\tau = 3$ ms (A); Stimulus 2, GWN with alpha function with $\tau = 10$ ms (B); Stimulus 3, Ornstein-Uhlenbeck process with $\tau = 40$ ms (C). Note that while both populations can represent the stimulus in A well, only population 2, the diverse population, can also represent the lower frequency stimuli in B and C. (D) Neuron GLM parameters for the populations in A-C. Top row indicates parameters for population 1 and bottom row for population 2 (green shades indicate different neurons). (E) Power spectra for the 3 stimuli in A-C (dotted, solid, dashed respectively). (F and G) Relative rankings of stimulus reconstruction accuracy (rmse) for all homogeneous (hom-, red) and 200 randomly sampled heterogeneous populations (het-, green) for stimuli 1 versus 2 (F) or 1 versus 3 (G). Populations in top right of graph indicate those which represent both stimuli accurately. Asterisks indicate populations highlighted in A-C. Note that hom- populations are among the bottom populations and are further from the unity line than het- populations. (H) Average rank of het- and hom- populations across 8 spectrally unique stimuli (see Methods). Het- populations are consistently ranked higher (more accurate) than hom- ones (p = .002, paired Wilcoxon). (I) Plot of generalizability, defined as the correlation of population ranks on stimulus pairs, for hom- and het- populations across all pairs of 8 stimulus types. Each dot corresponds to the generalizability between a pair of stimulus types (n = 28 total pairs). Het- populations are significantly more likely than hom- to generalize to novel kinds of stimuli (p = 1.5*10-4, paired Wilcoxon).

To extend this analysis we compared how 250 populations of randomly chosen five-neuron ensembles encoded stimuli with different frequency spectra (e.g. $1/f^{\alpha}$ noise with differing values of α, white noise, etc, n= 8 stimuli total, shown in 2.9). These stimuli were chosen to cover a wide range of input frequencies including the range of frequencies these neurons likely receive in vivo (Khan et al., 2012; Spors et al., 2006). We created homogeneous populations, each consisting of five copies of a single MC, and heterogeneous populations generated by randomly selecting five MCs from the recorded set with replacement. To compare population responses across stimulus spectra, we ranked the populations in order of increasing average reconstruction error for each kind of stimulus and compared ranks across different stimuli. Across pairs of stimulus types population ranks were correlated (2.3F, G, r = .80, .71 respectively), meaning that those populations that represent one stimulus well also represent other kinds of stimuli well (termed generalizability). Heterogeneous populations were better than homogeneous ones not just at encoding stimuli on average (2.3H), but also at generalizing across different stimuli (specific examples in 2.3F,G; summary in 2.3I). Thus the observed intrinsic diversity helps encode many kinds of stimuli, conferring representational robustness to MC populations.

2.4.4 Populations optimized for specific stimuli combine diversity with homogeneity.

Thus far, we have only considered sampling neurons randomly according to a particular rule (homogeneously versus heterogeneously). We next attempted to construct more optimal groups of neurons for encoding specific stimulus types. We liken this scenario to that of sister MCs associated with a single glomerulus, which receive inputs with a specific temporal structure (Carey and Wachowiak, 2011; Shusterman et al., 2011) based on olfactory receptor neuron (ORN) odorant binding kinetics, which differ across glomeruli and ORN subtypes (Nagel and Wilson, 2011). Would the best population for a given stimulus be

more diverse than selecting MCs at random from the physiologically-based set? Or would the best population be more homogeneous than random, perhaps allowing the responses of unreliable neurons to be improved upon by selecting neurons coding for redundant (i.e. degenerate) stimulus features? To answer these questions, we implemented a greedy search algorithm (Russell and Norvig, 2009) to build the best population of model MCs to encode a given stimulus by iteratively adding neurons one at a time such that the added neuron maximized the ability of the entire population to represent the stimulus (2.4A). While not guaranteed to find the global optimum, it is an efficient and intuitive method of finding neuron groups more informative than those generated through random sampling.

Visualizing the makeup of these greedy-search selected populations using dimensionality reduction (2.10) reveals that they reflect a balance between diversity - consisting of neurons with different properties - and homogeneity - often including multiple copies of selected neurons (2.4B,C and 2.11). Additionally, the stimulus type dictates the selection of specific neurons and the chosen level of population diversity. For example, the population selected to best encode a white noise stimulus (2.4C) was composed primarily of similar neurons with high firing rates whereas diversity in neuron properties was more important for encoding a more naturalistic stimulus with both rapidly and slowly-varying temporal components (2.4B). Using the greedy search algorithm to select populations for each of the 8 stimulus types, we quantified the diversity of these populations and of randomly sampled heterogeneous and homogeneous populations (2.4D). Surprisingly, greedy-search populations were on average 25% less diverse than heterogeneous ones when considering either stimulus filter and post-spike parameters. Furthermore, quantifying population diversity for MC groups selected to best encode different stimulus types reveals that they have varying levels of diversity (2.4E and 2.12), suggesting that population diversity should be preferentially tuned to the afferent stimulus distribution.

To ensure that the previous findings are not solely the result of the greedy selection

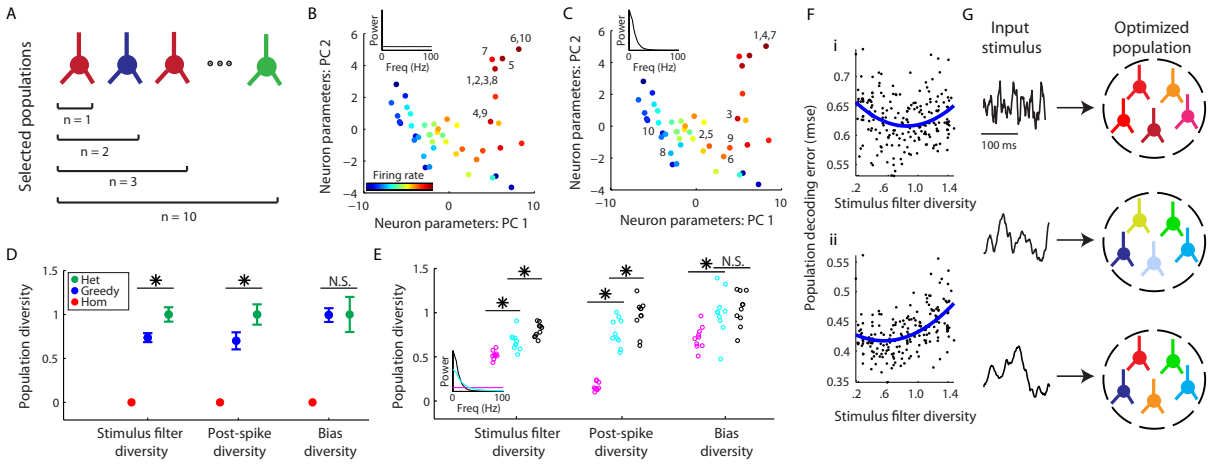

Figure 2.4: Populations optimized for stimulus representation combine homogeneity with diversity. (A) Cartoon of greedy-search algorithm to estimate the population which best represents a particular type of stimulus. Neurons were iteratively added, one at a time, to the current population of neurons such that the neuron chosen maximized the population's reconstruction accuracy. To allow for homogeneity, neurons could be added more than once (e.g. two red neurons). (B and C) Visualization of the population selected to best represent a white noise stimulus (B) or a low frequency stimulus (C). Graphs show neurons (as dots) projected into a 2-dimensional space using principal component analysis (PCs). Population sizes vary from n = 1 to n = 10, numbers next to dots correspond to algorithm iteration step when each neuron was added. Note that certain neurons are chosen multiple times and that stimulus type dictates the selected population diversity. (D) GLM parameter diversity of the greedy-search selected populations (blue) averaged over 8 different choices of stimulus spectra relative to homogeneous (red) and randomly sampled heterogeneous populations (green), n = 10 neurons per population. Asterisks indicate where greedy-search populations are significantly less diverse than heterogeneous ($p < .05$) and population diversity has been normalized to that of randomly sampled heterogeneous. Error bars indicate s.e.m (blue) and interquartile range (green). (E) Greedy-search population diversity for specific stimulus types. Colors indicate different stimulus types corresponding to inset power spectrum (magenta, stimulus as in B; cyan, OU process with $\tau = 10$ms; black, stimulus as in C), open circles indicate multiple runs of the greedy search algorithm (n = 10 per stimulus type), asterisks indicate significant differences in population diversity between stimulus types. (F) Population decoding error as a function of stimulus filter diversity for 200 randomly sampled populations (dots, n = 5 neurons per population) for stimulus 1 and 2 as in Fig. 3 (i, ii, respectively). Least-squares fits using a 2nd-order polynomial (blue) show that on average there is an intermediate level of stimulus filter diversity where decoding error is minimized (regression p-value $< .01$). (G) Cartoon showing that population diversity should be preferentially selected with respect to the specific incoming stimulus distribution.

process, we performed additional simulations by randomly constructing populations with differing amounts of diversity and examining the relationship between population diversity and decoding accuracy. As predicted from the greedy search results, we found evidence for a U-shaped relationship between decoding accuracy and population diversity (2.4F and 2.13), indicating that neural coding is optimized at intermediate levels of diversity. However, population size is also a relevant factor in the importance of population diversity, with diversity being more important to smaller populations than larger ones (2.14). This suggests that heterogeneity will be more important to populations in which the number of neurons devoted to representing a stimulus is relatively small. Furthermore, we found the benefit of neural variability to not be solely dependent upon a single GLM filter dimension (2.15), like the stimulus filter or bias term.

2.5 Discussion

Here we apply the framework of generalized linear models to study how cell-to-cell differences in intrinsic properties of olfactory bulb mitral cells influence stimulus encoding. To our knowledge this is the first application of this approach to quantifying cell-to-cell heterogeneity or population complexity. The statistical modeling approach that we have used accurately captures the neuronal properties determining spiking while avoiding overfitting. It also avoids making specific but difficult-to-verify claims about channel densities or properties that can arise from under-constrained Hodgkin-Huxley models (Bhalla and Bower, 1993). We show that diverse populations offer the advantages of more efficient encoding (defined in terms of information per cell or information per spike) and more robust coding of different kinds of stimuli, such as stimuli with wide ranges of spectral properties. This is because neurons encoding partially overlapping (i.e. degenerate) stimulus features can work together to overcome neural spike-generation noise and also encode more stimulus features together than separate. We also show that populations selected to best represent

stimuli with specific spectral properties have differing amounts of diversity, suggesting that population diversity should be selectively chosen with respect to the precise stimulus to be enocded. While variants of this framework have been used to model neural responses previously (including in single neuron modeling competitions (Gerstner and Naud, 2009; Jolivet et al., 2008)) this is the first use of this framework for describing the systematic biological differences among neurons and their impact on population coding. Given the generality of this framework, we believe that this methodology can similarly be extended to describe electrophysiological differences across neuron types and to develop hypotheses about the distinct roles of different neuron types throughout the brain.

One of the key advantages of this approach is that it allows us to use Bayesian stimulus decoding to ask how neuron-to-neuron differences in stimulus filtering and post-spike properties influence population coding of arbitrary stimuli. Bayesian decoding is advantageous because it offers an optimal, best-case view of neural encoding, making few assumptions that risk underestimating the complexity of the neural code (Pillow et al., 2008, 2010). While we explored the relationship between stimulus encoding in diverse and homogeneous populations in a previous study (Padmanabhan and Urban, 2010), performing stimulus reconstruction here allows the identification of the relative importance of variation in specific features of the sets of recorded neurons. This approach also allows us to investigate stimulus encoding in a more general context by simulating responses to arbitrary stimuli. An obvious advantage of simulation approaches is that we are not limited to only analyzing data that we are able to collect during recordings.

Our results make specific, testable predictions on the role of MC intrinsic diversity for encoding olfactory information. First, we show that when populations need to represent a variety of stimulus types, then intrinsic diversity facilitates generalizing representations across stimulus types. Secondly, when populations need to represent a single kind of stimulus and are allowed to selectively choose their level of variability, populations choose

a balance between complete homogeneity and diversity. That is, homogenizing the input received by a population of neurons should lead the population to be less diverse. This in silico finding is intriguing because it is consistent with recent experimental findings showing that sister MCs, receiving primary olfactory inputs from the same glomerulus and olfactory receptor subtype, are biophysically more similar to one another than sampling MCs at random (Angelo and Margrie, 2011). Furthermore, our work makes the additional hypothesis that the level of diversity across sister MCs should be adaptive with respect to the unique stimulus distribution that these neurons receive from their olfactory receptor subtype (Carey and Wachowiak, 2011; Nagel and Wilson, 2011). Therefore we predict that the levels of MC intrinsic diversity between sister MCs should be empirically different across glomeruli (2.4G).

We note that we made multiple assumptions here for the sake of computational tractability. Because our focus was to study the functional role of MC intrinsic diversity, we chose not to include any of the effects of neural connections like synapses between neurons in our experiments and simulations. Given that the olfactory bulb possesses extensive lateral circuitry (Giridhar et al., 2011) which has been shown to also diversify MC responses (Arevian et al., 2008; Dhawale et al., 2010), we expect that bulbar circuit activity will work in conjunction with intrinsic diversity in vivo to further diversify MC responses. Furthermore, when decoding we took the perspective that the best populations were those which resulted in the most faithful reconstruction of the stimulus. However, the biological solution dictating the actual amount of diversity may use alternative criteria for optimality. For example, in vivo, olfactory bulb MCs may seek to represent only odor-specific stimulus components or may try to maximize stimulus representation while also minimizing the number of spikes used to transmit the information (Weber et al., 2012). We chose to avoid assumptions about which features of ORN input are most important for MCs to represent and rather to take the agnostic view that MCs should try to represent the stim-

ulus in its entirety. Our approach, however, can readily be adapted to tasks that require representation of specific stimulus components. While these assumptions likely affect the quantitative details of our results, like specifying of the precise balance between diversity and feature similarity, our general finding that a precise stimulus-specific balance exists nevertheless likely holds.

We believe that our results generalize to other neural systems because this circuit motif in which multiple neurons receive highly correlated inputs occurs throughout the brain, including neocortex (Poulet and Petersen, 2008). Thus we predict that the observed degree of neuronal intrinsic variability plays a substantial role in tuning the output diversity (or redundancy) in these neurons' spiking responses and in improving stimulus encoding. Furthermore, our findings may in part explain the substantial informational redundancy found in neural populations throughout the brain (Puchalla et al., 2005; Schneider and Woolley, 2010). Given that the optimal networks here are neither maximally diverse nor maximally homogeneous, these results suggest similar design principles for other systems. By mixing diversity with neural feature similarity, complex systems can simultaneously maintain efficient functioning while remaining robust to uncertain events.

2.6 Materials and Methods

2.6.1 Neuron Recordings

Whole cell patch clamp recordings of mitral cells were obtained in vitro from mouse olfactory bulb slices using methods described previously (Padmanabhan and Urban, 2010). Mitral cells were identified under infrared differential inter¬ference contrast optics on the basis of their laminar position in the olfactory bulb and their morphology. All experiments were performed at 35 °C in standard Ringer's solution with excitatory (25 μM AP5 and 10 μM CNQX) and inhibitory (10 μM bicuculline) synaptic activity blocked.

Current-clamp recordings were performed while injecting neurons with a filtered white noise current stimulus. Noise traces were generated by convolving a 2.5-s white noise current with an alpha function of the form $t * exp(-t/\tau)$, where τ = 3ms. We chose this spectral structure as it generates reliable spiking in these neurons and corresponds to the time-scale of fast synapses afferent to MCs (Galán et al., 2008). Each neuron received one of a small number of stimuli generated via this method (most neurons received 1 of 3 stimulus templates) and was presented 40 stimulus repeats. The amplitude (variance) of the noise used was between 5% and 40% of the direct current (100-800 pA, σ = 20-80 pA) offset for each cell, with the majority of cells receiving 10-20% of the DC offset. The variance of the noise was selected as previously described (Galán et al., 2008), to induce reliable firing with¬out large input fluctuations. For all recordings, a 25 or 50 pA hyperpolarizing pulse was injected before stimuli were delivered to measure input resistance and membrane time constant, allowing us to track the stability of recordings over multiple trials. Only neurons whose firing patterns were stable across trials and fired a sufficient number of spikes in each trial (>5 Hz) were used in this study. Upon stimulation most neurons usually underwent a brief non-spiking adaptation period (111 +/- 14 ms) which was assessed visually and excluded from the analysis.

2.6.2 Model fitting

GLM models were fit and simulated using code provided by Jonathan Pillow (Pillow et al., 2008). Models consisted of a temporal stimulus filter k, a post-spike history filter h,and a constant bias term b. Stimulus and history filters were each represented using 10 spline-like cosine basis functions spaced logarithmically in time. The conditional intensity function of each neuron was modeled as $\lambda(t) = exp(kx + hr + b)$, where x denotes the stimulus and r is the recorded spike response of the neuron. Before fitting, stimuli were down-sampled to 1KHz and standardized by subtracting the DC component and dividing by the amplitude

of the stimulus noise. LNP models were fit using the spike triggered average stimulus as the linear filter and estimating the spike rate nonlinearity using 60 independent histogram bins.

Models were trained using all of the trials from the first 90% of the stimulus presentation and validated using the remaining 10%. Specifically, we validated the fit of our models by comparing real and model peri-stimulus time histograms (PSTHs) computed from the test stimulus set (i.e. stimuli not used in the training of the model). We simulated model spike trains using the GLM to probabilistically generate spikes elicited by the test stimulus. PSTHs were computed by summing spikes across trials and smoothing with a Gaussian filter of width $\sigma = 2$ ms. The similarity between real and model PSTHs was reported using Pearson's correlation coefficient. For visualization, MC rasters were randomized across trials.

To assess whether the GLM fitting procedure could also fit neuron responses to multiple stimulus types, we performed an additional set of experiments on mitral cells (n = 5 neurons) where each neuron was stimulated with both a high and low frequency stimulus (white noise convolved with an alpha function with $\tau = 3$ms and $\tau = 10$ms, respectively). We found that the GLM modeling procedure could sufficiently fit neuron responses to each of these stimulus types, indicating that the fitting procedure is not specific to the particular stimulus type used to generate stimulus evoked responses in this study (2.6).

2.6.3 Computation of neuronal statistics using GLM models.

We were interested in computing neuronal statistics like average firing rates and trial-to-trial reliability from the fitted GLM models. We computed these by simulating long experiments (2 minutes) of continuous stimulation and computing desired statistics based on these responses. We computed neuron reliability by stimulating each model neuron with multiple trials (n = 50) of the same stimulus and calculating reliability as the average

zero-lag correlation across trials using a bin size of 5ms.

To calculate to what extent neurons were driven by intrinsic (history plus bias) versus stimulus components (2.7B), we used the model to simulate spike trains while storing the stimulus and intrinsic currents which generated the spike trains. Here the stimulus driven component consists of the convolution of neuron's stimulus filter with the input stimulus whereas the intrinsic component is defined as the bias term plus convolution of neuron's spike train with its post-spike filter. We calculated the ratio of intrinsic to stimulus inputs as $< [stim]_+ > / < [intrinsic]_+ >$ where $[x]_+$ indicates selection of positive values of the currents and $< x >$ indicates the mean.

2.6.4 Stimuli generation for simulations.

We generated zero-mean Gaussian stimuli x with a defined temporal correlation structure and length n by first generating a signal autocorrelogram with the desired spectral structure. This autocorrelogram was used to define a Toeplitz $n * n$ covariance matrix C where the elements of C indicate the pairwise correlations between points of x. Correlated stimuli were then generated using the Cholesky decomposition to find a matrix L such that $C = L * L^T$, then multiplying L with a series of uncorrelated normal random variables of length n.

Here we chose 8 broadly different stimuli statistics: 3 stimuli generated via convolving white noise with an alpha function defined as $t * exp(-t/\tau)$, where τ = 3, 5, 10ms; 3 Ornstein-Uhlenbeck processes with τ = 10, 20, 40ms, which have flat followed by $1/f^2$-like frequency profiles; a pure white noise stimulus, with cutoff at 500 Hz; lastly, a naturalistic stimulus generated by combining an 8 Hz oscillatory stimulus with an Ornstein-Uhlenbeck process with τ = 10ms (displayed in 2.9).

2.6.5 Decoding

Decoding. We decoded the population spiking responses using the maximum a posteriori (MAP) estimator (Pillow et al., 2010), which finds the most probable stimulus given a particular population spike response. Stimuli x (typically of length .5 s, with sampling rate 1KHz) were decoded from simulated spike responses r by computing the mode of the posterior distribution, $\underset{x}{\text{argmax}} p(x|r)$, where $p(x|r) \propto (r|x)p(x)$ via Bayes' rule. Here $p(r|x)$ is the likelihood of a response given a stimulus and is given by the set of uncoupled neuron encoding models and $p(x)$ is a multivariate Gaussian prior specifying the specific stimulus autocorrelation structure (with covariance matrix, C, used to generate stimuli). Specifically, stimuli were decoded utilizing a recently described method(Pillow et al., 2010) which takes advantage of a convenient Gaussian approximation on the posterior distribution $p(x|r)$ and its log-concavity to exactly compute the maximum (i.e. the mode) of the posterior distribution via numerical optimization techniques. This method also provides an estimate of the uncertainty of the stimulus representation (2.8F,G). Matlab code for decoding and all other methods related to the simulation and analysis of spike trains generated from GLM models (detailed below) can be found at `https://github.com/stripathy/mitral_cell_diversity`.

2.6.6 Mutual information calculation

Mutual information calculation. We calculated the mutual information (Pillow et al., 2010) of the population response r about the stimulus x as $I(x;r) = H(x) - H(x|r)$. $H(x)$ denotes the entropy of the stimulus and is defined by the multivariate Gaussian stimulus prior $p(x)$ and $H(x|r)$ denotes the conditional entropy of the stimulus given the response and is estimated by approximating the posterior distribution $p(x \mid r)$ as a multivariate Gaussian $N(x_{map}, C)$ where the covariance matrix, C, is computed as a by-product of our decoding method. Here we utilize the fact that the entropy of a Gaussian with covariance matrix

C is $\ln \sqrt{((2\pi e)^n |C|)}$, where $|x|$ denotes matrix determinant and n is the dimension of the stimulus. Estimates of $I(x;r)$ were obtained by averaging $H(x|r)$ over responses elicited to multiple stimuli realizations (n = 50). Importantly, because this method estimates the entropy of the posterior distribution, it generally provides a better estimate of the mutual information than the commonly used lower-bound estimate of $I(x;r)$ obtained via the optimal linear estimator (Warland et al., 1997), especially when the neurons are non-linear and not well described by an LNP model.

We computed a normalized measure of the redundancy or synergy (Schneidman et al., 2003) of a pair of neurons $a;b$ relative to each of the neurons independently as $(I(x;a) + I(x;b) - I(x;a,b))/I(x;a,b)$. Positive values indicate informational redundancy while negative values indicate synergy.

To elaborate on our finding of synergistic pairs of neurons (2.2F), we note that due to computational constraints we can only decode stimuli of relatively short lengths (.5 seconds). Therefore we will tend to underestimate the information rates of neurons which fire at low firing rates. For example, when performing stimulus decoding to calculate the information rate of a single neuron with a very low firing rate, it may fire zero spikes during the time interval and thus encode no stimulus information. However, when considering two such neurons, the two will be much more likely to fire at least one spike between them, and thus encode some nonzero stimulus information. In this example, the case of a two-neuron pair would appear synergistic relative to a single neuron alone. Therefore, if we could simulate arbitrarily long stimulus presentations we would expect this apparent synergy effect to disappear.

2.6.7 Calculating population stimulus generalization.

Calculating population stimulus generalization. To calculate how well heterogeneous and homogeneous populations generalized across stimuli of differing types, we computed the

generalizability for each population type. Here generalizability is defined as $corr(ranks)_{stim1}, ranks_{stim2})$, or the correlation between population ranks on pairs of stimulus types.

2.6.8 GLM dimensionality reduction.

We chose to reduce the dimensionality of the space defined by neuronal GLM parameters using principal component analysis (Fig. S5) for visualization and further analysis. Principle components (PCs) were generated by first concatenating waveforms of stimulus, post-spike, and bias components across all neurons and standardizing before performing PCA. Post-spike and bias terms were transformed to units of log(Gain) before concatenating. The first 10 ms of post-spike filters were removed and not included in analysis.

2.6.9 Computing population diversity.

We calculated population diversity as the mean Euclidean distance of GLM parameters computed between all pairs of neurons in a population. We excluded the first 10 ms of the post-spike filters across neurons as most neurons were refractory during this period. The average diversity of heterogeneous populations was computed by averaging over 50,000 randomly sampled populations. When reporting the uncertainty in the diversity of randomly sampled populations (2.4D), we chose to show a measure of the population variance (interquartile range) as opposed to standard errors.

We sampled populations that varied greatly in their amount of diversity (from low to high; 2.4F and 2.13, 2.14) through implementing stratified sampling where we first sampled 2 million 5-neuron populations and then further sub-sampled this set to pick populations that varied uniformly in their diversity.

2.6.10 Eliminating diversity in a single GLM dimension

We constructed populations which had diversity eliminated along a single GLM dimension (stimulus, post-spike or bias) by modifying the neuron model parameters from the ones based on the recorded neurons (2.15). For example, to sample neurons where diversity in the stimulus filter had been eliminated, we set the stimulus filter for all neurons that of the mean stimulus filter computed over all neurons. We further ensured that mean of the firing rates across neurons were similar between the original and diversity-reduced populations.

2.7 Supplemental Figures

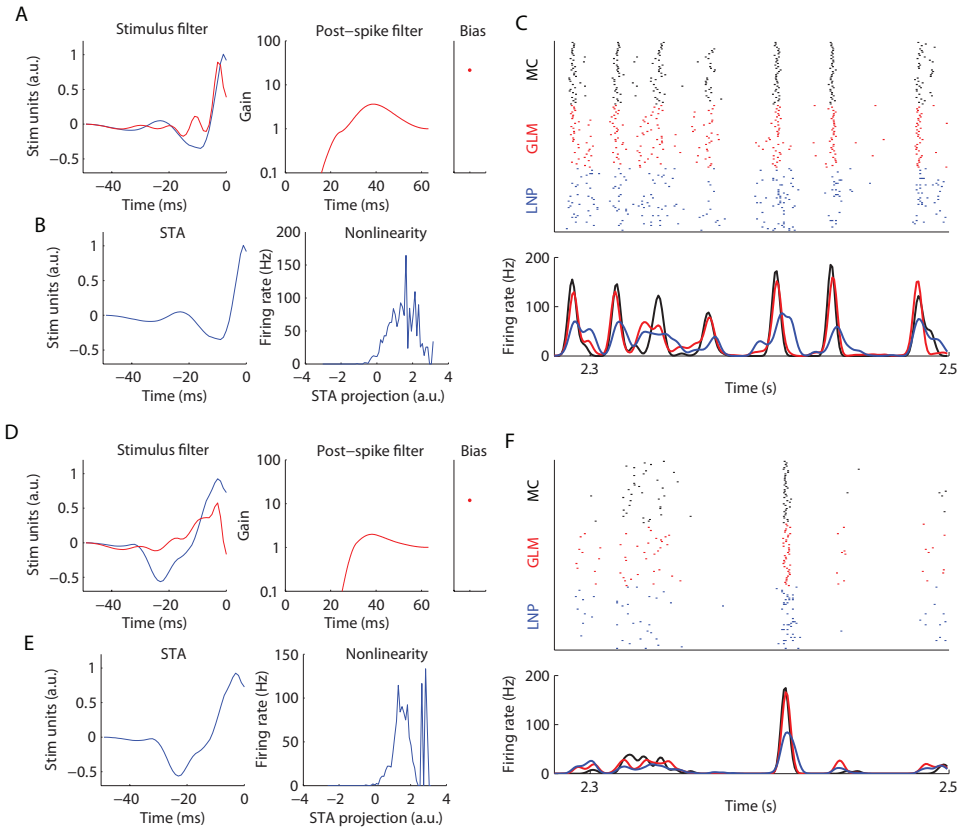

Figure 2.5: *Comparison of GLM and LNP model parameters and prediction accuracy.* (A) GLM parameters (red) and spike triggered average current (STA, leftmost panel, blue) for the neuron in Figure 1A. (B) LNP parameters for same neuron. (C) Experimental MC, GLM, and LNP rasters (top) and PSTH (bottom). Note that GLM spikes replicate the MC more precisely than the LNP model. (D-F) Same as A-C but for in neuron Figure 1B.

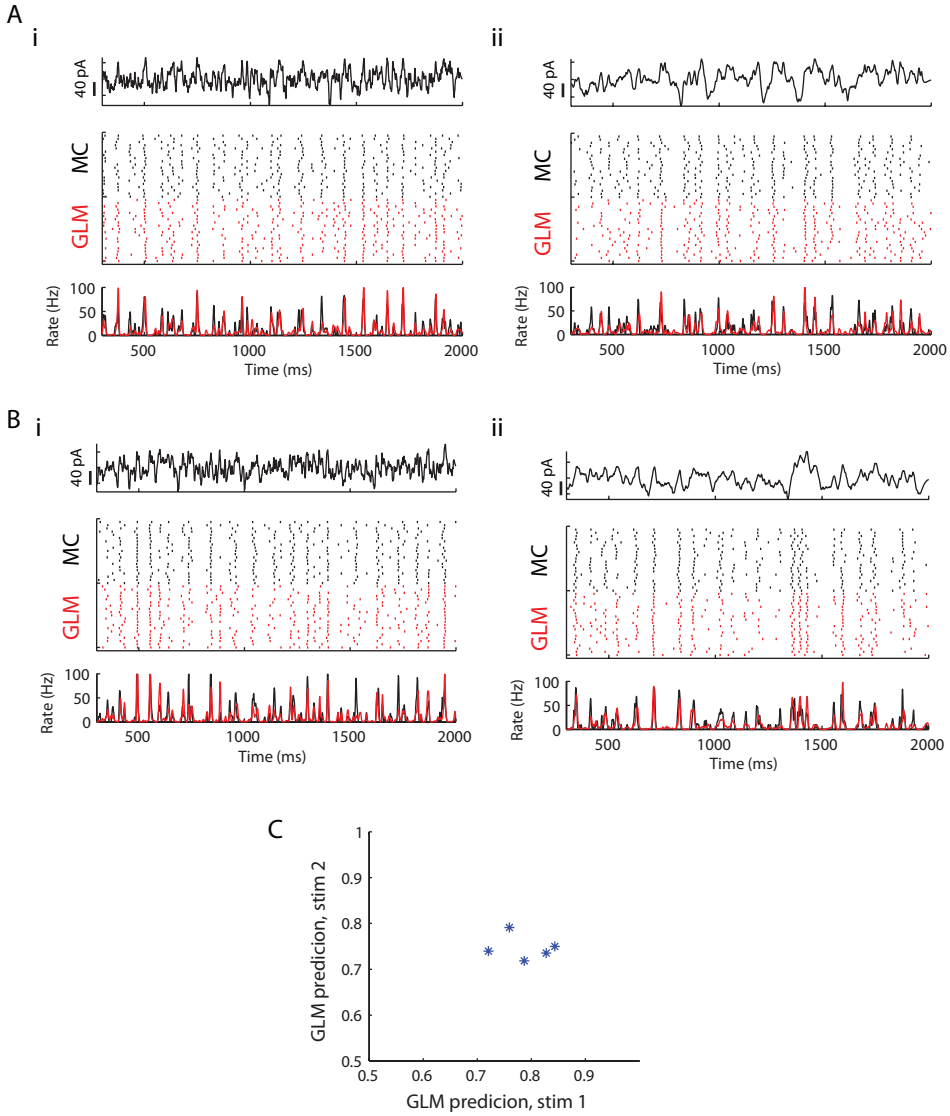

Figure 2.6: *GLM modeling procedure is effective at fitting neuron responses to different stimulus types and temporal correlation structures.* (A) Stimulus and spike rasters for a recorded mitral cell (MC, black) and simulated model responses (GLM, red) for a high frequency stimulus (i, white noise convolved with alpha function with $\tau = 3$ ms, stim 1) or a low frequency stimulus (ii, white noise convolved with alpha function with $\tau = 10$ ms, stim 2). Entire stimulus and spike response shown to illustrate temporal correlation difference between stimuli in i and ii. (B) Same as A but for a different mitral cell. (C) GLM models accurately predict neuron responses to each of these 2 stimulus types (n = 5 MCs). Prediction accuracy computed as the correlation coefficient between MC and model PSTH.

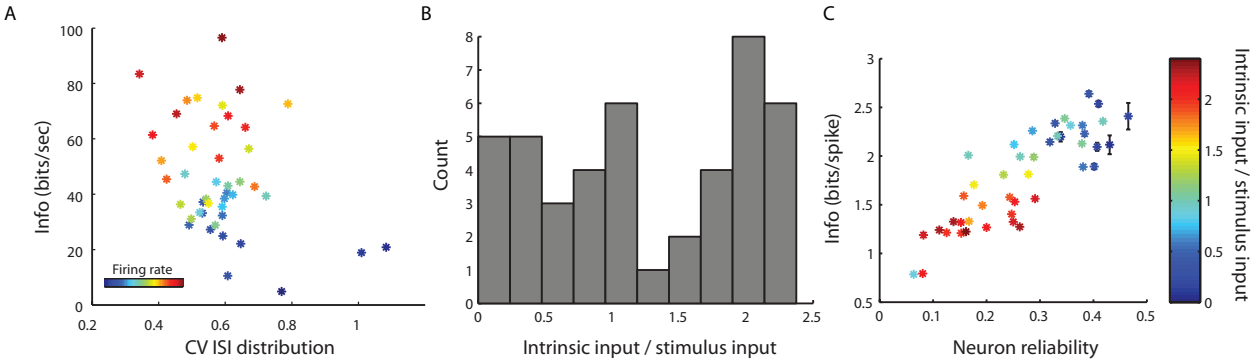

Figure 2.7: *Relationship between neuronal information rate and other spiking characteristics.* (A) Neuron information rate as a function of coefficient of variation of the inter-spike interval distribution (CV ISI distribution). Neurons with higher information rates tend to have lower CVs, indicating that they fire more regularly (i.e. less Poisson-like). (B) Histogram of the relative ratio of neuron spiking due to stimulus driven or intrinsic components (bias plus history term) as indicated from the GLM. Neurons vary from being driven entirely by stimulus to being driven primarily by intrinsic components. (C) Same as Figure 2D but neurons are colored by ratio of intrinsic to stimulus input (as in B). Neurons which are primarily stimulus driven are both more reliable and encode more information per spike.

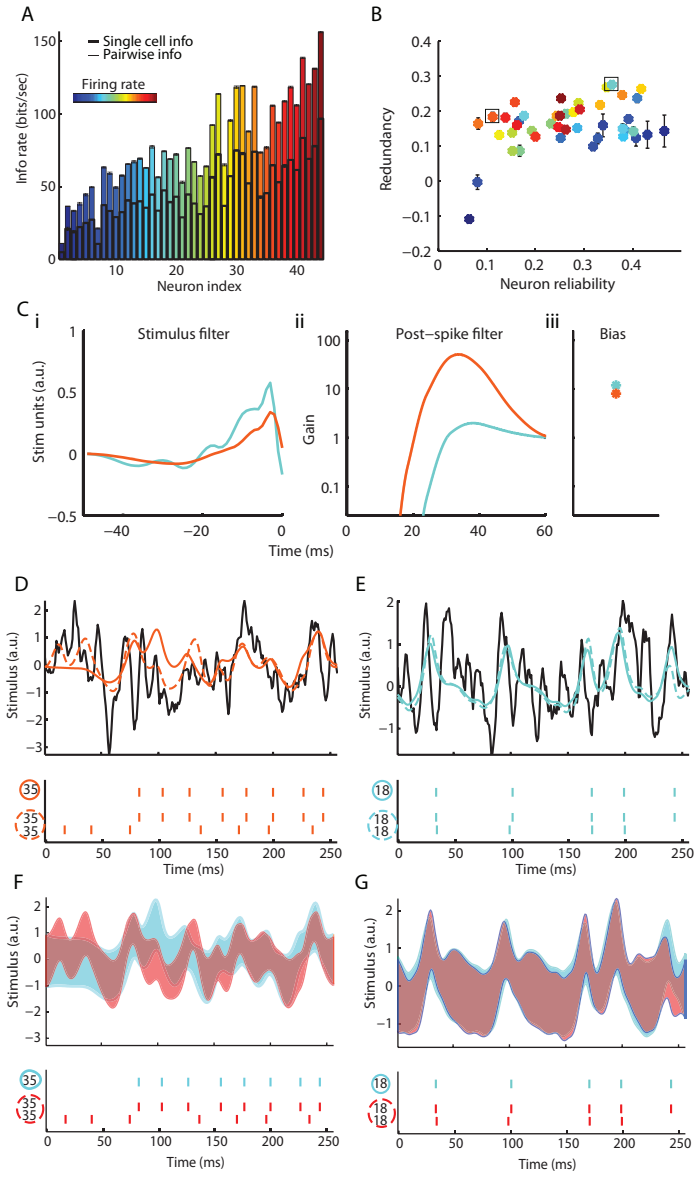

Figure 2.8: *Two copies of the same neuron are more informative than a single neuron alone.* (A) Stimulus-response mutual information for single neurons (thick line) and populations of two copies of the same cell (thin line). In all cases, two neurons convey more information than a single neuron alone. (B) Informational redundancy of homogeneous pairs versus single neuron populations plotted as a function of single neuron reliability. Boxes show neurons 18 and 35, which are also highlighted in Figure 2B. (C) GLM parameters for boxed neurons in B. These two neurons were chosen as an example because they have very different GLM parameters which cause them to spike differently in response to the same stimulus. Neuron 18 (cyan) fires primarily as a result of the stimulus whereas neuron 35 (orange) has a large post-spike rebound current causing it to fire relatively periodically. (D and E) Example stimulus and reconstructions for neurons 35 and 18. Note that for neuron 35, an unreliable neuron, there is an improved stimulus representation when another copy of 35 is added; this is effect is less pronounced for neuron 18, which fires more reliably. (F and G) Same as D and E but showing uncertainty in stimulus reconstruction. Blue shaded trace indicates neuron 18 and red trace indicates neuron 35. Stimulus not shown for clarity. Error bars represent 1 standard deviation of the Gaussian estimate of the decoded posterior distribution. For both neurons, the stimulus representation is less uncertain with neuron pairs versus single neurons, highlighting the benefit of redundancy or pooling representations over neurons.

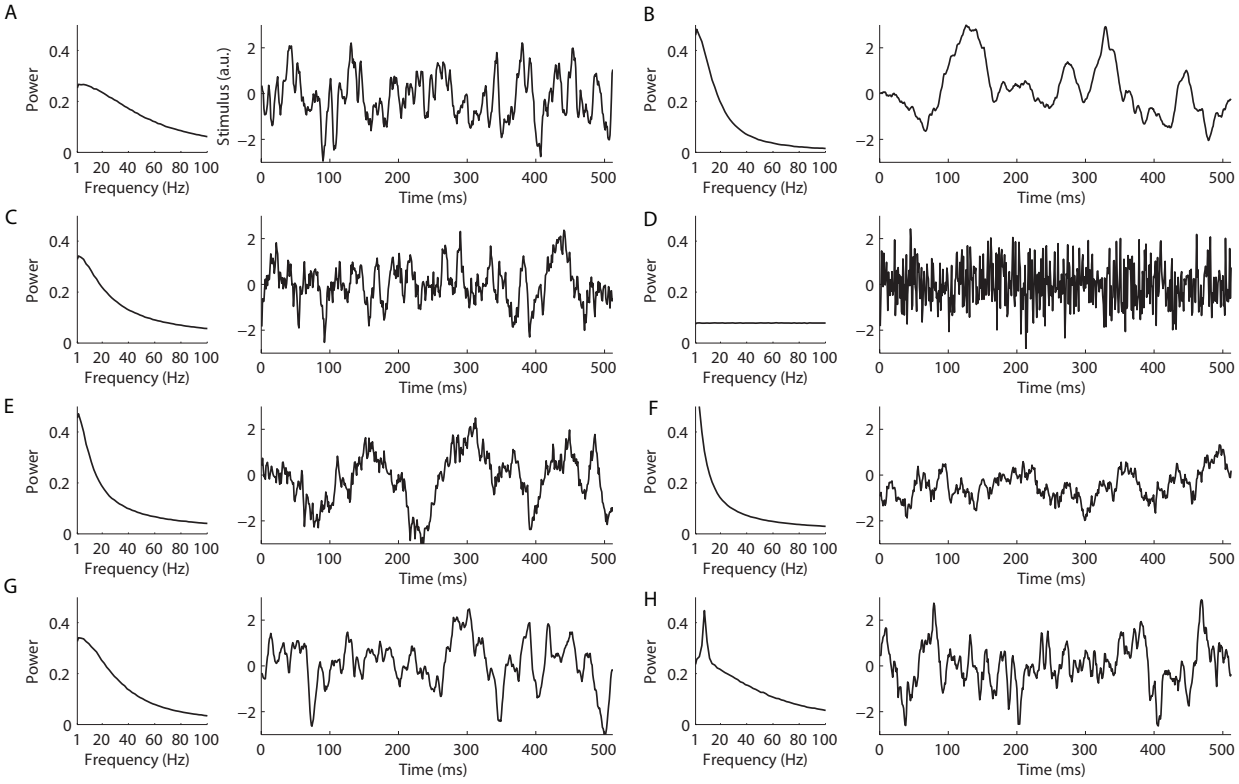

Figure 2.9: *Examples of the 8 stimulus statistics used in this study.* (A) White noise convolved with alpha function with $\tau = 3$ ms. (B) White noise convolved with alpha function with $\tau = 10$ ms. (C) Ornstein-Ulhenbeck process with $\tau = 10$ ms. (D) Pure white noise stimulus (max frequency = 500 Hz). (E) Ornstein-Ulhenbeck process with $\tau = 20$ ms. (F) Ornstein-Ulhenbeck process with $\tau = 40$ ms. (G) White noise convolved with alpha function with $\tau = 5$ ms. (H) Naturalistic sniffing stimulus generated by combining 8Hz oscillation with Ornstein-Ulhenbeck process with $\tau = 10$ ms.

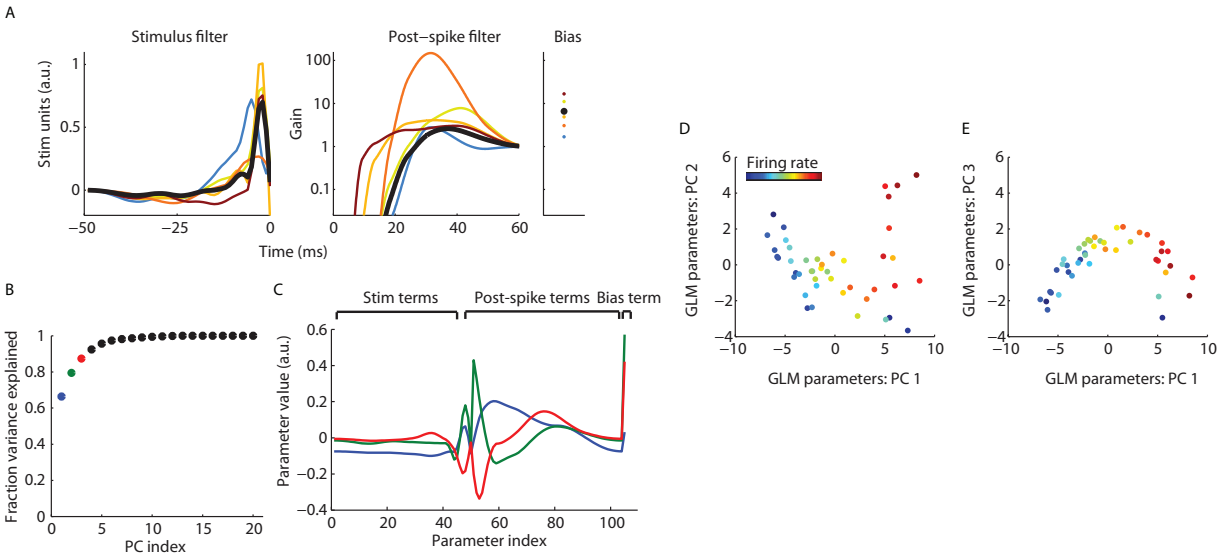

Figure 2.10: *Decomposition of neuronal GLM parameter space into a small number of principal components.* (A) GLM parameters for an example 5-neuron population (colors) and mean GLM parameters across all neurons (black). (B) Percent variance explained for each subsequent principle component (PC). The first 3 principal components explain 85% of variance among GLM parameters. (C) Visualization of the first 3 PCs computed from GLM parameters. (D and E) Projection of neurons (dots) into space defined by PCs 1 and 2 (D) or PCs 1 and 3 (E). This analysis plots neurons such that those with similar GLM parameters are plotted close to one another. The computed PCs largely reflect differences among post-spike and bias terms and to a lesser extent stimulus filters. Neurons with negative PC1 tend to have low firing rates, low amplitude stimulus filters (relative to the mean across neurons) and longer refractory periods with less of a tendency to burst 20-40 ms following a spike and vice versa for positive PC1. Neurons with high PC2 tend to have high baseline excitability, very short refractory periods and increased amplitude stimulus filters relative to the mean.

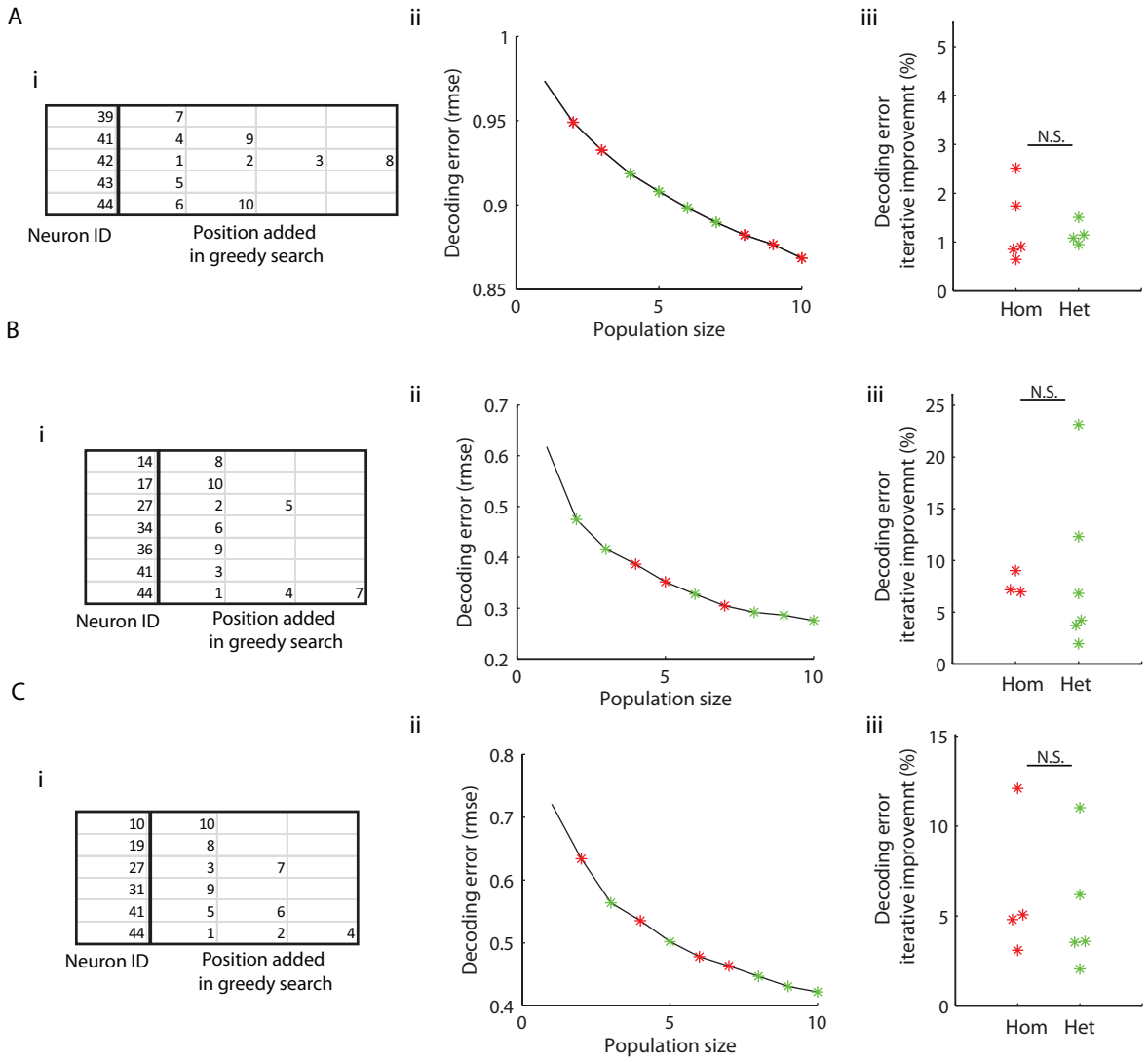

Figure 2.11: *Selection of homogeneous neurons during greedy search does not preferentially occur after coding improvements have saturated.* (A) (i) Detailed analysis of neurons added and their selected position during greedy search algorithm. Neuron population selected to best encode a white noise stimulus. (ii) Decoding error as a function of population size during greedy search optimization procedure. Colored asterisks indicate whether neuron added at nth iteration is a homogeneous neuron (red, i.e. a copy of selected neuron is already in the population) or heterogeneous (green). Same population as shown in i. (iii) Same data as ii, plotted as iterative improvement in decoding error when adding an additional neuron, broken down by if neuron added is a homogeneous neuron (red) or heterogeneous neuron (green). There is no significant difference ($p > .05$) in decoding error improvement between homogeneous and heterogeneous groups. (B and C) Same as A, but for a low frequency stimulus (B, white noise convolved with alpha function with $\tau = 10$ ms) or a high frequency stimulus (C, white noise convolved with alpha function with $\tau = 3$ ms).

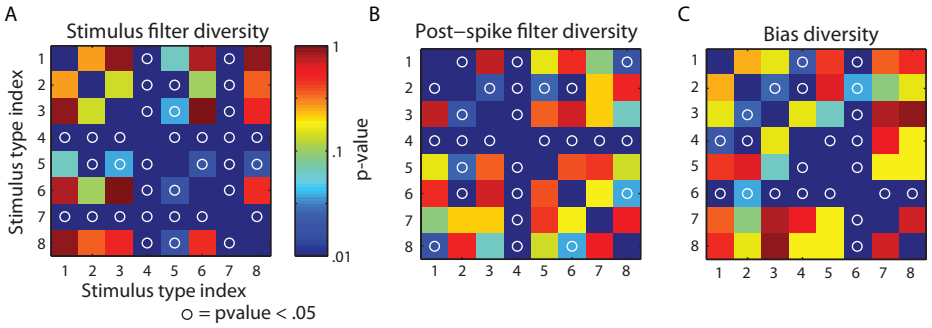

Figure 2.12: *Companion figure to Fig. 4E, showing that populations selected through the greedy search procedure to best represent different stimulus types tend to have different amounts of GLM parameter diversity.* (A-C) Pairwise comparisons between the amounts of population diversity resulting when populations are optimized to best represent different stimulus types. Population diversity values are first computed by running the greedy search algorithm multiple times for each stimulus type (n = 10). Colored squares indicate p-value for statistical test comparing population diversity between each pair of stimulus types. Circles between stimulus pairs indicates that populations are significantly different in terms of their diversity (Wilcoxon, $p < .05$). Stimuli are numbered as in Fig. S5: (1) White noise convolved with alpha function with $\tau = 3$ ms. (2) White noise convolved with alpha function with $\tau = 10$ ms. (3) Ornstein-Ulhenbeck process with $\tau = 10$ ms. (4) Pure white noise stimulus (max frequency = 500 Hz). (5) Ornstein-Ulhenbeck process with $\tau = 20$ ms. (6) Ornstein-Ulhenbeck process with $\tau = 40$ ms. (7) White noise convolved with alpha function with $\tau = 5$ ms. (8) Naturalistic sniffing stimulus generated by combining 8Hz oscillation with Ornstein-Ulhenbeck process with $\tau = 10$ ms.

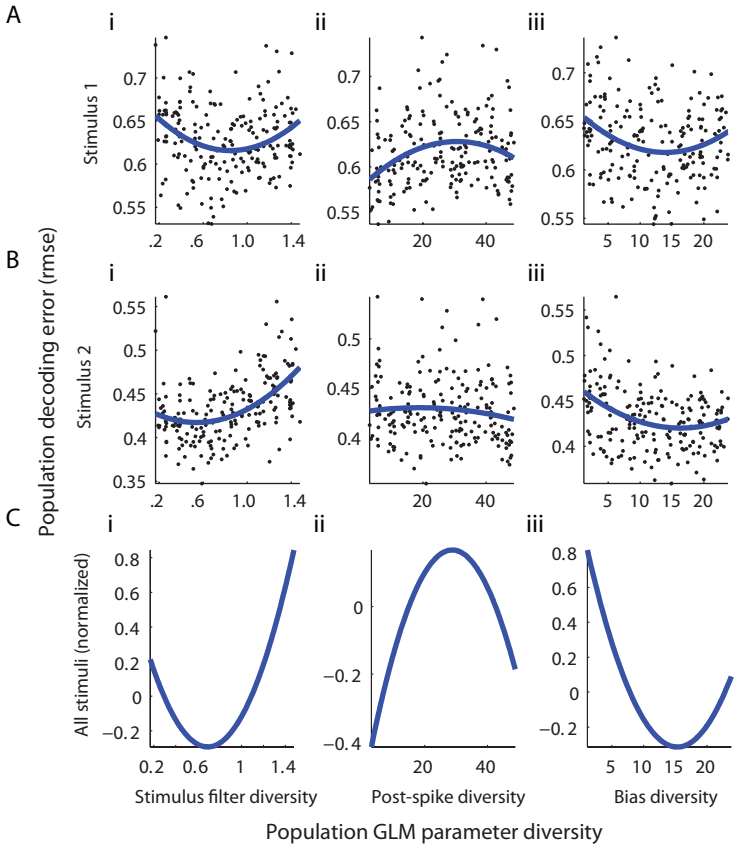

Figure 2.13: *Evidence for a U-shaped relationship between population diversity and decoding error.* (A and B) Decoding error for 5-neuron populations (black dots) as a function of population diversity along stimulus filters (i), post-spike filters (ii), and bias parameters (iii) for stimulus 1 (A, as in Figure 3A) or stimulus 2 (B, as in Figure 3B). For each i-iii, 200 heterogeneous populations were drawn such that populations of varying diversity (from super-diverse through sub-diverse) were sampled with equal probability (see Methods, n = 5 neurons per population). Blue line shows fit of a quadratic polynomial, used to test for expected U-shaped relationship. In all cases, the regression coefficient associated with the quadratic term of the polynomial fit was positive and significant ($p < .01$), except for Aii and Bii, indicating that reconstruction error is minimized at an intermediate values of stimulus filter and bias diversity. The reason why there does not appear to be a concave-up U-shaped relationship for post-spike filters is due to sampling confounds: low post-spike diversity populations tend to have higher firing rates than high post-spike diversity populations. (C) Same as A, but showing U-curves averaged across each of the 8 stimuli. In this case, the decoding error was first normalized to z-scores before performing the regression, allowing comparison across stimuli.

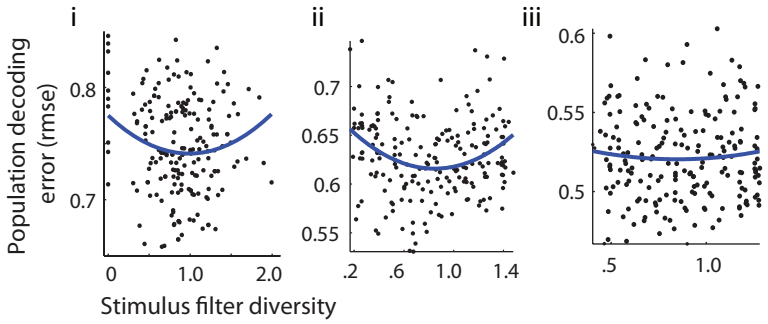

Figure 2.14: *Lack of a substantial U-shaped relationship between population diversity and decoding error for larger sized populations.* Decoding error for neuron populations (black dots) of size n = 2 neurons (i, data reanalyzed from Fig. 2), 5 neurons (ii, same as Fig. 4F), or 10 neurons (iii) per population as a function of population diversity along the GLM stimulus filter dimension for stimulus 1 (as in Figure 3A). For each i-iii, 200 heterogeneous populations were drawn such that populations of varying diversity (from super-diverse through sub-diverse) were sampled with equal probability (see Methods). Blue line shows fit of a quadratic polynomial, used to test for expected U-shaped relationship. The simplest explanation for why the observed U-shaped curve effect gets weaker with more neurons per population is that each of these larger populations have "saturated" in their ability to represent the stimulus. In this case, it matters less whether the populations are diverse (or not) because there are enough neurons in each population to effectively represent the stimulus.

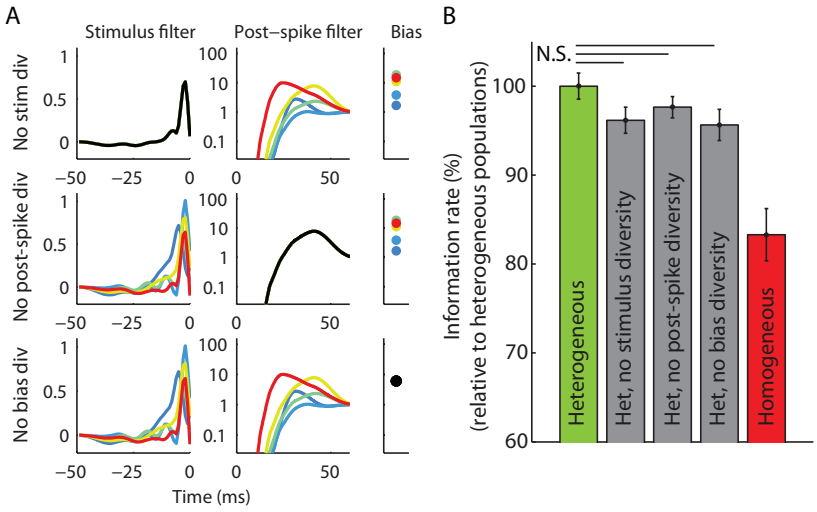

Figure 2.15: *Benefit of neuron variability does not depend on a single GLM model dimension.* (A) Example 5-neuron populations where population variability in a single GLM dimension (stimulus, post-spike, bias) has been eliminated (top, middle, bottom rows respectively). (B) Mean information rates for heterogeneous, homogeneous, and stimulus, post-spike, and bias reduced diversity populations. Information rates computed relative to heterogeneous populations (n = 5 neurons per population, 200 populations per condition, stimulus is high frequency stimulus, white noise convolved with alpha function with $\tau = 3$ ms). None of the reduced diversity populations were significantly different from random heterogeneous populations ($p > .05$, Wilcoxon, N.S.) indicating that the coding benefits of diversity do not rely upon a single GLM dimension. This figure suggests that the representational advantage of neuron variability is not specifically tied to any one of the three GLM dimensions.

Chapter 3

NeuroElectro: A Window to the World's Neurophysiology Data

3.1 Chapter Summary

In this chapter, I extend my study of neuron diversity of mitral cells as discussed in Chapter 2 to biophysical diversity of neuron types throughout the nervous system. Specifically, I describe building a database of basic electrophysiological properties on the majority of mammalian neuron types. Specifically, I focus on methods and strategies that I developed for obtaining this information from the existing research literature using a combination of automated text-mining and manual curation. I also describe the database and web interface developed for the storage and visualization of the electrophysiology dataset. The electrophysiology dataset produced from this work is then subsequently analyzed in chapter 4.

This chapter describes work that will be submitted for publication following the thesis defense. This work was undertaken primarily by myself; with contributions and extensive guidance from Rick Gerkin, a former post-doc in the lab and collaborator, who helped develop the relational database and built the hosting web server; and minor contributions

from Judith Savitskaya, a former undergraduate in the lab, who helped construct the electrophysiology ontology.

3.2 Abstract

The behavior of neural circuits is determined largely by the electrophysiological properties of their neurons. Understanding the roles and relationships of these properties requires the ability to first identify and catalog them. However, information about such properties is largely locked away in decades of closed access journal articles with heterogeneous conventions for reporting results, making it difficult to utilize the underlying data. We solve this problem with the NeuroElectro project: a Python library, RESTful API, and web application (at `http://neuroelectro.org`) for the extraction, visualization, and summarization of data concerning neurons' electrophysiological properties as found in this literature. Information is organized both by neuron type (using neuron definitions provided by NeuroLex) and by electrophysiological property (using a newly developed neurophysiology ontology). We describe the techniques and challenges associated with the automated extraction of tabular electrophysiological data and methodological metadata from journal articles. We further discuss strategies for how to best combine and organize data across these heterogeneous sources. NeuroElectro is a valuable resource for experimental physiologists looking to supplement their own data, for computational modelers looking to constrain their model parameters, and for theoreticians searching for undiscovered relationships among neurons and their properties.

3.3 Related Work

As discussed in Section 1.1.5, neurophysiology lacks a centralized resource where basic physiological measurements across both neuron types and studies are accessible for reference

and subsequent meta-analyses. This is in contrast to neuroanatomical connectivity, where information on connectivity between different brain regions has been compiled by experts at the Brain Architecture Management System project (BAMS) across hundreds of publications Bota et al. (2005). Parallel to this effort is the WhiteText Project, which addresses a complementary goal by algorithmically mining brain region connectivity statements from journal abstracts using biomedical natural language processing (bioNLP) methods (French et al., 2009, 2012). Similarly, in the domain of neuroimaging, the NeuroSynth Project has mined fMRI-based brain activation maps from published x,y,z coordinate data tables from thousands of neuroimaging publications (Yarkoni et al., 2011). These literature-based methods differ from projects such as NeuroMorpho.org (Parekh and Ascoli, 2013), which obtains neuron morphological reconstructions directly from investigators.

These projects are logically divided according to their methods for obtaining the source data: through the use of manual methods like expert curation or user contributions versus automated methods such as text-mining. Notably, these approaches differ in their scale and accuracy; while algorithmic methods can "scale-up" and be applied to arbitrary numbers of publications, they typically have a lower accuracy relative to human-curated content (French et al., 2009). This lower accuracy is often attributed to the rich lexical complexity of biomedical texts which often require considerable context and background knowledge to understand and parse (Dickman, 2003). Given these competing constraints of scale versus accuracy, a challenge has been developing approaches for obtaining neuroscience information that can effectively scale while remaining accurate.

Here, we built a custom infrastructure framework for extracting electrophysiological measurements for specific neuron types from published neurophysiology articles. These measurements included properties such as input resistance and resting membrane potential, as well as associated metadata (i.e., article-specific methodological details). Our methods combine algorithmic literature text-mining, drawing from the approach used by

NeuroSynth (Yarkoni et al., 2011) where neurophysiological measurements are primarily extracted from data tables, and manual curation, leveraging the background knowledge of domain experts. The resulting neurophysiology database, named NeuroElectro, can be interactively viewed and explored through a public web interface at `http://neuroelectro.org`.

3.4 Electrophysiological database construction

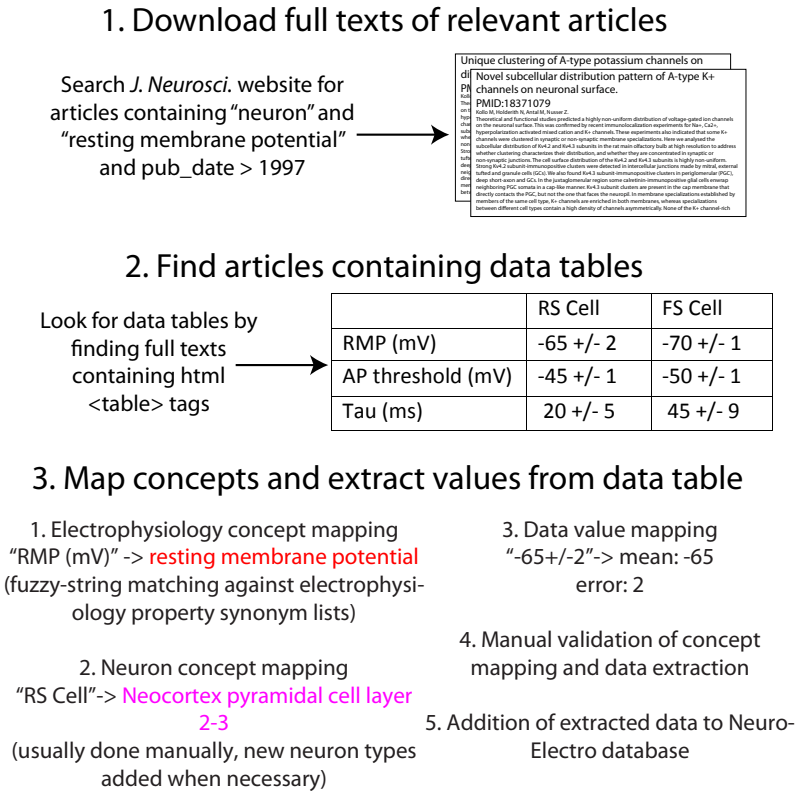

Figure 3.1: Illustration of text-mining workflow

3.4.1 Article identification

We obtained electrophysiological data from 15 neuroscience specific journals, which include: Journal of Neuroscience, Journal of Neurophysiology, and Journal of Physiology, European Journal of Neuroscience, Neuroscience, and Neuron (among others). We selected these

because 1) information on neuronal intrinsic physiology is often published in these articles; and 2) these articles often contain basic information on neuron electrophysiology that may not be explicitly published within other "higher-impact" journals.

In order to apply our text-mining methods to these articles, we first obtained approximately 92,000 full texts of articles directly from publisher websites. We identified potential articles that were likely to contain information relevant to neuron biophysics using the native search functions provided within the journal websites and only downloaded articles containing terms such as "input resistance" or "resting membrane potential". This pre-selection step prevented us from accessing articles that were irrelevant to our project, such as those on neuroimaging or neuroanatomical connectivity. Upon identifying candidate articles, we then downloaded the full text of each potentially-relevant article as HTML (or XML in the case of articles downloaded from the publisher Elsevier's text-mining API, which we then converted to HTML). We chose to work with HTML as opposed to PDF because HTML provides a machine-readable structured markup of the article's content, allowing us to easily identify relevant elements with the article – such as data tables and the methods section – using HTML-parsing algorithms (here we used the Beautiful Soup HTML-processing library implemented in Python: `http://www.crummy.com/software/BeautifulSoup/bs4/doc/`). Furthermore, because HTML is a single standard used by every publisher, we could write relatively generic HTML-processing algorithms applicable to content published across journals. However, our focus on using HTML limits us to relatively newer articles - typically those published after 1997 - because before this time most publications are only available as scanned PDF files.

We stored the text of each article in our database, and mapped each article to its corresponding entry in PubMed `http://www.ncbi.nlm.nih.gov`. Thus we could use PubMed-specific tools such as PubMed's numeric identifier system (i.e. 8-digit PubMed

IDs) as publisher-independent unique identifiers for each article within our database. The use of PubMed also gives us access to PubMed's excellent API (i.e., PubMed eutils, `http://www.ncbi.nlm.nih.gov/books/NBK25500/`), which provides the ability to query each article's MeSH terms (MEdical Subject Headings) and returns basic methodological information such as animal species and strain.

3.4.2 Electrophysiological property identification

Rationale for focusing on electrophysiological property extraction from data tables

In order to algorithmically extract information on neuron electrophysiology from these articles, we needed to first specify the data types of interest. Our preference was to obtain as much detailed information about neuron electrophysiology as possible: ideally raw data corresponding to recorded neuronal membrane voltage traces. In mining information from articles, we were presented with a few options (illustrated in Fig. 3.2), including extraction from: 1) the text of the article; 2) the figures of the article; or 3) data tables presented within the article. Given our preference to obtain data in their most raw form, we considered extraction of data from figures, e.g. scatter plots. However, converting article figure content, presented as an image, and digitizing it into something that can be further analyzed presents multiple challenges. Techniques and tools exist to digitize and parse figures, however substantial amounts of manual effort are required to employ them correctly.

Given the difficulty in automatically extracting raw voltage traces from figures, we instead focused on obtaining information about basic neuronal electrophysiological properties, such as input resistances and resting membrane potentials. Though this information is occasionally presented within the text of the article, these are often presented in complex sentence structures that are difficult to algorithmically parse. Therefore, we instead

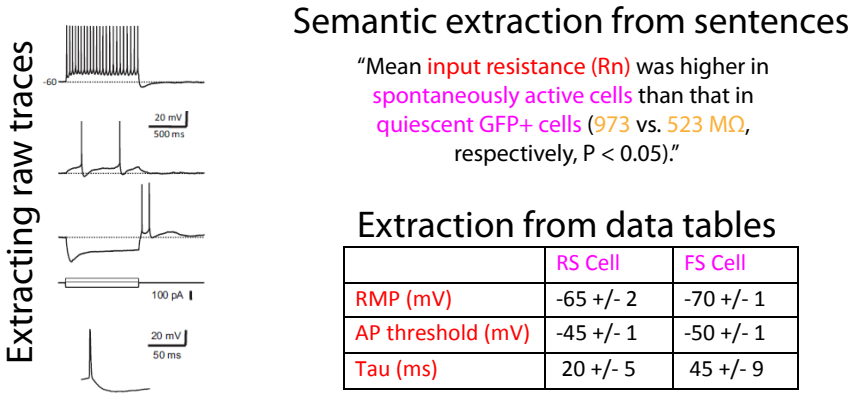

Figure 3.2: *Illustration of the sources within an article containing information relevant to neuron electrophysiology.* We chose to focus on data extraction from data tables.

chose to extract basic physiological information from data tables published within articles. This decision was primarily motivated by the relative ease in extracting information from structured data tables (Yarkoni et al., 2011). However, not all articles on neuronal electrophysiology contain data tables that succinctly present electrophysiological data; by our estimates only 5-10% of such articles contain data tables, which admittedly greatly limits the information that we can extract employing these methods.

Extracting information on electrophysiological properties

In extracting electrophysiological data, as mentioned in Section 1.1.2, we utilized the fact that there are sets of common, informal protocols that are applied to most neurons during intracellular recordings (Connors et al., 1982). For example, it is common to note each neuron's resting membrane potential when whole-cell access is achieved; to apply hyperpolarizing current injections for measurement of input resistance and membrane time constant; and to apply depolarizing current steps to evoke action potentials (spikes) and enable measurement of spiking characteristics like current and voltage threshold, spike width, and spike amplitude.

We developed an electrophysiological lexicon comprising 28 measurements that we found to be commonly reported in the literature (largely based on definitions found in

(Toledo-Rodriguez et al., 2004)). To account for subtle differences in terminology that authors use to refer to the same electrophysiological concept (e.g. resting membrane potential is often referred to as "rmp" as well as "V_{rest}"), we also identified a common list of synonyms to map to each concept. Together, these electrophysiological concepts and their synonyms define a formal ontology for electrophysiological concepts (Table 3.1).

To identify data corresponding to electrophysiological properties reported within a data table, we developed algorithms to search data table header elements and assess whether these elements corresponded to any of the electrophysiological concept synonyms in our ontology. We first identified table header elements by searching for table elements composed primarily of non-numeric characters. For each putative header element, we then used fuzzy string matching algorithms (implemented using the fuzzywuzzy library in Python: `https://github.com/seatgeek/fuzzywuzzy`), to assess the textual match between the header element and each of the electrophysiological synonyms. The fuzzy string matching algorithms employed leverage and combine a number of metrics, including assessing whether the pair of character strings: completely match; partially match; have substrings which completely or partially match; and have partially matching substrings, but which differ in substring order. If the table header and electrophysiological synonym match was higher than a specified value, the table header (as well as corresponding row or column of numeric values) was automatically indicated as corresponding to the electrophysiological concept.

We then manually corrected cases where these algorithms misassigned an electrophysiological concept. For example, a common algorithmic mis-assignment was the case when an author used the string "EPSP amplitude" to refer to the electrophysiological concept excitatory post-synaptic potential amplitude; in these cases, our algorithms incorrectly mapped this string to "spike amplitude" because the former concept is not in our current ontology. Quantifying the accuracy of electrophysiological identification in a subset of ar-

Table 1.
Comparison of electrophysiological properties in adult +/+ and stg/stg in deep layer cortical neurons

	+/+ Concept: Neocortex pyramidal cell layer 5-6	stg/stg
V_r, mV Concept: resting membrane potential	−74.4 ± 1.5 (25)	−73.7 ± 1.7 (27)
R_{in}, MΩ Concept: input resistance	170 ± 25 (20)	170 ± 13 (22)
Time constant, ms Concept: membrane time constant	26.9 ± 2.6 (14)	32.1 ± 3.1 (22)
AP overshoot, mV Concept: spike overshoot	37.0 ± 3.7	34.1 ± 3.14 [11–58] (23)

Figure 3.3: *Example data table illustrating mark-up and annotation of entities.* Markups in red and pink indicate electrophysiological and neuron type concepts and yellow indicates extracted data measurements. Example data table from Pasquale et al. (1997). Note that here the textual string "+/+" refers to the wild-type condition.

ticles which we manually validated through expert curation (335 articles total), we found that electrophysiological concepts were identified correctly with no supervision in 70% of cases (901 of 1292 total).

Accounting for differences in electrophysiological definitions across investigators

By focusing on textually matching the electrophysiological terms in each table to a list of electrophysiological concepts, we are implicitly assuming that electrophysiological properties are measured in the same way by authors across different articles. For example, the most common way that electrophysiologists use to measure a neuron's spike amplitude is to record from the neuron in current-clamp and then apply depolarizing currents at or near the neuron's rheobase to evoke spikes. Spike amplitude is then most commonly measured by calculating the difference between the neuron's voltage at spike threshold and spike peak for the first evoked spike (Toledo-Rodriguez et al., 2004). However, authors can also use different experimental protocols to measure the spike amplitude, like evoking spikes using current steps much greater than rheobase current ("protocol differences"). Additionally, the spike amplitude itself can be calculated in different ways, such as using the neuron's resting membrane potential as the baseline instead of the spike threshold ("calculation differences"). Furthermore, the value of spike amplitude that an author reports will also be affected by specific experimental conditions such as the animal species or age ("condition

differences").

When manually curating the text-mined content for the most commonly reported electrophysiological properties (resting membrane potential, input resistance, membrane time constant, spike half-width, spike amplitude, and spike threshold), we took care to account for and remove cases where the author had calculated an electrophysiological measurement using an inconsistent methodology (e.g., protocol or calculation differences). However we note that we could not identify all of these cases (in particular for spike amplitude, input resistance, and membrane time constant), in part because occasionally authors did not explicitly define how electrophysiological properties were calculated within their article. We note that in cases where we pool measurements which are measured using inconsistent protocols or calculations, this will tend to add noise and unexplained variance to our data set. Given these measurement inconsistencies, we provide our recommendations for how these electrophysiological properties should be reported in future investigations via our electrophysiology ontology (Table 3.1).

3.4.3 Neuron type identification

Using neuron types defined by NeuroLex

To extract physiological information specific to individual neuron types, we had to identify which neuron types were recorded in each article. However, in many cases uniquely identifying the neuron type(s) recorded in any given study and mapping these to a canonical "neuron type" is difficult. This difficulty arises in part because as mentioned in Section 1.1.4, investigators use different criteria for classifying neurons, including electrophysiological, morphological, or molecular characteristics (Ascoli et al., 2008).

We chose to use the externally sourced list of approximately 250 expert-defined neuron types and definitions provided by NeuroLex `http://neurolex.org`; (Larson and Martone, 2013; Hamilton et al., 2012)). Thus we did not have to perform the laborious task

of defining each of the neuron types found in the brain and instead could rely on the collective expertise provided by this community-generated resource. NeuroLex also provides synonyms for each neuron type, which we utilized to identify the neuron type(s) in each article. In cases where a neuron type was investigated in the literature across multiple articles but not indexed within NeuroLex (e.g. cerebellar nucleus neurons), we manually added this neuron type to our database's listing of neuron types and provided this neuron type to the NeuroLex neuron curators for incorporation (Larson and Martone, 2013). Our specific criteria for identifying each of the neuron types reflected in the database are given in Table 3.3.

Identifying specific mentions of neuron types within an article

Because of the complexity in unambiguously identifying neuron types, we primarily used manual methods to map the neuron types mentioned in each article to canonical neuron types. To aid the manual process of selecting neuron types, we used text-mining algorithms to provide a first-pass "best guess" of the most likely neuron type. Specifically, we used a bag-of-words approach (Aldous, 1985) on the article full text (i.e. ignoring the serial structure of the words in the document and looking only at the frequency of occurrence of each word within the document) and compared the word-frequency histogram to the listing of neuron synonyms provided by NeuroLex, ranking all neuron types by their likelihood of being considered within that article. We found that this simple bag-of-words approach was often sufficient to identify the neuron types associated within each article. Quantifying this method against the articles that we manually curated, we found that this approach accurately identified the neuron recorded from in each publication with an accuracy of 30% (120 of 399 total) and up to 55% when defining success as the studied neuron appearing as one of the top three neuron types suggested by the bag-of-words method.

The relatively low accuracy of this approach suggests considerable room for improve-

ment. For example, we note this approach was particularly ineffective when the neuron type investigated within an article was not already described in NeuroLex or when the neuron had an insufficient listing of associated common synonyms. Moreover, we did not consider the incorporation of common neuron type acronyms here (e.g., that olfactory bulb mitral cells are commonly referred to as "MCs"); doing so would likely would increase the accuracy of this approach in future iterations. We mapped whole rows or columns of a data table to specific neuron types, and also manually identified the rows or columns of such tables corresponding to data collected during normotypic or "control" conditions, where applicable. We note that our current neuron identification pipeline requires multiple manual steps, which limits the scalability of the current approach.

3.4.4 Extraction of electrophysiological data values

After identifying specific electrophysiological properties and neuron types reported in a data table (corresponding to row or column table headers), we then algorithmically extracted the data corresponding to the table intersection of these (Fig. 3.3). We developed custom string regular expressions (Thompson, 1968) to parse the string corresponding to the numeric data. Specifically, we found that data strings were often of the form: "XX +/- YY (ZZ)", where XX, YY, and ZZ refer to the mean, error term (either standard error of the mean or standard deviation, which we do not currently disambiguate but will do in future work), and sample size (i.e. the "n") respectively. Often the number of replicates or error measurement were not reported or were reported in alternative ways within the table. When designing our processing algorithms, we parsed data strings from right to left: first searching for data entities contained within parentheses, then for entities contained to the right of the +/- term, and finally the remaining term which we assumed to refer to the mean term. We also found that occasionally data were reported as "XX (LL - HH)" – where LL and HH indicate the lower and upper limits of a data range – and accounted for

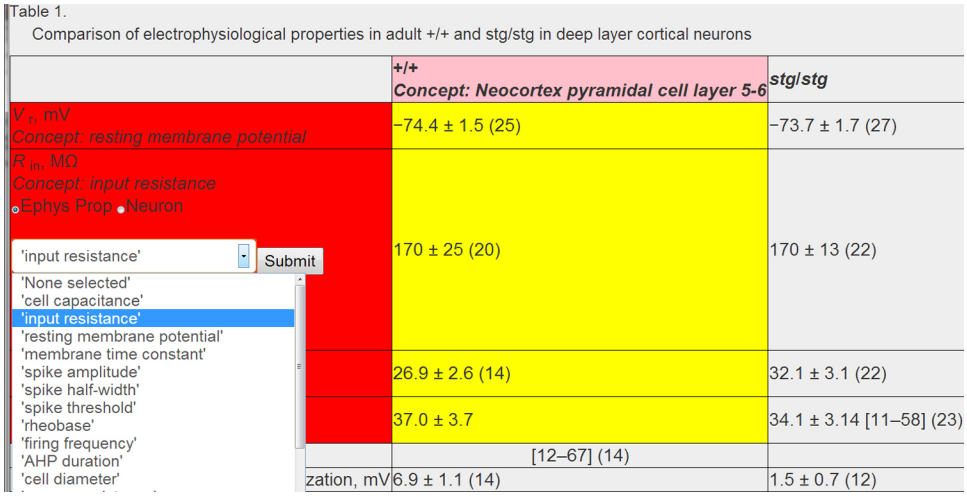

Figure 3.4: *Example of human validation annotation process.* All textual elements of table are enhanced using HTML and javascript to allow for assignment of neuron or electrophysiological concepts using drop down menus. Example data table from Pasquale et al. (1997).

these cases similarly.

We used regular expressions to identify entities such as digits, decimal signs, parentheses, and +/- signs. We then converted the individual data elements which were encoded as textual strings of digits to double precision decimal entities before storing these into our database. Our focus here was primarily on parsing the data mean value, but we also extracted and stored the error term and sample size where possible. Using these methods, we were able to extract 2344 electrophysiological values from 98 distinct neuron types within 335 articles.

3.4.5 Manual validation of automated data extraction

Following these automated concept identification and data extraction steps, we manually validated associated concepts and fixed incorrect concept mappings as necessary. We developed custom-HTML and javascript code to allow human users to graphically interact with downloaded HTML data tables and semantically annotate or "mark-up" entities within the table (Fig. 3.4). This code allows for textual based elements of the HTML table to be annotated using drop down menus and text fields.

3.4.6 Metadata identification

We identified information on article-specific experimental conditions like animal species or recording temperature by extracting this information – primarilty from each article's materials and methods section – using related semi-automated methodology. For each article, we found the methods section by developing custom HTML tag filters for each journal. For each metadata entity we wished to extract (species, animal strain, electrode type, preparation type, junction potential correction, animal age, recording temperature), we devised custom text searching methods to identify these based on combining regular expressions (Thompson, 1968) with PubMed MeSH terms (Table 3.2). In other words, rather than taking a machine-learning based approach and training classifiers (Mccallum, 2002; French et al., 2009), we took a rule-based approach and developed custom rules for identifying metadata entities. For example, to identify whether the electrode's liquid junction potential was corrected for in the study (Neher, 1992), we searched for whether the character string "junction potential" was mentioned within the methods section and if so, whether the sentence or phrase containing the term was explicitly negated (indicating that the junction potential was not corrected for). To identify distinct sentences, we used natural language processing tools provided within the Natural Language Tool Kit in Python (Bird et al., 2009).

Following automated identification of article metadata, we manually checked each article to ascertain that algorithmically-tagged metadata was identified correctly and, as before, we corrected misidentified content as necessary through the use of custom HTML forms. We found that the mean accuracy of algorithmic metadata assignment was approximately 50% (Fig. 3.5) and was typically lower for identifying continuous metadata (e.g., animal age or recording temperature) relative to nominal metadata such as species and electrode type.

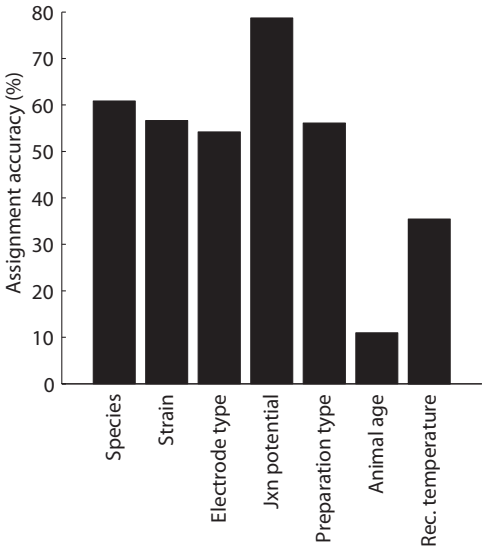

Figure 3.5: *Accuracy of metadata assignment using automated methods alone.*

3.4.7 Object models and relational database

We stored extracted data and metadata using a relational database implemented in MySQL (`http://dev.mysql.com/doc/refman/5.6/en/`) built from a Python Django object model (`https://www.djangoproject.com/`). The object model contains classes for a number of fields, including article full texts, electrophysiological properties, neuron types, synonyms, electrophysiological data values, and experimental metadata (among others; shown in Fig. 3.6). A useful feature of the relational nature of the database is that it enables linking between classes (e.g., representing that articles are written by authors who in turn can write multiple articles). This linking feature facilitates efficient and arbitrary querying of data; for example, querying for known electrophysiological data on olfactory bulb mitral cells recorded *in vitro* and published between the dates 2000 and 2004.

3.4.8 Web application

The primary results of the NeuroElectro database are viewable at `www.neuroelectro.org` where the data can be interactively explored. Furthermore, full API access to the electro-

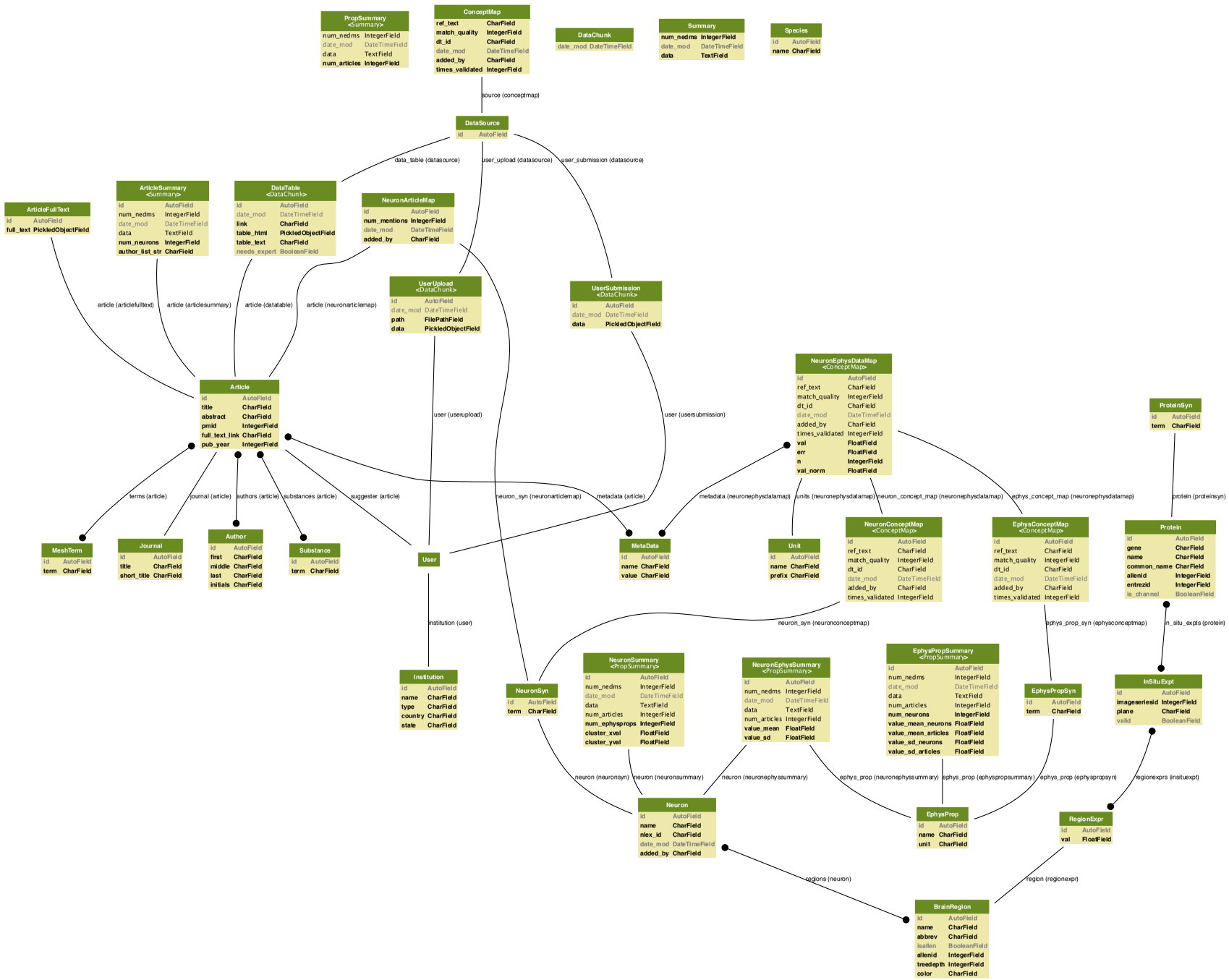

Figure 3.6: Illustration of NeuroElectro relational database schema

physiological data is described at `www.neuroelectro.org/api/docs/`. The web interface is organized around neuron types and electrophysiological properties. For example, each neuron type has its own webpage where extracted data corresponding to specific electrophysiological properties is graphically and interactively displayed. Users can thus visualize the mean and variability of electrophysiological values across papers, view references, and easily navigate to primary data from specific papers. Furthermore, users can view electrophysiological data across all of the neuron types in the database - putting phenotypic properties of a given neuron type into the larger context of other neuron types located throughout the nervous system.

The web application also contains preliminary features to allow website visitors to contribute to the NeuroElectro resource. For example, users can suggest articles which contain neurophysiological data which are not already in the database. Furthermore, we invite visitors to become "expert curators" for neurons of interest. In the future, we plan to build functionality that will allow investigators to upload raw and summary data, such as recorded voltage and current traces.

3.5 Discussion

3.5.1 Summary

We have developed, applied, and validated a methodology for extracting – from existing literature on cellular neurophysiology – measurements of basic biophysical properties from diverse neuron types throughout the nervous system. Currently, the NeuroElectro database contains 2344 manually curated electrophysiological measurements from 98 neuron types from 335 publications. Of these electrophysiological measurements, 2176 were obtained from 279 publications using the semi-automated approach described here. In addition, we machine-extracted and manually validated 1667 methodological conditions from these

publications.

3.5.2 Specific benefits provided by the semi-automated approach

One of the key advantages of the approach described here is that the automated pipeline adequately identifies publications which are likely to contain content relevant to our domain area (i.e., machine-readable measurements of neuronal biophysics). Thus a human needs only to manually curate the content first identified by the algorithms as being likely relevant, instead of having to identify the relevant content *de novo*. Moreover, the automated identification of neuron types in articles allows us to target manual curation effort to publications likely to contain data from specific neuron types, such as neurons that are currently underrepresented in the database.

Given our laboratory's focus on olfactory circuits (as illustrated in Chapter 2), we conducted a natural experiment to compare the efficacy of biophysical property extraction using these semi-automated methods versus traditional methods which do not make use of algorithmic text-mining pre-processing. In a seven-hour curation session, a member of our laboratory identified 91 electrophysiological measurements in 35 articles from 7 olfactory bulb neuron types using only prior knowledge on which articles were likely to contain such information. In a comparable seven-hour curation session using our semi-automated methods, a single curator identified 551 electrophysiological measurements from 70 articles across 40 neuron types throughout the nervous system.

3.5.3 Scalability of current approach

We note that multiple steps in our approach require manual intervention by a expert curator for electrophysiological measurements to be extracted with a negligible error rate. Namely, an expert curator needs to specify which neuron types are recorded from in each article and where data from the normotypic or "control" states of these neurons are textually

referenced within a data table. Moreover, given the current accuracy of the unsurpervised algorithmic assignment of electrophysiological concepts and experimental metadata (70% and 50% respectively), these also need to be manually validated and corrected as required by an expert. Because of the necessity of these manual steps for accurate data extraction, the scalability of our current approach is limited by our ability to manually curate this information. Despite this bottleneck, our total pipeline is still much faster and more effective than a purely manual one.

3.5.4 Extensions and improvements to the current semi-automated algorithms

We feel that among the most beneficial improvements to our current approach would be improving the accuracy of neuron type identification. Given the relative unspecificity of the current bag-of-words approach, a bioNLP classifier-based approach may prove more effective (Mccallum, 2002). Specifically, we propose adapting the methodology used by the WhiteText project for tagging brain regions mentioned in literature (French et al., 2009; French and Pavlidis, 2012) and first identifying spans of text likely to pertain to a neuron type before mapping these spans to a individual neuron type within the neuron ontology. Moreover, such an approach could easily incorporate techniques for expanding acronyms and abbreviations for neuron types (Okazaki and Ananiadou, 2006).

The approach described here is highly effective for extracting biophysical measurements presented within machine-readable data tables which are published within journal articles. However, the current requirement that these data tables exist in a machine parse-able format, such as HTML or XML, limits this approach from being directly applied to older manuscripts, which are only available as scanned images. However, existing approaches, such as optical character recognition technology (OCR) may be applied toward this problem (Ramakrishnan et al., 2012).

A more pressing issue with the current approach is its focus on extraction from data tables. While the approach is highly effective on data tables, the majority of neurophysiology publications do not summarize electrophysiological data within a published data table (approximately 5-10% of articles contain data tables). Instead, this information is usually presented within article text or within figure images. While automatically extracting measurements from figure images will likely prove challenging, we feel that our methods can be easily adapted to operate on article text, perhaps using a similar bioNLP classifier-based approach as suggested for neuron type identification. Furthermore, experimentalists who would like their data to be easily curated should consider using data tables within future publications.

3.6 Supplemental Tables

Table 3.1: Ontology describing electrophysiological properties Should also contain Neuroelectro Ids

Electrophysiological term	Definition (most common)	units (standard)
spike threshold	Voltage at which AP is initiated (as measured by looking at the rate of rise of Vm)	mV
AHP duration	Time to the maximal afterhyperpolarization following an AP	ms
membrane time constant	Time constant for the membrane to repolarize after a small hyperpolarizing current injection of fixed amplitude and duration	ms
FI slope	The slope of the current-discharge relationship from discharge threshold	Hz/nA
access resistance	Sum of the electrode resistance and the resistance at the electrode-cell junction	
sag ratio	Ratio between exponentially extrapolated voltage and steady-state voltage	
cell capacitance	Neuron capacitance, typically measured by dividing membrane time constant by membrane resistance	pF
resting membrane potential	Membrane potential at the onset of whole-cell recording	mV
input resistance	Input resistance at steady-state (steady-state of voltage response to current injection)	MΩ
spike width	Duration of AP, not explictly refered to as half-width	ms
slow afterhyperpolarization duration	Duration from first AP onset to minimum voltage, explictly refered to as slow	ms
fast afterhyperpolarization duration	Duration from first AP onset to minimum voltage, explictly refered to as fast	ms
ADP duration	Duration from first AP onset to maximum ADP	ms
spike overshoot	Difference between the peak of the action potential and 0 mV	mV
cell diameter	Diameter of the cell soma	μm
rheobase	Current threshold to discharge APs during a ramp depolarization. Minimum current required to fire an action potential.	pA
cell surface area	Cross-sectional area of the cell	$\mu\hat{\text{m}}\hat{2}$
spike half-width	Average time for first AP half amp to the same voltage during offset	ms
spike amplitude	Average amplitude of the first AP (measured from AP threshold to AP peak)	mV
Spontaneous firing rate	AP discharge rate in the absence of current injection or a stimulus	Hz
firing frequency	AP discharge rate	Hz
AHP amplitude	Amplitude from first AP onset to minimum voltage, not explicitly fast or slow	mV
slow afterhyperpolarization amplitude	Amplitude from first AP onset to minimum voltage, explictly refered to as slow	mV
fast afterhyperpolarization amplitude	Amplitude from first AP onset to minimum voltage, explictly refered to as fast	mV
ADP amplitude	Amplitude from first AP onset to maximum voltage, typically more depolarized that the resting membrane potential	mV

Table 3.2: Table of metadata attributes and values and summary of how values are extracted

Metadata Concept	Values	Accuracy (%)	Extraction method	Regular Expresssion	MeSH Term
Species		61	MeSH Term only		
	Rats				Rats
	Mice				Mice
	Guinea Pigs				Guinea Pigs
Electrode Type		54	MeSH Term + Regex		
	Patch-clamp			"whole cell" or "patch clamp"	Patch-Clamp Techniques
	sharp			"sharp electrode"	
Animal Strain		56	MeSH Term only		
	Fischer 344				Rats, Inbred F344
	Long-Evans				Rats, Long-Evans
	Sprague-Dawley				Rats, Sprague-Dawley
	Wistar				Rats, Wistar
	C57BL				Mice, Inbred C57BL
	BALB C				Mice, Inbred BALB C
Prep Type		56	MeSH Term + Regex		
	in vitro			"slice" or "in vitro"	
	in vivo			"in vivo"	
	cell culture			"culture"	Cell Culture Techniques
	model			"model"	Computer Simulation
Junction Potential		78	Regex		
	Not Corrected			"not... junction potential" or "no ...junction potential"	
	Corrected			"junction potential"	
Recording Temperature		35	Regex		
	Continuous value			Find digits near "record ... °C" or "experiment ... °C"	
	Room temperature			"record... room temperature" or "experiment ... room temperature"	
Animal Age		11	Regex		
	Continuous value			Find digits near: "P#-#" or "P#-P#" or "day old" or "age... week"	

Table 3.3: Neuron type defining criteria and NeuroLex and NeuroElectro IDs

Neuron Type	Defining Criteria	NeuroElectro Link	NeuroElectro ID	NeuroLex ID
Abducens nucleus motor neuron	All motoneurons in the abducens nucleus	http://neuroelectro.org/neuron/1/	1	nlx_16848
Amygdala basolateral nucleus pyramidal neuron	Pyramidal or principal cells in the BLA	http://neuroelectro.org/neuron/2/	2	nifext_152
Amygdala corticomedial nucleus pyramidal cell	Pyramidal cells in the cortical amygdala	http://neuroelectro.org/neuron/3/	3	nifext_151
Amygdaloid nucleus paracapsular intercalated cell	All neurons found in the Amygdaloid nucleus paracapsular intercalated nucleus	http://neuroelectro.org/neuron/4/	4	nlx_cell_100202
Basalis nucleus cholinergic neuron	Cholingergic neurons found in the nucleus basalis	http://neuroelectro.org/neuron/14/	14	nlx_cell_20090203
BNST beaded neuron	BNST neuron, explicitly referred to as not a projection neuron	http://neuroelectro.org/neuron/6/	6	BAMSC995
BNST common spiny neuron	BNST neuron, projection neuron	http://neuroelectro.org/neuron/7/	7	BAMSC991
Cerebellar nucleus cell	Any neuron found in any of the cerebellar nuclei	http://neuroelectro.org/neuron/208/	208	nlx_151895
Cerebellum Golgi cell	Explicitly referred to as Golgi cell in the cerebellum	http://neuroelectro.org/neuron/16/	16	sao1415726815
Cerebellum granule cell	Explicitly referred to as granule cell in the cerebellum	http://neuroelectro.org/neuron/21/	21	nifext_128
Cerebellum Purkinje cell	Explicitly referred to as Purkinje cell in the cerebellum	http://neuroelectro.org/neuron/18/	18	sao471801888
Cochlea hair cell inner	Inner hair cell in the cochlea	http://neuroelectro.org/neuron/27/	27	sao429277527
Cochlear nucleus (dorsal) cartwheel cell	Explicitly referred to as cartwheel cell	http://neuroelectro.org/neuron/29/	29	nifext_76
Cochlear nucleus (dorsal) pyramidal neuron	Explicitly referred to as pyramidal cell in DCN	http://neuroelectro.org/neuron/34/	34	nifext_74
Cochlear nucleus (ventral) bushy cell	Explicitly referred to as bushy cell	http://neuroelectro.org/neuron/37/	37	nlx_cell_20081201
Cochlear nucleus (ventral) multipolar T cell	Explicitly referred to as T cell	http://neuroelectro.org/neuron/38/	38	nifext_68
Cochlear nucleus (ventral) octopus cell	Explicitly referred to as octopus cell	http://neuroelectro.org/neuron/40/	40	nifext_72
Dentate gyrus basket cell	Basket cells found in the dentate gyrus or hilar region	http://neuroelectro.org/neuron/65/	65	nlx_cell_100201
Dentate gyrus granule cell	Granule cells in the dentate gyrus	http://neuroelectro.org/neuron/66/	66	nlx_anat_1008005
Dentate gyrus hilar cell	Interneurons found in the hilus	http://neuroelectro.org/neuron/67/	67	nlx_cell_20090727
Dentate gyrus mossy cell	Mossy cells found in the dentate gyrus or hilus	http://neuroelectro.org/neuron/68/	68	nlx_22799
Dorsal motor nucleus of vagus motor neuron	Any neuron found in the dorsal motor nucleus of vagus	http://neuroelectro.org/neuron/71/	71	nlx_38336
Dorsal root ganglion cell	Any neuron explicitly referred to as DRG neurons, all sizes	http://neuroelectro.org/neuron/72/	72	nifext_84
DRG temperature cell	Any neuron explicitly referred to as temperature senstive DRG neurons	http://neuroelectro.org/neuron/57/	57	nifext_90
Entorhinal cortex layer IV neuron	layer IV neuron in entorhinal cortex	http://neuroelectro.org/neuron/222/	222	nlx_cell_20090310
Globus pallidus principal cell	Any neuron in the globus pallidus	http://neuroelectro.org/neuron/78/	78	nifext_149
Hippocampus CA1 basket cell	Explicitly referred to as basket cell, PV+ cell, or FS cell in CA1	http://neuroelectro.org/neuron/82/	82	nlx_cell_091205
Hippocampus CA1 ivy neuron	Explicitly referred to as ivy cell in CA1	http://neuroelectro.org/neuron/210/	210	nlx_35220
Hippocampus CA1 neurogliaform cell	Explicitly referred to as neurogliaform cell in CA1	http://neuroelectro.org/neuron/83/	83	nifext_60
Hippocampus CA1 oriens lacunosum moleculare neuron	Explicitly referred to as OLM cell in CA1	http://neuroelectro.org/neuron/84/	84	nlx_cell_091206
Hippocampus CA1 pyramidal cell	Explicitly referred to as pyramidal cell in CA1	http://neuroelectro.org/neuron/85/	85	sao830368389
Hippocampus CA1 radiatum giant cell	Explicitly referred to as radiatum giant cell in CA1	http://neuroelectro.org/neuron/226/	226	
Hippocampus CA3 basket cell	Explicitly referred to as basket cell, PV+ cell, or FS cell in CA3	http://neuroelectro.org/neuron/230/	230	nlx_cell_091213
Hippocampus CA3 lacunosum moleculare neuron	Explicitly referred to as OLM cell in CA3	http://neuroelectro.org/neuron/229/	229	nlx_cell_091216
Hippocampus CA3 pyramidal cell	Explicitly referred to as pyramidal cell in CA3	http://neuroelectro.org/neuron/89/	89	nlx_36816
Hippocampus CA3 stratum radiatum giant cell	Explicitly referred to as radiatum giant cell in CA3	http://neuroelectro.org/neuron/232/	232	
Hippocampus CA3 trilaminar neuron	Explicitly referred to as trilaminar cell in CA3	http://neuroelectro.org/neuron/231/	231	
Hypoglossal nucleus motor neuron	Any motor neuron found in the hypoglossal nucleus	http://neuroelectro.org/neuron/91/	91	nlx_cell_100311
Hypothalamus oxytocin neuroendocrine magnocellular cell	Any magnocellular neurosecretory cell in the hypothalamus or supraoptic nucleus	http://neuroelectro.org/neuron/92/	92	nlx_416
Inferior colliculus neuron	Any neuron in the inferior colliculus	http://neuroelectro.org/neuron/207/	207	nlx_152522
Inferior olive neuron	Any neuron in the inferior olive	http://neuroelectro.org/neuron/219/	219	
Lateral amygdala projection neuron	A projection or pyramidal neuron in the lateral amygdala; not in the BLA	http://neuroelectro.org/neuron/211/	211	
Locus coeruleus NA neuron	Any noradrenergic neuron in the locus coeruleus	http://neuroelectro.org/neuron/94/	94	nlx_cell_20090202
Medial entorhinal cortex layer II stellate cell	Medial entorhinal cortex layer II stellate cell	http://neuroelectro.org/neuron/206/	206	nlx_36209
Medial entorhinal cortex layer III pyramidal cell	Medial entorhinal cortex layer III pyramidal cell	http://neuroelectro.org/neuron/218/	218	
Medial Nucleus of Trapezoid Body neuron	Any neuron found in the MNTB	http://neuroelectro.org/neuron/214/	214	nifext_79
Medial vestibular nucleus neuron	Any MVN neuron	http://neuroelectro.org/neuron/223/	223	
Neocortex basket cell	Explicitly referred to as basket cell, PV+ cell, or FS cell in any layer or subregion of neocortex	http://neuroelectro.org/neuron/99/	99	nifext_56
Neocortex bipolar neuron	Explicitly referred to as bipolar cell in any layer or subregion of neocortex	http://neuroelectro.org/neuron/101/	101	sao436474611
Neocortex chandelier cell	Explicitly referred to as chandelier cell in any layer or subregion of neocortex	http://neuroelectro.org/neuron/104/	104	nifext_57
Neocortex layer 4 stellate cell	Any excitatory cell in layer 4 of the neocortex; usually stellate cells in primary sensory areas	http://neuroelectro.org/neuron/107/	107	nifext_53
Neocortex Martinotti cell	Explicitly referred to as Martinotti cell or somatostatin positive or Low-threshold spiking cell in neocortex	http://neuroelectro.org/neuron/98/	98	nifext_55
Neocortex other cell	Any cell in neocortex that could not be classified to an existing neuron type	http://neuroelectro.org/neuron/109/	109	nifext_49
Neocortex pyramidal cell layer 2-3	Pyramidal cells of layer 2/3 of the neocortex; usually referred to as regular spiking cells	http://neuroelectro.org/neuron/110/	110	nifext_49
Neocortex pyramidal cell layer 5-6	Pyramidal cells of layer 5/6 of the neocortex; occassionally referred to as "deep" cells	http://neuroelectro.org/neuron/111/	111	nifext_50
Neostriatum cholinergic cell	Any cholinergic positive cell in the striatum	http://neuroelectro.org/neuron/115/	115	sao1866881837
Neostriatum gabaergic interneuron	Explicitly referred to as gabaergic cell, PV+ cell, or FS cell in striatum	http://neuroelectro.org/neuron/209/	209	nifext_143
Neostriatum medium spiny neuron	Explicitly referred to as MSNs in striatum	http://neuroelectro.org/neuron/117/	117	nifext_141
Neostriatum SOM/NOS cell	Somatostatin/NOS positive striatum cell	http://neuroelectro.org/neuron/113/	113	nifext_144
Neostriatum TH+ cell	Tyrosine hydroxylase positive cell in striatum	http://neuroelectro.org/neuron/114/	114	
Nucleus accumbens core neuron	Explicitly found in the nucleus accumbens core (usually MSNs)	http://neuroelectro.org/neuron/228/	228	nlx_151892
Nucleus accumbens medium spiny neuron	Any MSN found in the nucleus accumbens	http://neuroelectro.org/neuron/220/	220	
Nucleus accumbens shell neuron	Explicitly found in the nucleus accumbens shell (usually MSNs)	http://neuroelectro.org/neuron/227/	227	nlx_151893
Nucleus ambiguus motor neuron	Motor neuron found in the nucleus ambiguus	http://neuroelectro.org/neuron/120/	120	nlx_53276
Nucleus of the solitary tract principal cell	Projection neurons found in the nucleus of the solitary tract	http://neuroelectro.org/neuron/122/	122	nifext_100
Olfactory bulb (main) Blanes cell	Large neurons found in the olfactory bulb granule cell layer, usually referred to as Blanes cells or deep short-axon cells	http://neuroelectro.org/neuron/127/	127	nifext_124
Olfactory bulb (main) external tufted cell	Referred to as external tufted cells, any excitatory neurons with a cell body in the glomerular layer	http://neuroelectro.org/neuron/132/	132	nlx_82555
Olfactory bulb (main) granule cell	Granule cells found in the main olfactory bulb	http://neuroelectro.org/neuron/128/	128	nlx_407
Olfactory bulb (main) mitral cell	Cells in the mitral cell layer of the olfactory bulb	http://neuroelectro.org/neuron/129/	129	nlx_anat_100201
Olfactory bulb (main) periglomerular cell	Referred to as external tufted cells, any excitatory neurons with a cell body in the glomerular layer	http://neuroelectro.org/neuron/130/	130	nlx_cell_091202
Olfactory bulb (main) tufted cell (middle)	Referred to as middle tufted cells in main olfactory bulb	http://neuroelectro.org/neuron/131/	131	nifext_121
Olfactory cortex pyramidal cell	Explicitly referred to as pyramidal cells in the olfactory cortex	http://neuroelectro.org/neuron/135/	135	nifext_139
Olfactory cortex semilunar cell	Explicitly referred to as semilunar cells in the olfactory cortex	http://neuroelectro.org/neuron/136/	136	nlx_cell_091005
Paraventricular hypothalamic nucleus neurons	Any neuron in the paraventricular hypothalamic nucleus neurons	http://neuroelectro.org/neuron/217/	217	
Parvicellular reticular nucleus interneuron	Any interneuron in the Parvicellular reticular nucleus	http://neuroelectro.org/neuron/216/	216	
Spinal cord intermediate horn motor neuron sympathetic	Sympathetic motor neuron in spinal cord	http://neuroelectro.org/neuron/165/	165	nifext_109
Spinal cord ventral horn interneuron IA	IA interneuron in spinal cord	http://neuroelectro.org/neuron/169/	169	nifext_110
Spinal cord ventral horn interneuron II	II interneuron in spinal cord	http://neuroelectro.org/neuron/171/	171	nifext_112
Spinal cord ventral horn interneuron V2	V2 interneuron in spinal cord	http://neuroelectro.org/neuron/175/	175	nlx_cell_100207
Spinal cord ventral horn motor neuron alpha	Alpha motor neuron in spinal cord	http://neuroelectro.org/neuron/177/	177	sao1154704263
Subiculum pyramidal cell	Any pyramidal cell (all subtypes) in the subiculum	http://neuroelectro.org/neuron/182/	182	nlx_anat_1008012
Substantia nigra pars compacta dopaminergic cell	Substantia nigra pars compacta dopaminergic cell	http://neuroelectro.org/neuron/183/	183	nifext_145
Substantia nigra pars reticulata interneuron GABA	Substantia nigra pars reticulata cell, occassional referred to as projection neurons	http://neuroelectro.org/neuron/184/	184	nifext_147
Subthalamic nucleus neuron	Any neuron in the subthalamic nucleus	http://neuroelectro.org/neuron/233/	233	nlx_149137
Superior colliculus superficial layer neuron	Any neuron in superficial layers of the superior colliculus	http://neuroelectro.org/neuron/221/	221	BAMSC1123
Suprachiasmatic nucleus neuron	Any neuron in the suprachiasmatic nucleus	http://neuroelectro.org/neuron/213/	213	nlx_151894
Supratrigeminal nucleus interneuron	Interneurons in the supratrigeminal nucleus	http://neuroelectro.org/neuron/215/	215	
Thalamic reticular nucleus cell	Neurons in the thalamic reticulnar nucleus	http://neuroelectro.org/neuron/190/	190	nifext_45
Thalamus parafascicular nucleus neuron	Neurons in the thalamic parafascicular nucleus	http://neuroelectro.org/neuron/212/	212	
Thalamus relay cell	Neurons in the thalamus, not in the reticular or parafascicular nucleus, usually LGN or MGN or cortical projection neurons	http://neuroelectro.org/neuron/194/	194	nlx_cell_20081203
Trigeminal nucleus motor neuron	Motor neurons in the trigeminal nucleus	http://neuroelectro.org/neuron/199/	199	nlx_44081
Trigeminal nucleus principal cell	Principal cells in the trigeminal nucleus, not referred to as motor neurons	http://neuroelectro.org/neuron/200/	200	nifext_96
Trochlear nucleus motoneurons	Motor neurons in the trochlear nucleus	http://neuroelectro.org/neuron/225/	225	nlx_70345
Ventral tegmental area dopamine neuron	Any dopaminergic neuron found in the VTA	http://neuroelectro.org/neuron/203/	203	nlx_cell_20090305

Chapter 4

A literature-based brain-wide analysis of the electrophysiological diversity of mammalian neurons

4.1 Chapter Summary

In this chapter, I analyze the NeuroElectro database that was constructed in Chapter 3. Specifically, I describe analyses I have performed towards understanding the electrophysiological diversity of neuron types throughout the brain. Because the NeuroElectro database is constructed from data published across hundreds of papers, I found that I needed to account for the fact that these data were collected under different experimental conditions. Using simple regression models, I was able to systematically account for a large fraction of across-experiment measurement variability. I then further analyzed the dataset for the presence of unknown relationships among neuron types on the basis of differences in basic electrophysiological properties. By analyzing neuron types, I show that 1) much of the biophysical differences among neurons is explained by neuron size (i.e. small neurons versus large neurons) and 2) there are approximately 7 functionally-distinct neuron "super-classes"

based on electrophysiological differences. Moreover, I show that these super-classes also correlate with corresponding gene expression differences observed in these neurons.

This chapter describes work that will be submitted for publication shortly following the thesis defense. This work was primarily completed by myself, with substantial intellectual contributions from Nathan Urban, Rick Gerkin, and Shawn Burton. I am also very grateful to Shawn and Matt Geramita in particular, for collecting electrophysiological data to help validate data contained within the NeuroElectro database.

4.2 Abstract

For decades, electrophysiologists have recorded and characterized the biophysical properties of a rich diversity of neuron types. This diversity of neuron types is critical for generating functionally important patterns of brain activity and implementing neural computations. Identifying specific roles for these neuron types, or even determining what these types are remains challenging, however, because the vast collection of electrophysiological data remains scattered throughout the literature. Here, we describe the creation of an expansive and interactive public collection of electrophysiological properties, NeuroElectro, available at `www.neuroelectro.org`. NeuroElectro was initially populated through text-mining and manual curation and contains information about biophysical properties (such as resting membrane potential and input resistance) of 98 neuron types as reported in 326 studies published between 1984 and 2013. Capitalizing on the statistical power offered by this unprecedented collection of electrophysiological data, we show that knowledge of a few key experimental conditions (e.g., recording temperature and animal age) accounts for a large fraction of the variability in values reported across studies. After adjusting for differences in experimental conditions, we find that neurons across the brain can be divided among $\sim$7 fundamental biophysical types, each with a corresponding functional role. These $\sim$7 "super-classes" of neurons align well with previous classification schemes based on neurochemical

and morphological properties. Finally, we take initial steps toward understanding the mechanistic origins of biophysical diversity by integrating NeuroElectro with a brain-wide gene expression dataset.

4.3 Introduction

Electrophysiologists have recorded and published vast amounts of quantitative data about the biophysical properties of neurons across many years of studies. Compared to other fields, however, little progress has been made in compiling and cross-analyzing this data, let alone collecting or depositing measurements or raw traces (Akil et al., 2011). Thus, for example, it is difficult to determine whether a Purkinje cell responds more like a CA1 pyramidal cell or a cortical basket cell without first collecting new data, even though thousands of recordings have been made from these cell types across hundreds of laboratories. By analogy to genetics, imagine if genes needed to be re-sequenced every time an investigator wanted to examine their homology. The lack of a centralized collection of neuron biophysical properties is thus a barrier to comparison and generalization of results across neuron types, and routinely leads to unnecessary replication of experiments and the overall slowing of progress (Akil et al., 2011). Moreover, recent proposals for large scale electrophysiological analyses across brain areas (Human/Blue Brain Project (Markram, 2006); BRAIN initiative (Alivisatos et al., 2012; Insel et al., 2013)) will require such a repository if the data are to be effectively shared and used by the community.

Our specific goals in building a structured repository of neuron biophysical properties are threefold. First, to generate a data-driven *"parts list"* of the brain, providing scientists efficient access to available data on the properties of different neurons types. Currently, these data are impossible to obtain without substantial effort and expertise. Use of such a parts list will further help standardize both the characterization of new neuron types and the comparison of properties between control and manipulated animals. Second, to

aid in the discovery of new knowledge that is, like *"buried treasure"*, available in the vast neuroscience literature but largely inaccessible because it is not compiled, organized, and searchable. With such a resource a scientist can rapidly and efficiently generate or test hypotheses that may have otherwise gone unnoticed (Akil et al., 2011; Voytek and Voytek, 2012). Third, to encourage and enable scientists to *"share data"*, as well as to link their data to existing and emerging resources as they are generated (e.g., the Allen Brain Atlas (Lein et al., 2007)). These linkages will enhance the utility of each individual electrophysiological study and thereby further accelerate discovery in the field. A database of neuron biophysical properties and the linkages that it supports are particularly important given that large scale efforts to map neuronal activity across the brain are in their early stages (Insel et al., 2013; Kandel et al., 2013).

Here, we describe the construction and validation of a public database that aggregates information on key biophysical properties and the experimental conditions under which they were collected for the majority of mammalian neurons in the brain. Our methods use a combination of automated text-mining (French et al., 2009; Ambert and Cohen, 2012) and expert manual curation to extract relevant information from the existing literature. After populating the database, we assess how certain experimental conditions systematically influence electrophysiological measurements across neuron types. We then explore the emergence of both intuitive and unexpected groups of neuron types according to commonalities in their biophysical properties.

Though NeuroElectro is not yet comprehensive, the framework provides a shared infrastructure for data that can be used to facilitate comparison of neuron types across studies and laboratories. Clearly, NeuroElectro will become more useful as more data are included, but we demonstrate that powerful inferences can be generated even in its current state. Just as the GenBank database (Benson et al., 2013) and BLAST algorithm (Altschul et al., 1990) have enabled biologists to infer protein function and binding patterns from compar-

isons of genetic sequences (Zhou, 2004; Flower and Attwood, 2004; Bairoch and Apweiler, 2000), we envision the development of analogous tools that enable electrophysiologists to compare neurons on the basis of their biophysical properties and infer shared computational function. These features would complement existing resources, such as NeuronDB and CellPropDB, which allow similar searching of neuron types based on specified shared physiological features such as somatodendritic ionic current distributions (Crasto et al., 2007). Critically, extension of such tools will also facilitate the linkage of electrophysiological data with neuron morphology (Parekh and Ascoli, 2013) and gene expression (Lein et al., 2007; Wichterle et al., 2013) to further our understanding of the fundamental link between neuronal biophysical properties and computational roles.

4.4 Results

4.4.1 Generating a brain-wide database of neuronal biophysical properties

The utility of comparing electrophysiological data across neuron types increases with the amount of data considered. As much of this electrophysiological data already exists within the literature, we chose to "mine" the data from the text of published papers. While forgoing the difficulties of recording from multiple neuron types and brain areas, a data-mining approach is not without its own challenges, such as accounting for inconsistencies in published neuron naming schemes (Ascoli et al., 2008) and experimental conditions. However, many of these limitations can be overcome by capitalizing on the redundancy of published values and the presence of informal community-based reporting standards (Ascoli et al., 2008; Toledo-Rodriguez et al., 2004), providing a unified dataset of sufficient quality for use in subsequent meta-analyses.

As described in detail in Chapter 3, from approximately 92,000 published articles from

15 journals containing electrophysiological data, we mined information on basic biophysical properties (e.g., resting membrane potential and input resistance) of 98 distinct types of normotypic (i.e., "wild-type") neurons (based on definitions from `http://neurolex.org`; (Johnston and Wu, 1995; Hille, 2001; Shepherd, 2003; Stuart et al., 2007; Hamilton et al., 2012; Larson and Martone, 2013)) from 328 articles. Our mining strategy follows a two stage process (Fig. 4.1A; detailed in Section 4.7). Briefly, we first examined the text of articles from neuroscience-specific journals (e.g., the Journal of Neuroscience, the Journal of Neurophysiology, etc.) and developed automated text-mining algorithms (French et al., 2009; Ambert and Cohen, 2012) to extract content related to biophysical properties and experimental conditions. We focused on developing algorithms for extracting data from formatted data tables because this increased the reliability with which we could extract relevant data (Dickman, 2003; Yarkoni et al., 2011). Algorithms focused on data within text and figures will be considered in future work. Next, we manually curated the automatically extracted data, taking care to fix incorrectly identified content such as mislabeled neuron types. Given the text mining approach that we used, manual curation was still necessary, with 66% of electrophysiological concepts (1397 of 2102 total) and 30% of neuron type mentions (120 of 399 total) identified correctly using automated methods alone.

A sample of the resulting data is shown in Fig. 4.1 and the dataset in its entirety can be interactively explored and downloaded through our web interface at `http://neuroelectro.org`. The sample data in Fig. 4.1 reflects known features of these neurons; for example, cortical basket cells have narrow action potentials (Markram et al., 2004) and striatal medium spiny neurons rest at hyperpolarized potentials (Kreitzer, 2009). The composition of the entire NeuroElectro dataset also reflects biases in the literature: CA1 pyramidal cells and cortical neurons are among the best studied neurons, while most neuron types are characterized by 5 or fewer articles (Supp. Fig. 4.7A). Additionally, authors are more likely to report only a subset of specific properties within an article, such as resting

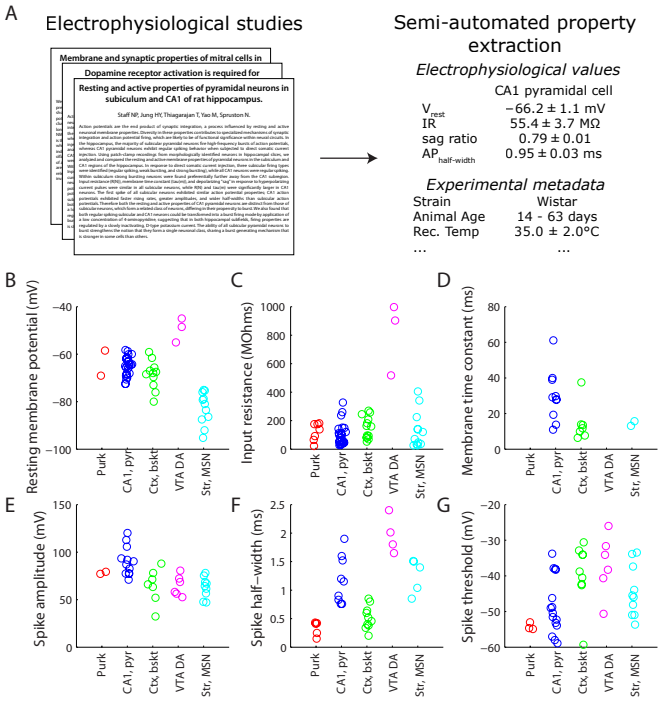

Figure 4.1: Schematic of NeuroElectro database construction and example electrophysiological measurements. A) Semi-automated text-mining algorithms were developed and applied to neurophysiology journal articles to extract neuron-specific electrophysiological measurements and corresponding experimental metadata. B-G) Example electrophysiological measurements extracted from the research literature for cerebellar Purkinje cells, CA1 pyramidal cells, cortical basket cells, ventral tegmental area dopaminergic cells, and striatal medium spiny neurons (abbreviated as Purk; CA1, pyr; Ctx, bskt; VTA, DA; and Str, MSN respectively). Each circle denotes the value of the mean electrophysiological measurement reported within an article.

membrane potential and input resistances (Supp. Fig. 4.7B). Moreover, more sparsely reported properties, such as action potential (spike) afterhyperpolarization amplitude, tend to be calculated in different ways across articles. In light of these reporting confounds, we have focused our current analyses on six commonly and reliably reported biophysical properties: resting membrane potential, input resistance, membrane time constant, spike half-width, spike amplitude, and spike threshold (abbreviated as V_{rest}, R_{input}, τ_m, AP_{hw}, AP_{amp}, AP_{thr}, respectively).

4.4.2 Experimental metadata helps explain the observed variance among electrophysiological measurements

Our literature-based approach relies on pooling information across multiple articles, which has the inherent advantage of distilling the "consensus" view of several expert investigators. However, data collected by different investigators under different experimental conditions may not be directly comparable. For example, input resistances tend to decrease as animals age (e.g., see (Zhu, 2000; Okaty et al., 2009; Kinnischtzke et al., 2012)), thereby rendering data that include animals of different ages more variable. Because our data are randomly sampled from the literature, these relationships between experimental conditions (the "metadata") and electrophysiological measurements (the "data") should also be reflected within our dataset (e.g., Fig. 4.2B). By annotating each electrophysiological measurement in our database with a corresponding set of experimental metadata (using methods section text-mining and manual curation as described in Chapter 3; Fig. 4.2A), we were able to address the following three questions. First, can experimental metadata be used to reduce the variability of data reported across studies? Second, what is the influence of specific experimental conditions (e.g., recording temperature and electrode type) on measurements of biophysical properties? Third, what is the residual variability in reported values after differences in several experimental conditions have been accounted

for?

We used linear regression models to characterize the relationship between electrophysiological measurements and experimental metadata. We first asked to what extent the variability observed among electrophysiological measurements could be explained by neuron type alone (i.e., how consistent are measurements of the same neuron type from study to study). We found that V_{rest} was reported fairly consistently (Fig. 4.2C; adj. $R^2 = 0.6$; i.e. 60% of the variability in V_{rest} across cells was explained by cell type). However, most properties, such as τ_m and AP_{thr}, had measurements which differed greatly across studies recording from the same neuron type (adj. $R^2 < 0.25$). Thus, there exists a high degree of "noise" or variance unexplained by neuron type in the literature among these electrophysiological data.

We found in many cases, however, that experimental metadata could significantly explain the variability in reported electrophysiological data (Figs. 4.2D-F, summary in G). For example, knowing whether neurons were recorded using patch versus sharp electrodes explained a substantial fraction of the observed variance in R_{input}, with sharp electrodes yielding on average 100 MΩ lower R_{input} than patch electrodes (Fig. 4.2D). Thus, the dataset inherently reflects a historical controversy of the late 20^{th} century when the patch-clamp technique was first introduced and large discrepancies were observed in R_{input} measurements made with patch vs. sharp electrodes. Consistent with our dataset, sharp electrode recordings were found to systematically underestimate R_{input} (Spruston and Johnston, 1992; Staley et al., 1992). Moreover, the dataset quantitatively reflects a number of other qualitatively known relationships between experimental conditions and electrophysiological measurements, such as an inverse correlation between animal age and τ_m (Fig. 4.2E) and a correlation between AP_{thr} and liquid junction potential correction (Fig. 4.2F).

Collectively, integrating experimental metadata with neuron type accounted for considerably more measurement variability than neuron type alone (Fig. 4.2C; details in Section

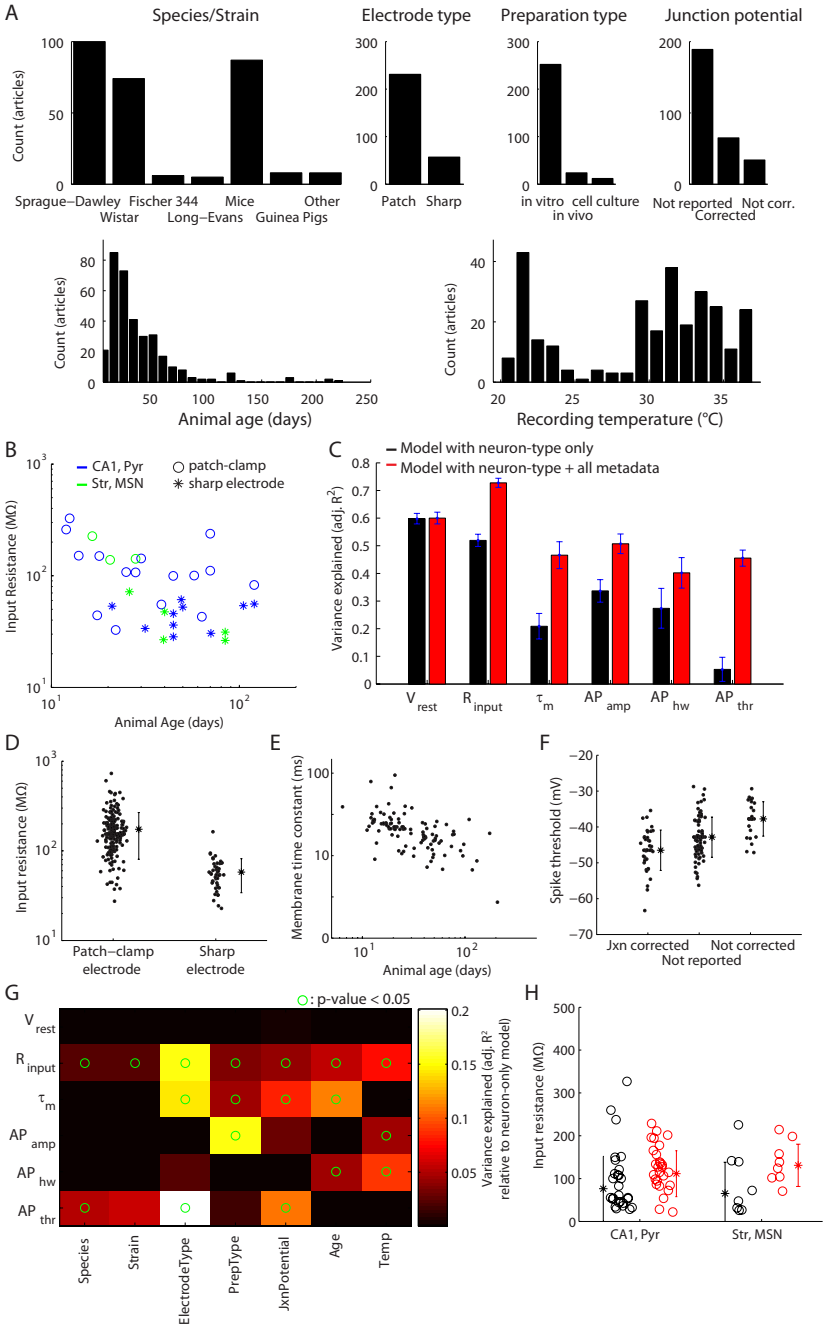

Figure 4.2: Experimental metadata helps explain variability in reported electrophysiological data across studies. A) Histograms of the observation frequencies of different kinds of experimental metadata. B) Example data showing how measured values of R_{input} varies as a function of recording electrode type and animal age. C) Variance explained by statistical models for each electrophysiological property when only neuron type information is used (black) and when neuron type plus all metadata attributes are used (red). Error bars indicate standard deviations, computed from 90% bootstrap resamplings of the entire dataset. D-F) Example relationships between specific metadata predictors and variation in electrophysiological properties. Dots show electrophysiological measurements after accounting for neuron type specific differences. Panel F refers to correction of junction potential ("jxn"). G) Influence of individual metadata predictors in helping explain variance in specific electrophysiological properties. Heatmap values indicate relative improvement over the model that includes neuron type information only. Circles indicate where regression model including metadata attribute was statistically more predictive than model with neuron type information alone ($p < 0.05$; ANOVA). H) Example electrophysiological data before (black) and after using statistical models to adjust for differences in metadata among electrophysiological measurements (red). After adjustment of electrophysiological measurements for metadata differences, values become less variable and less skewed.

4.7). Thus, we used our regression models to statistically adjust each electrophysiological measurement (e.g., Fig. 4.2H) to help account for systematic differences in recording practices across neuron types and to improve subsequent between-neuron comparisons. As a caveat, we note that there still exists a substantial amount of variance in reported electrophysiological measurements even after integrating the experimental metadata. This variance likely reflects: (1) within-type neuronal variation (e.g., Padmanabhan and Urban (2010); Angelo et al. (2012), as described in Chapter 2); (2) additional experimental conditions not taken into consideration (e.g., recording solution contents); and (3) differences in analysis methods across investigators. Because of these confounding factors, the ability of experimental metadata to account for electrophysiological data variability reported here should thus be viewed as a lower-bound, and should improve as more data are included.

Given the potential sensitivity of this analysis to the size of the data set being used, we next asked whether small reductions in the dataset substantially altered our metadata adjustments. In other words, how crucial is a large database of biophysical properties to understanding the systematic relationship between experimental conditions and electrophysiological measurements? To answer this, we examined how the regression results changed after randomly subsampling a portion of the dataset. We found that the statistical regression algorithms could typically tolerate a loss of 20-50% of the total dataset without a corresponding loss in the predictive ability of metadata for explaining variability in electrophysiological measurements (Supp. Fig. 4.10). Thus, only with a dataset of sufficient size (practically exceeding >100 examples across multiple neuron types), such as that contained within NeuroElectro, can we quantitatively learn systematic relationships between experimental conditions and electrophysiological measurements.

4.4.3 Targeted recordings yield measurements consistent with the NeuroElectro dataset

Thus far, we have demonstrated that our text-mined dataset is internally consistent and, through inclusion of experimental metadata, reflects heuristic knowledge of the field. To more explicitly examine the validity of our dataset, we next performed a series of targeted recordings from a subset of the best-studied neuron types, including hippocampal CA1 pyramidal cells (Fig. 4.3A), layer 5-6 neocortical pyramidal cells (Fig. 4.3B), main olfactory bulb mitral cells (Fig. 4.3C), and neocortical basket cells (Fig. 4.3D). After adjusting the text-mined dataset to reflect the recording conditions used in our laboratory (details in Section 4.7), we observed close agreement between the NeuroElectro dataset and each of the neuron types and biophysical properties that we recorded (Fig. 4.3E). We further note that this close agreement even extended to specific properties for which only 2-3 values were present in the database (e.g., mitral cell AP_{amp}). We thus conclude that the text-mined data and metadata populating NeuroElectro accurately reflects the electrophysiological properties of real neurons across the brain.

4.4.4 Investigating brain-wide correlations among biophysical properties

We next performed a series of analyses on our validated brain-wide electrophysiology dataset with the goal of gaining insights into the relationships between biophysical properties and diverse neuron types. We first looked for correlations between biophysical properties. Though several studies have previously examined this topic (Padmanabhan and Urban, 2010; Aizenman et al., 2003; Toledo-Rodriguez et al., 2004), they were typically focused on one or a small number of neuron types from a single brain region. In contrast, we now ask whether electrophysiological relationships hold across a large number of the

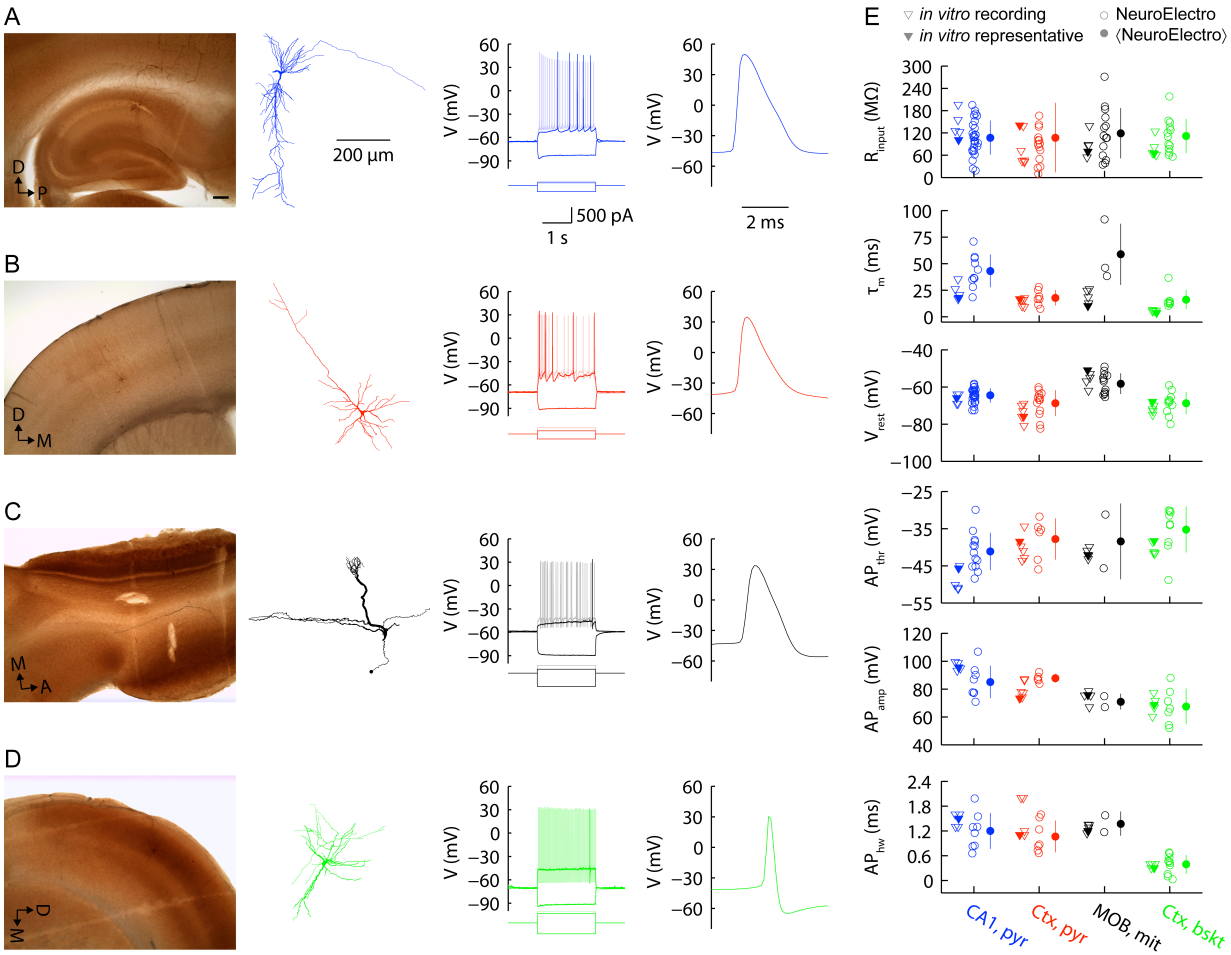

Figure 4.3: Validation of NeuroElectro database measurements with collection of raw data. A) Representative targeted recording of a hippocampal CA1 pyramidal cell, showing anatomical position and morphological reconstruction (left), response to hyperpolarizing and depolarizing rheobase and suprathreshold step current injections (middle), and action potential waveform (right). Anatomical scalebar: 200 μm. B-D) Same as A for: layer 5-6 neocortical pyramidal cell (B), main olfactory bulb mitral cell (C), and neocortical basket cell (D). E) Summary of targeted *in vitro* recordings and comparison to text-mined, metadata-adjusted values from NeuroElectro. Abbreviations: dorsal (D), posterior (P), medial (M), anterior (A).

neuron types in the brain.

We observed a number of significant correlations among neuron-averaged electrophysiological properties (examples in Figs. 4.4A,B; summary in C). These observations included intuitive correlations expected *a priori*, such as a positive correlation between R_{input} and τ_m. However, we also observed correlations more difficult to explain via first principles of neural biophysics, including a positive relationship between R_{input} and AP_{amp} and a positive relationship between V_{rest} and AP_{hw}.

Using dimensionality reduction (via a probabilistic form of principal component analy-

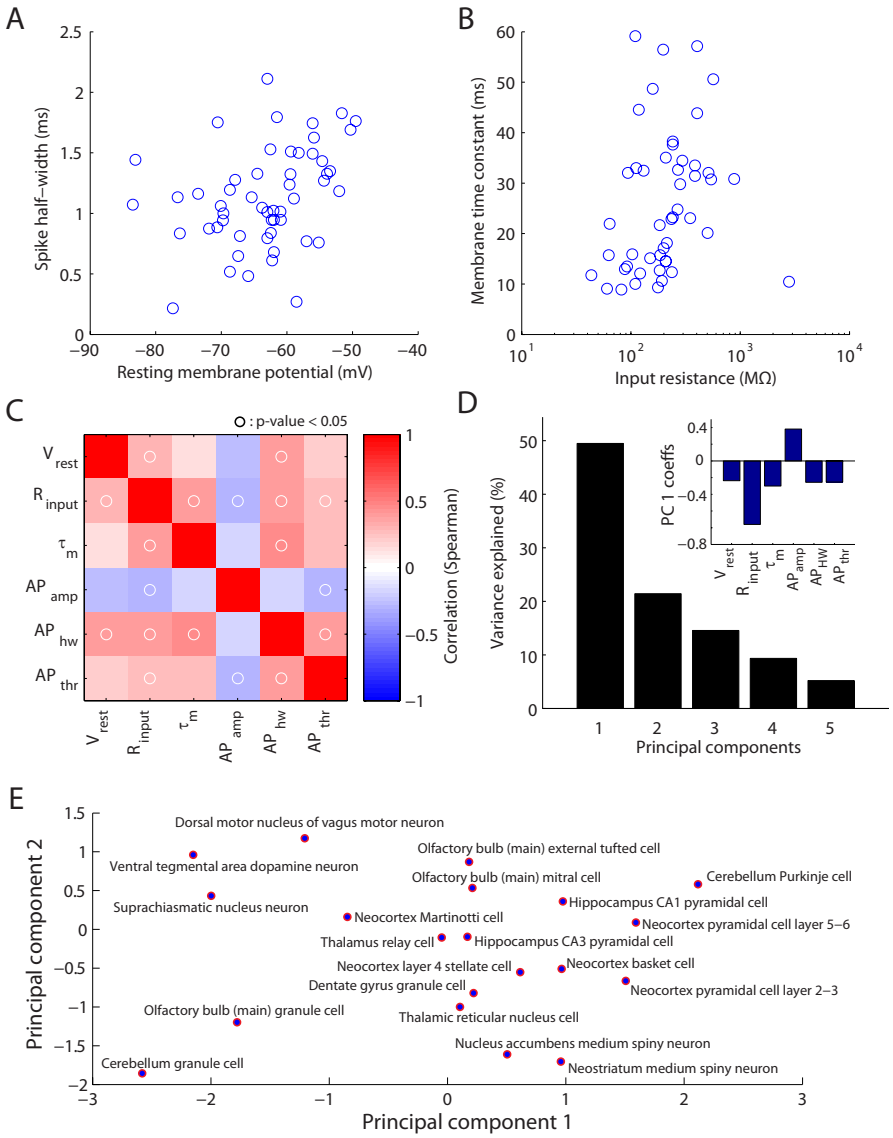

Figure 4.4: Exploring pairwise correlations of electrophysiological properties. A,B) Example data showing pairwise correlations among electrophysiological properties. Each data point corresponds to measurements from a single neuron type (after averaging observations collected across multiple studies and adjusting for experimental condition differences). C) Correlation matrix of electrophysiological properties (Spearman's correlation). Circles indicate where correlation of electrophysiological properties was statistically significant ($p<0.05$). D) Variance explained across principal components of electrophysiological correlation matrix. Inset shows the coefficients corresponding to the first principal component, defining the dimension of maximal variation among neuronal electrophysiological measurements. E) Projection of neuron types onto space defined by first two principal components. Only neurons with electrophysiological measurements defined by at least 4 articles (total) plotted for visual clarity. Note that the first principal component qualitatively reflects differences in neuron electrotonic size.

sis, pPCA, to account for missing electrophysiological measurements; see Section 4.7), we found that a single principal component could explain 50% of the variance in the pooled dataset (Fig. 4.4D). This principal component qualitatively reflects the difference between electrotonically large and small neurons (Fig. 4.4E), where large neurons (like Purkinje cells and cortical projection neurons) tend to be correlated with low R_{input} and tall, narrow spikes relative to more compact neurons (like cerebellar and olfactory bulb granule cells). The ability of this principal component to account for such a large fraction of the total variance indicates that correlations in biophysical properties across the majority of neurons in the brain are substantial and that electrophysiological properties are highly predictive and correlated with one another. This finding further suggests that subsequent analyses of biophysical properties may be facilitated by considering such approaches for dimensionality reduction. In other words, a scientist may need to only measure only 2 of these 6 biophysical properties to be able to infer the other 4 properties (and perhaps even additional properties, such as spike train burst phenotypes).

4.4.5 Biophysical similarity identifies approximately 7 neuron super-classes

We next investigated whether we could use our dataset to identify known and unknown similarities between distinct neuron types on the basis of their biophysical properties. For example, just as fast spiking interneurons populate multiple brain areas (e.g., neocortex and hippocampus (Markram et al., 2004; Martina et al., 1998)), are there other sets of neurons that share physiological properties and potential computational functions?

We performed a hierarchical clustering analysis of the neuron types within our dataset, where for each pair of neurons we assessed their similarity by comparing the set of 6 basic biophysical properties defined above. Here, we chose to be agnostic about the relative importance of each biophysical property (e.g., (Toledo-Rodriguez et al., 2004; Druckmann

et al., 2012)) and weighted properties based solely on their variance across studies (e.g., see: Fig. 4.2C). We further mitigated measurement noise by focusing on neurons defined by at least 3 different articles and with no more than 2 of the 6 biophysical properties missing (as in Fig. 4.4D, we used pPCA to account for unobserved values when missing).

The results of our clustering analysis are displayed in Fig. 4.5 (see Supp. Fig. 4.11B for a full distance matrix of neuron type biophysical similarity). To assess the statistical significance of each dendrogram subtree (i.e., how strongly each subtree was supported by the data), we used the procedure of bootstrap resampling from phylogenetics (Felsenstein, 2004; Suzuki and Shimodaira, 2006) to assess the probability of observing each subtree across multiple resamplings of the data matrix shown in Fig. 4.5. Here, in each bootstrap sample, the data matrix is resampled by randomly sampling columns corresponding to individual biophysical properties with replacement (per the convention in phylogenetics (Felsenstein, 2004); see Section 4.7 for additional details).

We note that several previously established neuron classes emerged from this analysis, validating our unbiased clustering approach. For example, we observed two distinct clusters composed of large glutamatergic projection neurons defined by low R_{input} and high AP_{amp}. These two clusters differed in their V_{rest} and τ_m; with neurons such as olfactory bulb mitral cells and CA3 pyramidal cells resting more depolarized relative to deep layer cortical neurons or CA1 pyramidal cells. Likewise, GABAergic fast-spiking basket cells from the neocortex and hippocampus were also closely clustered, providing further validation of this approach.

At least 4 novel statistically significant neuron super classes also emerged from our clustering analysis. First, we observed a cluster composed of smaller projection neurons defined by very hyperpolarized V_{rest}, including glutamatergic layer 2/3 neocortical pyramidal cells and dentate gyrus granule cells and GABAergic medium spiny neurons from the dorsal and ventral striatum. Intriguingly, a number of these neurons are hypothesized to

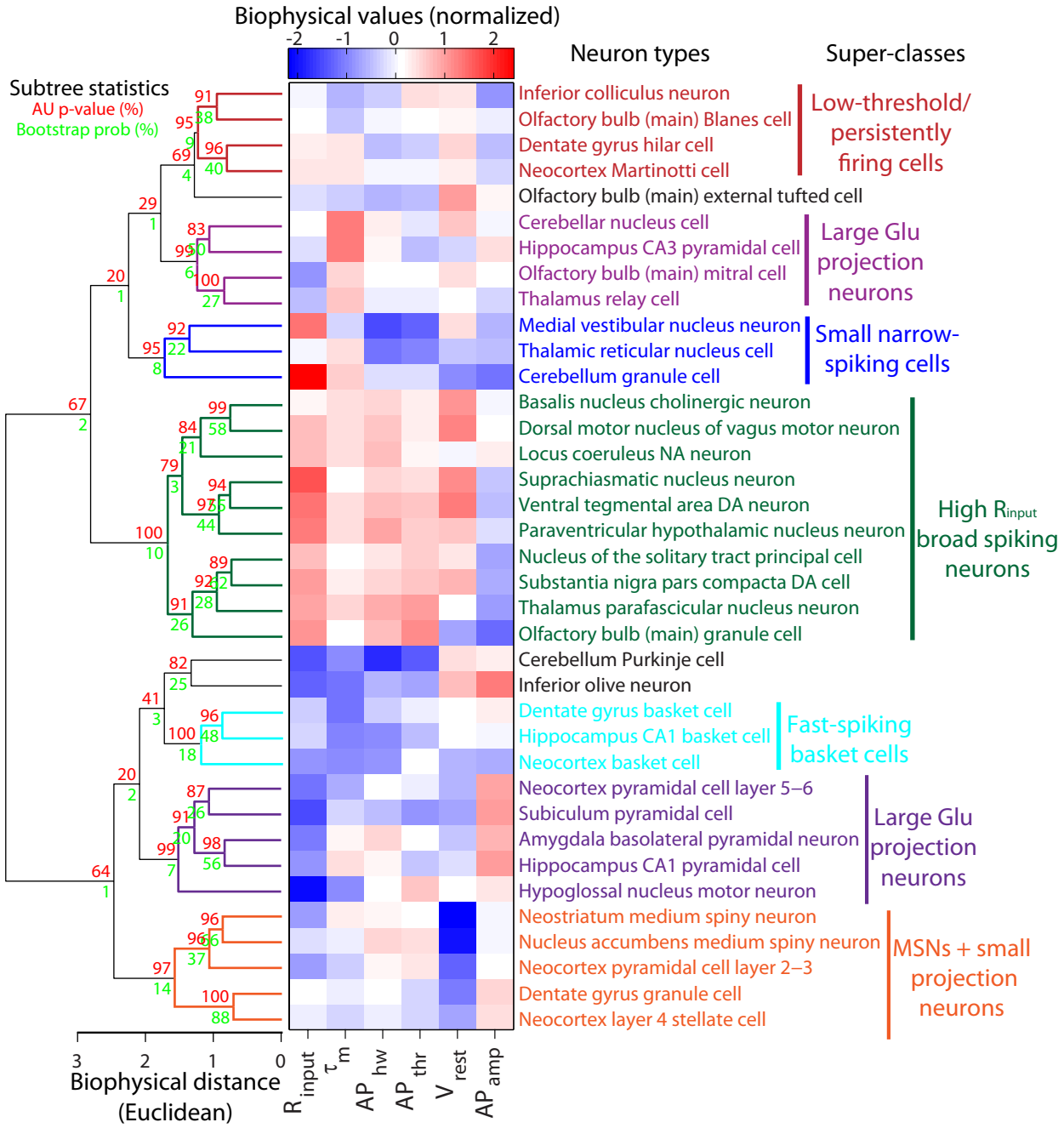

Figure 4.5: Hierarchical clustering of neuron types on the basis of biophysical similarity. Neuron types sorted in order of biophysical similarity (similiarity indicated by levels of dendrogram; dendrogram linkages computed using Ward's method and Euclidean distances). Heatmap values indicate observed neuron-specific electrophysiological measurements, red (blue) values indicate large (small) values relative to mean across neuron types. Statistical consistency of dendrogram subtrees calculated via bootstrap resampling (green values indicate percentage of bootstrap resamples containing subtree; red values indicate approximately unbiased (AU) p-values; p-values and bootstrap percentages rounded to nearest integer for visualization; $n = 10,000$ bootstrap resamples). Dendrogram subtrees with AU p-value greater than p > 95% are grouped into neuron super-classes indicated by text coloring (and are otherwise black). Only neuron types with measurements defined by at least 3 articles and with no more than 2 (of the 6 total) biophysical properties not observed were used in this analysis (probabilistic PCA used to impute unobserved measurements).

represent information sparsely in cell assemblies and exhibit up- and downstates (Kreitzer, 2009; Larimer and Strowbridge, 2010; Barth and Poulet, 2012). Second, we observed a large cluster of high R_{input}, large AP_{hw} neurons with depolarized V_{rest} and AP_{thr} composed of neurons from the substantia nigra, ventral tegmental area, hypothalamus, and brainstem (Fig. 4.5). While markedly diverse in their combined neurochemistry (including dopaminergic, GABAergic, and peptidergic neurotransmitters), many of these these neuron types nevertheless exhibit similar activity patterns comprised of spontaneous "pacemaker"-like firing rates (Stern, 2001; Tateno and Robinson, 2011), an emergent property predicted by their distinctively depolarized V_{rest} (Fig. 4.5). Third, we observed a cluster of neurons containing olfactory bulb Blanes cells, dentate gyrus hilar cells, and neocortical Martinotti cells that were uniquely defined by a depolarized V_{rest} and low AP_{thr}. Consistent with their low effective spiking threshold, these neurons are known to act as "first-responders" within their larger circuit and regulate the overall excitability of other neurons (Pressler and Strowbridge, 2006; Larimer and Strowbridge, 2010; Fanselow and Connors, 2010). Lastly, we observed a high R_{input} cluster composed of cerebellar granule cells and neurons from the medial vestibular and thalamic reticular nuclei that shares features similar to basket cells, namely short AP_{hw} and depolarized AP_{thr}.

Additionally, we note that across the entire dataset, we observed a qualitative correspondence between biophysical similarity and gross anatomical position (e.g., hindbrain neurons were generally more similar to one another than to neocortical neurons), suggesting that shared precursor lineage yields similar biophysical properties (Gage, 2000; Ohtsuki et al., 2012). In total, clustering of brain-wide neurons by just 6 biophysical properties thus revealed several novel insights about biophysical similarities and possible computational roles. We note that this biophysical parameter based-approach for clustering neuron types differs from alternative grouping methods, such as those based on somatodendritic distribution of ionic conductances and morphological properties(Migliore and Shepherd, 2002,

2005). To elaborate, we find that the major dimensions along which neurons vary here are based on neuronal electronic size and membrane depolarization at rest whereas Migliore and Shepherd (2005) found that neurons primarily varied based on dendrite thickness and presence or absence of dendritic back propagating action potentials. This indicates that, in contrast to BLAST searches in GenBank across a unidimensional inventory of sequence similarly (Altschul et al., 1990; Benson et al., 2013), searches across neurons are across a multidimensional inventory which can yield different shared motifs. A future challenge will be to incorporate these and other approaches into a consensus multidimensional representation of neuron diversity.

4.4.6 Differences in gene expression predict differences in biophysical properties

In our final experiments, we sought to better understand the mechanistic origin of these ∼7 neuronal super-classes and, more generally, the biophysical differences across brain-wide neurons. To do so, we took advantage of a publicly available gene expression dataset provided by the Allen Brain Atlas (Lein et al., 2007) to test how biophysical differences correlate with transcriptional differences. Such an analysis may allow us to understand which of the ∼20,000 genes in the genome are responsible for specific biophysical differences among neurons. Furthermore, this "genome-wide" approach to understanding the basis of diverse biophysical phenotypes could complement existing single-gene or single-current approaches provided by targeted gene-knockouts and channel-specific pharmacology (Coetzee et al., 1999).

At present, there are not brain-wide gene expression datasets at the resolution of individual neuron types. However, the Allen Brain Atlas does contain data on the expression of most genes in the genome and a number of non-coding RNAs at the resolution of individual brain regions and often within individual cellular lamina (Fig. 4.6A). Thus, for brain re-

gions and cellular lamina composed of relatively homogeneous populations of neuron types, we were able to extract gene expression data for individual neuron types. For brain regions containing a heterogeneous population of neuron types, we considered only the most common neuron type. For example, because layer 2/3 of the neocortex is composed of ∼80% pyramidal cells (Douglas and Martin, 2004), we tested the gene expression data of layer 2/3 as a potential predictor for the biophysical properties of layer 2/3 pyramidal cells, but did not examine other neocortical cells such as basket cells. While yielding only an approximation of neuron type-specific gene expression, this approach nevertheless makes possible the best *comprehensive* analysis of brain-wide electrophysiological and transcriptional data until datasets at higher cellular resolution become available. Consequently, the following results identify a lower bound on the predictive relationship between genome-wide gene expression and biophysical properties.

To quantify the relationship between biophysical and transcriptional differences, we first calculated the pairwise differences among neuron types based on electrophysiological differences reported in NeuroElectro (Fig. 4.6C) and based on trascriptional differences among voltage-gated ion channel genes (Fig. 4.6D). We then compared the collection of resulting pairwise differences to determine how well gene expression differences predict biophysical differences. We note that a perfect match between these sets of pairwise differences is not expected for several reasons. For example, in addition to the limited resolution of our transcriptional dataset, differences in mRNA measurements do not directly reflect differences in protein abundance (Miller et al., 2011; Coetzee et al., 1999). Further, biophysical properties may be influenced by combinatorial patterns of ion channel expression (Prinz et al., 2004; Marder and Taylor, 2011), in addition to different cell morphologies (Mainen and Sejnowski, 1996). Despite these caveats, however, we observed a surprising degree of correlation between brain-wide biophysical differences and voltage-gated ion channel expression differences (Pearson's $r = 0.36$; $p < 6 \times 10^{-5}$, Mantel's Test).

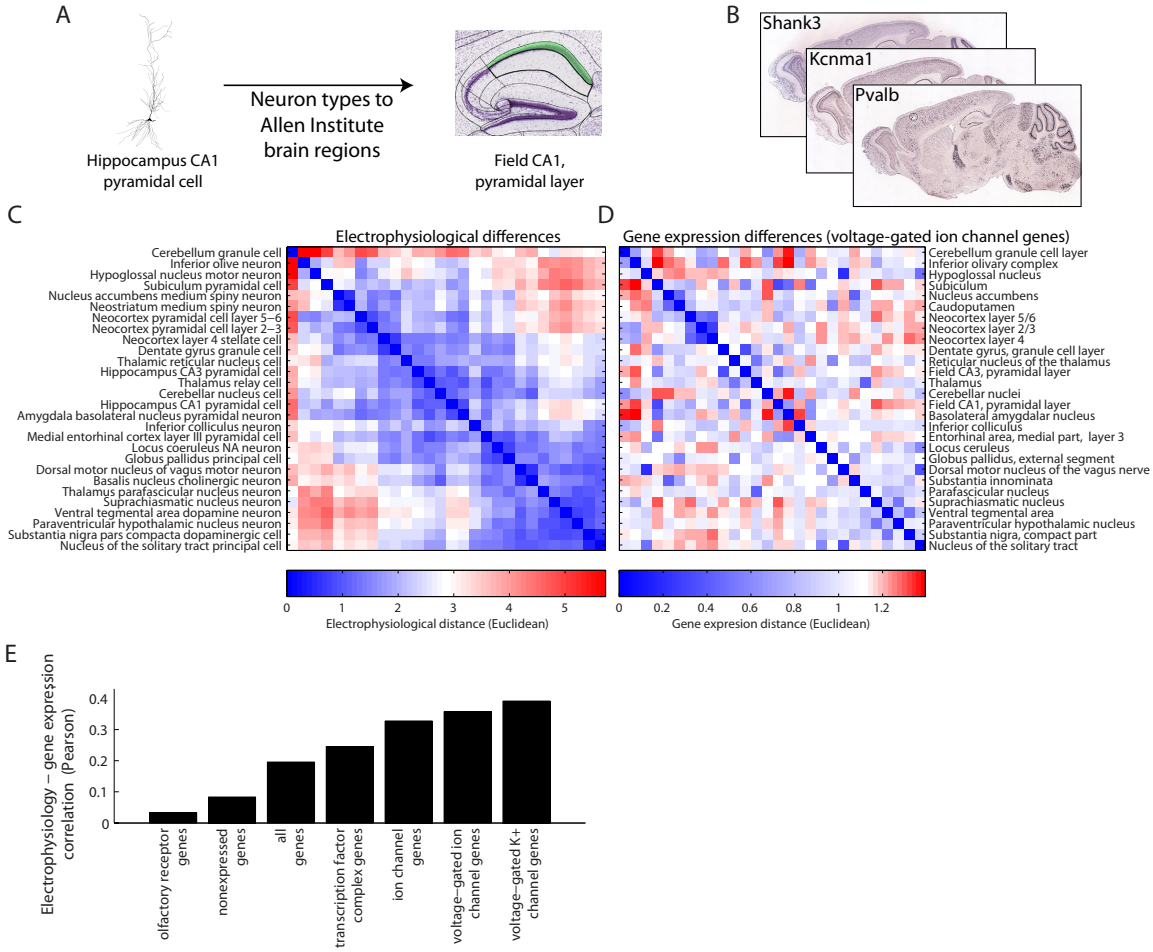

Figure 4.6: Correlation of differences in neuronal electrophysiology with differences in gene expression. A) Illustration of mapping between neuron types and the brain regions (from the Allen Institute mouse brain atlas) in which they are contained. B) Example Allen Institute *in situ* hybridization experiments showing the gene expression profiles of Shank3, Kcnma1 (KCa1.1), and Pvalb (parvalbumin). C) Distance matrix of electrophysiological differences among neuron types for neurons which could be uniquely mapped to a corresponding brain region in the Allen Institute atlas. Heatmap values indicate electrophysiological similarity; blue (red) indicates similarity (dissimilarity). D) Same as C, but computed for gene expression differences when considering the set of genes coding for voltage-gated ion channels. Brain regions corresponding to neuron types sorted as in C. E) Correlation of distance matrices defined by electrophysiological differences and gene expression differences, when considering different sets of genes defined by Gene Ontology functional categories.

We then repeated this analysis using other gene classes defined by the Gene Ontology project (Ashburner et al., 2000), such as transcription factors, synaptic scaffolding proteins, etc., to determine if other genes less intuitive than voltage-gated ion channels underlie brain-wide biophysical differences. Select examples from this analysis are shown in Fig. 4.6E, while the full analysis and summary of the top 100 predictive gene classes are provided in Supp. Fig. 4.12 and Supp. Table 4.1, respectively. As expected, classes of genes with minimal expression in the brain demonstrated low levels of correlation with neuronal biophysical differences (Fig. 4.6E). Such gene classes included olfactory receptor genes (which are primarily expressed in the nasal epithelium (Buck and Axel, 1991)) and a class of ∼3,000 genes and non-coding RNAs assessed by the Allen Institute to exhibit negligible expression in the brain (Lein et al., 2007). In contrast, gene classes known to underpin electrophysiological properties, such as ion transporters and voltage-gated ion channels genes (Toledo-Rodriguez et al., 2004), were among the classes mostly strongly correlated with biophysical differences. Surprisingly, other gene classes not known to directly underpin electrophysiological properties, including a number of synapse-specific and neurotransmitter receptor gene classes, also exhibited high levels of correlation with neuronal biophysical differences (Supp. Fig. 4.12; Supp. Table 4.1). These findings motivate more direct investigation of the relationship between these surprisingly predictive gene classes and basic biophysical properties.

4.5 Discussion

We have created a database, NeuroElectro, that compiles the results of a large number of electrophysiological publications to help gain insight into the brain's *"parts list"*. We believe that NeuroElectro will accelerate, through the discovery of *"buried treasure"* in the literature, the elucidation of both the mechanistic bases and functional consequences of neuron-to-neuron biophysical differences and similarities. NeuroElectro also provides

a framework for *data sharing*, and will thus facilitate large-scale efforts to understand neuronal properties and brain function, such as the US federal BRAIN initiative (Alivisatos et al., 2012; Insel et al., 2013) and the EU Human/Blue Brain Project (Markram, 2006). Critically, NeuroElectro is readily expandable, even with electrophysiologists continuing their current practices of data collection and reporting. That is, NeuroElectro does not impose any top-down requirements on how data are collected, analyzed, or reported. In this way, we see NeuroElectro as standing as a community-based alternative or complement to large-scale single-institute efforts to determine the properties of all neuron types in the brain (e.g., the Human/Blue Brain Project (Markram, 2006)).

4.5.1 Summary

Obtaining a full understanding of the biophysical diversity of neuron types across the brain will require a comprehensive collection of electrophysiological data. Here, we have taken the first steps toward this goal by using semi-automated text-mining algorithms to link neuron types with electrophysiological data scattered throughout the vast neuroscience literature. The resulting centralized database of neuron-specific electrophysiological data is now publicly available through an interactive web-interface at `neuroelectro.org`.

The utility of the raw data aggregated in this database is limited by the variability in reported values across studies. However, the statistical power harbored in this unprecedented collection of electrophysiological data enabled us to learn systematic relationships between experimental conditions and biophysical properties. With this knowledge, we were able to adjust data values mined across multiple studies to account for differences in basic experimental conditions, yielding a unified dataset amenable to brain-wide comparisons. The rest of the current study was devoted to exploring a subset of these brain-wide comparisons.

We observed that, across the brain, neuronal biophysical properties exhibit several intu-

itive relationships, such as a positive correlation between AP_{thr} and AP_{amp}, as well as more unintuitive relationships, such as a positive correlation between V_{rest} and AP_{hw}. Further, by comparing biophysical properties between neuron types, we uncovered several known and unknown "super-classes" of neuron types projected to exhibit similar functionality. For example, we identified a class of hippocampal and olfactory bulb neuron types capable of persistent bistable activity – an emergent property attributable, in part, to a uniquely depolarized V_{rest}, hyperpolarized AP_{thr}, and relatively low R_{input}. Lastly, we quantified the relationship between electrophysiological and transcriptional differences across numerous gene classes, providing a mechanistic basis for the biophysical divergence of distinct neuron types across the brain.

4.5.2 Strengths and weaknesses of our literature-based text-mining approach

As described in Chapter 3, the use of text-mining algorithms (Ambert and Cohen, 2012) greatly facilitated our ability to aggregate electrophysiological data across a large number of neuron types into NeuroElectro. These algorithms provide an efficient, first-pass scan of the literature by identifying putatively relevant articles and tagging electrophysiological concepts and entities within those articles via simple text-matching rules. Importantly, such text-mining algorithms greatly augment the ability of human experts to quickly annotate articles for neuron type-specific electrophysiological information. That is, for a human expert, validating the findings of such text-mining algorithms is more efficient than manually locating and identifying those results *de novo*.

However, our text-mining approach comes with a number of potential confounds that are not typically present when investigators record data within their own laboratory. Foremost, we are only able to extract data which investigators choose to include in a machine-readable format in published articles. Moreover, because there are not community-adopted

standards for how electrophysiological measurements should be determined, the meaning of a term like "spike amplitude" may differ from article to article. Where possible, we have attempted to manually account for these cases by excluding or normalizing measurements which were calculated in distinct ways or reported in distinct units (e.g., R_{input} measured in GΩ vs. MΩ).

In addition to these potential confounds, our text-mining approach also introduces subtle biases into our dataset. Our text-mining algorithms require articles to be formatted as HTML; thus, we cannot currently mine articles published before ∼1997 because they are typically only available as scanned PDFs. Further, we primarily focused our efforts on extracting electrophysiological data from formatted data tables (e.g., of 2,253 total validated measurements, 93% come from data tables). This greatly limits the amount of extractable information – by our estimates, ∼10% of relevant articles contain information in a structured data table. However, semi-automated extraction from tables is substantially more accurate than from article text or figures (Dickman, 2003). Moreover, our analyses demonstrate that we have extracted sufficient data to support each of our main results (Supp. Fig. 4.10). Each of these limitations could be overcome in subsequent efforts, however, through: 1) improved methods for data extraction from text and figures, 2) by increasing manual curation efforts over a larger team of human experts, and 3) by obtaining and automatically processing raw data from investigators, such as voltage traces collected in a standardized format.

Additionally, our text-mining approach requires mapping of each extracted datum to a canonical neuron type. Because investigators use multiple criteria for classifying neurons (Ascoli et al., 2008), we chose to use the community-generated expert-defined list of neurons provided by NeuroLex (`http://neurolex.org/wiki/Vertebrate_Neuron_overview`, (Hamilton et al., 2012; Larson and Martone, 2013)). While these definitions currently "lump" rather than "split" neuron types (e.g., "neocortex layer 5/6 pyramidal

neuron"), these definitions will evolve as community input accumulates. Accordingly, we have built the mapping of data to neuron type in NeuroElectro to be highly flexible (as described in Chapter 3, allowing NeuroElectro to similarly evolve to match updates in neuron type definitions.

4.5.3 The utility of a public brain-wide database of electrophysiological properties

In its current form, the NeuroElectro project can provide to the community a number of distinct functions: a valuable resource for experimental physiologists looking for references to compare their data with existing data; a repository for computational modelers looking for parameters to help constrain their models; and a knowledgebase for theoreticians searching for undiscovered relationships among neurons and their properties. It also presents a useful way of surveying the literature containing such descriptions of neuronal properties and can serve a pedagogical role in the training of scientists new to electrophysiology and in exposing the general public to the rich diversity of neurons.

Moving forward, we plan to extend the NeuroElectro database and expand upon the available analysis functionalities provided via the public web interface. Specifically, we intend to further mine the research literature and accurately capture a greater number of electrophysiological measurements and include data collected from neuron types recorded under non-normotypic conditions (e.g., under pharmacological manipulation or from genetically modified animals). Moreover, we plan to engage the research community for aid in curating the machine-mined content and allow researchers to upload and share existing datasets. Given these expanded data, we intend to enhance the existing web interface with additional features, such as making data sortable by experimental conditions or neurological disease states and allowing for clustering of neuron types based on arbitrary sets of physiological features.

In the longer term, our vision is to develop NeuroElectro into a resource much like the genetics tools GenBank and BLAST (Altschul et al., 1990; Benson et al., 2013) and allow searching for neuron types on the basis of biophysical similarity. For example, a user could enter a data-based query like recorded voltage traces or summary measurements from a little-studied neuron type and be returned a list of biophysically similar neuron types plus associated references. Thus by studying these better-characterized neurons, the user could gain insights into potential computational roles or channel mechanisms underlying her neuron.

Furthermore, through integrating NeuroElectro with other databases, such as on gene transcription (Lein et al., 2007; Wichterle et al., 2013), computational models (Migliore et al., 2003), or morphology (Parekh and Ascoli, 2013), these tools could help investigators infer how the expression of particular genes or the presence of specific morphological features gives rise a neuron's unique biophysical phenotype. We feel that this approach, which explicitly links together the work of the community of investigators, increases the reach and impact of any one publication and has the potential to greatly increase the rate of progress in the field. While there is much to be done for this vision to be realized, as our dataset grows in quantity and quality we believe that the utility of NeuroElectro will lead physiologists to wonder how they worked without it.

4.6 Acknowledgments

We thank Rob Kass, Gordon Shepherd, William Cohen, Etienne Sibille, Anne-Marie Oswald, and Aryn Gittis for helpful discussions and comments on the manuscript. We additionally thank Aryn Gittis for supplying the parvalbumin reporter mouse for targeted neocortical basket cell recordings, and we thank Greg LaRocca for excellent technical support. We are especially grateful to all of the investigators whose collected data are represented within the NeuroElectro database. We also thank the Allen Institute for Brain

Science for making their gene expression datasets and brain atlases publicly available and for funding S.J.T. to attend their 2012 summer hackathon. We thank the academic journal publishers (in particular, Elsevier and Wiley and Highwire) for allowing us access to their full-texts for text-mining. This work was supported by a National Science Foundation Graduate Research Fellowship and a R. K. Mellon Foundation Fellowship (to S.J.T.), an Achievement Rewards for College Scientists Foundation Fellowship and National Institute On Deafness and Other Communication Disorders (NIDCD) National Research Service Award F31DC013490 (to S.D.B.), and NIDCD Grant R01DC005798 (to N.N.U.).

4.7 Methods

4.7.1 Electrophysiological database construction: Overview

We built a custom infrastructure framework for extracting from neuron type-specific electrophysiological measurements, such as R_{input} and AP_{hw}, as well as associated metadata (Figure 4.1 as described in Chapter 3).

4.7.2 Data analysis

Metadata incorporation

To account for the fact that the data were collected under different experimental conditions, we considered the influence of specific metadata attributes which we obtained from the article's methods section through text mining and manual curation. Specifically, we considered the effect of: animal species, animal strain (here we distinguished between strains of rats but not different genetic strains of mice), electrode type (sharp versus patch-clamp), preparation type (*in vitro*, *in vivo*, cell culture), junction potential correction (explicitly corrected, explicitly not corrected, not reported in manuscript), animal age (in days; where animal weight was reported instead of animal age, we manually converted weights reported

in grams to age in days using weight to age conversion tables provided by animal vendors, e.g., `http://www.criver.com/` and `http://www.harlan.com/`), and recording temperature (we assigned reports of "room temperature" recordings to 22℃ and *in vivo* recordings to 37℃). For the purposes of statistical adjustment of metadata (described below), where metadata attributes were not reported or were unidentifiable within an article (which were rare for these attributes), we used mean (or mode) imputation for continuous (or categorical) metadata attributes (Little and Rubin, 2002).

We used statistical models to account for the influence between experimental metadata and measured electrophysiological values. Specifically, we modeled the relationship between electrophysiological measurements and experimental metadata as $\vec{y} = \beta X$ where $\vec{y}$ denotes the vector of electrophysiological measurements corresponding to a single property across all articles (e.g. V_{rest}); X denotes the regressor matrix where rows denote the metadata attributes associated with a single measurement y_i, e.g. $\vec{x_i} = [x_{NeuronType,i}, x_{Species,i}, x_{Strain,i}, ...]$; and β are the regression coefficients denoting the relative contribution of each metadata attribute. We log_{10} transformed measurements of R_{input}, τ_m, AP_{hw}, and animal age to normalize values because these varied across multiple orders of magnitude and/or to enforce that these values remain strictly positive.

When combining the influence of multiple metadata attributes into a single regression model (Fig. 4.2D), we wished to use powerful and flexible models to capture the relationship between metadata and measurement variance while also mitigating the tendency of more complex statistical models to overfit the data. Thus when fitting statistical models, we used stepwise regression methods (implemented as LinearModel.stepwise in MATLAB) to add model terms one-by-one and added terms until either the model's AIC or BIC was optimized (Mitchell, 1997). Furthermore, for each electrophysiological property, we selected the potential model complexity from a set of candidate models (i.e. models that included terms for only: constant, linear, purely quadratic, interaction, interaction + quadratic)

and whether to use either AIC or BIC, two examples of penalized information criteria used to optimize the tradeoff between model fit and complexity, as a model stopping criteria. We selected model complexity and stopping criterion using 10-fold cross-validation and minimizing the sum of squared errors on out of sample data (Supp. Fig. 4.9).

We found that for all electrophysiological properties BIC was chosen as the optimal stopping criterion as opposed to the less conservative AIC measure. Furthermore, usually linear-only models were chosen after cross-validation (a notable exception is spike threshold, where a purely quadratic model had the optimal predictive power). We note that because simple linear-only models were chosen via cross-validation here suggests that we may not have a sufficient amount of electrophysiological measurements and experimental metadata to fit more complex models, such as those that consider interactions between experimental metadata attributes. For example, though evidence suggests that the age dependence of input resistance varies across different types of neurons (e.g., see: (Zhu, 2000; Kinnischtzke et al., 2012)), our statistical models were not able to uncover this relationship following cross-validation. Thus the R^2 values shown in Fig. 4.2C should perhaps be treated as a lower-bound. In reporting the variance explained by different models, we used adj. R^2 to compare between models differing in their number of parameters.

After fitting metadata regression models for each electrophysiological property, we then adjusted each electrophysiological measurement to its predicted value had it been measured in an environment described by the population mean metadata value (or mode for categorical metadata attributes). For example, since the majority of data were recorded using patch-clamp electrodes, we then adjusted measurements made using sharp electrodes to their predicted value had they been recorded using patch-clamp electrodes.

To assess the robustness of the fit of the regression models, we reran the regression analysis on different versions of the dataset where the data were randomly subsampled. Here, we note that in each of these runs of subsampled datasets, we did not rerun the

cross-validation procedure to pick between AIC or BIC or model complexity for purposes of computational tractability.

Electrophysiology property correlation and Neuron type similarity analysis

For analysis of electrophysiological and neuronal correlations, we first pooled data by averaging measurements collected within the same neuron type. We then identified each neuron type with its vector of electrophysiological measurements. We quantified correlations between pairs of electrophysiological properties using Spearman's correlation, which assesses the rank-correlation and allows for detection of relationships that are monotonic but not necessarily linear.

To quantify how much variance across electrophysiological properties could be explained by subsequent principal components (PCs), we needed to first account for missing measurements within our dataset. For example, some neurons did not have a reported measurement for τ_m or AP_{thresh} within our dataset. To address this issue of missing data (Little and Rubin, 2002), we used pPCA, a modification of traditional PCA that is robust to missing data. To further mitigate the problem of missing data, in this analysis we only considered neuron types that were defined by at least 3 different articles and with no more than 2 of the 6 total electrophysiological properties missing; after this filtering step, less than 10% of total electrophysiological observations were missing.

To quantify the electrophysiological similarity of neuron types, we calculated the pairwise Euclidean distances between pairs of neuron types defined by the vector of 6 electrophysiological properties and used a dendrogram analysis to sort neuron types on the basis of electrophysiological similarity. Missing or unobserved electrophysiological measurements were imputed using pPCA. The dendrogram, D, denoting the hierarchical similarity among neuron types was constructed using linkages computed by Ward's minimum variance method. We used multiscale bootstrap resampling to assess the statistical significance of

subtrees of D using the pvclust package in the language R (Suzuki and Shimodaira, 2006).

The pvclust multiscale bootstrap resampling algorithm is sketched as follows: specifically, given an $n \times p$ data matrix M (here, n refers to neuron types and p refers to the 6 electrophysiological properties), pvclust first generates a number of bootstrapped versions of M through randomly sampling columns from M with replacement (here, 10000 bootstrap samples were used). For each bootstrapped data matrix, M_i, a dendrogram D_i was generated through hierarchical clustering. Next, for each subtree in the original dendrogram D, the analysis assesses how often the same subtree appears across the bootstrapped dendrograms $D_{1:10000}$ (this is referred to as the bootstrap probability in Figure 4.5). Here, subtree equality is defined by subtrees that share identical tree topology and neuron membership but does not assess equality of branch lengths. Lastly, because the bootstrap probability is known to be a downwardly biased measure for determining subtree probability (Felsenstein, 2004), pvclust corrects for this downward bias by performing the entire bootstrap procedure multiple times at a number of scales by resampling M to have differing numbers of columns (here, we use 3 through 9 columns in M). This allows for the bootstrap probability to be corrected, yielding the approximately unbiased p-value for each subtree (referred to as the AU p-value in Figure 4.5).

4.7.3 Gene expression analysis

We obtained a dataset on genome-wide whole brain gene expression from the Allen Institute for Brain Science (`http://brain-map.org`). Specifically, using the Allen Institutes's public API (`http://help.brain-map.org/display/mousebrain/API`), we obtained the institutes's dataset on brain-wide gene expression as measured in the adult mouse brain through series of in situ hybridization (ISH) experiments (Lein et al., 2007). Though this data comes from whole-brain assays of gene expression (as opposed to being at the resolution of individual neuron types), this whole-brain data has been registered to a com-

mon neuroanatomical atlas via image processing algorithms, allowing for the average gene expression of individual brain regions and cellular lamina to be quantified. Thus we obtained these brain region averaged datasets corresponding to each gene and ISH experiment ($\sim$20,000 unique genes and mRNA probes; $\sim$26,000 total ISH experiments). We note that we only obtained data corresponding to genes and mRNA probes which passed internal quality controls. We further note that we chose to use the adult mouse gene expression dataset (Lein et al., 2007) as opposed to the developing mouse dataset because of the higher anatomical resolution of the adult atlas relative to the developing mouse atlases (Henry and Hohmann, 2012). Since the Allen Institute dataset provides an estimate of average gene expression at the resolution of brain regions and cell lamina whereas the NeuroElectro dataset provides electrophysiological information at the resolution of neuron types we mapped NeuroElectro neuron types to Allen Institute brain regions (Fig. 4.6). This allows us to approximate neuron type gene expression with brain region gene expression.

In analyzing this data, we used the expression energy measure of regional gene expression and used a log_2 transform to normalize this data. We then a performed Euclidean distance analysis, where for every pair of brain regions we asked how similar these regions were based on their patterns of gene expression (i.e. analogous to our analysis of neuron type similarity based on electrophysiological properties). We performed this analysis using different classes of genes (composed of at least 10 genes), e.g. classes corresponding to ion channels or and transcription factors, using the Gene Ontology (Ashburner et al., 2000).

Our analysis yielded a distance matrix defined by neuron types on the basis of electrophysiological properties and a distance matrix defined by brain regions on the basis of gene expression similarity. To quantify the degree of similarity between these two matrices, i.e. the correspondence between similarity based on genes and similarity based on electrophysiological properties we calculated the Pearson correlation coefficient between the two distance matrices (French et al., 2011). Here, we only considered the upper diagonal ele-

ments. We statistically quantified the significance of this correlation using a Mantel's test, a special permutation test for assessing similarity of pairs of distance matrices (Mantel, 1967).

4.7.4 Electrophysiology

Animals

Hippocampal CA1 recordings were conducted using postnatal day (P)15-16 M72-GFP mice (Potter et al., 2001). Layer 5/6 neocortical pyramidal cell recordings were conducted using P16-18 M72-GFP and Thy1-YFP-G mice (Feng et al., 2000). Main olfactory bulb mitral cell recordings were conducted using P15-18 M72-GFP, Thy1-YFP-G, and C57BL/6 mice. Neocortical basket cell recordings were conducted using a P26 parvalbumin reporter mouse, resulting from a cross between Pvalb-2A-Cre (Allen Institute for Brain Science) and Ai3 (Madisen et al., 2010) lines. A total of 9 mice of both sexes were used in this study. All experiments were completed in compliance with the guidelines established by the Institutional Animal Care and Use Committee of Carnegie Mellon University.

Slice preparation

Mice were anesthetized with isoflurane and decapitated into ice-cold oxygenated dissection solution containing (in mM): 125 NaCl, 25 glucose, 2.5 KCl, 25 $NaHCO_3$, 1.25 NaH_2PO_4, 3 $MgCl_2$, 1 $CaCl_2$. Brains were rapidly isolated and acute slices (310 μm thick) prepared using a vibratome (5000mz-2; Campden). Slices recovered for 30 min in $\sim$37℃ oxygenated Ringer's solution that was identical to the dissection solution except for lower Mg^{2+} concentrations (1 mM $MgCl_2$) and higher Ca^{2+} concentrations (2 mM $CaCl_2$). Slices were then stored in room temperature oxygenated Ringer's solution until recording. Parasagittal slices were used for hippocampal recordings. Parasagittal and coronal slices were used for cortical recordings. Horizontal slices were used for olfactory bulb recordings.

Recording

Slices were continuously superfused with 37℃ oxygenated Ringer's solution during recording. Cells were visualized using infrared differential interference contrast video microscopy. Hippocampal CA1 pyramidal cells and layer 5/6 neocortical pyramidal cells were identified by their large soma size and pyramidal shape, and soma position within their respective cell body layers. Neocortical basket cells were identified by expression of YFP fluorescence. Main olfactory bulb mitral cells were identified by their large cell body size and position within the mitral cell layer. Whole cell recordings were made using electrodes (final electrode resistance: 5.8±1.1 MΩ, $\mu \pm \sigma$) filled with (in mM): 120 K-gluconate, 2 KCl, 10 HEPES, 10 Na-phosphocreatine, 4 Mg-ATP, 0.3 Na_3GTP, 0.2 EGTA, 0-0.25 Alexa Fluor 594 (Life Technologies), and 0.2% Neurobiotin (Vector Labs). Cell morphology was reconstructed under a 100X oil-immersion objective with Neurolucida (MBF Bioscience). No cells included in this dataset exhibited gross morphological truncations. Mitral cells were recorded in the presence of CNQX (10 μM), DL-APV (50 μM), and Gabazine (10 μM) to limit the influence of spontaneous synaptic long-lasting depolarizations on measurement of biophysical properties (Carlson et al., 2000). Data were low-pass filtered at 4 kHz and digitized as 10 kHz using a MultiClamp 700A amplifier (Molecular Devices) and an ITC-18 acquisition board (Instrutech) controlled by custom software written in IGOR Pro (WaveMetrics). The MultiClamp Pipette Offset operation was used to correct for liquid junction potentials before each recording, and solutions were not changed during the course of the recording. Pipette capacitance was neutralized and series resistance (13.4±2.7 MΩ, $\mu \pm \sigma$; range: 8.4-19.6 MΩ) was compensated using the MultiClamp Bridge Balance operation and frequently checked for stability during recordings. After determination of each cell's native V_{rest}, current was injected to normalize V_{rest} to -65, -70, -58, and -70 mV for hippocampal CA1 pyramidal cells, layer 5/6 neocortical pyramidal cells, main olfactory bulb mitral cells, and neocortical basket cells, respectively, before determination of other

biophysical properties.

Analysis

V_{rest} was determined immediately after cell break in. τ_m was calculated from a single-exponential fit to the initial membrane potential response to a hyperpolarizing step current injection. R_{input} was calculated as the slope of the relationship between a series of hyperpolarizing step current amplitudes and the steady-state response of the membrane potential to injections of those step currents. To determine action potential properties of each neuron, a series of 2 s-long depolarizing step currents was injected into the neuron. The first action potential evoked by the weakest suprathreshold step current (i.e., the rheobase input) was used to determine the action potential properties of the neuron. AP_{thr} was calculated as the first point where the membrane potential derivative exceeded 20 mV/ms. AP_{amp} was measured from the point of threshold crossing to the peak voltage reached during the action potential. This amplitude was then used to determine AP_{hw}, calculated as the full action potential width at half maximum amplitude of the action potential.

4.8 Supplemental Figures

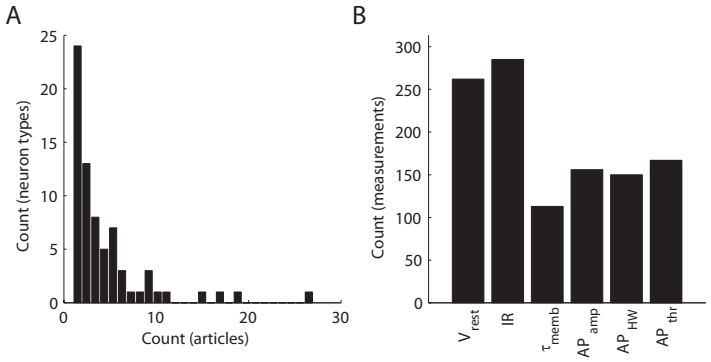

Figure 4.7: Count of distribution neuron types and electrophysiological properties represented in NeuroElectro database. A) Frequency histogram of distribution of neuron types versus number of articles containing information about each neuron type. B) Count of unique measurements for the 6 electrophysiological properties explored in this manuscript.

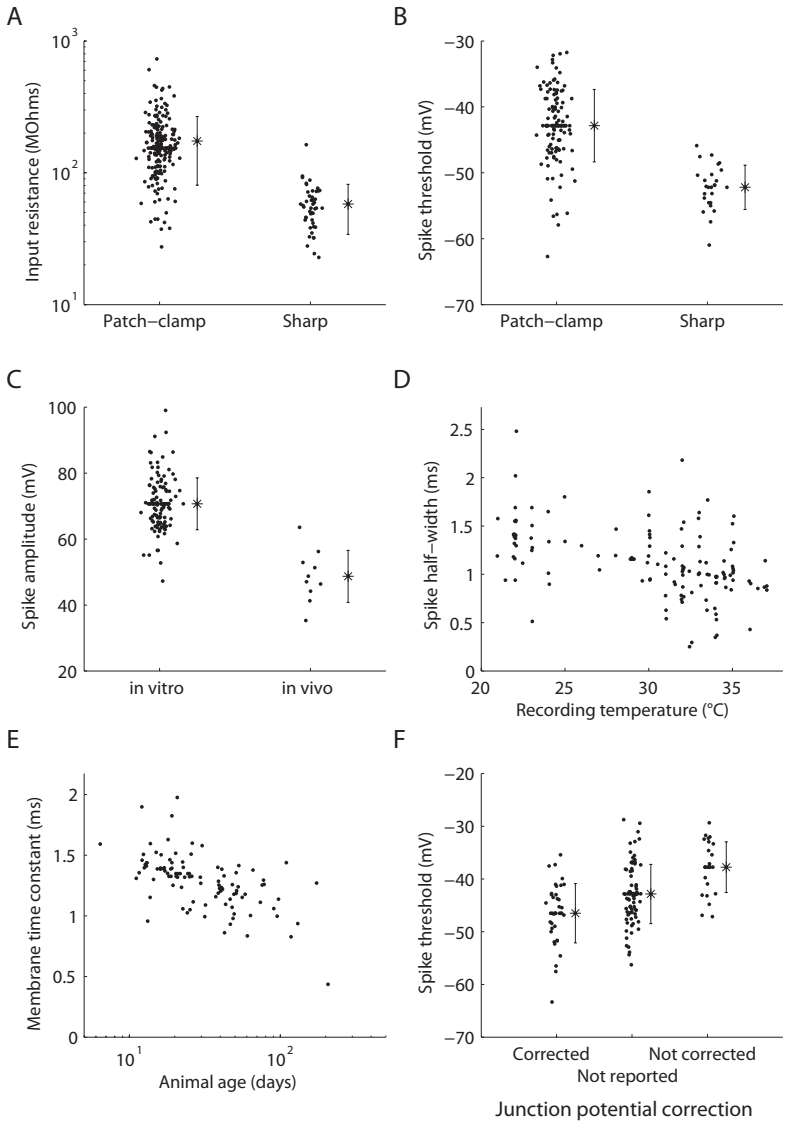

Figure 4.8: Predictive power of metadata attributes for explaining variability in specific electrophysiological properties. A-F) Influence of specific metadata attributes on selected electrophysiological properties. Dots indicate electrophysiological measurement after accounting for the influence of neuron-type specific differences in electrophysiological measurments.

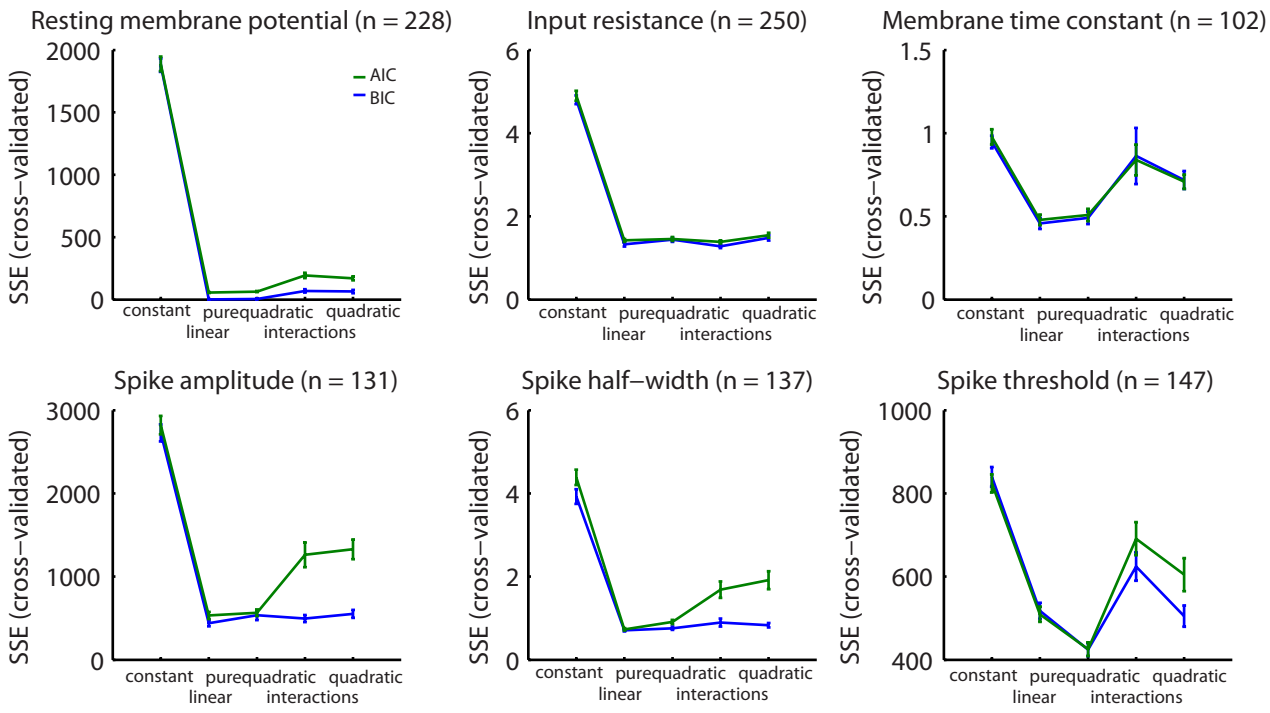

Figure 4.9: Model-selection and cross-validation for assigning optimal statistical model complexity for specific electrophysiological properties. Panels show electrophysiological properties and lines show cross-validated error (as sum of squared error) for fitting statistical models of varying complexity (on abscissa) and using either AIC (green) or BIC (blue) for model selection. Panel title indicates the number of distinct measurements for each electrophysiological property.

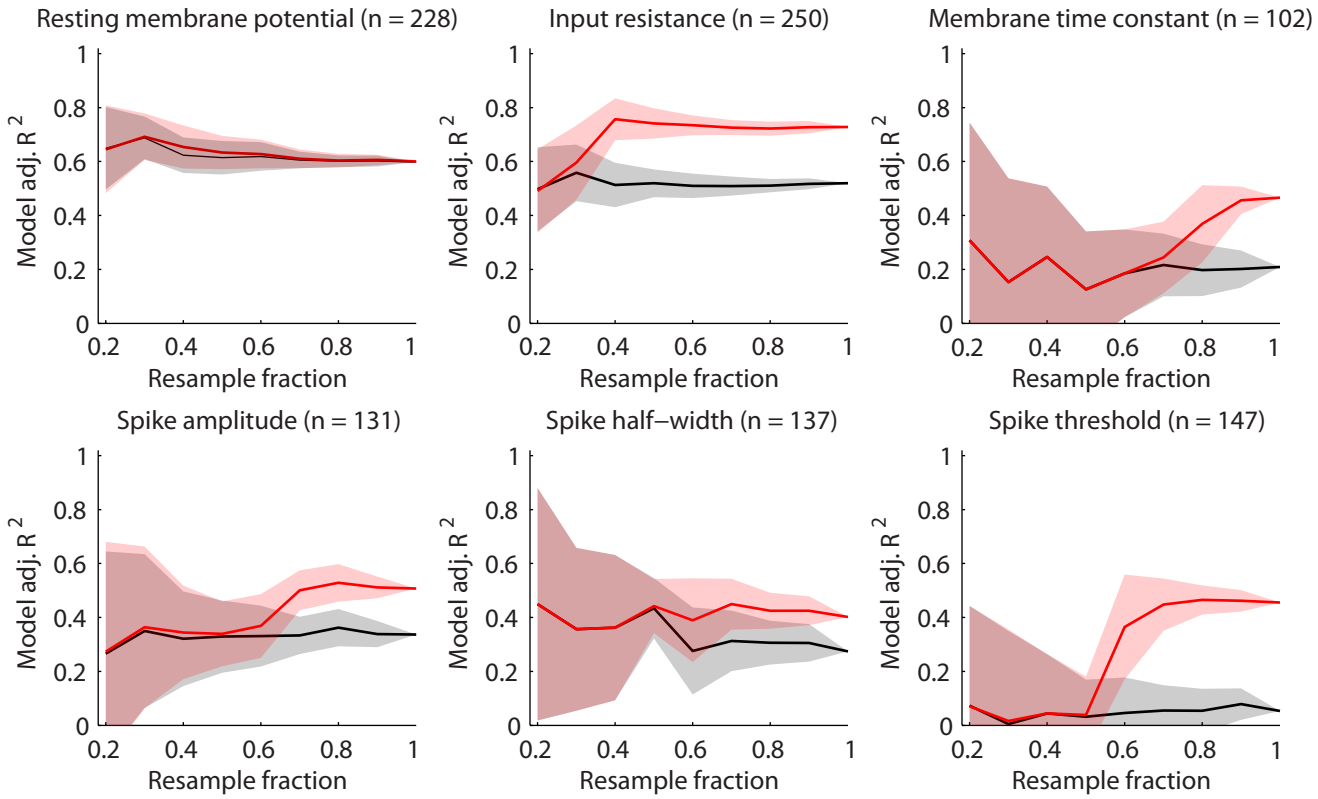

Figure 4.10: Influence of amount of data in dataset and predictive power of metadata for explaining electrophysiological measurement variability. Panels show electrophysiological properties and lines show explanatory power of statistical model when using neuron type information only (black) or neuron type + all metadata (red) as a function of subsampling the original dataset to smaller sizes. Shaded lines indicate standard deviations when subsampling the dataset 25 times per subsampling size. Note that as dataset is subsampling to smaller sizes explanatory power of model including metadata is not greater than model including neuron type information only.

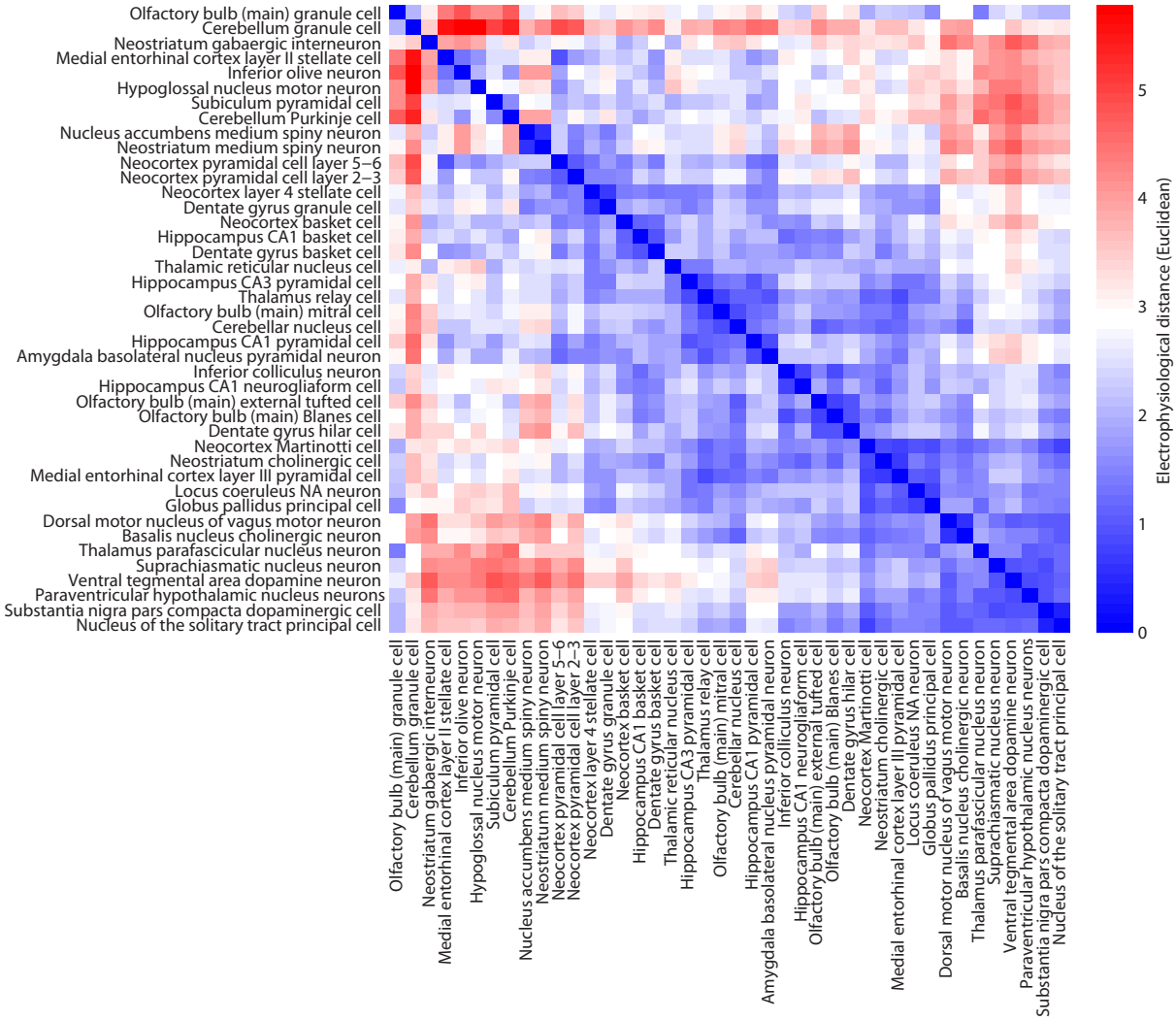

Figure 4.11: Electrophysiological distance matrix of neuron types based on pairwise electrophysiological differences.

Table 4.1: Top 100 gene ontology gene classes which are most correlated with biophysical differences among neuron types.

Index	Ephys-gene correlation	Gene ontology class name	GO ID	number genes
1	0.49913799	negative regulation of oligodendrocyte differentiation	48715	12
2	0.465501172	N-methyl-D-aspartate selective glutamate receptor complex	17146	10
3	0.43893821	establishment of nucleus localization	40023	11
4	0.430445672	regulation of synaptic plasticity	48167	100
5	0.429507439	Rac GTPase binding	48365	23
6	0.426569184	angiotensin receptor binding	31701	11
7	0.425357171	centrosome localization	51642	11
8	0.425252725	type 1 angiotensin receptor binding	31702	10
9	0.421952244	ionotropic glutamate receptor activity	4970	18
10	0.420095334	adipose tissue development	60612	18
11	0.419403771	MAP kinase phosphatase activity	33549	12
12	0.418074423	negative regulation of glial cell differentiation	45686	25
13	0.4180122	Cul3-RING ubiquitin ligase complex	31463	16
14	0.415132168	inactivation of MAPK activity	188	16
15	0.413957933	MAP kinase tyrosine/serine/threonine phosphatase activity	17017	11
16	0.413684152	extracellular-glutamate-gated ion channel activity	5234	17
17	0.410591229	regulation of neuronal synaptic plasticity	48168	46
18	0.409591218	regulation of cation channel activity	2001257	33
19	0.40871743	regulation of transmembrane transporter activity	22898	84
20	0.408706763	positive regulation of cardiac muscle hypertrophy	10613	15
21	0.408657874	regulation of synaptic transmission	50804	211
22	0.404242269	negative regulation of Notch signaling pathway	45746	15
23	0.403316481	neuron-neuron synaptic transmission	7270	50
24	0.401273347	postsynaptic membrane	45211	186
25	0.398605318	dendritic shaft	43198	43
26	0.398037756	negative regulation of ERK1 and ERK2 cascade	70373	28
27	0.392499028	synaptic membrane	97060	217
28	0.392125246	protein phosphatase inhibitor activity	4864	27
29	0.392103023	voltage-gated potassium channel activity	5249	74
30	0.392064356	dendrite	30425	395
31	0.391245681	regulation of ion transmembrane transporter activity	32412	78
32	0.390407449	phosphatase inhibitor activity	19212	30
33	0.390274114	regulation of long-term neuronal synaptic plasticity	48169	31
34	0.389991889	potassium channel complex	34705	52
35	0.388479872	sphingolipid biosynthetic process	30148	31
36	0.386626963	negative regulation of neural precursor cell proliferation	2000178	18
37	0.384949166	regulation of excitatory postsynaptic membrane potential	60079	41
38	0.384322048	voltage-gated potassium channel complex	8076	51
39	0.383579373	protein phosphatase binding	19903	81
40	0.380887788	PDZ domain binding	30165	102
41	0.379858443	dendritic spine	43197	155
42	0.379858443	neuron spine	44309	155
43	0.379840665	postsynaptic density	14069	108
44	0.379840665	dendritic spine head	44327	108
45	0.379188658	oligodendrocyte development	14003	26
46	0.378447761	ionotropic glutamate receptor signaling pathway	35235	23
47	0.377697086	regulation of dendrite morphogenesis	48814	47
48	0.37710019	sarcoplasm	16528	57
49	0.376053512	activation of NF-kappaB-inducing kinase activity	7250	15
50	0.375533061	protein tyrosine/serine/threonine phosphatase activity	8138	35
51	0.375398393	synaptic vesicle exocytosis	16079	21
52	0.374822387	peripheral nervous system development	7422	54
53	0.374474383	behavior	7610	428
54	0.373927266	potassium ion transmembrane transport	71805	83
55	0.373815709	voltage-gated cation channel activity	22843	121
56	0.373746375	potassium channel activity	5267	110
57	0.373672152	extracellular ligand-gated ion channel activity	5230	65
58	0.373276592	regulation of platelet-derived growth factor receptor signaling pathway	10640	11
59	0.372561473	delayed rectifier potassium channel activity	5251	28
60	0.372336137	learning or memory	7611	164
61	0.370890788	cognition	50890	176
62	0.370736119	reproductive behavior	19098	76
63	0.370348115	musculoskeletal movement	50881	22
64	0.369982778	regulation of postsynaptic membrane potential	60078	47
65	0.369639218	cellular response to vascular endothelial growth factor stimulus	35924	22
66	0.369554773	Ras protein signal transduction	7265	83
67	0.369360104	locomotory behavior	7626	164
68	0.369302326	regulation of potassium ion transmembrane transport	1901379	11
69	0.368227647	NIK/NF-kappaB cascade	38061	16
70	0.36711519	calcium ion transmembrane transporter activity	15085	104
71	0.36710319	sarcoplasmic reticulum	16529	49
72	0.366877854	positive regulation of erythrocyte differentiation	45648	16
73	0.366801853	regulation of ion transmembrane transport	34765	231
74	0.366092512	positive regulation of glucose import	46326	29
75	0.365892065	regulation of astrocyte differentiation	48710	25
76	0.365037834	potassium ion transmembrane transporter activity	15079	127
77	0.364975611	cation:cation antiporter activity	15491	19
78	0.364756497	cardiac muscle hypertrophy	3300	17
79	0.363903599	peptide hormone processing	16486	14
80	0.363194258	positive regulation of nitric-oxide synthase activity	51000	11
81	0.363005367	synapse	45202	545
82	0.362675141	vesicle targeting	6903	13
83	0.361875132	positive regulation of RNA splicing	33120	12
84	0.361502683	actin filament-based movement	30048	29
85	0.361490683	ion gated channel activity	22839	273
86	0.360955122	mitochondrial fragmentation involved in apoptotic process	43653	10
87	0.360580451	negative regulation of systemic arterial blood pressure	3085	13
88	0.359844443	Rho protein signal transduction	7266	38
89	0.359615996	synaptic vesicle transport	48489	53
90	0.359559551	barbed-end actin filament capping	51016	10
91	0.359378215	regulation of dendrite development	50773	73
92	0.358277759	protein deubiquitination	16579	54
93	0.358185313	muscle hypertrophy	14896	19
94	0.357846643	chemical homeostasis within a tissue	48875	14
95	0.357739086	voltage-gated ion channel activity	5244	166
96	0.357172857	positive regulation of ion transmembrane transporter activity	32414	26
97	0.356244847	muscle system process	3012	137
98	0.355997733	ionotropic glutamate receptor complex	8328	43
99	0.354981278	costamere	43034	15
100	0.354719497	positive regulation of astrocyte differentiation	48711	10

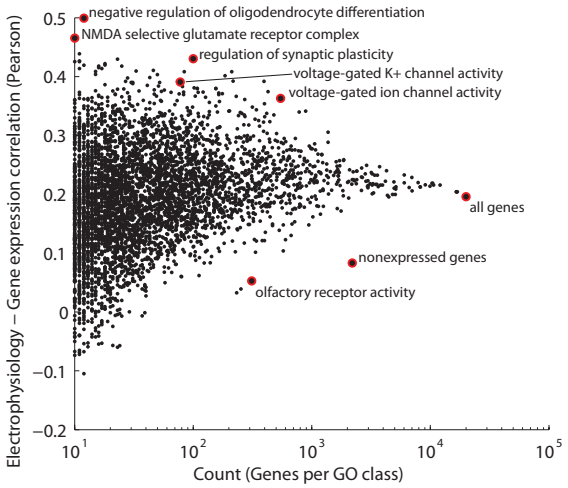

Figure 4.12: Scatterplot of all gene ontology (GO) classes and correlation with electrophysiological distance matrix. Only GO classes containing at least 10 genes were used. Selected examples highlighted in plot.

Chapter 5

Conclusions and Future work

In this concluding chapter I review the main methodological and scientific contributions of this dissertation, and discuss future work that will help develop the ideas explored here even further. Specifically, I revisit and address the 3 motivating questions of this thesis:

1. How do neurons transform their inputs to outputs? How should this transformation be described?
2. How should electrophysiological differences among neurons be quantified?
3. How should researchers communicate results on the electrophysiological findings from neurons?

5.1 Summary

As a computational neuroscientist in an experimental laboratory studying neural circuits, my primary contribution was to develop methodologies and approaches for better understanding the extent, scale, and computational significance of neuronal biophysical diversity. These contributions came in two areas - understanding the consequences of cell-to-cell variability and in creating and analyzing the most comprehensive database of neurophysiological properties ever known.

Specifically, in my first project studying the biophysical variability of olfactory bulb mitral cells (MCs, described in Chapter 2), I developed and applied generalized linear models (GLMs) to help describe and quantify the complex biophysical profiles and input/output functions of single MCs. These models facilitate comparisons between neurons and provide an intuition for how neurons differ among functionally relevant dimensions, such as in their refractory or burst responses following a spike. Moreover, these models provide a computational framework for studying neuronal intrinsic biophysics using "single-cell" approaches, which treat individual neurons as unique (as opposed to the common practice of "lumping" neurons into types and subtypes). Using GLMs, I showed that MCs differ in their stimulus preferences, threshold responses, and in their post-spike behavior, varying from tonic to burst firing. Moreover, there did not appear to be distinct subtypes of MCs but instead varied continuously as a population between these phenotypes.

A key advantage to using GLMs for describing the intrinsic biophysics of single neurons is that the GLMs allowed us to use stimulus decoding methods to study how biophysical properties observed in single neurons and across groups of neurons influence computational function. Thus we could ask how groups of MCs, which differed in their biophysical variability from one group to the next, encoded common afferent stimuli. This methodology allowed us to make predictions on the optimal structure and variability of MC populations from the standpoint of stimulus representation. I found that variability among simulated sister mitral cells (mitral cells which receive primary inputs from a single glomerulus and the same olfactory receptor neuron subtype) allows the population of mitral cells to better and more efficiently represent afferent input into the population. Specifically, when the computational objective is to represent stimuli of a single type (e.g., with the same spectral structure like from a single ORN subtype), I found that the optimal population should be neither maximally diverse nor homogeneous. In summary, the GLM methodology presents a very general approach for capturing neuronal biophysics and linking specific biophysical

phenotypes to computational roles, which could in principle be applied to many other neuron types.

After studying the detailed properties of olfactory bulb mitral cells, I became interested in extending this approach to other neuron types. Just as I had studied how mitral cells differ from one another, I wanted to know how mitral cells as a group differed from neurons elsewhere throughout the brain. Moreover, I wished to address this question using approaches that would best make use of data that had already been collected across decades of neurophysiological investigation, as opposed to collecting data anew within our lab. Thus I became interested in the prospect of "mining" information on the biophysical properties of different neuron types directly from the research literature.

I developed and applied methods for extracting from the cellular neurophysiology research literature measurements of basic biophysical properties across the major neuron types in the mammalian nervous system (described in Chapter 3). Unlike the single-cell high resolution voltage traces and spike train data from the earlier mitral cell project, the data I could most reliably extract were investigator-reported measurements that summarized the average properties of a number of recorded neurons of the same type. In extracting this data, I found the need to further develop additional tools and resources, such as an ontology describing basic electrophysiological properties as well as methods for extracting experimental conditions in which experiments were conducted (i.e. experimental metadata). Moreover, in developing the database, because of its size and complexity I incurred substantial challenges in even navigating and visualizing the dataset. Thus, I built a web interface for the data, `neuroelectro.org`, and specifically developed features to allow interactive exploration of the dataset, such as visualizing electrophysiological measurements and publications corresponding to a specific neuron type. Realizing that this resource could be of use to the greater community, we decided from an early stage to make the web interface and corresponding data publicly available on the Internet.

Having compiled this unprecedented dataset of biophysical properties of 98 neuron types from 326 publications (using automated methods to sort through over 92,000 articles from 15 journals), I wanted to analyze the NeuroElectro dataset in similar ways to my earlier analysis of mitral cell biophysical variability; namely, I wished to better understand neuron type biophysical diversity. For example, do neuron types throughout the brain vary continuously in their biophysical properties or are there distinct categories or clusters of biophysically-related neuron types?

Analyzing the pooled NeuroElectro dataset presented unique challenges; for example, unlike the mitral cell dataset, the NeuroElectro data were collected by hundreds of investigators who in turn use diverse experimental methodologies and preparations. Thus if two neuron types differed in their biophysical measurements, I could not attribute whether these were due to real differences between these neurons or were simply due to how the experiments were conducted. This would be especially problematic if investigators studying specific types of neurons systematically use different experimental conditions. By utilizing the experimental metadata I had collected from each publication, I developed and fit multivariate linear regression models which accounted for the influence of experimental conditions on biophysical measurements. I found that knowledge of specific experimental conditions was significantly predictive of the observed variance among electrophysiological measurements collected across different publications. Moreover, the resulting regression models were consistent with known relationships between neuronal biophysics and experimental conditions, for example, that sharp electrode recordings underestimate measurements of input resistance relative to patch clamp electrodes (Spruston and Johnston, 1992; Staley et al., 1992).

Another methodological issue with the NeuroElectro dataset was its piecemeal nature, since I could only reliably extract measurements that authors explicitly chose to include in a formatted data table. Thus I was faced with a "missing data" problem: how could I

analyze across publications and neuron types with incomplete datasets? Namely, standard techniques like dimensionality reduction using principle component analysis and hierarchical clustering usually require that there are no missing or unobserved measurements. My approach to this problem was to pool data across publications corresponding to the same neuron type and then use imputation to fill in missing values based on the existing data and the overall correlation structure of biophysical properties.

Having normalized the dataset of biophysical properties across neuron types, I could then study similarities among neuron types on the basis of their biophysical properties. I found that biophysical properties are very correlated with one another and that the primary axis of neuron diversity is qualitatively related to neuronal electrotonic size. Further analyzing neuron differences, I compared neuron types based on their vector of basic biophysical properties and sorted them hierarchically based on biophysical similarity. Utilizing methods for comparing and quantifying hierarchies from phylogentic analysis (Felsenstein, 2004), I found that the neuron types fit into one of approximately 7 neuron super-classes including large glutamatergic projection neurons and a class of neurons defined by high input resistances and depolarized membrane potentials and diverse neurochemical release. Interestingly, these neuron super-classes differ from previous classifications based on somatodendritic distribution of ion channel currents (Migliore and Shepherd, 2002, 2005).

Beyond this descriptive analysis of biophysical diversity, I wished to better understand the mechanistic basis of neuronal biophysical diversity. Through mapping neuron types in NeuroElectro (i.e., neurons defined by NeuroLex (Shepherd, 2003; Larson and Martone, 2013)) to brain regions in the Allen Institute mouse brain atlas (Lein et al., 2007), I could use the Allen Institute's genome-wide transcription datasets as an approximate estimate for neuron type-specific gene expression. Thus I could search for classes of genes whose expression highly correlated with neuronal biophysical differences. I found that the genes that were expected to be significantly correlated with neuronal biophysics, such as genes

encoding voltage-gated ion channels and ion transporters, were among the most predictive gene classes. However, I also observed additional highly-predictive gene classes which we did not expect *a priori*, such as NMDA receptor coding genes, genes known to regulate synaptic plasticity, and genes which regulate oligodendrocite differentiation. These unexpected gene classes are exciting because they present novel hypotheses which can be tested empirically in future experiments.

To summarize NeuroElectro, I view the resource and methodology as a novel approach for explicitly bridging results across the numerous subfields of cellular neurophysiology. Beyond its role as an index for helping physiologists locate references on specific neuron types or providing data and parameters to computational modelers, I believe that the power of NeuroElectro is in tying many results together and subsequently using the sum total to help generate hypotheses and drive novel experiments.

5.2 Limitations of current approaches and discussion of potential solutions

Many of the scientific results from Chapter 2 on the computational implications of mitral cell diversity rely on the assumption that GLMs sufficiently model the response properties of individual MCs. While GLMs adequately capture many features of MC activity, one feature the GLMs insufficiently model is temporally precise and reliable trial-to-trial MC spiking evoked across multiple trials of the same "frozen noise" stimulus. In ongoing work, I have been collaborating with Wanjie Wang, a CMU statistics PhD student, to adapt the GLM models to better capture this important facet of MC activity. Moreover, another important question with the GLM approach is its generality: are the recovered parameters stable over the course of the experiment? do they generalize from *in vitro* to *in vivo*? These questions are important and need to be further explored before the GLM methodology can

be expanded into other contexts.

As described in Chapters 3 and 4, one of the main limitations of the NeuroElectro project is that the data are collected across a number of investigators in the absence of formally-defined standards for data collection and reporting (however, there are informal standards such as the Petilla terminology (Ascoli et al., 2008)). For example, neuronal properties such as action potential amplitude tend to be collected and calculated using different methods by investigators. Thus when measurements collected using inconsistent methodologies are mined into the NeuroElectro database but have not been appropriately flagged by a human curator, these measurements will tend to add unexplained variance or "noise" into the pooled dataset. This issue of reconciling multiple investigator methodologies is a common challenge faced by all literature-based approaches, such as NeuroSynth and WhiteText (Yarkoni et al., 2011; French et al., 2012). While I feel that one component to the solution for this problem is further metadata extraction (e.g., keeping track of how specific properties are calculated within each article), I feel that a complementary approach is to spark a greater conversation about data standards and collection practices within the larger community. To this end, I have been working with collaborators from Elsevier Research Data Services (`http://researchdata.elsevier.com/`) and scientific reproducibility experts (Vasilevsky et al., 2013) on improved methods for tracking data and metadata in cellular neurophyisology.

5.3 Future Work

Moving forward, my goal is to work towards integrating the two somewhat disparate projects of the my thesis: the first, a high-resolution description and functional analysis of the intrinsic variability inherent to a single neuron type; the second, a low-resolution analysis of neuron diversity across the major neuron types throughout the brain. Specifically, since the mitral cell project was conducted within a single lab, we could optimize

experimental design, data collection, and analysis for making specific hypotheses on the optimal structure of biophysical variability among mitral cells. Contrastingly, for my second project where I developed and populated the NeuroElectro database from existing literature, for technical reasons I was limited in data access because I could only collect coarse descriptions of neuronal biophysics and confirm them in a few specific cases. For example, given that I was relying on literature-published values as the source of data, I could only extract parameters such as neuronal resting membrane potentials and input resistances as opposed to richer descriptions of neuronal activity like voltage traces and spike trains. Thus my subsequent analyses of brain-wide neuronal intrinsic diversity are likely limited by the relatively low-resolution of the collected data.

In the future, my goal is to see these two approaches begin to merge. Specifically, through my role as the creator and administrator of NeuroElectro, I am in a unique position to make suggestions to the greater community for how neurophysiological data should be collected and stored. For example, if investigators wished to voluntarily upload and share their raw data through NeuroElectro, I could make suggestions for what kinds of data would be preferred for optimal utility and visibility. Moreover, I could ensure that the uploaded data be amenable for fitting general and powerful computational neuron models, like the GLMs I employed for studying mitral cell variability. Thus, as I asked how mitral cell biophysics influence stimulus encoding, one could use these submitted data to construct models to ask how the biophysical properties of diverse neuron types influence stimulus coding.

Furthermore, there is increasing interest in producing a comprehensive catalog of the neuron types throughout the brain. For example, the first goal of the NIH portion of the US federal BRAIN initiative for fiscal year 2014 is to "generate a census of cell types" (`www.nih.gov/science/brain/ACD_BRAIN_interimreport_executivesummary.htm`). Though it is presently unclear what constitutes such a census, NeuroElectro in its current state arguably

makes considerable progress towards such a goal. Moreover, such a census will likely bridge multiple aspects of neuron physiology to include information on gene expression patterns, intrinsic biophysics, morphology, and connectivity. As NeuroElectro makes explicit connections with additional modalities (as illustrated in Chapter 4 for gene expression), the power of the resource to drive additional hypotheses should only increase. For example, my current efforts include integrating links to and from NeuroElectro with existing resources such as NeuronDB (Crasto et al., 2007) and NeuroLex (Larson and Martone, 2013) and allowing dynamic query of NeuroElectro data via an API (`http://neuroelectro.org/api/docs/`). Additionally, NeuroElectro at present only contains information about neuron types in their control or unperturbed states. Given the involvement of specific neuron types in neurological disorders (Marín, 2012) or in active versus passive mental states (Gentet et al., 2010, 2012), cataloging how neuronal properties change in these perturbed states will become critical for understanding the functional roles of diverse neurons.

Bibliography

Aizenman, C. D., Huang, E. J., and Linden, D. J. (2003). Morphological correlates of intrinsic electrical excitability in neurons of the deep cerebellar nuclei. *Journal of neurophysiology*, 89(4):1738–1747. PMID: 12686564. 4.4.4

Akil, H., Martone, M. E., and Van Essen, D. C. (2011). Challenges and opportunities in mining neuroscience data. *Science*, 331(6018):708 –712. 4.3

Aldous, D. J. (1985). Exchangeability and related topics. In Hennequin, P. L., editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, number 1117 in Lecture Notes in Mathematics, pages 1–198. Springer Berlin Heidelberg. 3.4.3

Alivisatos, A. P., Chun, M., Church, G. M., Greenspan, R. J., Roukes, M. L., and Yuste, R. (2012). The brain activity map project and the challenge of functional connectomics. *Neuron*, 74(6):970–974. 4.3, 4.5

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410. PMID: 2231712. 4.3, 4.4.5, 4.5.3

Altschuler, S. J. and Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell*, 141(4):559–563. PMID: 20478246. 2.3

Ambert, K. H. and Cohen, A. M. (2012). Text-mining and neuroscience. *International review of neurobiology*, 103:109–132. PMID: 23195123. 4.3, 4.4.1, 4.5.2

Angelo, K. and Margrie, T. W. (2011). Population diversity and function of hyperpolarization-activated current in olfactory bulb mitral cells. *Sci. Rep.*, 1. 2.3, 2.4.1, 2.5

Angelo, K., Rancz, E. A., Pimentel, D., Hundahl, C., Hannibal, J., Fleischmann, A., Pichler, B., and Margrie, T. W. (2012). A biophysical signature of network affiliation and sensory processing in mitral cells. *Nature*, 488(7411):375–378. PMID: 22820253. 1.1.4, 4.4.2

Antal, M., Eyre, M., Finklea, B., and Nusser, Z. (2006). External tufted cells in the main olfactory bulb form two distinct subpopulations. *The European journal of neuroscience*, 24(4):1124–1136. PMID: 16930438. 1.1.2

Arevian, A. C., Kapoor, V., and Urban, N. N. (2008). Activity-dependent gating of lateral inhibition in the mouse olfactory bulb. *Nat Neurosci*, 11(1):80–87. 2.5

Ascoli, G. A., Alonso-Nanclares, L., Anderson, S. A., Barrionuevo, G., Benavides-Piccione,

R., Burkhalter, A., Buzsaki, G., Cauli, B., DeFelipe, J., Fairen, A., Feldmeyer, D., Fishell, G., Fregnac, Y., Freund, T. F., Gardner, D., Gardner, E. P., Goldberg, J. H., Helmstaedter, M., Hestrin, S., Karube, F., Kisvarday, Z. F., Lambolez, B., Lewis, D. A., Marin, O., Markram, H., Munoz, A., Packer, A., Petersen, C. C. H., Rockland, K. S., Rossier, J., Rudy, B., Somogyi, P., Staiger, J. F., Tamas, G., Thomson, A. M., Toledo-Rodriguez, M., Wang, Y., West, D. C., and Yuste, R. (2008). Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nature Reviews Neuroscience*, 9(7):557–568. WOS:000256929300015. 1, 1.1.2, 1.1.4, 3.4.3, 4.4.1, 4.5.2, 5.2

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29. PMID: 10802651. 4.4.6, 4.7.3

Atick, J. J. and Redlich, A. N. (1993). Convergent algorithm for sensory receptive field development. *Neural Computation*, 5(1):45–60. 2.4.2

Azouz, R. and Gray, C. M. (1999). Cellular mechanisms contributing to response variability of cortical neurons in vivo. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 19(6):2209–2223. PMID: 10066274. 2.3

Badel, L., Lefort, S., Berger, T., Petersen, C., Gerstner, W., and Richardson, M. (2008). Extracting non-linear integrate-and-fire models from experimental data using dynamic I–V curves. *Biological Cybernetics*, 99(4):361–370. 1.1.3

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48. PMID: 10592178. 4.3

Barth, A. L. and Poulet, J. F. A. (2012). Experimental evidence for sparse firing in the neocortex. *Trends in neurosciences*, 35(6):345–355. PMID: 22579264. 4.4.5

Bean, B. P. (2007). The action potential in mammalian central neurons. *Nature Reviews Neuroscience*, 8(6):451–465. 1.1.2

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). GenBank. *Nucleic acids research*, 41(Database issue):D36–42. PMID: 23193287. 1.1.5, 4.3, 4.4.5, 4.5.3

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E F, J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology*, 112(3):535–542. PMID: 875032. 1.1.5

Bhalla, U. S. and Bower, J. M. (1993). Exploring parameter space in detailed single neuron models: simulations of the mitral and granule cells of the olfactory bulb. *Journal of Neurophysiology*, 69(6):1948–1965. PMID: 7688798. 1.1.3, 2.5

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly, Beijing; Cambridge [Mass.]. 3.4.6

Borst, A. and Haag, J. (2001). Effects of mean firing on neural information rate. *Journal of Computational Neuroscience*, 10(2):213–221. 2.4.2

Bota, M., Dong, H.-W., and Swanson, L. W. (2005). Brain architecture management system. *Neuroinformatics*, 3(1):15–48. PMID: 15897615. 1.1.5, 3.3

Brette, R., Rudolph, M., Carnevale, T., Hines, M., Beeman, D., Bower, J. M., Diesmann, M., Morrison, A., Goodman, P. H., Harris, Frederick C, J., Zirpe, M., Natschläger, T., Pecevski, D., Ermentrout, B., Djurfeldt, M., Lansner, A., Rochel, O., Vieville, T., Muller, E., Davison, A. P., El Boustani, S., and Destexhe, A. (2007). Simulation of networks of spiking neurons: a review of tools and strategies. *Journal of computational neuroscience*, 23(3):349–398. PMID: 17629781. 1.1.3

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65(1):175–187. PMID: 1840504. 4.4.6

Bug, W. J., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A. R., Larson, S. D., Rubin, D., Shepherd, G. M., Turner, J. A., and Martone, M. E. (2008). The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, 6(3):175–194. PMID: 18975148. 1.1.5

Butts, D. A., Weng, C., Jin, J., Yeh, C.-I., Lesica, N. A., Alonso, J.-M., and Stanley, G. B. (2007). Temporal precision in the neural code and the timescales of natural vision. *Nature*, 449(7158):92–95. 2.3

Carey, R. M. and Wachowiak, M. (2011). Effect of sniffing on the temporal structure of Mitral/Tufted cell output from the olfactory bulb. *The Journal of Neuroscience*, 31(29):10615 –10626. 2.4.4, 2.5

Carlson, G. C., Shipley, M. T., and Keller, A. (2000). Long-lasting depolarizations in mitral cells of the rat olfactory bulb. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 20(5):2011–2021. PMID: 10684902. 4.7.4

Carrasquillo, Y. and Nerbonne, J. M. (2013). IA channels: Diverse regulatory mechanisms. *The Neuroscientist: a review journal bringing neurobiology, neurology and psychiatry*. PMID: 24106264. 1.1.1

Coetzee, W. A., Amarillo, Y., Chiu, J., Chow, A., Lau, D., McCORMACK, T., Morena, H., Nadal, M. S., Ozaita, A., Pountney, D., Saganich, M., De Miera, E. V.-S., and Rudy, B. (1999). Molecular diversity of k+ channels. *Annals of the New York Academy of Sciences*, 868(1):233–255. 1.1.1, 4.4.6, 4.4.6

Connors, B. W. and Gutnick, M. J. (1990). Intrinsic firing patterns of diverse neocortical neurons. *Trends in neurosciences*, 13(3):99–104. PMID: 1691879. 1.1.4

Connors, B. W., Gutnick, M. J., and Prince, D. A. (1982). Electrophysiological properties of neocortical neurons in vitro. *Journal of neurophysiology*, 48(6):1302–1320. PMID: 6296328. 1, 1.1.1, 1.1.2, 1.1.4, 3.4.2

Crasto, C. J., Marenco, L. N., Liu, N., Morse, T. M., Cheung, K.-H., Lai, P. C., Bahl,

G., Masiar, P., Lam, H. Y. K., Lim, E., Chen, H., Nadkarni, P., Migliore, M., Miller, P. L., and Shepherd, G. M. (2007). SenseLab: new developments in disseminating neuroscience information. *Briefings in bioinformatics*, 8(3):150–162. PMID: 17510162. 1.1.1, 1.1.5, 4.3, 5.3

Davison, A. P., Feng, J., and Brown, D. (2003). Dendrodendritic inhibition and simulated odor responses in a detailed olfactory bulb network model. *Journal of neurophysiology*, 90(3):1921–1935. PMID: 12736241. 1.1.3

De Schutter, E. and Bower, J. M. (1994). An active membrane model of the cerebellar purkinje cell. i. simulation of current clamps in slice. *Journal of neurophysiology*, 71(1):375–400. PMID: 7512629. 1.1.3

DeFelipe, J. (1997). Types of neurons, synaptic connections and chemical characteristics of cells immunoreactive for calbindin-D28K, parvalbumin and calretinin in the neocortex. *Journal of chemical neuroanatomy*, 14(1):1–19. PMID: 9498163. 1.1.4

Depienne, C., Trouillard, O., Saint-Martin, C., Gourfinkel-An, I., Bouteiller, D., Carpentier, W., Keren, B., Abert, B., Gautier, A., Baulac, S., Arzimanoglou, A., Cazeneuve, C., Nabbout, R., and LeGuern, E. (2009). Spectrum of SCN1A gene mutations associated with dravet syndrome: analysis of 333 patients. *Journal of Medical Genetics*, 46(3):183–191. PMID: 18930999. 1, 1.1.1

Dhawale, A. K., Hagiwara, A., Bhalla, U. S., Murthy, V. N., and Albeanu, D. F. (2010). Non-redundant odor coding by sister mitral cells revealed by light addressable glomeruli in the mouse. *Nature Neuroscience*. PMID: 20953197. 2.5

Dickman, S. (2003). Tough mining. *PLoS Biol*, 1(2):e48. 3.3, 4.4.1, 4.5.2

Douglas, R. J. and Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, 27(1):419–451. PMID: 15217339. 4.4.6

Druckmann, S., Berger, T. K., Schürmann, F., Hill, S., Markram, H., and Segev, I. (2011). Effective stimuli for constructing reliable neuron models. *PLoS Comput Biol*, 7(8):e1002133. 1.1.3

Druckmann, S., Hill, S., Schürmann, F., Markram, H., and Segev, I. (2012). A hierarchical structure of cortical interneuron electrical diversity revealed by automated statistical analysis. *Cerebral cortex (New York, N.Y.: 1991)*. PMID: 22989582. 1.1.2, 4.4.5

Ermentrout, B. (2002). *Simulating, analyzing, and animating dynamical systems: a guide to XPPAUT for researchers and students*. Society for Industrial and Applied Mathematics, Philadelphia. 1.1.3

Fanselow, E. E. and Connors, B. W. (2010). The roles of somatostatin-expressing (GIN) and fast-spiking inhibitory interneurons in UP-DOWN states of mouse neocortex. *Journal of neurophysiology*, 104(2):596–606. PMID: 20538767. 4.4.5

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates, Sunderland, Mass. 4.4.5, 4.7.2, 5.1

Feng, G., Mellor, R. H., Bernstein, M., Keller-Peck, C., Nguyen, Q. T., Wallace, M., Nerbonne, J. M., Lichtman, J. W., and Sanes, J. R. (2000). Imaging neuronal subsets

in transgenic mice expressing multiple spectral variants of GFP. *Neuron*, 28(1):41–51. PMID: 11086982. 4.7.4

Field, G. D. and Chichilnisky, E. J. (2007). Information processing in the primate retina: circuitry and coding. *Annual Review of Neuroscience*, 30:1–30. PMID: 17335403. 1.1.4

Flower, D. R. and Attwood, T. K. (2004). Integrative bioinformatics for functional genome annotation: trawling for g protein-coupled receptors. *Seminars in cell & developmental biology*, 15(6):693–701. PMID: 15561589. 4.3

French, L., Lane, S., Xu, L., and Pavlidis, P. (2009). Automated recognition of brain region mentions in neuroscience literature. *Frontiers in Neuroinformatics*, 3:29. 1.1.5, 3.3, 3.4.6, 3.5.4, 4.3, 4.4.1

French, L., Lane, S., Xu, L., Siu, C., Kwok, C., Chen, Y., Krebs, C., and Pavlidis, P. (2012). Application and evaluation of automated methods to extract neuroanatomical connectivity statements from free text. *Bioinformatics*, 28(22):2963–2970. PMID: 22954628. 1.1.5, 3.3, 5.2

French, L. and Pavlidis, P. (2012). Using text mining to link journal articles to neuroanatomical databases. *The Journal of Comparative Neurology*, 520(8):1772–1783. 3.5.4

French, L., Tan, P. P. C., and Pavlidis, P. (2011). Large-scale analysis of gene expression and connectivity in the rodent brain: insights through data integration. *Frontiers in Neuroinformatics*, 5:12. 4.7.3

Friedrich, R. W. (2006). Mechanisms of odor discrimination: neurophysiological and behavioral approaches. *Trends in Neurosciences*, 29(1):40–47. PMID: 16290274. 2.4.2

Gage, F. H. (2000). Mammalian neural stem cells. *Science*, 287(5457):1433–1438. PMID: 10688783. 4.4.5

Galán, R. F., Ermentrout, G. B., and Urban, N. N. (2008). Optimal time scale for spike-time reliability: theory, simulations, and experiments. *Journal of Neurophysiology*, 99(1):277–283. PMID: 17928562. 2.6.1

Galán, R. F., Fourcaud-Trocmé, N., Ermentrout, G. B., and Urban, N. N. (2006). Correlation-induced synchronization of oscillations in olfactory bulb neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 26(14):3646–3655. PMID: 16597718. 1.1.3

Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., Halavi, M., Kennedy, D. N., Marenco, L., Martone, M. E., Miller, P. L., Müller, H.-M., Robert, A., Shepherd, G. M., Sternberg, P. W., Van Essen, D. C., and Williams, R. W. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics*, 6(3):149–160. PMID: 18946742. 1.1.5

Gentet, L. J., Avermann, M., Matyas, F., Staiger, J. F., and Petersen, C. C. H. (2010). Membrane potential dynamics of GABAergic neurons in the barrel cortex of behaving mice. *Neuron*, 65(3):422–435. PMID: 20159454. 5.3

Gentet, L. J., Kremer, Y., Taniguchi, H., Huang, Z. J., Staiger, J. F., and Petersen, C. C. H. (2012). Unique functional properties of somatostatin-expressing GABAergic neurons in mouse barrel cortex. *Nature neuroscience*, 15(4):607–612. PMID: 22366760. 5.3

Gerstner, W. and Naud, R. (2009). How good are neuron models? *Science*, 326(5951):379–380. 2.5

Giocomo, L. M., Zilli, E. A., Fransén, E., and Hasselmo, M. E. (2007). Temporal frequency of subthreshold oscillations scales with entorhinal grid cell field spacing. *Science (New York, N.Y.)*, 315(5819):1719–1722. PMID: 17379810. 1.1.1

Giridhar, S., Doiron, B., and Urban, N. N. (2011). Timescale-dependent shaping of correlation by olfactory bulb lateral inhibition. *Proceedings of the National Academy of Sciences*. 2.3, 2.5

Glaser, J. R. and Glaser, E. M. (1990). Neuron imaging with neurolucida–a PC-based system for image combining microscopy. *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, 14(5):307–317. PMID: 2224829. 1.1.5

Gleeson, P., Crook, S., Cannon, R. C., Hines, M. L., Billings, G. O., Farinella, M., Morse, T. M., Davison, A. P., Ray, S., Bhalla, U. S., Barnes, S. R., Dimitrova, Y. D., and Silver, R. A. (2010). NeuroML: a language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS computational biology*, 6(6):e1000815. PMID: 20585541. 1.1.5

Hamilton, D. J., Shepherd, G. M., Martone, M. E., and Ascoli, G. A. (2012). An ontological approach to describing neurons and their relationships. *Frontiers in Neuroinformatics*, 6:15. 1, 1.1.4, 1.1.5, 3.4.3, 4.4.1, 4.5.2

Harmar, A. J., Hills, R. A., Rosser, E. M., Jones, M., Buneman, O. P., Dunbar, D. R., Greenhill, S. D., Hale, V. A., Sharman, J. L., Bonner, T. I., Catterall, W. A., Davenport, A. P., Delagrange, P., Dollery, C. T., Foord, S. M., Gutman, G. A., Laudet, V., Neubig, R. R., Ohlstein, E. H., Olsen, R. W., Peters, J., Pin, J.-P., Ruffolo, R. R., Searls, D. B., Wright, M. W., and Spedding, M. (2009). IUPHAR-DB: the IUPHAR database of g protein-coupled receptors and ion channels. *Nucleic Acids Research*, 37(suppl 1):D680–D685. PMID: 18948278. 1.1.1

Heintz, N. (2004). Gene expression nervous system atlas (GENSAT). *Nature Neuroscience*, 7(5):483–483. 1.1.4

Helmstaedter, M., Briggman, K. L., Turaga, S. C., Jain, V., Seung, H. S., and Denk, W. (2013). Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174. PMID: 23925239. 1.1.4

Henry, A. M. and Hohmann, J. G. (2012). High-resolution gene expression atlases for adult and developing mouse brain and spinal cord. *Mammalian genome: official journal of the International Mammalian Genome Society*, 23(9-10):539–549. PMID: 22832508. 4.7.3

Hille, B. (2001). *Ion Channels of Excitable Membranes*. Sinauer Associates Inc 2001-07,

3rd edition edition. 1.1.1, 1.1.3, 4.4.1

Hines, M. L. and Carnevale, N. T. (1997). The NEURON simulation environment. *Neural computation*, 9(6):1179–1209. PMID: 9248061. 1.1.3

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544. PMID: 12991237. 1.1.1, 1.1.3

Huang, J. and Zeng, H. (2013). Genetic approaches to neural circuits in the mouse. *Annual Review of Neuroscience*, 36(1):183–215. PMID: 23682658. 1.1.4

Imam, F. T., Larson, S. D., Bandrowski, A., Grethe, J. S., Gupta, A., and Martone, M. E. (2012). Development and use of ontologies inside the neuroscience information framework: A practical approach. *Frontiers in genetics*, 3:111. PMID: 22737162. 1.1.5

Insel, T. R., Landis, S. C., and Collins, F. S. (2013). The NIH BRAIN initiative. *Science*, 340(6133):687–688. PMID: 23661744. 4.3, 4.5

Izhikevich, E. M. (2010). *Dynamical systems in neuroscience: the geometry of excitability and bursting*. MIT Press, Cambridge, Mass.; London. 1

Johnston, D. and Wu, S. M.-s. (1995). *Foundations of cellular neurophysiology*. MIT Press, Cambridge, Mass. 4.4.1

Jolivet, R., Schürmann, F., Berger, T. K., Naud, R., Gerstner, W., and Roth, A. (2008). The quantitative single-neuron modeling competition. *Biological Cybernetics*, 99(4-5):417–426. PMID: 19011928. 2.5

Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R., and Koch, C. (2013). Neuroscience thinks big (and collaboratively). *Nature Reviews Neuroscience*, 14(9):659–664. 1.1.6, 4.3

Kass, R. E. and Ventura, V. (2001). A spike-train probability model. *Neural Computation*, 13(8):1713–1720. 1.1.3, 2.3, 2.4.1

Kawaguchi, Y. and Kubota, Y. (1997). GABAergic cell subtypes and their synaptic connections in rat frontal cortex. *Cerebral Cortex*, 7(6):476–486. PMID: 9276173. 1.1.4

Khan, A. G., Sarangi, M., and Bhalla, U. S. (2012). Rats track odour trails accurately using a multi-layered strategy with near-optimal sampling. *Nature Communications*, 3:703. 2.4.3

Kinnischtzke, A. K., Sewall, A. M., Berkepile, J. M., and Fanselow, E. E. (2012). Postnatal maturation of somatostatin-expressing inhibitory cells in the somatosensory cortex of GIN mice. *Frontiers in neural circuits*, 6:33. PMID: 22666189. 4.4.2, 4.7.2

Koch, C. (1999). *Biophysics of computation: information processing in single neurons*. Oxford University Press, New York. 1, 1.1.3

Kreitzer, A. C. (2009). Physiology and pharmacology of striatal neurons. *Annual Review of Neuroscience*, 32(1):127–147. PMID: 19400717. 4.4.1, 4.4.5

Lampl, I. and Yarom, Y. (1997). Subthreshold oscillations and resonant behavior: two manifestations of the same mechanism. *Neuroscience*, 78(2):325–341. PMID: 9145790.

1.1.1

Larimer, P. and Strowbridge, B. W. (2010). Representing information in cell assemblies: persistent activity mediated by semilunar granule cells. *Nature neuroscience*, 13(2):213–222. PMID: 20037579. 4.4.5

Larson, S. D. and Martone, M. E. (2009). Ontologies for neuroscience: what are they and what are they good for? *Frontiers in Neuroscience*, 3:1. 1.1.5

Larson, S. D. and Martone, M. E. (2013). NeuroLex.org: an online framework for neuroscience knowledge. *Frontiers in Neuroinformatics*, 7:18. 1.1.4, 1.1.5, 3.4.3, 4.4.1, 4.5.2, 5.1, 5.3

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J., Chen, L., Chen, L., Chen, T.-M., Chi Chin, M., Chong, J., Crook, B. E., Czaplinska, A., Dang, C. N., Datta, S., Dee, N. R., Desaki, A. L., Desta, T., Diep, E., Dolbeare, T. A., Donelan, M. J., Dong, H.-W., Dougherty, J. G., Duncan, B. J., Ebbert, A. J., Eichele, G., Estin, L. K., Faber, C., Facer, B. A., Fields, R., Fischer, S. R., Fliss, T. P., Frensley, C., Gates, S. N., Glattfelder, K. J., Halverson, K. R., Hart, M. R., Hohmann, J. G., Howell, M. P., Jeung, D. P., Johnson, R. A., Karr, P. T., Kawal, R., Kidney, J. M., Knapik, R. H., Kuan, C. L., Lake, J. H., Laramee, A. R., Larsen, K. D., Lau, C., Lemon, T. A., Liang, A. J., Liu, Y., Luong, L. T., Michaels, J., Morgan, J. J., Morgan, R. J., Mortrud, M. T., Mosqueda, N. F., Ng, L. L., Ng, R., Orta, G. J., Overly, C. C., Pak, T. H., Parry, S. E., Pathak, S. D., Pearson, O. C., Puchalski, R. B., Riley, Z. L., Rockett, H. R., Rowland, S. A., Royall, J. J., Ruiz, M. J., Sarno, N. R., Schaffnit, K., Shapovalova, N. V., Sivisay, T., Slaughterbeck, C. R., Smith, S. C., Smith, K. A., Smith, B. I., Sodt, A. J., Stewart, N. N., Stumpf, K.-R., Sunkin, S. M., Sutram, M., Tam, A., Teemer, C. D., Thaller, C., Thompson, C. L., Varnam, L. R., Visel, A., Whitlock, R. M., Wohnoutka, P. E., Wolkey, C. K., Wong, V. Y., Wood, M., Yaylaoglu, M. B., Young, R. C., Youngstrom, B. L., Feng Yuan, X., Zhang, B., Zwingman, T. A., and Jones, A. R. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176. 1.1.2, 1.1.5, 1.1.6, 4.3, 4.4.6, 4.4.6, 4.5.3, 4.7.3, 5.1

Lisman, J. E. (1997). Bursts as a unit of neural information: making unreliable synapses reliable. *Trends in neurosciences*, 20(1):38–43. PMID: 9004418. 1.1.2

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition*. Wiley-Interscience, 2 edition. 4.7.2, 4.7.2

Llinás, R. R. (1988). The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. *Science (New York, N.Y.)*, 242(4886):1654–1664. PMID: 3059497. 1, 1.1.1, 1.1.2, 1.1.4

Ma, Y., Hu, H., Berrebi, A. S., Mathers, P. H., and Agmon, A. (2006). Distinct subtypes of somatostatin-containing neocortical interneurons revealed in transgenic mice. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 26(19):5069–5082. PMID: 16687498. 1.1.4

Madisen, L., Zwingman, T. A., Sunkin, S. M., Oh, S. W., Zariwala, H. A., Gu, H., Ng, L. L., Palmiter, R. D., Hawrylycz, M. J., Jones, A. R., Lein, E. S., and Zeng, H. (2010). A robust and high-throughput cre reporting and characterization system for the whole mouse brain. *Nature neuroscience*, 13(1):133–140. PMID: 20023653. 4.7.4

Mainen, Z. F. and Sejnowski, T. J. (1996). Influence of dendritic structure on firing pattern in model neocortical neurons. *Nature*, 382(6589):363–366. PMID: 8684467. 4.4.6

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2):209–220. PMID: 6018555. 4.7.3

Marder, E. and Taylor, A. L. (2011). Multiple models to capture the variability in biological neurons and networks. *Nat Neurosci*, 14(2):133–138. 1.1.1, 1.1.3, 2.3, 2.4.1, 4.4.6

Markram, H. (2006). The blue brain project. *Nat Rev Neurosci*, 7(2):153–160. 1.1.5, 1.1.6, 4.3, 4.5

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews. Neuroscience*, 5(10):793–807. PMID: 15378039. 1, 1.1.2, 1.1.2, 1.1.4, 4.4.1, 4.4.5

Marsat, G. and Maler, L. (2010). Neural heterogeneity and efficient population codes for communication signals. *Journal of Neurophysiology*. PMID: 20631220. 2.3

Marsat, G. and Pollack, G. S. (2006). A behavioral role for feature detection by sensory bursts. *The Journal of Neuroscience*, 26(41):10542–10547. PMID: 17035539. 1.1.2

Martina, M., Schultz, J. H., Ehmke, H., Monyer, H., and Jonas, P. (1998). Functional and molecular differences between voltage-gated k+ channels of fast-spiking interneurons and pyramidal neurons of rat hippocampus. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 18(20):8111–8125. PMID: 9763458. 4.4.5

Marín, O. (2012). Interneuron dysfunction in psychiatric disorders. *Nature Reviews Neuroscience*, 13(2):107–120. 5.3

Mccallum, A. (2002). MALLET: a machine learning for language toolkit. 3.4.6, 3.5.4

Mensi, S., Naud, R., Pozzorini, C., Avermann, M., Petersen, C. C. H., and Gerstner, W. (2012). Parameter extraction and classification of three cortical neuron types reveals two distinct adaptation mechanisms. *Journal of Neurophysiology*, 107(6):1756–1775. 1.1.3

Migliore, M., Morse, T. M., Davison, A. P., Marenco, L., Shepherd, G. M., and Hines, M. L. (2003). ModelDB. *Neuroinformatics*, 1(1):135–139. 1.1.3, 1.1.5, 4.5.3

Migliore, M. and Shepherd, G. M. (2002). Emerging rules for the distributions of active dendritic conductances. *Nature reviews. Neuroscience*, 3(5):362–370. PMID: 11988775. 1.1.2, 1.1.4, 4.4.5, 5.1

Migliore, M. and Shepherd, G. M. (2005). Opinion: an integrated approach to classifying neuronal phenotypes. *Nature reviews. Neuroscience*, 6(10):810–818. PMID: 16276357. 1.1.2, 1.1.2, 4.4.5, 5.1

Miller, M. N., Okaty, B. W., Kato, S., and Nelson, S. B. (2011). Activity-dependent

changes in the firing properties of neocortical fast-spiking interneurons in the absence of large changes in gene expression. *Developmental neurobiology*, 71(1):62–70. PMID: 21154910. 4.4.6

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1 edition. 4.7.2

Momin, A. and Wood, J. N. (2008). Sensory neuron voltage-gated sodium channels as analgesic drug targets. *Current opinion in neurobiology*, 18(4):383–388. PMID: 18824099. 1.1.1

Moore, C. I., Carlen, M., Knoblich, U., and Cardin, J. A. (2010). Neocortical interneurons: From diversity, strength. *Cell*, 142(2):189–193. 1, 1.1.4, 2.3

Nagel, K. I. and Wilson, R. I. (2011). Biophysical mechanisms underlying olfactory receptor neuron dynamics. *Nat Neurosci*, 14(2):208–216. 2.4.4, 2.5

Narayanan, N. S., Kimchi, E. Y., and Laubach, M. (2005). Redundancy and synergy of neuronal ensembles in motor cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 25(17):4207–4216. PMID: 15858046. 2.3, 2.4.2

Neher, E. (1992). [6] correction for liquid junction potentials in patch clamp experiments. In Bernardo Rudy, editor, *Methods in Enzymology*, volume Volume 207, pages 123–131. Academic Press. 3.4.6

Neher, E. and Sakmann, B. (1976). Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, 260(5554):799–802. 1.1.1

Nelson, S. B., Sugino, K., and Hempel, C. M. (2006). The problem of neuronal cell types: a physiological genomics approach. *Trends in Neurosciences*, 29(6):339–345. 1

Ohtsuki, G., Nishiyama, M., Yoshida, T., Murakami, T., Histed, M., Lois, C., and Ohki, K. (2012). Similarity of visual selectivity among clonally related neurons in visual cortex. *Neuron*, 75(1):65–72. 4.4.5

Okaty, B. W., Miller, M. N., Sugino, K., Hempel, C. M., and Nelson, S. B. (2009). Transcriptional and electrophysiological maturation of neocortical fast-spiking GABAergic interneurons. *The Journal of Neuroscience*, 29(21):7040–7052. PMID: 19474331. 4.4.2

Okazaki, N. and Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–3095. PMID: 17050571. 3.5.4

Oswald, A.-M. M., Chacron, M. J., Doiron, B., Bastian, J., and Maler, L. (2004). Parallel processing of sensory input by bursts and isolated spikes. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 24(18):4351–4362. PMID: 15128849. 1.1.2

Padmanabhan, K. and Urban, N. N. (2010). Intrinsic biophysical diversity decorrelates neuronal firing while increasing information content. *Nature Neuroscience*, 13(10):1276–1282. 1.1.2, 1.1.4, 2.3, 2.4.1, 2.4.1, 2.4.2, 2.5, 2.6.1, 4.4.2, 4.4.4

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network (Bristol, England)*, 15(4):243–262. PMID: 15600233. 1.1.3

Parekh, R. and Ascoli, G. (2013). Neuronal morphology goes digital: A research hub for cellular and system neuroscience. *Neuron*, 77(6):1017–1038. 1, 1.1.2, 1.1.4, 1.1.5, 3.3, 4.3, 4.5.3

Pasquale, E. D., Keegan, K. D., and Noebels, J. L. (1997). Increased excitability and inward rectification in layer v cortical pyramidal neurons in the epileptic mutant mouse stargazer. *Journal of Neurophysiology*, 77(2):621–631. PMID: 9065835. 3.3, 3.4

Pillow, J. W., Ahmadian, Y., and Paninski, L. (2010). Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains. *Neural Computation*, 23(1):1–45. 2.3, 2.4.2, 2.4.2, 2.5, 2.6.5, 2.6.6

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999. 2.3, 2.4.1, 2.4.2, 2.4.2, 2.5, 2.6.2

Potter, S. M., Zheng, C., Koos, D. S., Feinstein, P., Fraser, S. E., and Mombaerts, P. (2001). Structure and emergence of specific olfactory glomeruli in the mouse. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 21(24):9713–9723. PMID: 11739580. 4.7.4

Poulet, J. F. A. and Petersen, C. C. H. (2008). Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice. *Nature*, 454(7206):881–885. 2.5

Pressler, R. T. and Strowbridge, B. W. (2006). Blanes cells mediate persistent feedforward inhibition onto granule cells in the olfactory bulb. *Neuron*, 49(6):889–904. PMID: 16543136. 4.4.5

Prinz, A. A., Bucher, D., and Marder, E. (2004). Similar network activity from disparate circuit parameters. *Nat Neurosci*, 7(12):1345–1352. 1.1.3, 4.4.6

Puchalla, J. L., Schneidman, E., Harris, R. A., and Berry, M. J. (2005). Redundancy in the population code of the retina. *Neuron*, 46(3):493–504. PMID: 15882648. 2.3, 2.5

Rajakulendran, S., Schorge, S., Kullmann, D. M., and Hanna, M. G. (2007). Episodic ataxia type 1: a neuronal potassium channelopathy. *Neurotherapeutics: the journal of the American Society for Experimental NeuroTherapeutics*, 4(2):258–266. PMID: 17395136. 1

Rall, W. and Shepherd, G. M. (1968). Theoretical reconstruction of field potentials and dendrodendritic synaptic interactions in olfactory bulb. *Journal of Neurophysiology*, 31(6):884–915. PMID: 5710539. 1.1.3

Ramakrishnan, C., Patnia, A., Hovy, E., and Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. *Source code for biology and medicine*, 7(1):7. PMID: 22640904. 3.5.4

Ramón y Cajal, S. (1995). *Histology of the nervous system of man and vertebrates*. Oxford University Press, New York. 1.1.4

Ranjan, R., Khazen, G., Gambazzi, L., Ramaswamy, S., Hill, S. L., Schürmann, F., and Markram, H. (2011). Channelpedia: an integrative and interactive database for ion

channels. *Frontiers in Neuroinformatics*, 5:36. 1.1.1, 1.1.5

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition. 2.4.4

Schneider, D. M. and Woolley, S. M. N. (2010). Discrimination of communication vocalizations by single neurons and groups of neurons in the auditory midbrain. *J Neurophysiol*, 103(6):3248–3265. 2.3, 2.4.2, 2.5

Schneidman, E., Bialek, W., and Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 23(37):11539–11553. PMID: 14684857. 2.3, 2.4.2, 2.6.6

Schoppa, N. E. and Westbrook, G. L. (2001). Glomerulus-specific synchronization of mitral cells in the olfactory bulb. *Neuron*, 31(4):639–651. PMID: 11545722. 2.4.2

Schwartzkroin, P. A. (1975). Characteristics of CA1 neurons recorded intracellularly in the hippocampal in vitro slice preparation. *Brain research*, 85(3):423–436. PMID: 1111846. 1.1.1

Shepherd, G. M., editor (2003). *The Synaptic Organization of the Brain*. Oxford University Press, USA, 5 edition. 4.4.1, 5.1

Shusterman, R., Smear, M. C., Koulakov, A. A., and Rinberg, D. (2011). Precise olfactory responses tile the sniff cycle. *Nat Neurosci*, 14(8):1039–1044. 2.4.2, 2.4.2, 2.4.4

Siegert, S., Cabuy, E., Scherf, B. G., Kohler, H., Panda, S., Le, Y.-Z., Fehling, H. J., Gaidatzis, D., Stadler, M. B., and Roska, B. (2012). Transcriptional code and disease map for adult retinal cell types. *Nature Neuroscience*, 15(3):487–495. 1.1.4

Singer, T., McConnell, M. J., Marchetto, M. C. N., Coufal, N. G., and Gage, F. H. (2010). LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends in neurosciences*, 33(8):345–354. PMID: 20471112. 1

Slee, S. J., Higgs, M. H., Fairhall, A. L., and Spain, W. J. (2005). Two-dimensional time coding in the auditory brainstem. *J. Neurosci.*, 25(43):9978–9988. 2.3, 2.4.1

Spors, H., Wachowiak, M., Cohen, L. B., and Friedrich, R. W. (2006). Temporal dynamics and latency patterns of receptor neuron input to the olfactory bulb. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 26(4):1247–1259. PMID: 16436612. 2.4.2, 2.4.3

Spruston, N. and Johnston, D. (1992). Perforated patch-clamp analysis of the passive membrane properties of three classes of hippocampal neurons. *Journal of Neurophysiology*, 67(3):508–529. PMID: 1578242. 1.1.6, 4.4.2, 5.1

Staley, K. J., Otis, T. S., and Mody, I. (1992). Membrane properties of dentate gyrus granule cells: comparison of sharp microelectrode and whole-cell recordings. *Journal of neurophysiology*, 67(5):1346–1358. PMID: 1597717. 4.4.2, 5.1

Stern, J. E. (2001). Electrophysiological and morphological properties of pre-autonomic neurones in the rat hypothalamic paraventricular nucleus. *The Journal of Physiology*, 537(1):161–177. PMID: 11711570. 4.4.5

Stocks (2000). Suprathreshold stochastic resonance in multilevel threshold systems. *Physical Review Letters*, 84(11):2310–2313. PMID: 11018872. 2.3

Strogatz, S. H. (2000). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview Press, Cambridge, MA. 1.1.3

Stuart, G., Spruston, N., and Häusser, M. (2007). *Dendrites*. Oxford University Press, Oxford; New York. 4.4.1

Stuart, G. J., Dodt, H. U., and Sakmann, B. (1993). Patch-clamp recordings from the soma and dendrites of neurons in brain slices using infrared video microscopy. *Pflügers Archiv*, 423(5-6):511–518. 1.1.1

Sugino, K., Hempel, C. M., Miller, M. N., Hattox, A. M., Shapiro, P., Wu, C., Huang, Z. J., and Nelson, S. B. (2006). Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature Neuroscience*, 9(1):99–107. 1, 1.1.4

Suzuki, R. and Shimodaira, H. (2006). Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542. PMID: 16595560. 4.4.5, 4.7.2

Tateno, T. and Robinson, H. P. C. (2011). The mechanism of ethanol action on midbrain dopaminergic neuron firing: a dynamic-clamp study of the role of ih and GABAergic synaptic integration. *Journal of Neurophysiology*, 106(4):1901–1922. PMID: 21697445. 4.4.5

Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Commun. ACM*, 11(6):419–422. 3.4.4, 3.4.6

Tkacik, G., Prentice, J. S., Balasubramanian, V., and Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences*, 107(32):14419 –14424. 2.3, 2.4.2

Toledo-Rodriguez, M., Blumenfeld, B., Wu, C., Luo, J., Attali, B., Goodman, P., and Markram, H. (2004). Correlation maps allow neuronal electrical properties to be predicted from single-cell gene expression profiles in rat neocortex. *Cerebral Cortex*, 14(12):1310 –1327. 1.1.1, 1.1.2, 3.4.2, 3.4.2, 4.4.1, 4.4.4, 4.4.5, 4.4.6

Tripathy, S. J., Padmanabhan, K., Gerkin, R. C., and Urban, N. N. (2013). Intermediate intrinsic diversity enhances neural population coding. *Proceedings of the National Academy of Sciences*, 110(20):8248–8253. PMID: 23630284. 2.1

Vasilevsky, N. A., Brush, M. H., Paddock, H., Ponting, L., Tripathy, S. J., LaRocca, G. M., and Haendel, M. A. (2013). On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*, 1:e148. 5.2

Verhagen, J. V., Wesson, D. W., Netoff, T. I., White, J. A., and Wachowiak, M. (2007). Sniffing controls an adaptive filter of sensory input to the olfactory bulb. *Nature Neuroscience*, 10(5):631–639. PMID: 17450136. 2.4.2

Voytek, J. B. and Voytek, B. (2012). Automated cognome construction and semi-automated hypothesis generation. *Journal of neuroscience methods*, 208(1):92–100. PMID: 22584238. 4.3

Wang, Y., Toledo-Rodriguez, M., Gupta, A., Wu, C., Silberberg, G., Luo, J., and Markram, H. (2004). Anatomical, physiological and molecular properties of martinotti cells in the somatosensory cortex of the juvenile rat. *The Journal of Physiology*, 561(1):65–90. PMID: 15331670. 1.1.4

Warland, D. K., Reinagel, P., and Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, 78(5):2336–2350. PMID: 9356386. 2.4.1, 2.6.6

Weber, F., Machens, C. K., and Borst, A. (2012). Disentangling the functional consequences of the connectivity between optic-flow processing neurons. *Nature Neuroscience*, 15(3):441–448. 2.5

Wichterle, H., Gifford, D., and Mazzoni, E. (2013). Mapping neuronal diversity one cell at a time. *Science*, 341(6147):726–727. PMID: 23950522. 4.3, 4.5.3

Woody, C. D. and Gruen, E. (1978). Characterization of electrophysiological properties of intracellularly recorded neurons in the neocortex of awake cats: a comparison of the response to injected current in spike overshoot and undershoot neurons. *Brain research*, 158(2):343–357. PMID: 709370. 1.1.2, 1.1.4

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat Meth*, 8(8):665–670. 1.1.5, 3.3, 3.4.2, 4.4.1, 5.2

Zhou, H.-X. (2004). Improving the understanding of human genetic diseases through predictions of protein structures and protein-protein interaction sites. *Current medicinal chemistry*, 11(5):539–549. PMID: 15032602. 4.3

Zhu, J. J. (2000). Maturation of layer 5 neocortical pyramidal neurons: amplifying salient layer 1 and layer 4 inputs by ca2+ action potentials in adult rat tuft dendrites. *The Journal of physiology*, 526 Pt 3:571–587. PMID: 10922009. 1.1.6, 4.4.2, 4.7.2

Zuberi, S. M., Eunson, L. H., Spauschus, A., Silva, R. D., Tolmie, J., Wood, N. W., McWilliam, R. C., Stephenson, J. P. B., Kullmann, D. M., and Hanna, M. G. (1999). A novel mutation in the human voltage-gated potassium channel gene (kv1.1) associates with episodic ataxia type 1 and sometimes with partial epilepsy. *Brain*, 122(5):817–825. PMID: 10355668. 1