EXPLORING SOURCES OF VARIABILITY IN ELECTROPHYSIOLOGY DATA OF MAMMALIAN NEURONS

by

Dmitry Tebaykin

B.Sc., The University of British Columbia, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2016

Abstract

Recently, there has been a major effort by neuroscientists to systematically organize and integrate vast quantities of brain data. However, electrophysiological properties have been shown to be sensitive to experimental conditions, thus directly comparing them between experiments could lead to inconsistent results. Here, I characterize the general effects of experimental solution composition differences on the reported ephys measurements. For that purpose, I employ text-mining, supplemented with manual curation to gather experimental solution information from published neurophysiological articles. I integrate the extracted information into the existing NeuroElectro database, which contains the electrophysiology, neuron type and experimental conditions information (temperature, electrode type, animal age, etc.) from the above neuroscientific literature. Exploring commonly used experimental solution recipes, I found the effect of solution compositions of explaining variance in electrophysiological properties to be small, relative to the amount of the existing ephys variability. Then, I created models for predicting the variability of ephys properties commonly reported by neurophysiologists, using the available experimental conditions information. These models can be used to remove a portion of the ephys variance when comparing results from different experiments, generally making such comparisons more reliable. To validate their performance, I adjusted a portion of NeuroElectro data to experimental conditions used by Allen Institute for Brain Science and compared the respective ephys properties before and after the adjustment.

Preface

Under the supervision of Dr. Paul Pavlidis, I conducted and authored the work presented henceforth.

A version of this work will be submitted to a peer reviewed journal for publication. Dmitry Tebaykin, Shreejoy J. Tripathy, Brenna Li, Paul Pavlidis. Experimental solution compositions assist in modeling the variability of reported neuron electrophysiology properties (in preparation).

Shreejoy J. Tripathy and I co-developed large parts of NeuroElectro.org, including ephys data aggregation, normalization and data export. Additionally, we extended the design and implementation of the NeuroElectro database. Brenna Li, James Liu, Patrick Savage, Nathalie Binnion, Kerrie Tsigounis, Ryan Sefid, Dawson Born, Ellie Hogan and Athanasios Kritharis formed the NeuroElectro curation team, dramatically increasing the number of curated articles in the database and helping to validate the performance of the text-mining algorithms.

Table of Contents

List of Tables

List of Figures

Glossary

ACSF	Artificial CerebroSpinal Fluid
AHP	After-HyperPolarization
AIBS	Allen Institute for Brain Science
AICc	corrected Akaike Information Criterion
AP	Action Potential
ATP	Adenosine TriPhosphate
BAPTA	1,2-Bis(o-AminoPhenoxy)ethane-N,N,N′,N′-Tetraacetic Acid
Ca	Calcium
Cl	Chloride
Cs	Cesium
E	Reversal potential
EDTA	EthyleneDiamineTetraacetic Acid
EGTA	Ethylene Glycol-bis(β-aminoethyl ether)-N,N,N',N'-Tetraacetic Acid
G	Ion permeability
GTP	Guanosine TriPhosphate
HEPES	4-(2-HydroxyEthyl)-1-PiperazineEthaneSulfonic acid
ISI	Inter-Spike Interval
K	Potassium
KNN	K-Nearest Neighbour
Mg	Magnesium
Na	Sodium

NE	NeuroElectro
NN	Neuron Name, synonym for neuron type
NT	Neuron Type
RF	Random Forest
Rin	Input Resistance
RMP	Resting Membrane Potential
SVM	Support Vector Machines

Acknowledgements

First, I would like to thank my supervisor, Dr. Paul Pavlidis, whose guidance helped me at every step of the way.

I would like to express my sincere appreciation for my thesis committee, comprising Dr. Yu Tian Wang and Dr. Jason Snyder, for their time and effort in reviewing my work. Special thanks to Dr. Sohrab Shah, the examination chair.

To postdoctoral fellows, Dr. Shreejoy J. Tripathy and Dr. Lilah Toker, thank you for your valuable guidance, advice and patience throughout my project. I would also like to thank past and present Pavlidis lab members for their undying support. I wish to thank Dr. Richard C. Gerkin for suggesting Random Forests for modeling the effects of solutions on ephys variability.

I am grateful to the undergraduate students, whose passion for Neuroscience led them to joining the NeuroElectro curation team.

Finally, thanks to faculty members, staff members and funding agencies of the Canadian Institute of Health Research, Strategic Training Program in Bioinformatics for making this experience a fulfilling one.

Dedication

For bridging differences in scientific inquires.

Chapter 1:

Introduction

Electrophysiological (ephys) recordings are widely used for characterizing neuron function. The field of electrophysiology focuses on studying electrical properties of neurons, their action potential and synaptic activity. Many different cell types in the brain possess different intrinsic electrophysiological properties that enable them to perform crucial and highly specific functions.

Electrophysiology as a field is moving towards larger kinds of data analyses trying to not only understand one neuron type in isolation, but to study many kinds of neurons simultaneously (Kandel et al., 2013). For example, the first goal of the US NIH BRAIN project is to generate a "census of cell types" (Jorgenson et al., 2015), involving neuron comparison using genetic, morphological and electrophysiological characteristics. One way to address this is by aggregating vast amounts of already published neuroscientific data. However, combining and comparing electrophysiology data across labs directly and on a large scale is challenging, because such data is often collected under different experimental conditions. It is generally thought that subtle variation in experimental conditions introduces certain variation into the corresponding ephys measurements. Therefore, comparing data across differently designed experiments without accounting for variability introduced by experimental conditions could lead to incorrect or inconsistent results, so the goal of my project is to enable the reliable comparison and aggregation of electrophysiological data across different experiments.

1.1 History and early neurophysiological mechanisms

From Greek, neurophysiology translates as the logic of neuron physics. It studies the intrinsic properties of neurons and their interactions. In neurons, action potentials (APs) are the primary communication protocols that propagate signal from one cell to the next. APs are relatively short events during which the neuron's membrane potential depolarizes to the AP peak value and then eventually repolarizes back to the resting membrane potential.

A series of five landmark papers published in 1952 by Hodgkin and Huxley unveiled many of the basic mechanisms that govern neuron electrophysiology, providing neurophysiologists with the initial sodium-potassium mechanism of neuron action potentials. At rest, a typical neuron maintains a high concentration of potassium and a low concentration of sodium and chloride ions inside relative to the outside. These concentrations are maintained by Na / K pumps that keep moving sodium ions outside of the cell and potassium inside, expending ATP. The concentration differences across the neuron membrane cause sodium (E_{Na}) / potassium (E_K) reversal potentials, calculated using the Nernst equation (Hille, 1984), to be respectively very high / similar, relative to the resting membrane potential. Additionally, ionic driving forces across the neuron membrane are calculated as a difference between their reversal potentials and the membrane potential, thus at rest sodium has a strong inward driving force, while potassium has a slight outward driving force.

When a sufficiently strong stimulus causes a neuron's membrane potential to increase to the point of its AP threshold, the action potential gets triggered. During the neuron action potential, voltage-gates sodium channels in the cell membrane open, causing the sodium ion permeability (G_{Na}) across the cell membrane to increase dramatically, allowing Na^+ ions to flood inside the neuron due to the high driving force of sodium (Hille, 1984). As the AP approaches its peak, both sodium driving force and permeability decrease, because the neuron's membrane potential is closer to E_{Na} and a large portion of the sodium channels close. At that point, potassium voltage-gated channels open and its large outward driving force is complemented with the dramatic rise in its permeability (G_K) across the membrane. Because of potassium ions rapidly leaving the cell and further potassium channels opening in response to the influx of calcium ions during the AP, the neuron's membrane potential hyperpolarizes to values below its normal resting potential (Hille, 1984). Then, the neuron membrane eventually restores to its resting potential. Neurons cannot fire a second AP immediately following the first one because of the absolute refractory period, during that time sodium channels are in their inactivated state and cannot be opened regardless of the membrane potential.

Since ionic balance is crucial for neurons to function, the brain must maintain specific ion concentrations inside and outside the neurons. That is accomplished by cerebrospinal fluid (CSF), which was originally discovered by Emanuel Swedenborg in 1741. However, its chemical composition was accurately described only in 1912 by William Mestrezat (Hajdu, 2003). The current medical physiology text-books define CSF composition as follows, (in mM, converted): "142 Na^+, 2.5 K^+, 1.3 Ca^{++}, 0.8 Mg^{++}, 124 Cl^-, 3.9 glucose" (Hall, 2015). Since action potentials are functions of changing membrane potential due to ionic concentration changes (among other

things), neuronal electrophysiology characteristics must depend on the ion concentrations inside and outside of the cell.

1.2 Typical methodologies in electrophysiological experiments

To provide some context for the methodology used in intracellular electrophysiology, a typical *in vitro* experiment involves: extracting the brain of an anesthetized animal and cutting thin slices from the brain; letting the slices recover in a bath of a carefully designed solution; transferring a designated slice to a recording chamber, where the ephys measurements are taken. In the recording chamber a brain slice is continuously perfused with the recording (external, extracellular, ACSF) solution at a constant temperature. Finally, a recording electrode is attached to the neuron, allowing for injection of electrical current and the quantification of electrophysiological parameters. The electrode also contains the internal solution (intracellular, pipette), which in the case of patch-clamp electrodes completely dialyzes the cell and replaces its intracellular milieu.

Sharp and patch-clamp electrodes are the common electrode types used by the neurophysiological community. The sharp electrode procedure involves inserting a fine microelectrode inside the neuron to measure membrane potential. The electrode is typically filled with multimolar (> 1M) potassium (Hille, 1984). The Patch-clamp technique was developed by Erwin Neher and Bert Sakmann, who received the Nobel Prize in Physiology/Medicine in 1991,

“for their discoveries concerning the function of single ion channels in cells”. The patch-clamp electrode resembles a micropipette with a large tip, which is pressed up to the cell membrane and, through suction, it forms a tight (high resistance) seal with the neuron without severely damaging it. Pipette solutions used for patch-clamp electrodes are more diverse than those used for sharp electrodes, however, they tend to use more physiological concentrations of potassium, on the scale of 120-150 mM (Hall, 2015).

Neurophysiologists commonly use two clamping techniques during ephys recordings: voltage clamp and current clamp. The voltage clamp technique involves setting the neuron membrane potential to a chosen value, which provides the opportunity to record the amount of ionic current that crosses the cell membrane at the specified voltage. On the other hand, the current clamp technique involves injecting specific amounts of current into the neuron through the recording electrode. Here, the minimal amount of current required to cause an AP is termed rheobase (Hille, 1984). While it is difficult to do voltage clamp recordings using a sharp electrode (electrode typically has high resistance values and the cell is damaged in the process of inserting the electrode), the patch-clamp technique is commonly used during both voltage and current clamp recordings. The patch-clamp voltage clamp recordings are performed using cesium instead of potassium in their electrode solutions. This is done to block potassium channels, removing the noise that they could introduce into the experiment data (Isenberg, 1976).

1.3 The search for causes of variance in reported electrophysiology values

The common practice among neurophysiologists is to only analyze data that they have collected themselves. This is largely because it is generally thought that subtle variation in experimental conditions introduces certain variation into the corresponding measurements.

Many neurophysiologists address the task of exploring the effects of experimental conditions on neuron ephys properties using experimental electrophysiology techniques (Armentia and Sah, 2004; Kim and Connors, 2012; Lee et al., 2005). In the past, ephys data has been shown to be sensitive to experimental conditions. For example, differences in animal ages, especially during development (Suter et al., 2013); or varying extracellular Ca^{2+} concentrations (Aivar et al., 2014) result in changes in electrophysiological properties of neurons. However, this experimental approach is limited to varying a single condition and studying one or several neuron types at a time. Therefore, it is unclear how well the discovered relationships between electrophysiology properties and experimental conditions would generalize to other neuron types, animal species, ages and other confounding factors that typically remained fixed throughout each experiment.

To the best of my knowledge, there are no comprehensive and systematic analysis (or meta-analyses) of the effects of experimental conditions on electrophysiological measurements. However, there are papers that explore the effects of specific experimental setups (Aghajanian and Rasmussen, 1989; Moyer and Brown, 1998). Additionally, the effects of specific

experimental solution components on neuron survival rates were discussed by several papers (MacGregor et al., 2001; Richerson and Messer, 1995; Tanaka et al., 2008).

Previously, my colleague, Shreejoy Tripathy designed and created NeuroElectro, an online database that contains text-mined and curated population mean electrophysiological measurements, neuron type and experimental setup information from normal control samples of published neuroscientific studies (Tripathy et al., 2015). Using a large-scale meta-analysis method, he showed that animal age, recording temperature, electrode type choices significantly explain the study-to-study variance in reported ephys values.

Since a typical electrophysiological experiment uses carefully designed solutions inside and outside the measured neurons, I hypothesized that study-to-study ephys variability could be partially explained by the experimental setup (metadata) differences, focusing on the recording and pipette solution compositions. To test my hypothesis, I employed a combination of text-mining and curation approaches to extract experimental solutions used in published neurophysiological articles. Then, I integrated my solution extraction algorithms into the NeuroElectro database. After exploring the external and internal solution recipes commonly used by electrophysiologists, I applied univariate linear models to uncover the effects of solutions on the measured ephys values. These initial models proved ineffective, which prompted me to use a non-linear multiple regression approach.

1.4 Machine learning multiple regression approaches

Towards the end of my statistical analysis, I turn to multiple regression approaches for modeling variability in reported electrophysiological properties of neurons. For that purpose, I examined the benefits and drawbacks of several commonly used supervised multiple regression models: K-nearest neighbor, Neural Networks, Support Vector Machines, Random Forest. A supervised regression algorithm relies on having a training dataset that contains both feature information and the target prediction value. Then the algorithm learns a function that allows it to map the target values to certain combinations of the input feature values.

KNN is a non-parametric model that would be predicting ephys values based on the K closest articles, where 'closest' would be defined by metadata (Altman, 1992). This approach, however, does not extrapolate its predictions to metadata combinations it has not encountered before. Since my dataset is semi-sparse (especially for the rarely reported ephys properties), this approach would struggle to predict the ephys values.

Supervised and unsupervised Neural Networks, specifically deep learning, is a powerful regression algorithm. It is probably one of the most talked about regression algorithms nowadays, as it is being used by Google for image/voice recognition and interpreting streams of sensorimotor data (Jaderberg et al., 2016). However, the drawback of neural networks is that they require huge amounts of data when compared to the number of features (Linoff and Berry, 2011). Since that is not the case with my dataset, I had to search for other algorithm types.

Support Vector Machines were used in the previous analysis that found certain basic metadata features to be significantly correlated with the variability of ephys properties (Tripathy et al., 2015). However, SVMs are linear models that suffer from overfitting on the outliers (Cortes and Vapnik, 1995). I compared the performance of SVMs to my models of choice and showed their relative instabilities during the 10-fold cross-validation testing (Appendix A, Figure 3). Similarly to SVMs, logistic regression approaches (lasso and ridge regression) are prone to the effects of outliers when modeling relatively small datasets (Le Cessie and Van Houwelingen, 1992; Tibshirani, 1996).

Finally, Random Forest is a supervised non-linear multiple regression approach that relies on ensembles of Decision Trees, generated from small samples of the training dataset with controlled variance, to predict the target ephys values (Breiman, 2001). Each decision tree gives its best prediction of what the ephys value should be given a set of metadata values, then these predictions are combined across all decision trees to produce a final predicted value. Random Forests are resilient to outliers (bad Decision trees are discarded) and they perform well on small datasets (Liaw and Wiener, 2012).

In the next chapters, I present the research methods used, followed by the findings and a discussion of potential implications.

Chapter 2:

Exploring experimental solution recipes extracted from published papers via automated text-mining and manual curation

2.1 Methods

I used the existing NeuroElectro (www.neuroelectro.org) database as a starting point for determining experimental solution recipes that could influence the variability in reported electrophysiological measurements. NeuroElectro stores and data-mines thousands of Neuroscience articles that may contain neurophysiological data (Tripathy et al., 2015). These articles are downloaded in full-text HTML format and text-mined for means +/- standard errors of the commonly reported electrophysiological properties, for example: resting membrane potential, input resistance, action potential spike half-width and amplitude (for a full list of ephys properties refer to http://neuroelectro.org/ephys_prop/index/). Additionally, NeuroElectro text-mines basic experimental conditions (metadata) from Methods sections of the full-text articles: recording temperature; junction potential and offset; animal species, strain, age and weight; preparation type (*in vivo*, *in vitro*, cell culture, etc.), electrode type. At the start of my project, the text-mined ephys data and metadata was curated by my colleague and the original creator of NeuroElectro, Shreejoy Tripathy. During the curation process, he assigned a neuron type to each article, based on an expert-defined list of neuron types provided by NeuroLex.org (Larson and

Martone, 2013). Neuron instances reported in articles that could not be curated unambiguously to a single type were curated to the general neuron type “other”.

To achieve the goals of my project, I extended the existing NeuroElectro functionality by introducing experimental solutions text-mining algorithms. Additionally, I implemented a new curation interface that enabled metadata curation inside ephys data tables extracted from published papers. Then I assisted in developing the new NeuroElectro curation protocol and creating the NeuroElectro curation team. Next, I explored experimental solution composition recipes that are commonly used in neurophysiological articles. Then, I used the curated neuron types and metadata information to model the study-to-study variability of commonly reported ephys properties. Finally, I validated the proposed models by shifting ephys data stored in NeuroElectro to the experimental conditions used by Allen Institute for Brain Science and comparing the corresponding ephys properties. In the following sections, I further elaborate on the details of each step.

2.2 Text-mining and curating electrophysiology-relevant chemical solutions

2.2.1 Specifics of the text-mining algorithm

To automatically find the most likely solutions used during the recording, I used the existing HTML articles in the NeuroElectro database and extended the implemented basic metadata text-

mining infrastructure to extract sentences that contain experimental solution information. The difficulty of this task lies in the fact that ephys recordings are often performed in slices, thus ephys papers generally report more than just the two solution types used during the recordings (extracellular and pipette). The most common *other* solution types include: cutting, storage and incubation. The text-mining algorithm keeps track of them, but I do not directly use the *other* solutions in my analysis.

Recording chamber (external, extracellular, ACSF) and pipette (internal, intracellular, electrode) solutions used during electrophysiological experiments were the most promising experimental conditions, in terms of explaining study-to-study ephys variability, that were not yet tracked by NeuroElectro. I define the solution extraction task as a 4-step process: 1) Locate the Methods section of the target article; 2) ranking each sentence in the Methods section on how likely it is to contain a solution; 3) identifying the solution type for each solution-containing sentence; and 4) extracting individual compound concentrations from external and internal solutions (Examples: Na+, K+, HEPES, etc).

The first step of locating the given article's Methods section was implemented in the original version of NeuroElectro and re-used here. It relies on regular expressions to check all section headers of an article (section headers are defined by article's HTML formatting) and return the most likely candidate for the Methods section. I assume that all the relevant solutions metadata information can be found in the Methods section of each article.

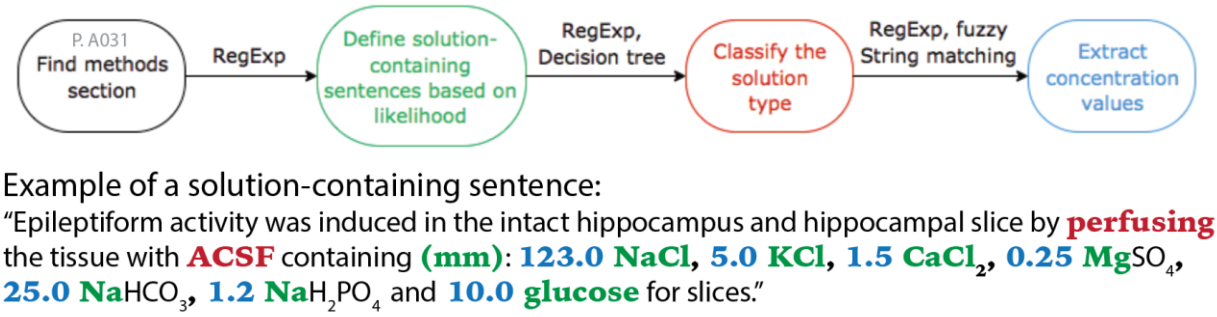


Figure 1: Solutions text-mining is a 4-step process. The initial step of finding methods sections was already implemented in NeuroElectro. Tools that were used to transition between steps are mentioned above arrows. Colors represent different processing steps and link to the targeted text. Sentence extracted from Derchansky et al. 2008.

The second step is carried out using a combination of regular expressions and a decision tree: each sentence is assigned a score based on whether it contains the ions of interest (Ca, Mg, Na, K, Cl) and has a general solution-describing structure. Specifically, a typical solution-containing sentence mentions the compound concentration units (mM or μM), lists a series of chemical compounds separated by commas or other delimiters and might end with a pH measurement.

Once a sentence has been identified as solution-containing, the algorithm uses regular expressions to check the sentence for key words that define external (recording, perfusing, extracellular, ACSF), internal (pipette, electrode, intracellular) and other (incubation, storage, cutting, dissecting, ice bath) solutions. If no key words have been found within the solution-containing sentence, the search is first expanded one sentence at a time by up to 3 sentences before and then 1 after the solution-containing sentence. External solutions can often be referred to as "the same as storage solution" or "ACSF used for dissecting the brain", meaning that the same solution can be used for multiple steps of an ephys experiment. Therefore, I assume that a missing explicit reference to an external solution implies that the last-mentioned storage or

cutting solution was also used for electrophysiological recordings. Empirically, incubation solutions do not get re-used as extracellular solutions in the recording chambers.

Finally, my text-mining algorithm extracts solution concentration values by identifying the location of each compound of interest in solution sentences using regular expressions. It splits up the solution sentence into pieces (fragments) using one of the compound separators, semicolon having precedence over comma due to the "(in mM): NaCl, 135; $CaCl_2$, 2" notation. Keywords "and", "or" are also used as fragment separators. Then, each fragment contains a single compound and a concentration value, apart from the first and the last fragments that include the parts of the sentence before and after the solution recipe, respectively. Next, if a targeted compound is located within a fragment (using regular expressions), the algorithm searches for the closest positive number that is not a part of a chemical formula, i.e. the "2" in "$CaCl_2$, 5 mM" is not recognized as a concentration value, even though it is closer to both calcium and chloride mentions, but the "5" is. After the fragments have been parsed in this manner, each compound's concentrations are summed up to obtain the total concentration in the solution. The algorithm accounts for element valence: "2 mM $CaCl_2$" would be parsed as 2 mM of Ca^{2+} and 4 mM of Cl^-, even if the compound is fully spelled out (disodium sulfate instead of Na_2SO_4 or sodium creatine instead of Na_2-creatine). Complete dissociation for all chemical compounds is assumed here, because the algorithm does not have access to dissociation constants of each solution component at the specified temperature, which tends to differ from one article to the next. These total concentration values are then stored in the NeuroElectro database.

After implementing and testing the major ion concentration extraction algorithm, I decided to extend it to extract several commonly mentioned experimental solution components. These include: glucose (dextrose), EGTA, EDTA, cesium, HEPES, BAPTA, ATP, GTP. To achieve this goal, I designed new regular expression for each of them and added them to the list of compounds to extract. This also served the purpose of evaluating the difficulty of extending the text-mining algorithm to more chemical components.

To evaluate each step of my text-mining algorithm, I (with the help of the NeuroElectro curation team) manually curated a set of 100 randomly chosen NeuroElectro articles. Each step of the algorithm was evaluated using recall, precision and F_1-score metrics (Van Rijsbergen, 1979).

Text-mining tables of HTML articles for ephys properties and Methods sections for experimental conditions is done in C-Python using the following libraries: NLTK, conversion of imported HTML articles to Python data structures (Bird, 2006); RE, regular expressions (default Python package); Numpy, scientific computing methods (Van Der Walt et al., 2011); FuzzyWuzzy, partial String matching (https://github.com/seatgeek/fuzzywuzzy). The process of text-mining for ephys properties and basic metadata (temperature, animal age, species, etc.) has not been significantly adjusted since the previous NeuroElectro paper (Tripathy et al., 2015).

2.2.2 Manual curation methodology

The NeuroElectro curation protocol follows text-mining with two rounds of curation by trained undergraduates (Figure 2): the first curator's task is to identify the types of neurons reported in the article as similar as possible to the author's neuron type descriptions; assign a NeuroElectro neuron type that is most closely represents the authors definition; record experimental conditions and ephys properties missed by text-mining. The main goal of the second round of curations is to validate all the annotations. Both rounds of curations check the text-mining output and they must be performed by different students, without collaboration. In our analysis, we only use the data that has been put through both rounds of curations. The detailed curation protocol can be found in Appendix B, Figure 1.

To support the NeuroElectro curation team's efforts, I developed a new curation interface using JavaScript. The old interface was difficult to scale and including metadata information proved to be a challenge due to Python implementation restrictions. Additionally, it was using sub-optimal data transfer protocols between the server and client, required only 1 curation step to be performed at a time, had a confusing visual design and did not allow curators to delete their annotations. These issues were addressed during the implementation of the new curation interface. It enabled the curation of experimental conditions within ephys data tables (Figure 3).

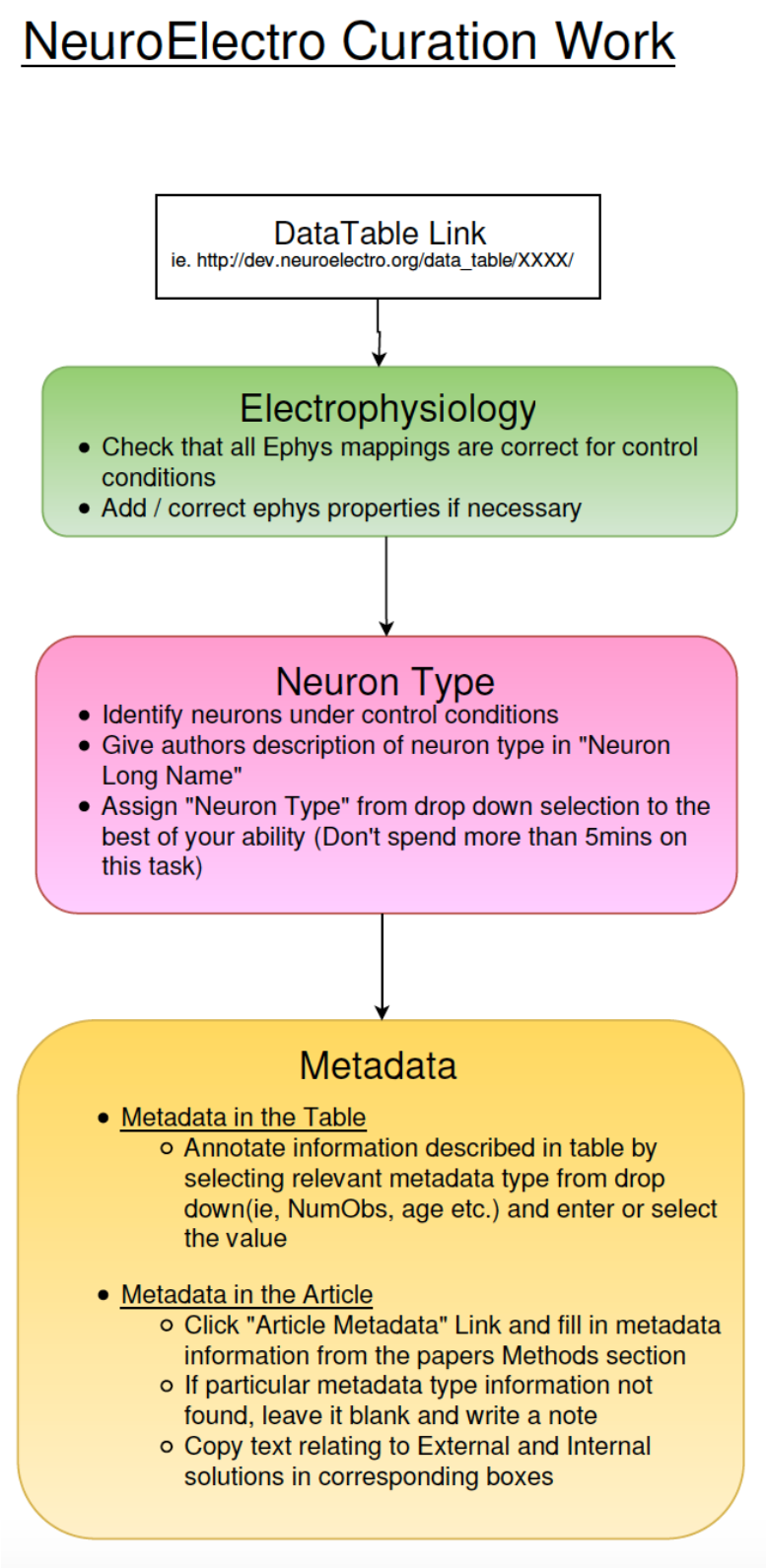


Figure 2: Overview of NeuroElectro curation protocol. Credit to Brenna Li for creating this figure.

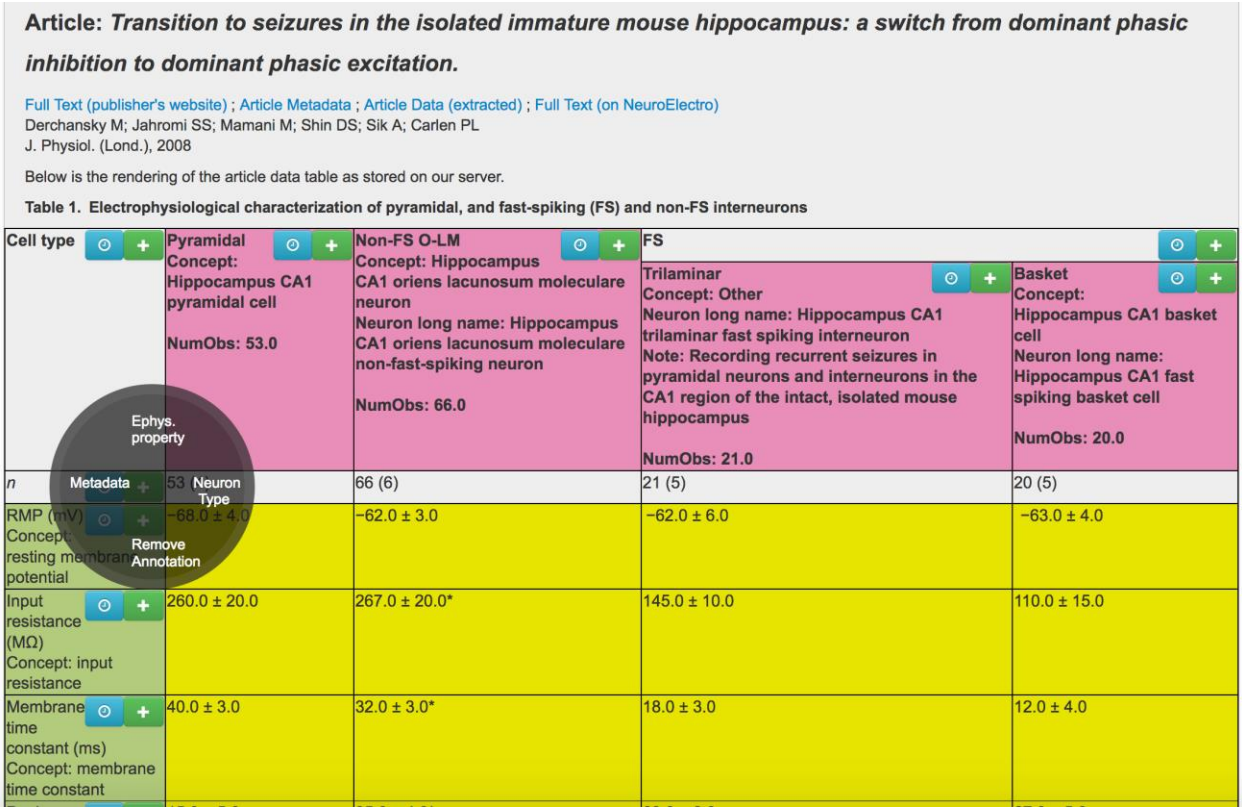


Figure 3: New NeuroElectro curation interface. The green plus sign button allows to select the annotation options: ephys property, neuron type, metadata or remove all annotations from the cell. The Blue clock button shows the history of annotations for the cell (Credit: Shreejoy Tripathy).

The new interface enables curators to dynamically update and delete annotations. Each table cell can only be annotated with either neuron type or ephys property mention, but not both. However, metadata information can be added to any table cell. The entire data table can be annotated at once through a concept of 'staging' annotations before submitting them to the server with a single button click (Appendix B, Figure 2). The new curation interface automatically scales with inclusion of any additional ephys properties, neuron types or metadata types to the NeuroElectro database. When the curated table is submitted to NeuroElectro, all cells that are located on the intersections of annotated neuron type mentions and electrophysiology properties get processed as the reported ephys values of the corresponding neuron types.

Ephys properties are commonly measured and reported using slightly different definitions. For example, AP amplitude can be reported as the difference between AP peak and AP threshold, or AP peak and resting membrane potential. NeuroElectro has a system in place that enables the curation team to annotate such cases separately, wherever possible. Then, automated algorithms standardize them to a single baseline. Article experimental conditions were annotated on a per-article basis, except for papers that report ephys measurements under different experimental conditions that NeuroElectro can track. With regards to solution recipes, cesium-based solutions (associated with voltage-clamp experiments) were generally avoided during the curation process, because NeuroElectro focuses on extracting electrophysiological properties that can only be measured during current-clamp experiments.

2.2.3 Data and code availability

The python code used for text-mining and preprocessing is incorporated into the NeuroElectro codebase and can be found on https://github.com/neuroelectro/neuroelectro_org in assign_metadata.py file. The most up-to-date CSV data spreadsheet can be found at http://neuroelectro.org/static/src/article_ephys_metadata_curated.csv. The R files with the data wrangling, analysis and model creation are stored in https://github.com/dtebaykin/neuronephys.

2.2.4 Statistical exploration of experimental solution recipes

2.2.4.1 Data preprocessing

The curated and standardized ephys values, neuron names and metadata are aggregated into a single CSV file of the following format: each line of the file corresponds to a unique combination of an ephys table in a neuroscientific article and a neuron type reported in that table. For example, an article with 1 ephys table that provide ephys information about 4 neuron types will have 4 data rows in the CSV data spreadsheet. Each row of data contains information about the article (PubMed ID, title, year published, authors, etc.), NeuroElectro and author-defined neuron types, ephys properties found in the article ephys table and all metadata we could gather from the Methods section of the article. Each ephys property is stored as it is reported in the article: mean +/- standard deviation or standard error and number of measurements. The ephys properties get checked against a dictionary of allowed values per ephys property type (RMP cannot be positive, AP amplitude cannot be negative, etc.). The ephys values that violate these rules either get flagged for inspection or automatically corrected. The ephys property flagging and correction algorithm was developed by Shreejoy Tripathy.

Little preprocessing is performed on the metadata entries since some of them are pre-defined categorical variables (species, strain, electrode type, preparation type and junction potential). Like ephys properties, continuous metadata variables are checked against possible value thresholds: age and weight cannot be negative, recording temperature must be within a specified

range (Implemented by: Shreejoy Tripathy). No preprocessing steps were performed for experimental solution concentrations; the total concentration values are reported in the data spreadsheet as they are stored in the NeuroElectro database.

The CSV data spreadsheet was imported into a local installation of RStudio (R version 3.3.0). Several filtering and processing steps were required to clean up the data since my work focuses on the effect of metadata and, more specifically, solutions on the resulting ephys values. Subsequently, I have filtered out articles that did not have any solutions associated with them in our database. Possible reasons include: the solutions used were described in another paper and only cited in the article of interest, requiring more curation work to deal with these cases; the solutions were missed by both text-mining and curation efforts; or the solutions were not reported by the authors.

The most important experimental conditions preparation steps include: assigning a default compound concentration and reversing junction potential corrections. The first step involves assigning an arbitrarily small default concentration of 10^{-6} millimoles (mM) for each ion concentration that was not mentioned in the solution sentence. I calculate reversal potentials in my analysis and the data would otherwise be too sparse. Ideally, the second preparation step would be done in the other direction – correcting RMP and AP threshold values for junction potentials in the articles that did not do so themselves. As a reference, liquid junction potential is the voltage difference between two solutions that are in contact with each other. In the case of neurophysiology, the external and internal solutions interact with each other via the cell membrane, forming a liquid junction potential that affects membrane voltage measurements.

Among articles stored in NeuroElectro, only 46% report the junction potential value and 24% correct for it. Thus, I decided that it would be more appropriate to reverse the corrections than attempt to impute unreported junction potentials based on experimental solutions used. I assume articles that do not report a junction potential correction did not perform a correction. It was important to address the junction potential problem because we must standardize important ephys properties like resting membrane potential and AP threshold.

2.2.4.2 Exploration of common solution recipes

Initially, I explored the distributions of experimental solution compound concentrations in R using basic plotting and data wrangling tools. Attempting to find trends in the relative simultaneous major ion concentration changes of internal and external solutions, I calculated reversal potentials of major ions using the Nernst equation (1) (Hille, 1984). Sodium example:

$$E = \frac{RT}{zF} ln \frac{[Na]_{outside}}{[Na]_{inside}}, \quad (1)$$

where R denotes the universal gas constant (R = 8.314 $J*K^{-1}*mol^{-1}$), T – temperature in Kelvin, z – ion valence, F – Faraday constant (F = 96485 $C*mol^{-1}$). Inside the logarithm is the ratio of external and internal ion concentrations.

Next, I temporarily filtered my data for Patch-clamp electrodes to identify the trends in their experimental solution recipes. Using the information about the five major ion concentrations from the filtered dataset, I performed principal component analysis to pinpoint largest differences in the recipes. Next, I used all experimental solution concentrations data and performed hierarchical clustering. The results of this analysis led me to the identification of several ‘schools of thought’ when it comes to preparing extracellular and intracellular solutions for Patch-clamp experiments.

2.3 Results

This section provides the following information: evaluation of the automated text-mining and curation approach; exploration of commonly used experimental solution recipes among neurophysiologists.

2.3.1 Evaluation of the text-mining and curation data extraction pipeline

I developed a novel text-mining algorithm for extracting experimental solutions from methods sections of published articles (see Methods). I evaluated this pipeline on a gold standard fully manually curated set of 100 articles randomly chosen from NeuroElectro.

Briefly, the text-mining algorithm identifies sentences of methods sections from neurophysiology articles stored in NeuroElectro that contain extracellular and intracellular solutions. Next, the algorithm extracts the concentration values of compounds that are commonly included into solution recipes. Major ion concentrations are calculated by summing the concentrations of compounds they are present in (valence considered if provided), for example: 151.25 mM of Na and 133 mM of Cl are extracted from “in mM: NaCl, 124; KCl, 5; NaH2PO4, 1.25; MgSO4, 2; CaCl2 2; NaHCO3, 26 and dextrose, 10” (Agmon and Connors, 1991). It is important to note that the curation protocol involves verification of external and internal solution sentences, but not the correctness of concentration extraction.

First, I evaluated the accuracy of experimental solution-containing sentences identification (**External solution identification** and **Internal solution identification**). Second, I compared these experimental solutions to the ones annotated by NeuroElectro curators (**External solution sentence curation** and **Internal solution sentence curation**). Third, I evaluated compound concentration extraction using a stringent criterion: if even one ion or compound concentration was extracted incorrectly, the entire solution was counted as incorrectly parsed (**Major ions concentration extraction** and **Other compounds concentration extraction**). I found it essential for downstream analysis to optimize this concentration value extraction step, since solution concentrations were not further manually curated.

To evaluate text-mining and curation performance, I calculated precision and recall. In this context, recall represents the fraction of articles where the corresponding task yielded results. Similarly, precision is the fraction of recalled articles where the task was performed correctly. To give an example for **External solution identification**: recall is the fraction of external solution sentences that were tagged as solution-containing sentences, precision shows how many of those sentences had their type assigned as "external". I used the F_1 score as a measure of each tasks accuracy, it was calculated as a function of precision and recall of each step in the text-mining and curation process.

Task	Precision	Recall	F_1 score	Major error causes
External solution identification	**0.94**	**0.97**	**0.95**	**Multiple internal/external solutions, ambiguous solution compositions**
External solution sentence curation	**0.99**	**0.99**	**0.99**	
Internal solution identification	**0.88**	**0.97**	**0.92**	
Internal solution sentence curation	**0.98**	**0.99**	**0.98**	
Major ions concentration extraction	**0.96**	**0.98**	**0.97**	**Typos, inconsistent compounds listings, edge cases**
Other compounds concentration extraction	**0.98**	**0.73**	**0.84**	**Typos, limited chemical vocabulary**

Table 1: Solutions text-mining and curation performance. A set of 100 NeuroElectro articles was fully curated for the correctness of external and internal solutions data extraction pipeline. Solution identification was evaluated separately from concentration values extraction.

I identified several common causes of errors in the text-mining and curation process. The **solution type identification** algorithm and trained curators often struggled with complex articles – the more electrophysiological experiments reported in a single article, the harder it was to identify the correct solutions for each experiment. Another common source of errors was introduced by multiple solutions mentions in a single sentence: "Electrophysiology Patch electrodes were … filled with two internal solutions consisting of the following (in mM): 1) 140 $KMeSO_4$, 10 KCl, 10 HEPES, 4 Mg_2ATP, and 0.4 Na_3GTP *or* 2) 130 $KMeSO_4$, 10 KCl, 10 HEPES, 10 BAPTA, 4 Mg_2ATP, and 0.4 Na_3GTP." (Wu et al., 2004, emphases added). It is

difficult even for trained curators to correctly separate the two internal solutions listed in one sentence (Only the first solution should be identified as internal for this example, because it was the one used to record ephys properties under control conditions). Generally, the sentence gets parsed as a single internal solution, effectively doubling several chemical concentration values, which eventually allowed us to flag it for re-curation.

The **Compound concentration extraction** algorithm had difficulties handling idiosyncratic solution descriptions. Specific examples include: 1) first part of the solution in the beginning of the sentence and the other part in the end, or in a different sentence entirely; 2) compounds are separated by commas, except for one or two that are separated by special symbols (Example: semicolon); 3) typos (Typo examples: using "Ci" for chloride instead of Cl, "phosphocreatinine" instead of phosphocreatine); 4) chemicals spelled-out informally (Example: calcium chloride instead of calcium dichloride). The relatively low recall of the **other compounds concentration extraction** task can be explained by difficulty of identifying such compounds and their respective concentration values in text, especially when they are fully spelled-out (Examples: *N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid for HEPES and ethylene glycol-bis (β-aminoethyl ether)-*N,N,N′,N*′-tetraacetic acid for EGTA).

The text-mining algorithm is robust enough to be applied to the entirety of the articles contained within the NeuroElectro database (nearly 100,000 articles). However, NeuroElectro lacks an algorithm for fully automated text-mining of ephys properties and neuron types. Consequently, in the next steps of my analysis I use solutions from articles that have been manually curated.

2.3.2 Analysis of experimental solution recipes used by neurophysiologists

The first step to understanding the effect of solutions on electrophysiological variance is determining the magnitude of variance within solutions themselves. If most labs use very similarly designed artificial cerebrospinal fluids (ACSFs) and pipette solutions, is unlikely to be a source of substantial variance in ephys measurements. Initially, my approach was to extract the solution constituents that are consistently present and are known to contribute directly and strongly to electrophysiological processes (Hille, 1984). These include: sodium (Na), potassium (K), magnesium (Mg), chloride (Cl) and calcium (Ca). The total ion concentrations are calculated by summing up the concentrations of each compound where that ion is present. For example, the sodium concentration of 157.2 mM and 141 mM of chloride are extracted from the following recording solution: "130 mm NaCl, 3 mm KCl, 1.25 mm NaH_2PO_4, 26 mm $NaHCO_3$, 2 mm $MgCl_2$, 2 mm $CaCl_2$ and 10 mm glucose oxygenated with 95% O_2/5% CO_2, pH 7.2–7.4, 290–310 mOsm." (André et al., 2010).

Because electrophysiologists attempt to mimic their extracellular solutions after cerebrospinal fluid, it is not surprising that similar major ion and common compounds concentrations tend to be used throughout the neuron electrophysiology community. Therefore, I expected external solutions to use relatively similar recipes, the differences in which would not account for significant portions of ephys study-to-study variability. I observed the following general trends throughout the literature: external solutions use ~150 mM of sodium and ~130 mM of chloride

with small amounts (1-3 mM) of magnesium and calcium (Figure 4). The potassium concentration is commonly kept very close to 0 mM; however, I identified a subset (~10%) of articles that include 5-6 mM of K into their artificial cerebrospinal fluid (ACSF) composition. On the other hand, there is a clear distinction between internal solutions used by patch-clamp and sharp electrodes: the former commonly uses ~140 mM of potassium with a wide variety of chloride concentrations (0-200 mM), magnesium (1-8 mM) and sodium (0-50 mM); while the latter tends to contain several moles of potassium, typically paired up with acetate, methylsulfate or chloride.

Next, I examined the distributions of the 26 extracted major ions and other compounds concentrations in experimental solution recipes. I excluded chemical compounds that were used less than ten times from this recipe analysis. The concentration values of major ions (Na, K, Cl, Ca, Mg) used in extracellular solutions generally follow approximately normal distributions (Figure 4A). However, the other common compounds concentrations are not normally distributed across the recipes: glucose is primarily used at a concentration of 10 mM, with the rest of the recipes increasing it up to 40 mM; and HEPES is usually not included into ACSF, but in ~5% of the recipes it was present at a concentration of 10 mM. In many cases, HEPES was included into external solutions of cell culture ephys experiments, however, it was also included into dissociated neuron experiments (Gittis and Lac, 2007) and two-photon guided *in vivo* whole-cell recordings (Chen et al., 2015).

The recipes of intracellular solutions vary more than the compositions of extracellular solutions. Potassium concentration values are almost uniformly distributed, except for the large peaks at

130, 135 and 140 mM. (Figure 4B). Cesium concentration values closely follow potassium's distribution, since voltage-clamp recordings utilize cesium-based recipes to block K channels.

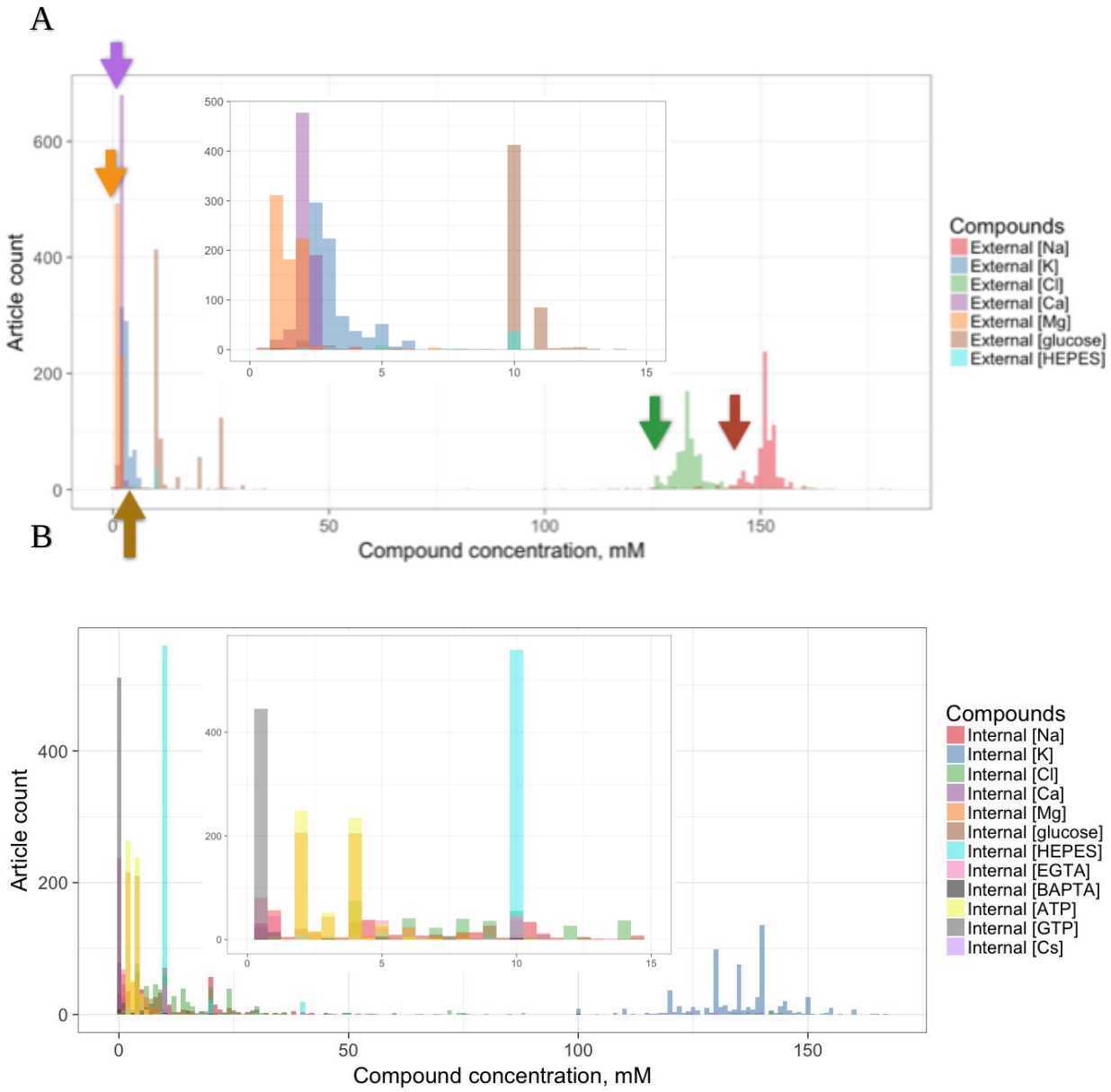


Figure 4: Chemical compositions of experimental solutions. Data from 731 curated Patch-clamp solutions. Histograms of compounds that are commonly found in: A) External (extracellular, ACSF) and B) internal (pipette, electrode) solutions. The ion concentrations were calculated by summing concentrations of their respective compounds, assuming complete dissociation. Histogram bin width is set to 1 mM on the main plots and to 0.5 mM on the 0-15 mM histograms. Arrows denote CSF composition as described in medical literature: "142 Na^+, 2.5 K^+, 1.3 Ca^{++}, 0.8 Mg^{++}, 124 Cl^-, 3.9 glucose" (Hall, 2015).

There is a lot of variation in internal sodium and chloride concentrations, though both distributions are skewed towards the 0-30 mM range. Magnesium and ATP are predominantly used at 2 mM and 4 mM concentrations. Additionally, HEPES, GTP and EGTA are consistently included into pipette solutions at the respective concentrations of 10 mM, 0-1 mM and 0-11 mM. Finally, calcium, glucose and BAPTA are rarely included into the intracellular solutions at small concentrations (Gall et al., 2003; Goldfarb et al., 2007; Prestori et al., 2008). It is important to remember that the concentration values extraction algorithm performs at roughly 90% accuracy, and while the above concentration value distributions represent the true summary of recipes used by electrophysiologists, some of the edge cases (internal Ca, glucose and BAPTA concentrations) have slightly inflated numbers.

The above exploration gives the impression of a great deal of at least minor variability in solution makeups. A quantitative analysis confirms this: considering only the five major ions, there are 358 (49%) different external and 482 (66%) different internal solutions in my data, with 603 (82%) papers using unique combinations of the two – out of 731 possible patch-clamp solution recipes. The most frequent ACSF recipe was used 62 times, (in mM): 151.25 Na, 2.5 K, 133.5 Cl, 1 Mg, 2 Ca. This recipe was most commonly used by the Spruston lab: 6 times over the course of almost 20 years: "ACSF consisted of 125 mM NaCl, 2.5 mM KCl, 25 mM $NaHCO_3$, 1.25 mM NaH_2PO_4, 1 mM $MgCl_2$, 2 mM $CaCl_2$, and 25 mM dextrose" (Cembrowski et al., 2016; Cooper et al., 2003; Golding et al., 2005; Graves et al., 2012; Lübke et al., 1998; Staff et al., 2000). While there was little consistency overall, the differences between recipes were generally

minor. For example, external Na varied from 150 mM to 153 mM (interquartile range; see Figure 4).

Patch-clamp intracellular solutions are even more diverse: out of N recipes from 731 papers, only 55 were used twice, 24 – 3 times, 2 – 4 times, 4 – 5 times, 2 – 6 times and single recipes were used 7 and 8 times, the other 481 recipes were unique. The pipette solution that was used 7 times contained, (in mM): 120 K, 6 Cl, 4 Mg; and the one that was used 8 times, (in mM): 140 K, 14 Cl, 4 Mg. Among the patch-clamp articles, 41 recipes for both solutions were shared between 2 articles, 8 recipes – 3 articles, 4 recipes – 4 articles and 1 recipe was the same in 6 articles.

Out of the 128 curated Sharp electrode articles: 5 ACSF recipes were used 2 times, 5 – 3 times, 3 – 4 times, 1 – 5 times. This most common recipe was, (in mM): 151.25 Na, 3 K, 131 Cl, 2 Mg, 2 Ca. The pipette solutions of Sharp electrodes are less diverse: 3 recipes were used twice, 1 – 4 times, 2 – 5 times, 2 – 14 times (1 M and 4 M of K), 1 – 19 times (3 M of K) and 1 – 34 times (2 M of K).

To explore possible patterns in recipe creation for recording and pipette solution, I used principal component analysis supplemented by hierarchical clustering to identify trends in recipe creation for recording and pipette solutions (Appendix A, Figures 1 and 2).

Chemical	Internal Solution, mM		External Solution, mM	
	Low Na, Cl (N = 353)	High Na, Cl (N = 183)	High Mg (N = 276)	Low Mg (N = 371)
Na	0 - 10	15 - 50	140 – 160	
K	120 - 150		1 - 5	
Cl	0 - 30	15 - 50	125 - 140	
Cs	0		0	
Mg	0 - 10 (81% in 2-4 range)		2 – 2.5	1 - 1.5
Ca	0 - 1 (95% use 0 mM)		2 - 3	
HEPES	5 - 15 (96% use 10 mM)		0-10 (89% use 0 mM)	
EGTA	0 - 10 (87% use 0 mM)		0	
ATP	5 - 10		0	
GTP	0 - 10 (~95% use 0 mM)		0	
glucose	0		0 - 25	

Table 2: Summary of trends in electrode and recording solution designs. In this general trend analysis, outlier recipes were not considered. Number of articles analyzed: 703 Patch-clamp, *in vitro* studies performed on rats, mice or guinea pigs. N is the number of articles with the specified solution composition.

No obviously distinct clusters presented themselves, suggesting that electrophysiologists use similar recipes with slight variations, within biologically reasonable concentration values. However, two trends were identifiable among the Patch-clamp solution recipes: internal solutions could be separated into those with low Na, Cl concentrations and high Na, Cl (Table 2); external solutions can be split by their relatively low and high Mg concentrations. It is important to note that cesium-based solutions (associated with voltage-clamp experiments) were generally avoided during the curation process, because NeuroElectro curation heavily prioritizes current-clamp experiments, as it is focused on ephys properties such as action potential characteristics.

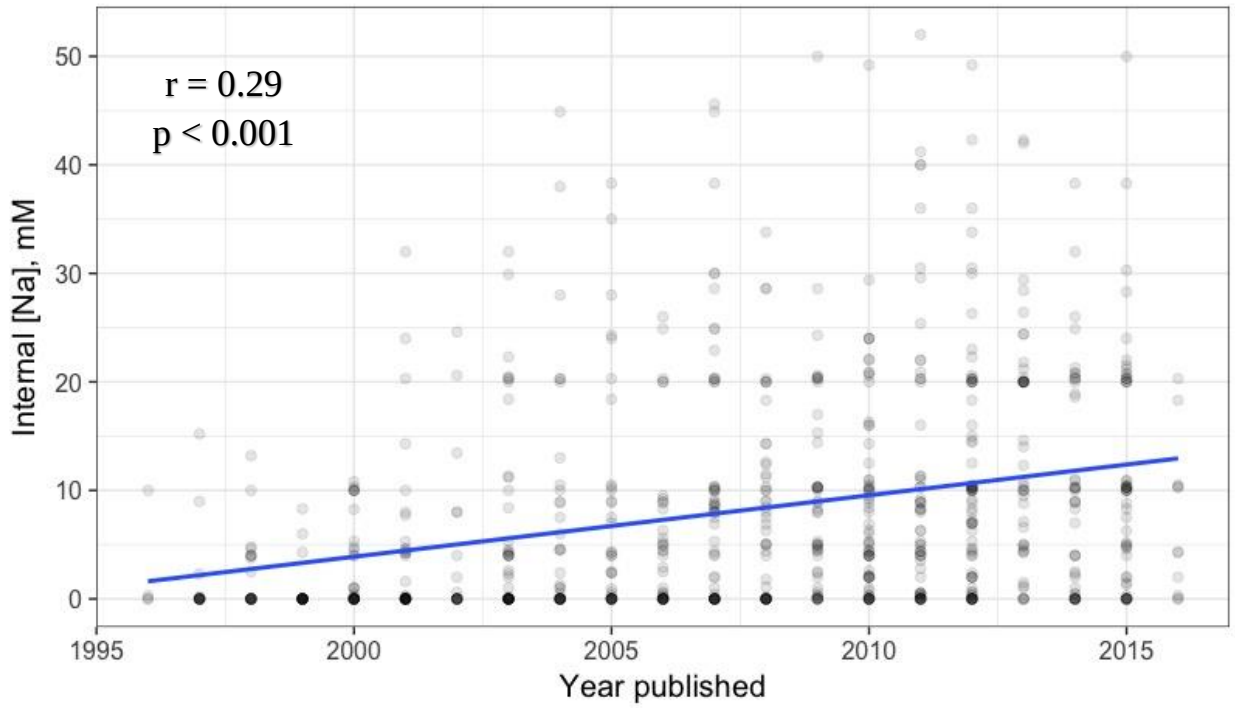


Figure 5: Internal sodium concentration increased in 2003. Boxes represent solution concentrations from articles published in the corresponding year (X-axis). The blue line is a linear fit between internal sodium concentration and publication year. Internal sodium concentration significantly increases throughout the years (r = 0.29, p < 0.001).

An interesting variation in Patch-clamp pipette solution recipes is internal sodium concentrations increase throughout the years (Figure 5). It seems to be caused by the introduction of 10-20 millimoles per litre of Na_2-phosphocreatine to internal solutions, which became popular in mid-2000s. The implication is that recipes do change over time and it is entirely possible that a single lab or a small set of labs can start new trends in the designs of solution recipes.

This exploration of the distributions of solution components and verifying that neurophysiologists tend to use at least slightly different solutions, I proceeded to the task of determining whether these solution component variations help to further explain the variability of commonly reported ephys properties.

Chapter 3:

Explaining study-to-study variability of electrophysiological properties using experimental conditions

3.1 Methods

3.1.1 Modeling the effects of experimental conditions on the variability in electrophysiological properties

Here, I use 11 most commonly reported electrophysiological properties (in order of abundant to sparse): input resistance, RMP, AP threshold, AP amplitude, AP half-width, membrane time constant, AHP amplitude, rheobase, maximum firing frequency, cell capacitance, adaptation ratio and ignore the rest due to their extreme sparsity. My statistical pipeline does not depend on the number of ephys properties analyzed, but it does require a reasonable number of articles reporting them, otherwise the resulting models would have insignificant explanatory power. With more articles being added to NeuroElectro, my analysis can be applied to more ephys properties without major changes to the algorithms. For a full and most up-to-date list of ephys properties visit neuroelectro.org/ephys_prop/index/.

3.1.1.1 Constructing univariate linear models

My initial approach for modeling the relationships between experimental conditions (metadata) and the variability of ephys properties was to use univariate linear regression models. They are reliable, simple to implement and comprehend. The drawbacks and faults of linear models are well-documented; they are sensitive to outliers, overfitting and they are restricted to modeling linear interactions, thus non-linear relationships would likely be insignificant (Freedman, 2009; Yan, 2009). The built-in R function for linear models was used for this task.

Since I modeled each compound concentration and ephys property pair separately, I ran into a multiple comparisons problem, which states that a set of statistical inferences performed simultaneously increase the false discovery rates (Chandler, 1995; Miller, 1981). The solution was to apply a multiple testing correction algorithm to the obtained p-values to account for performing hundreds of similar tests. I used the Bonferroni correction approach, which reduces the significance threshold by a factor of comparisons made, thus the significance threshold for p-value decreased from 0.05 to $0.05 / 286 = 1.75 * 10^{-4}$ (Bonferroni, 1936; Dunn and Goldstein, 1959; Savin, 1984).

3.1.1.2 Multiple regression approach

When considering using multiple solutions features to model ephys properties, the first intuitive model to use was the Goldman-Hodgkin-Katz equation that predicts resting membrane potential from the recording temperature and external/internal concentrations of Na, Cl and K (2).

$$V_m = \frac{RT}{F} ln \frac{P_{Na}[Na^+]_{out} + P_K[K^+]_{out} + P_{Cl}[Cl^-]_{in}}{P_{Na}[Na^+]_{in} + P_K[K^+]_{in} + P_{Cl}[Cl^-]_{out}}, \quad (2)$$

where V_m denotes resting membrane potential and P denotes ion permeability across the cell membrane. Ionic permeabilities were approximated with the default text-book values: $P_{Na} = 0.05, P_{Cl} = 0.45, P_K = 1$ (Hille, 1984).

Before delving into multiple regression model selection, I established the training and testing datasets, because testing a model using the same data it was trained on causes overfitting. To that end, I used a 10-fold cross-validation technique (Kohavi, 1995). I randomly separated the data spreadsheet into 10 folds, the models would be trained on 9 out of 10 folds and tested on the remaining single fold. Each of the 10 folds would get a chance of being the testing fold, that way I can estimate the robustness of the models. Since one article can have multiple data rows in the NeuroElectro spreadsheet (one per reported neuron type), the only rule for fold separation was that each fold must contain unique articles (by PubMed ID). If that was not the case, my models could be learning to predict the article's PubMed ID instead of ephys measurements using metadata as features.

After examining several multiple regression models (section 1.3), I used the Random Forest regression algorithm for modeling ephys properties with multiple features simultaneously. Specifically, I used the randomForest implementation (RandomForest package in R, version 4.6-12) for constructing the regression models that predict ephys properties given experimental condition features, and cforest implementation (party package in R, version 1.0-25) for feature importance ranking. The randomForest implementation cannot handle categorical variables with more than 53 different values, only one variable failed to meet that criteria – neuron type (NT) which currently has 115 unique entities in NeuroElectro. My workaround to this problem was to expand the NT column into a matrix where each neuron type is a column and each row contains a 1 for the NT mentioned in the respective article and 0 otherwise. This is a common solution used by regression approaches to model continuous variables with categorical features.

I created six different models that enable me to explore the performance of solution components when predicting ephys properties, compared to neuron types, basic metadata and their combinations. These models are: neuron type only, basic metadata, solutions metadata, basic metadata + neuron type, solutions metadata + neuron type, and all features together (neuron type, basic and solutions metadata). The hypothesis is that if solutions provide valuable information to the models, they should perform reasonably well on their own and improve the basic metadata + neuron type model performance. That result would be observed, if the all features model had the best performance.

To study the effect of reducing the number of available samples on the model performance, I used the all metadata features model to predict input resistance with reducing the number of available samples by 100 with each iteration.

3.1.1.3 Incorporating only the highest predictive features into each ephys property model

After creating the initial models for predicting ephys properties with experimental conditions, I decided to find the best combinations of experimental conditions for each commonly reported ephys property. For that, I used cforest's feature importance ranking and the corrected Akaike Information Criterion (AICc).

Cforest implementation of the random forest algorithm uses conditional inference trees as its base learners, instead of decision trees, making it less prone to assigning inappropriately high importance to correlated features (Strobl et al., 2008). However, cforest performs slightly worse when used for regression modeling than randomForest, because that task is not fully optimized yet (https://cran.r-project.org/package=party).

Having ordered the features by importance for each ephys property, I used AICc to choose the optimal number of top performing features. The information criterion theory refers to estimating the amount of information lost when using statistical models to predict the process that generates the target data. AIC is meaningless for comparison of model performances for different ephys properties or models trained on different data. However, it is useful for choosing the optimal

number of features to include into a given model. AIC depends on the model's performance and the features that were used to create it (3).

$$AIC = 2k - 2\ln(L), \tag{3}$$

where k denotes the number of features, L denotes maximum likelihood for the model. AIC tends to underestimate the information loss on datasets where the number of samples is not several orders of mangitude greater than the number of features used to train the models (Anderson and Burnham, 2002; Burnham and Anderson, 2004). This is particularly important for less popular ephys properties. AICc adds a correction term for limited datasets (4) to the Akaike Information Criterion.

$$AICc = AIC + \frac{2k(2k+1)}{n-k-1}, \tag{4}$$

where n denotes the number of samples in the dataset. AICc has several assumptions: the sample elements must be nearly independent and their underlying distribution must be unimodal, neither badly skewed, not heavy tailed (Anderson and Burnham, 2002). Both assumptions hold for the NeuroElectro data: we can treat articles as independent and the underlying distributions of ephys properties are expected to be approximately normal. The last step in the model creation is the calculation of Random Forest's maximum likelihood, since the algorithm does not provide one automatically. However, the mean squared errors can be calculated using the observed ephys

values and the predicted values. Using them for maximum log-likelihood calculation I obtain the following formula (5).

$$\ln(L) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} MSE * n, \quad (5)$$

where σ denotes standard deviation of the ephys property in the training data, n is the number of samples, MSE is the mean squared error between observed and predicted ephys values. Substituting formula (5) into (3), and their result into (4), I get (6).

$$\text{AICc} = 2k + n\log(2\pi\sigma^2) + \frac{1}{\sigma^2} MSE * n + \frac{2k(2k+1)}{n-k-1}. \quad (6)$$

Models with the lowest AICc should have the optimal performance when predicting the ephys property.

To choose the optimal models, I split the available data into 90% and 10% portions (per ephys property). The larger portion of the data was used to rank all features by their importance. Next, I was using top X (where X ranged from 1 to 33) features and the 90% portion of the data to create 33 models that predict R_{in}. Each consecutive model had 1 more feature than the previous. For each model an AICc was calculated using the remaining 10% portion of the data. As before, the 90%/10% split was performed 10 times so that each 10% of the data had a chance to be in the testing set. AICc depends on the amount of available data, so it cannot be used to compare the

model performance of different ephys properties to each other. Finally, features that were chosen at least 90% of the time were included into the models for each ephys property.

3.1.2 Validating proposed ephys property models

To validate the performance of the models built with the selected features, I compared them to the previously created models that use distinct feature sets: neuron type only, basic metadata, solutions, and their combinations. This was done with 10-fold cross-validation of RandomForest.

Next, I showcase the ability of my models to reduce the variability in reported ephys measurements, making the results more comparable to the targeted experiment's conditions. For that purpose, I adjusted the ephys data stored in NeuroElectro to the experimental conditions used by Allen Institute for Brain Science and compared the respective ephys properties (http://celltypes.brain-map.org/). The experimental conditions baseline shifting formula was developed by Shreejoy Tripathy (7).

$$NE_{adj} = NE_{obs} - NE_{pred} + AIBS_{shift}, \quad (7)$$

where NE_{adj} denotes the shifted NeuroElectro ephys values, NE_{obs} – default ephys values stored in NeuroElectro, NE_{pred} – ephys values from NeuroElectro, predicted with the models, trained using 10-fold cross-validation without seeing the values of the articles they are trying to predict in this step, and $AIBS_{shift}$ is the ephys values predicted with NeuroElectro models, using AIBS

experimental conditions. The observed and adjusted ephys values from NeuroElectro were then compared to the reported AIBS ephys values.

3.2 Results

My main goal was to measure the impact of experimental solution recipes on the results of electrophysiological experiments. I performed an analysis of 882 published intracellular neurophysiology articles to explore common experimental solution compositions and to discover the general effects of experimental solutions on neuronal electrophysiology. I then extend my analysis to include previously known sources of ephys variability (Examples: animal species, age, type of electrode used, recording temperature) and compare their relative impact. Finally, I propose models for several commonly reported ephys properties that allow adjusting the ephys values from one set of experimental conditions to another. I validate these models using a new dataset provided by Allen Institute for Brain Science.

3.2.1 Dataset overview

In NeuroElectro, we have gathered electrophysiology, neuron type data and experimental conditions (metadata) from text-mined and manually curated neuroscience articles. NeuroElectro does not have access to the original raw experimental measurements (i.e. voltage traces), instead the ephys values are curated as population means with standard errors and number of samples (Tripathy et al., 2014). The dataset primarily contains ephys data reported under normal control conditions (control samples, defined by the original paper), enabling comparisons across articles.

Entity name	Quantity (rows of data)	Entity name	Quantity (rows of data)
Unique PubMed ID	882	**Solutions metadata:**	
		External [Na]	1471
Electrophysiological properties:		**External [K]**	1466
Input Resistance (R_{in}, rin)	1435	**External [Cl]**	1486
Resting Membrane Potential (rmp)	1314	**External [Mg]**	1478
Action Potential Threshold (apthr)	935	**External [Ca]**	1479
Action Potential Amplitude (apamp)	990	**External [Cs]**	2
Action Potential Half-Width (aphw)	980	**External [glucose]**	1446
AfterHyperPolarization Amp. (ahpamp)	687	**External [HEPES]**	84
Membrane Time Constant (τ, tau)	682	**External [EGTA]**	6
Adaptation Ratio (adratio)	308	**External [EDTA]**	0
Rheobase (rheo)	303	**External [BAPTA]**	0
Cell Capacitance (cap)	258	**External [ATP]**	4
Maximum AP Frequency (maxfreq)	229	**External [GTP]**	4
		Internal [Na]	1119
Neuron Type	1588	**Internal [K]**	1466
		Internal [Cl]	1340
Basic metadata:		**Internal [Mg]**	1217
Species	943	**Internal [Ca]**	244
Strain	887	**Internal [Cs]**	60
Electrode Type	943	**Internal [glucose]**	46
Preparation Type	943	**Internal [HEPES]**	1241
Recording Temperature	1511	**Internal [EGTA]**	797
Animal Age	1388	**Internal [EDTA]**	3
Animal Weight	272	**Internal [BAPTA]**	30
Junction Potential	1588	**Internal [ATP]**	1193
Junction Offset	551	**Internal [GTP]**	1083

Table 3: Summary of data stored in NeuroElectro database. Color highlights: green – 11 most commonly reported ephys properties, yellow – neuron type mentions defined by NeuroLex, orange – basic metadata, blue – recording (external) and pipette (internal) solutions metadata. Data extracted on: 25.09.2016

For my analysis, I use NeuroElectro data from a set of 882 curated articles. The NeuroElectro curation team has identified 1588 neuron type mentions in the collected articles. For each of

these mentions, ephys data and metadata information can be reported or missing from the article's methods section. The most consistently reported information is recording temperature. Junction potential contains the full 1588 entries because it is a categorical variable with one of the options being "Unreported", so it is not regarded as missing data by NeuroElectro. Each data entry in NeuroElectro is annotated with one of 120 neuron types that are defined by the extended dictionary, originally provided by NeuroLex.org. The full list of NeuroElectro neuron types can be found here: http://neuroelectro.org/neuron/index/.

An ephys property can be reported multiple times in the same article (once per measured neuron type), resulting in the total number of measured properties (Examples: R_{in}, RMP) exceeding the number of articles. NeuroElectro only contains data that authors choose to publish, therefore some ephys properties are not reported as consistently (Examples: rheobase, capacitance, maximum firing frequency). My analysis focuses on the top 11 commonly reported ephys properties (EPs), shown in table 1. The full up-to-date list of ephys properties can be found here: http://neuroelectro.org/ephys_prop/index/.

One of the challenges of comparing values of ephys properties across studies stems from inconsistent definitions. For example, action potential spike amplitude can be measured from resting membrane potential (Cui et al., 2011; Perkowski and Murphy, 2011) or AP threshold (Novkovic et al., 2015). The adaptation ratio, defined in NeuroElectro as the ratio of durations between early and late APs inter-spike intervals (ISI) in an AP train, is even less standardized. It can be reported as a ratio of first / last ISI (Nassar et al., 2015; Scorza et al., 2011), a ratio of last / first ISI (Novkovic et al., 2015; Zhou et al., 2015), a percentage (Fujiwara-Tsukamoto et al.,

2004; Zaitsev et al., 2009), or 1 – ratio first / last ISI (Derchansky et al., 2008; Lamsa et al., 2007). In each of these cases, NeuroElectro curators have standardized these ephys measurements for the different baselines. However, there are many other reporting inconsistencies that the curation team has not been able to address. These examples simply outline the types of problems in attempting to aggregate electrophysiological data that go above and beyond the effects of experimental conditions metadata.

Here, I distinguish experimental conditions (metadata) stored in NeuroElectro into two types: basic (recording temperature, animal age, species, etc.) and solutions metadata (pipette and extracellular concentrations of ions and compounds). Typically, all metadata is curated once per article and then copied into all rows of data extracted from that article. The exceptions are articles that alter experimental conditions between measurements. There are 4 basic metadata types in NeuroElectro (preparation type, animal weight, junction potential and junction offset) that are not used directly in my analysis. Briefly, I only use *in vitro* studies for modeling ephys properties, animal weights get converted to animal age, junction potentials and junction offsets are used to standardize RMP and AP threshold values before the analysis (see Methods for more details).

3.2.2 Assessing study-to-study electrophysiological variability

Electrophysiology values might have relatively large intrinsic cell-to-cell variability, which could conceal the effects of experimental conditions on the results (Tripathy et al., 2015). To assess this, I considered whether between-experimental variance for a single neuron type is greater than within-study variance. If the within-study variance was higher, the meta-analysis approach would likely yield inconclusive results.

In the context of a single experiment, the scientist measuring RMPs of hippocampus CA1 pyramidal neurons expects to observe values that are approximately normally distributed, with a sample mean providing an estimate of the population mean. If experimental conditions do not introduce significant variance when comparing ephys properties across studies, then multiple electrophysiology studies should report similar ranges of values while measuring from one neuron type. Figure 3 shows mean +/- standard error of the mean for three relatively common neuron types in NeuroElectro, after correcting for junction offset. Disregarding several outliers, SEMs do not cover the whole range of reported RMP means. In the case of hippocampus CA1 pyramidal cells the mean RMPs range from -73 mV to -53 mV with an average SEM of 1.7 mV. There is an even greater spread in the reported mean resting membrane potentials in Martinotti cells (from -73 mV to -48 mV with a mean SEM of 2.6 mV) and medium spiny neurons (-95 mV to -61 mV, mean SEM of 2.8 mV), still with relatively small standard errors. These data do not support the hypothesis that different electrophysiology experiments report ephys values from the same normal distribution (ANOVA P-value of $4.04*10^{-15}$ for RMPs of hippocampus CA1

pyramidal neurons). I found that, other ephys properties behave very similarly to RMP (not shown). Thus, the hypothesis that ephys measurements are unaffected by experimental

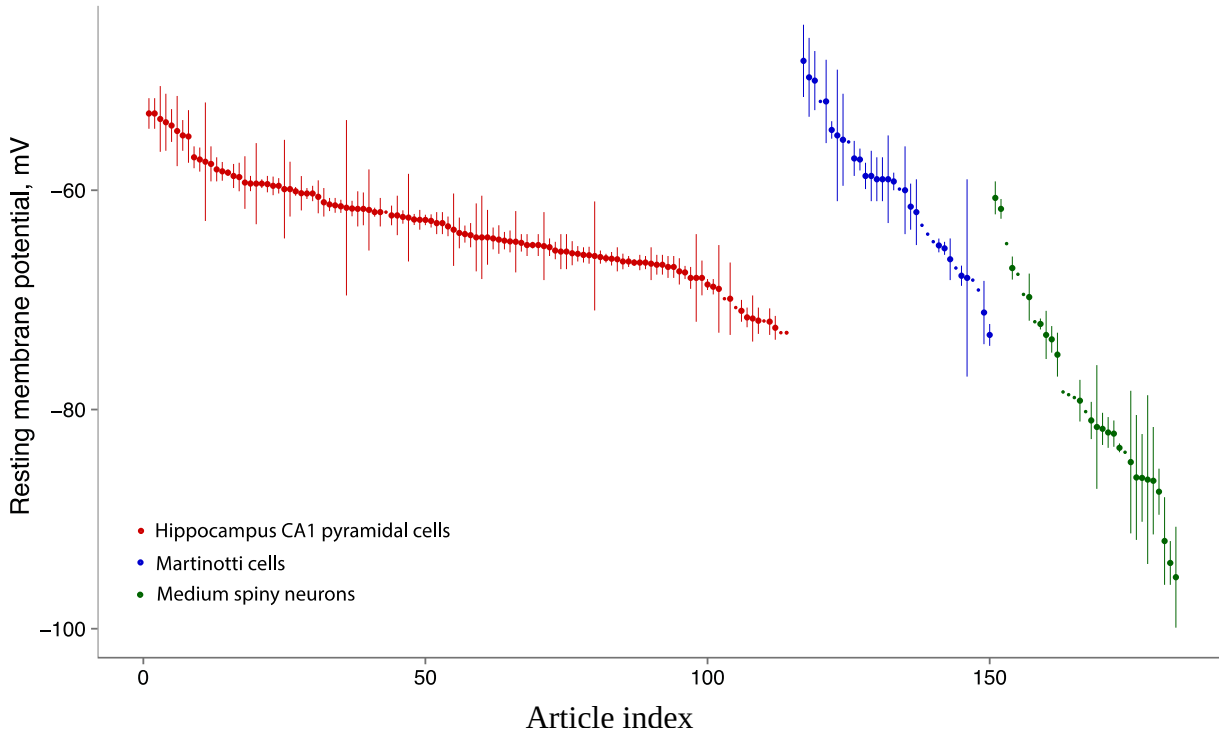


Figure 6: Electrophysiological variability is higher between experiments than within experiments. Resting potentials of hippocampal CA1 pyramidal neurons, neocortex Martinotti cells and Striatum medium spiny neurons, across articles in NeuroElectro. Each point and line is an RMP mean +/- SEM, reported by an article.

conditions must be false. These inter-study differences must be partially due to differences in experimental procedures.

Since the RMP means for a single neuron type reported in different articles are highly unlikely to originate from the same normal distribution, I hypothesize that there are factors contributing to the high variability of resting membrane potentials when compared across labs. This argument holds for other electrophysiological properties. In fact, certain experimental conditions (animal

species, age, electrode type, recording temperature) have been previously shown to be systematically correlated with variance in ephys measurements (Tripathy et al., 2015). This analysis motivated my consideration of experimental solution compositions as a potential explanation for inter-study variance.

3.2.3 Modeling electrophysiological properties with experimental metadata

3.2.3.1 Explaining electrophysiological variance using single solution components

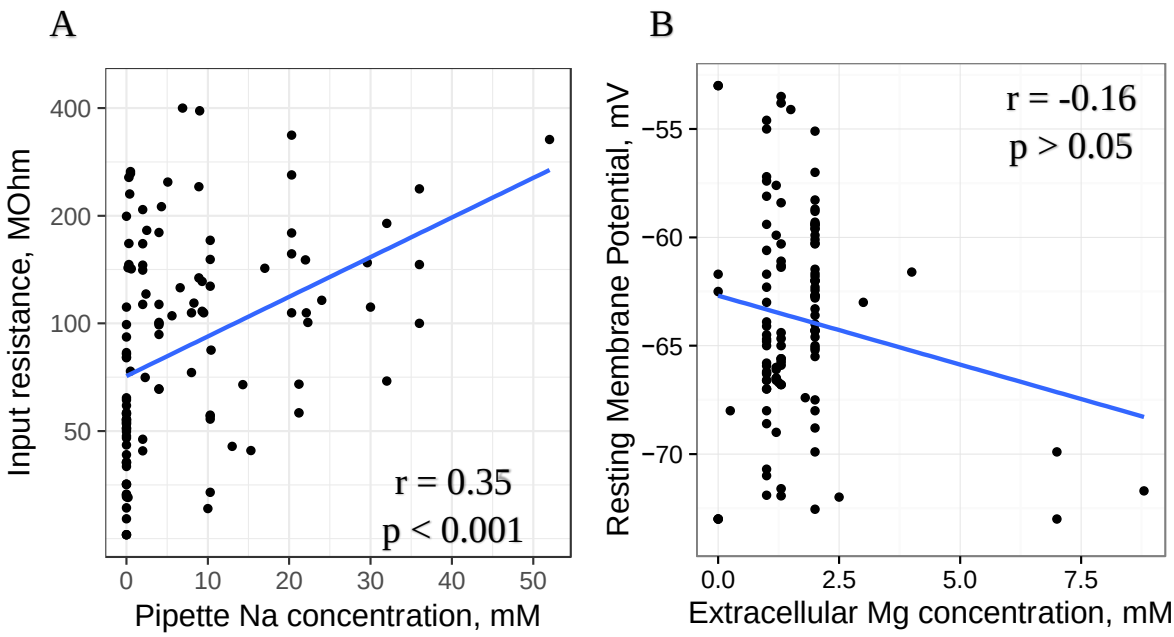


Figure 7: Univariate relationships between electrophysiological properties and solution concentrations. Each point is a mean ephys value reported by an article for Hippocampus CA1 pyramidal neurons. Blue line is the best univariate linear fit for the data, grey area shows 95% confidence interval for the linear fit. A) Input resistance increases with internal sodium concentration (r = 0.35, p < 0.001). B) RMP linear model is driven by 3 outliers in the 7-7.5 mM range of external magnesium concentration, rendering its results insignificant.

The simplest approach to studying the effects of major ion concentrations on electrophysiological properties is to treat the ions one at a time. This approach ignores the possibility of more complicated interactions, and cannot handle multiple cell types. However, it can identify strong correlations between specific solution components and ephys properties in a single cell type. Two examples of this univariate approach, when applied to hippocampus CA1

pyramidal neuron type: input resistance significantly correlates with internal sodium concentration (Figure 7A). As an example of a negative finding, the relationship between resting membrane potential and external magnesium concentration is driven by outliers (Figure 7B).

Systematically applying this univariate linear modeling approach to each neuron type, I did not find significant relationships between ephys data and individual ion or compound concentrations (FDR < 0.05). Confounding effects of other factors (age, species, electrode type, other solution components) likely mask true correlations if any exist. Searching for articles that have the same methods apart from a single solution component was not a feasible approach due to sparsity of the dataset. There are too few articles that use the same experimental conditions except one, at that point the limited sample size would render the analysis statistically underpowered. Subsequently, I considered a multiple regression approach that incorporates the influence of several experimental parameters on the same ephys property simultaneously.

3.2.3.2 Multiple regression approach for modeling variability in electrophysiological properties

Building on the regression approach developed previously, I hypothesized that it should be possible to model the effects of solution parameters on the resulting ephys measurements. To that purpose, I used a Random Forest machine learning algorithm to construct regression models relating ephys properties to metadata features (described in detail in Methods). The models were designed to simultaneously capture the effects of neuron type, solution composition information and basic metadata like species, age, temperature, electrode type (Table 3). I chose Random Forest over the classic linear regression approach because it is a non-linear model that empirically better handles statistical overfitting (which would cause a failure to generalize well to unseen test data) when using datasets with many features relative to sample size (Breiman, 2001). All models were trained and tested using 10-fold cross-validation, with performance summarized by R^2. An R^2 value of 1 means that the model was 100% correct in all predictions (which is essentially unattainable). An R^2 value of 0 means that the predictions are as accurate as using the mean value of the training ephys data for the predictions of test samples. A negative R^2 means that the model performs worse than the mean because of overfitting to the training data. Additionally, I define a 'baseline' R^2 value (calculated to be -0.30), which is generated by randomly shuffling ephys values. It serves as a lower bound for the worst predictions that could be made when the model is essentially predicting noise. When my models consistently achieve positive R^2 values, they should be used for predicting ephys values (instead of using the mean value). I consider models with negative (but above the lower bound, figure 9) R^2 values capable

of explaining a small amount of ephys variability, however they should not be used for predicting ephys values.

To compare the effect of solutions metadata to basic metadata when modeling the variability in ephys properties, I designed several models: neuron type only, neuron type + basic metadata, solutions only, neuron type + solutions, basic metadata only and all three sets of features combined. I expected the all features model to have the best performance since it has access to the information other models lack. My expectation for the solutions metadata models to be overall less successful than the basic metadata models, but the solutions + neuron type models to outperform the basic metadata + neuron type models, meaning that solutions are less correlated with neuron type than basic metadata.

Applying the Random Forest algorithm to my data, first I used a model that related all metadata features to input resistance (Figure 8A). The predicted values are the model's best estimates of what the observed ephys values should be given the experimental conditions from each article. In general, the predicted ephys values have less variance than the observed ones. That behaviour is expected, because the models can only partially predict the ephys variance.

The next step was to evaluate the relative contributions of neuron type, basic metadata and solutions when predicting ephys properties, starting with input resistance (Figure 8B, metadata details listed in table 3). Since the folds are assigned to articles in a random fashion, the performance of the models in each fold is slightly different. However, all 6 models are run using

the data from the same 10 folds, there is no reshuffling of data between different models. Judging by the model performances, solution features help to predict input resistance.

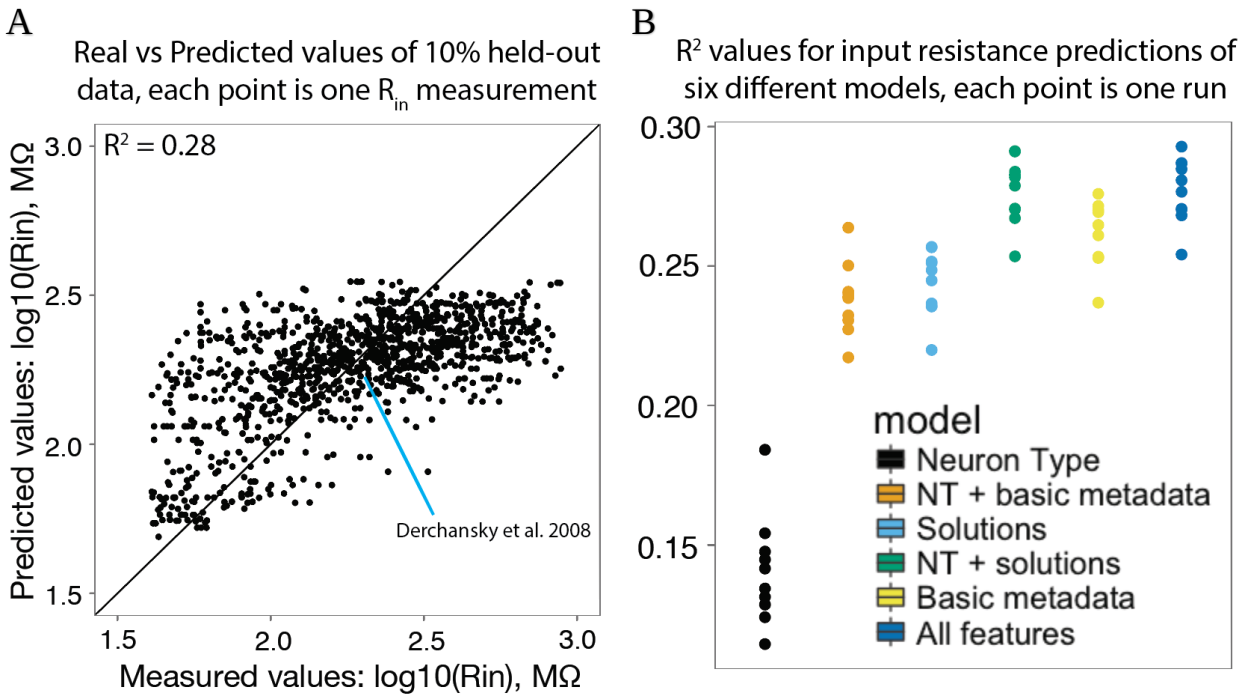


Figure 8: Multivariate regression models can predict ephys properties. Predictions are performed on held-out data (10x cross-validation). A) Each point is an input resistance value, reported by an article and predicted by a model using all metadata features (1 fold). B) Comparison of 6 different models for input resistance, each model uses a different set of features. Briefly, *Neuron Type* (NT) indicates a model using neuron type information only, *basic metadata* refers to information like animal age, recording temperature, etc., *solutions* refer to the use of internal and external solution concentrations, and *all features* refers to the combined set of metadata.

All features and the neuron type + solution features perform very similarly and better than neuron type + basic metadata. However, solutions on their own perform worse than basic metadata. It could mean that neuron type and basic metadata features provide similar information to the models, whereas solutions explain additional variance in input resistance. Neuron types alone cannot predict input resistance values as well as in conjunction with basic and solution features.

Expanding input resistance modeling to 11 commonly reported ephys properties, I evaluated the effectiveness of each model type in predicting them (Figure 9).

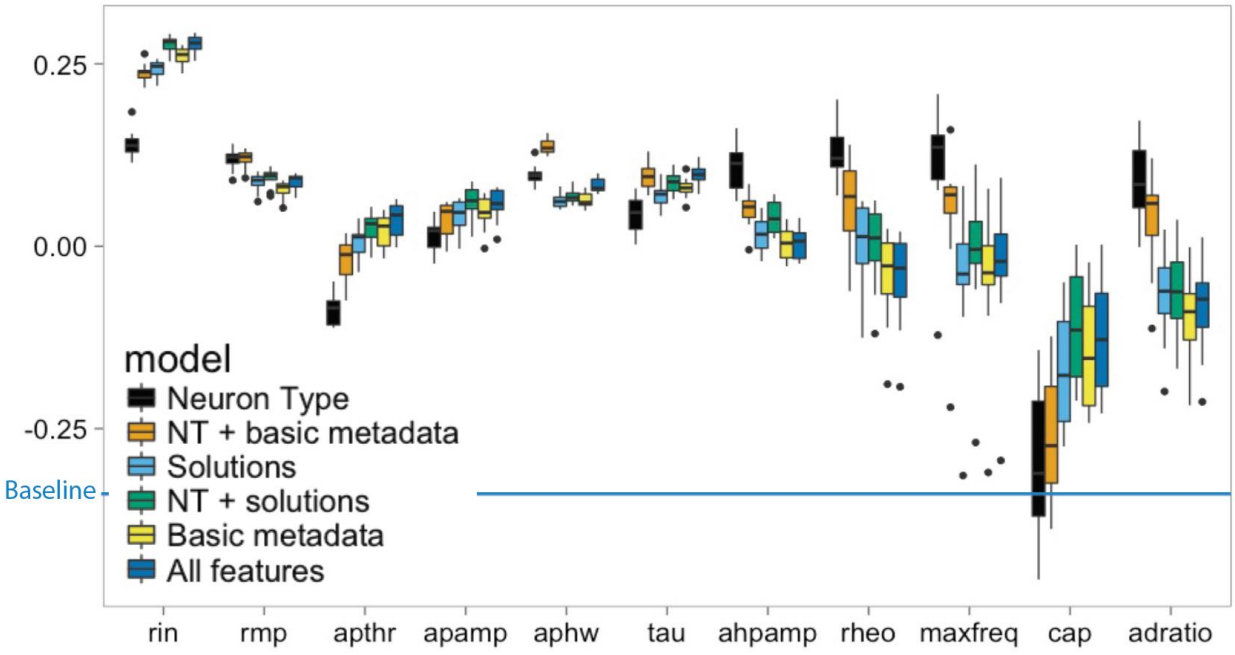


Figure 9: Comparison of models featuring basic and solutions metadata. Random Forest models with different feature sets (legend) predict commonly reported ephys properties. Baseline is the lower bound for model performance. Each boxplot represents R^2 values of 10 runs for that model. The number of data rows per property decreases from left to right.

Here, the null-model (using ephys property means as predictions) has an R^2 of 0. The figure can be interpreted as: the models that perform better than the null-model in all 10 folds can capture a significant amount of the ephys variance; the models with R^2 values between baseline (-0.3) and 0 can explain some amount of variance in the ephys property, but not enough to outperform the null-model. Input resistance and resting membrane potential models can explain a significant amount of their respective variance, however, in most cases the models are only slightly better (if at all) than simply taking an average of the observed values and using that as an estimate for the

ephys property. Interestingly, the 4 out of 5 properties on the right-hand side of the plot (AHP amplitude, rheobase, maximum firing frequency and adaptation ratio) get the best predictions out of neuron type only model. Additionally, ephys properties with less available data (Ordered from left to right: abundant to sparse) have much less stable performance levels. These effects are likely to be artifacts of not having enough data to sufficiently train the multiple regression Random Forest models. I observe a general increase of 0.2-0.5 in the predictive power of my models when comparing to the baseline, implying that, in most cases, the models can partially explain the variability in ephys values. On average, solutions contribute less to the overall model predictive power than neuron name or basic metadata, however, in some cases they substantially increase all features models performances (R_{in}, AP_{thr}, AP_{amp}). Therefore, solutions can contribute different information than neuron name or basic metadata when modeling ephys properties.

To formalize the effect of available data on a model's performance, I quantify the effect of varying the number of data points that a model can use to predict an ephys property. I used input resistance as the ephys property with the most available data and ran a model that uses all metadata features on a subset of available articles. There is a strong correlation between the number of articles for an ephys property and the R^2 values of a model that predicts it (Figure 10). This fact is reassuring, as models performance is highly likely to improve with more articles being added to NeuroElectro. It might also mean that ephys properties that currently have less than 300 entries and are not predicted reliably could improve their models performances drastically. Inevitably, the value of adding new articles will decrease, but we are not at that stage yet.

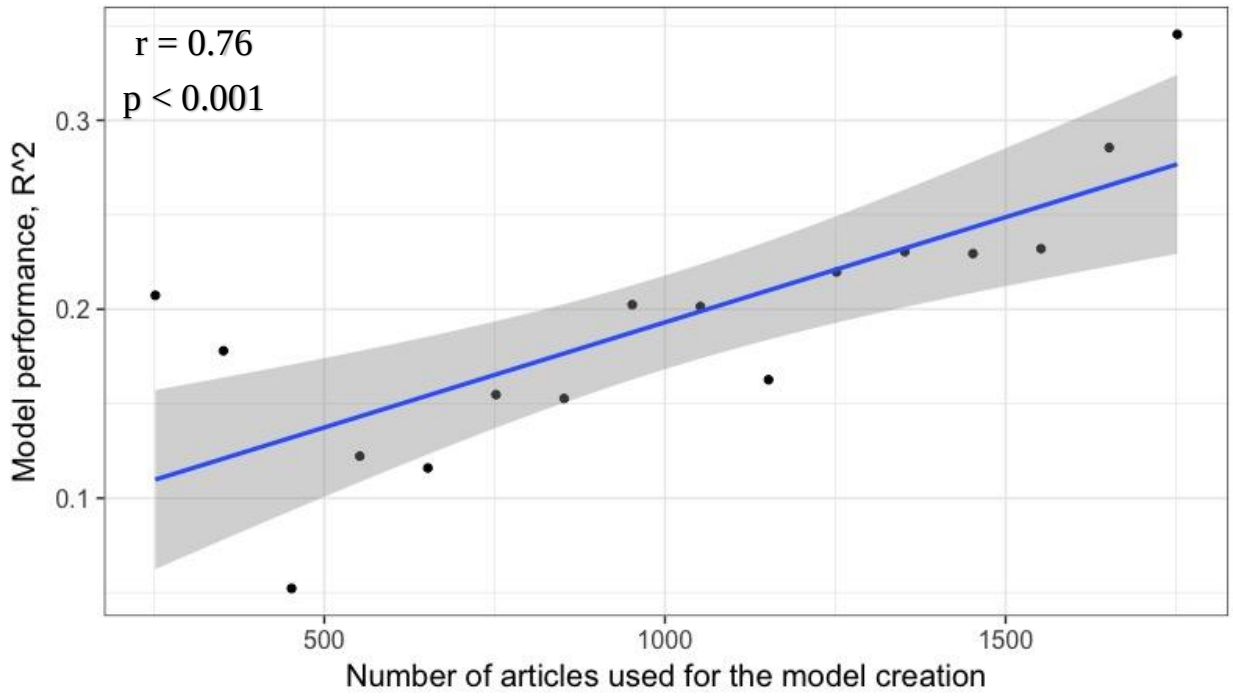


Figure 10: Multivariate model performance improves with N. R^2 performance for predicting input resistance with N rows of data. Blue line is the linear model best fit; grey region represents 95% confidence interval for the fitted line.

Additionally, I evaluated the performance of the GHK equation for modeling membrane potentials of neurons at rest by comparing its predictions based on recording temperature and Na, Cl, K concentrations to the observed ephys values. Strikingly, the GHK model essentially failed to predict the RMP values of hippocampal CA1 neurons ($R^2 < 0$).

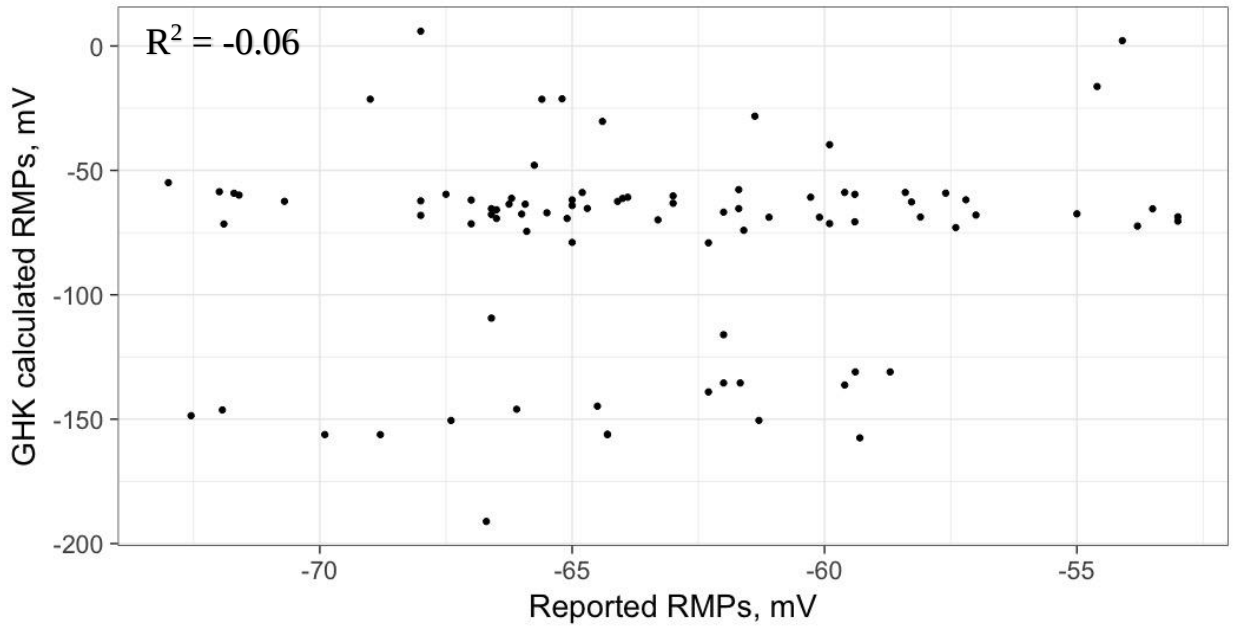


Figure 11: Predicting RMPs of hippocampus CA1 pyramidal cells with the GHK equation. Each point is a mean RMP reported by a single article in NeuroElectro. GHK calculated RMPs refer to the usage of experimental metadata stored in NeuroElectro for the prediction of resting membrane potentials.

A possible explanation is that using generic membrane permeability values (Hille, 1984) did not provide a reasonable substitution for the actual Na, K and Cl ionic permeability values of CA1 neurons. Another possibility is that the noise introduced by other experimental conditions when measuring neuronal resting membrane potentials prevents the GHK equation from accurately estimating their values.

3.2.4 Optimizing multiple regression models for predicting specific electrophysiological properties

3.2.4.1 Selection of the most predictive experimental conditions per property

The final step of my project was to create models using only the important features, which might be different for each ephys property. For example, solution features represent an aggregate of 22 different features: 5 major ions and 6 commonly used compounds per internal and external solutions. What if some of these concentration features are important when predicting ephys properties, while the rest only introduce noise? Next step of my project was to determine which features (solutions and basic metadata) should be used to model each ephys property. Briefly, I used Random Forest internal variable importance tool (Strobl et al., 2008) and Akaike information criterion with a correction for finite sample sizes (AICc). It assigns a score to each model based on its performance, adjusted for the number of features used and the amount of data that is available. The goal is to choose the optimal number of top features to model each ephys property.

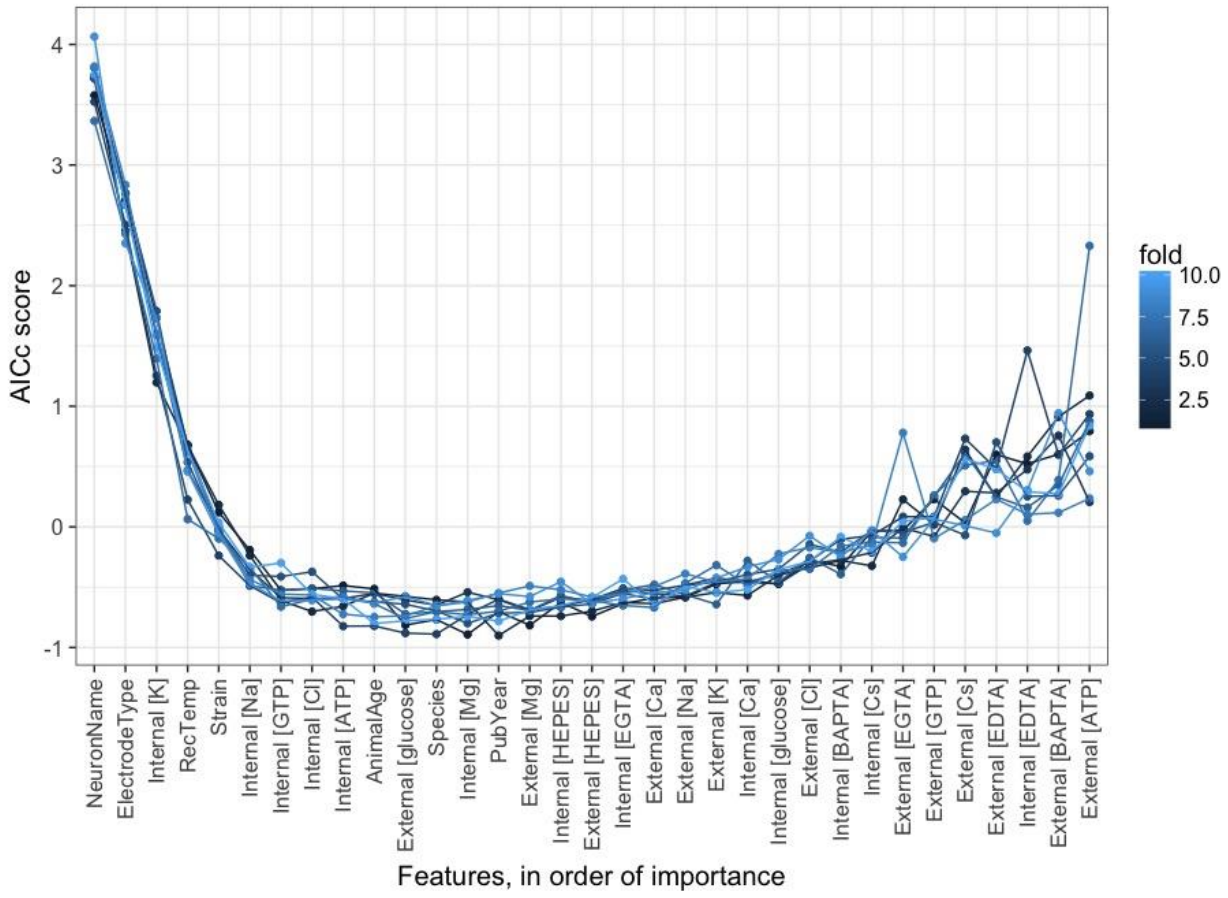


Figure 12: Model comparison using AICc score. One run of 10-fold cross-validation, each line is an AICc curve (per fold) calculated by adding top X (from 1 to all 33) features to the model that predicts input resistance. Model with the lowest AICc score is the best performing one. Metadata features are ordered from high to low based on their importance, calculated by cforest (X-axis).

Here, I use AICc to evaluate input resistance models with regards to each other (Figure 12). The lower AICc is, the better the model. Adding the first 6 features results in a big drop off, meaning that those features should always be included into the best input resistance model. After that, the AICc curve shifts up and down, depending on the fold. Finally, adding any features after external calcium and sodium would hurt the model rather than help it (at least at the current amount of data, this might change when more articles are added to NeuroElectro). The last 7 features

illustrate the amount of instability and noise bad features can add to a model (the effect of overfitting), making its performance vary greatly between different folds.

Generalizing the above Random Forest variable importance ranking complemented by AICc approach for input resistance model selection, I applied the same algorithm to the 11 of the most abundant electrophysiological properties in NeuroElectro. I have also performed this entire procedure 10 times for each ephys property to ensure I was getting stable results. Ten runs of the 10-fold cross-validation are summarized in Figure 13.

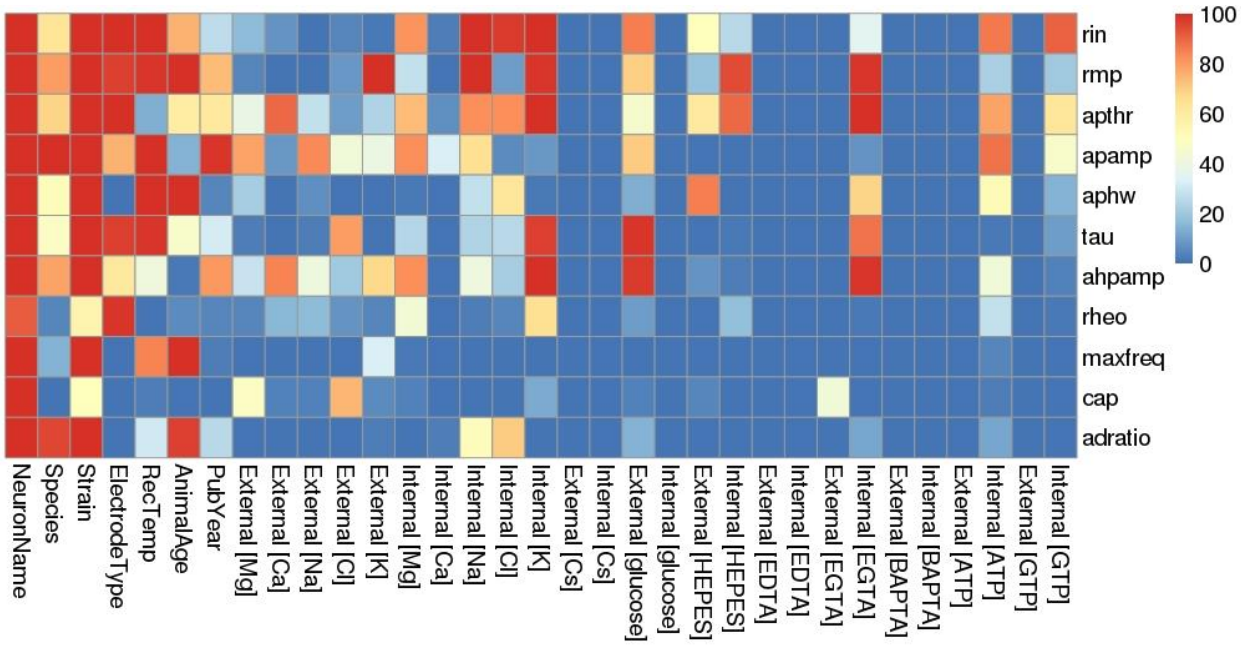


Figure 13: Feature importance, based on the frequency of inclusion into the models. Ephys properties and metadata features are listed vertically and horizontally, respectively. Color represents the number of times the feature was chosen for the ephys property's model (from 0 to 100 times). NeuronName stands for neuron type.

Neuron type and strain features are almost always chosen for modeling each ephys property. The species feature is included less frequently because most of its contribution is covered by strain,

which is also more informative, because 93% of the data in NeuroElectro comes from the experiments performed on mice or rats. A few solution features get consistently included for at least 1 ephys property as well: external K for RMP, internal Na for R_{in} and RMP, etc. It reinforces the earlier observation that certain solution components are important when predicting specific electrophysiological properties. Surprisingly, the publication year was an important feature when modeling AP amplitude (Figure 14). I hypothesize that this effect can be explained by changes in the AP amplitude calculation or measurement protocols.

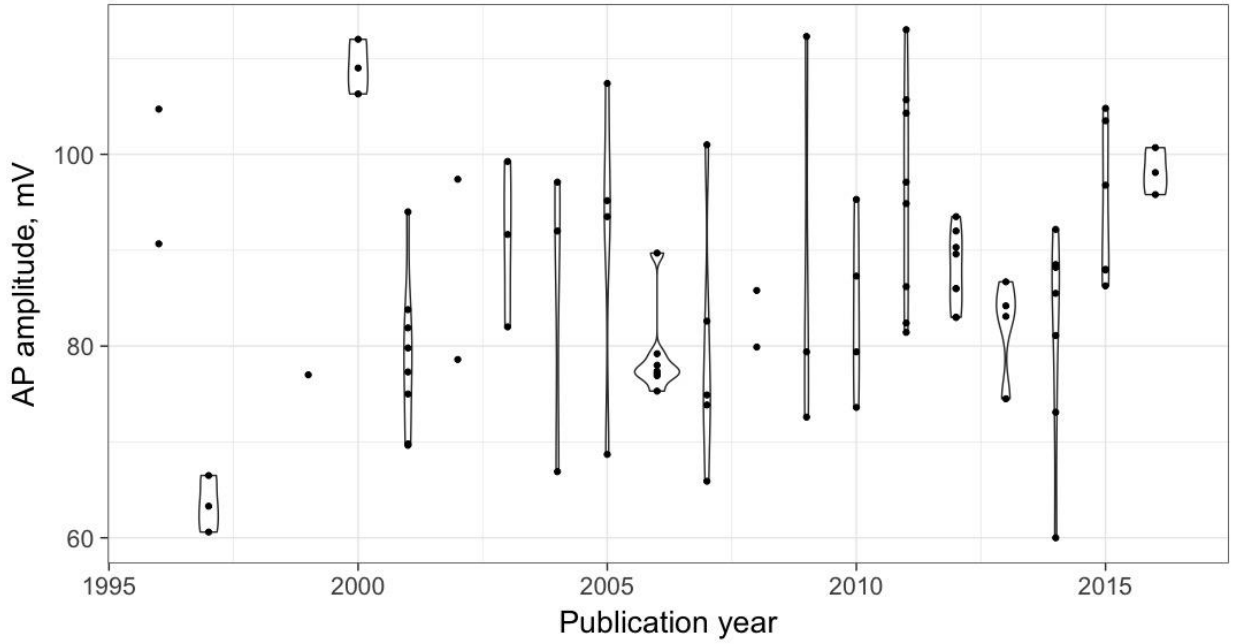


Figure 14: Reported action potential amplitudes of CA1 pyramidal cells vary with time. Each point is a population mean AP_{amp} value of Hippocampus CA1 pyramidal cells, reported by a single article published in the corresponding year. Violins outline the distributions of values for each year.

Another important aspect of the feature selection heatmap is that very few cells are colored yellow. It means that most features are either robustly good or robustly poor when predicting the corresponding ephys property. The poor performance of certain solutions features (BAPTA, EDTA, Cs, etc.) can be linked to their sparsity. If a solution component is included into <5% of

recipes, it is unlikely to be predictive of electrophysiological variance between studies stored in NeuroElectro. However, NeuroElectro curation generally targets current-clamp experiments, naturally causing low Cesium (a common voltage-clamp compound) inclusion rates. When a feature gets included into the best model <50% of the times, its performance is unstable, likely due to overfitting. Between 50% and 90% inclusion is the uncertain area where the feature might not be important enough to be included all the time but it does provide some useful information. The final models were created using features that are included in >90% of the best models to minimize overfitting.

3.2.4.2 Validating optimized models with NeuroElectro and Allen Institute Cell Types data

Having created the models for all ephys properties, I needed to compare their performance to the models used previously (Figure 9). To achieve that, I once again employ 10-fold cross-validation and calculate R^2 values for each model, substituting the neuron type + solutions model for the best features model (Figure 15). My models achieve the highest performance levels (or on par with other models) when predicting R_{in}, AP_{thr}, AP_{amp}. However, they fall short of "no solutions" model when predicting RMP and AP_{hw}. The differences in model performances are small and could be explained by randomness of splitting the data into 10 folds.

After comparing the feature-selected models to the basic ones, I decided to apply these new models to data unused in the fitting process or cross-validation, from the Allen Institute for Brain Science. Only AIBS neuron types that could be definitively assigned to a NeuroLex cell type were used in this analysis. Since AIBS data was produced during a set of experiments in a single

lab – all ephys property measurements were aggregated into mean values per neuron type, because NeuroElectro stores reported means, not individual measurements.

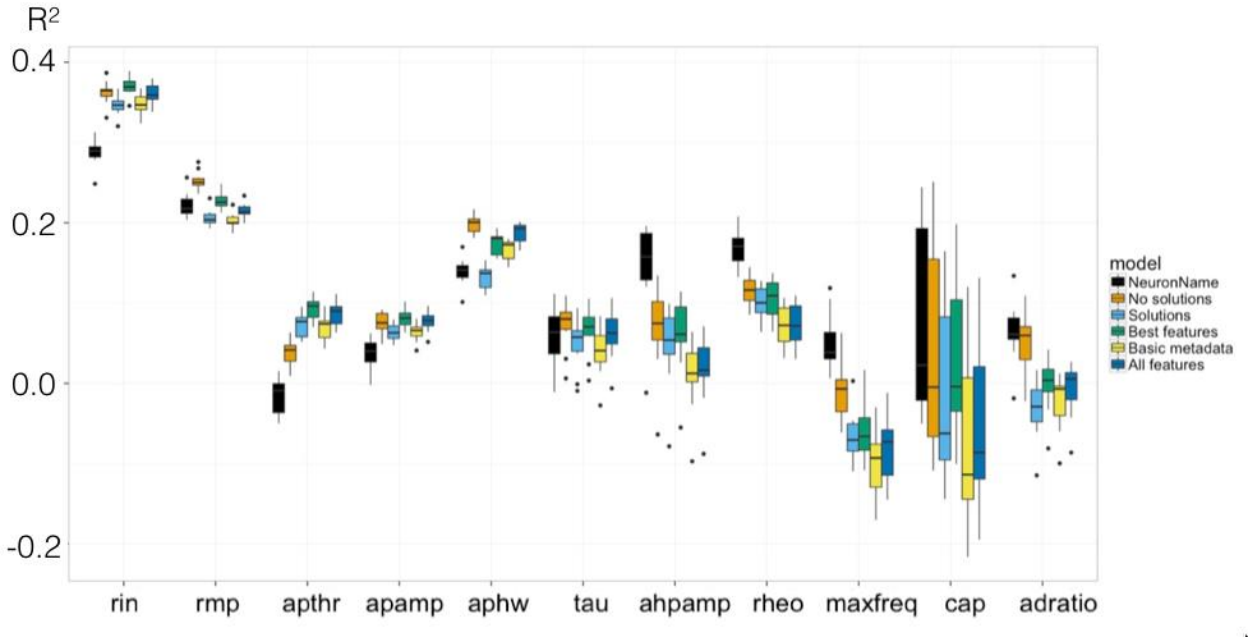


Figure 15: Comparison of feature-selected models to basic models. The best model (per AICc) for each property is shown in green color. Variable refers to the list of commonly reported ephys properties. R^2 value on the y-axis represents each model's performance.

To evaluate the feature-selected models performance, I used them to 'shift' NeuroElectro ephys data (NE) to AIBS experimental conditions baseline (See Methods for details). Briefly, I first predicted and removed ephys variance from NE data that could be explained by the models, then I added ephys variance introduced by AIBS experimental conditions.

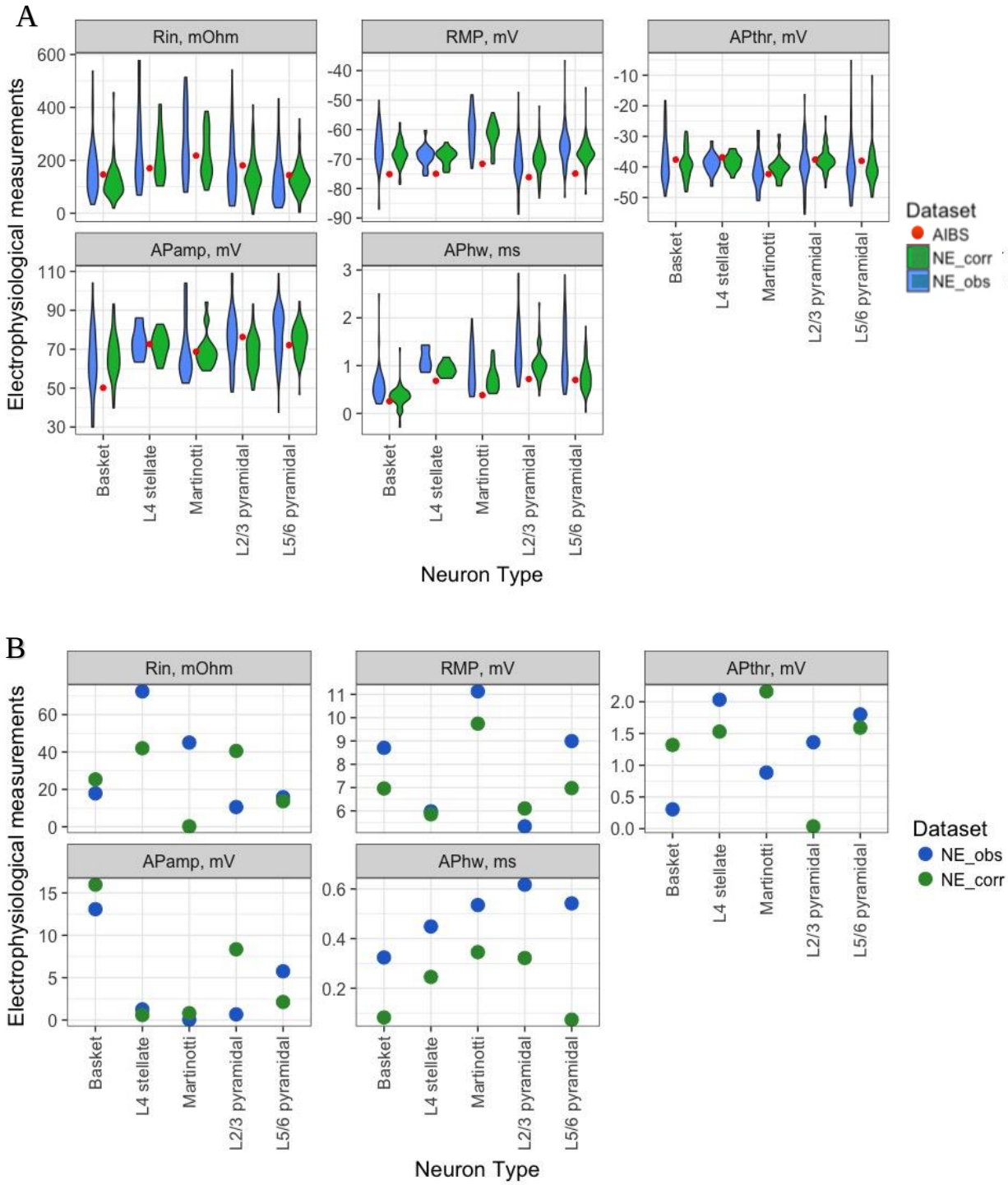


Figure 16: Adjusting NeuroElectro data to AIBS conditions. All AIBS neuron types come from neocortex. A) Violin plots of NeuroElectro data (blue), Allen Institute for Brain Science data (red), adjusted NeuroElectro data using the feature-selected models (green). Each point is a mean ephys property value reported by an experiment. B) Absolute differences between NeuroElectro raw and adjusted ephys values when compared to AIBS ephys means. The model correction tends to squeeze NeuroElectro data around the mean and bring it closer to AIBS value.

If the models work, the adjusted NE data should be more closely distributed around AIBS electrophysiological measurements. Figures 16A and 16B support that claim. Generally, the corrected NeuroElectro ephys values have less variance, which is the result of removing the explained variance from each reported ephys value, and their means are in most cases closer to the AIBS mean ephys values than raw NeuroElectro data.

Chapter 4:

Discussion and conclusion

In my thesis work, I have researched potential sources of study-to-study variability in measurements of neuronal electrophysiological properties. For that purpose, I performed a meta-analysis of neuron electrophysiology data and metadata gathered from published neuroscientific articles. Employing a combination of text-mining and curation techniques, I integrated experimental solutions used in these neurophysiological experiments into NeuroElectro. Then, I provided the most comprehensive exploration of the extracellular and intracellular solutions chemical compositions available to date. Additionally, I examined the relationships between experimental conditions and reported electrophysiological measurements. Finally, I proposed models for each commonly reported ephys property that allow partially correcting the ephys values, based on their experimental conditions.

The main finding of my work is that experimental solutions follow similar recipes, yet it is possible to use them, supplemented with basic metadata and neuron type information, to partially model electrophysiological properties of neurons. My research builds on the previous meta-analysis of ~300 neuroscientific papers, that focused on discovering the significant correlations between basic experimental conditions and study-to-study ephys variability (Tripathy et al., 2014). In this final section, I discuss my text-mining algorithms, my findings and their

implications with respect to the field of electrophysiology, noting certain limitations of the current study and suggesting directions for future work.

4.1 Discussion

4.1.1 Challenges in gathering experimental solution information using text-mining and curation

Extracting targeted words, phrases and sentences from texts written in natural language often proves to be a challenging task (Hirschberg and Manning, 2015). Exceptionally so, when the task is text-mining data from Biomedical literature (Pletscher-Frankild et al., 2015; Uzuner and Stubbs, 2015). In my opinion, two types of text-mining algorithms have the potential to succeed: general algorithms that focus on co-occurrence of terms (Examples: protein-protein interactions, linking gene expression changes to different phenotypes) if they are trained on a huge amount of data; and algorithms that extract very specific types of information (Fleuren and Alkema, 2015). The latter algorithm type does not require vast amounts of training data before it can achieve its task. My goal was to identify sentences with a unique structure, listing several chemical compounds near each other. That peculiarity provided the target required for developing a text-mining algorithm. I trained the algorithm on a manually curated set of 60 articles, which is by no means a large dataset, however it provided enough examples of solution-containing sentences for the text-mining algorithm to achieve ~90% accuracy rates. Therefore, highly targeted text-mining on a small scale can be successful.

While the solution-containing sentences identification step of the algorithm performed its task with high accuracy, assigning the correct type to each solution sentence proved to be more

difficult. The task to distinguish sentences that refer to extracellular, electrode and other (cutting, storage, incubation) solutions in simple cases was trivial. It was efficiently addressed by creating a dictionary of terms that refer to each solution type. However, articles that described several ephys experiments done under different conditions posed a significant challenge. For example, simultaneously reported patch-clamp and sharp electrode recordings, or current-clamp and voltage-clamp recordings in the same paper (and the same methods section). As before, the identification of solution types was generally correct, but the logic for choosing the right external and internal solution per experiment type was not included into the current implementation of the text-mining algorithm. That task was deemed too difficult for an automated text-mining algorithm to perform, because it proved to be challenging even for trained human curators.

Despite these challenges, my analysis of solution recipes confirmed several strict rules in internal solution designs: voltage-clamp experiments use cesium instead of potassium to block inwardly rectifying potassium channels (Rang and Dale, 2003); current-clamp experiments consistently use 120-140 mM of potassium; sharp electrode solutions can be distinguished by their relatively high concentrations of potassium (1-4 M). Using this background information, it might be possible to utilize internal solution compositions for the task of assigning solution sentences to the correct experiments. However, the merits of improving the solutions text-mining algorithm rely on other data (ephys, neuron types and basic metadata) being available for the text-mined articles. Therefore, until robust text-mining algorithms are implemented to collect such data, there is little profit to be gained for the time spent on improving the solutions extraction algorithm.

Throughout the course of this project, I helped to assemble and train the NeuroElectro curation team. It proved to be an invaluable asset in expanding the NeuroElectro database: over the course of 2 years the database grew from ~300 curated articles to nearly 900. Assisting the curators in their task, I developed an online curation interface, integrated into NeuroElectro (Figure 3). Its detailed description, curation speed and quality improvement metrics will be included into the future NeuroElectro paper. NeuroElectro primarily stores data from ephys tables that was reported under normal control conditions, meaning that with the old interface there was no way to fully annotate ephys experiments that studied the effects of experimental condition (temperature, animal age, solution compositions, etc.) changes on the ephys properties. The new interface enabled the addition of metadata types tracked by NeuroElectro (Table 3) to the columns of ephys data tables. Thus, the new curation interface not only assisted in increasing the number of curated neurophysiology articles stored in the NeuroElectro database, but it also enabled us to gather more data from papers that alter experimental conditions during the experiment.

One of the most important rules the NeuroElectro curation team had to follow was the "15-minute rule". It states that if a curator is spending more than 15 minutes on curating a single article, they should instead skip it, because with almost 100,000 articles available for curation more data could be gained from several simple-to-curate articles than from a single complicated paper. Anecdotally, the curation quality tends to decrease with the increase in article's complexity. Because an article had to be curated inside the 15-minute window, the chemical compound concentrations extraction step was not included into the normal curation protocol. Thus, it was designed to be accurate enough (F_1 score of 0.97 for major ions and 0.84 for other

compounds) to enable the downstream analysis. The main error causes are misspellings and inconsistencies. The concentration extraction algorithm cannot accommodate for any mistakes in the spellings of major ions or other compound abbreviations, because of how short their names are, a single letter variation could mean an entirely different compound. The algorithm assumes that the separator (comma, semicolon) that is used to split up the first few compound concentrations would be consistently used throughout the recipe, but that is not always the case. Standardizing the practices of reporting experimental conditions used in electrophysiological experiments could increase the effectiveness of text-mining them in the future.

4.1.2 Trends in experimental solution recipes

Major ion (Na, K, Cl, Ca, Mg) concentrations used in extracellular solution recipes resemble normal distributions, which could mean that, historically, several different labs measured these ionic concentrations in cerebrospinal fluids of animal brains (most of NeuroElectro data comes from mice and rats). Then, throughout the years the ion concentrations were tweaked slightly around the reported values. Another possibility is that students tend to inherit solution recipes from their supervisors, occasionally tweaking the major ion concentrations slightly. This implies that in the beginning of neuron electrophysiology there was a common ancestor who designed the first ACSF. A likely candidate for the role could be Sydney Ringer, who in 1882-1885 determined that the recipe for the physiological solution perfusing a frog's beating heart must contain sodium, potassium and calcium in certain proportions (Hille, 1984). Naturally, the true reason could be a combination of the two proposed explanations, or something else entirely.

Only two other compounds, besides the major ions, were detected in ACSF recipes: glucose and HEPES. Glucose is very consistently included into extracellular solutions, albeit at different concentrations, because neurons become irreversibly damaged if deprived of glucose for extended periods of time (Burdakov et al., 2005; Routh et al., 2004). HEPES is generally used for its pH buffering properties, however only a small subset of papers (~10%) use it externally, other papers adjust the pH by adding small amounts of a strong base or acid into ACSF (Boehlen et al., 2013; Koyama and Appel, 2006).

Electrode solution compositions do not share the trends of extracellular solution recipes. Electrophysiologists tend to agree that including tiny amounts of GTP is good for the well-being of the cells. However, other compounds that are routinely included into pipette solutions are used at two or more different concentration levels. The abundance or varying recipes might be explained by the need to tailor electrode solutions to the specific requirements of each experiment, even when using similar clamping techniques. Very few papers agree what the internal sodium and chloride concentrations should be, as they are almost uniformly distributed between 0 mM and 50 mM (skewed towards 0 mM).

The different 'schools of thought' represent the largest patch-clamp solution recipes trends I could identify. The reasons behind the extracellular Mg concentration separation into 1-1.5 mM or 2 – 2.5 mM bins remain unknown. I hypothesize that this effect could be an artifact of recipes being inherited through generations of electrophysiologists. On the other hand, in mid/late-2000's electrophysiologists started to consistently add phosphocreatine to their internal solutions (Pilarski et al., 2011; Yang et al., 2013) and, since Na_2-phosphocreatine is a relatively

inexpensive way to fulfill that goal when compared to K_2-phosphocreatine (71.5 USD/gram versus 334 USD/gram, on EMD Millipore website), internal sodium concentrations started to increase. Surprisingly, my analysis has indicated a small positive correlation between internal sodium and chloride, implying that internal chloride concentrations also increased with time (not significantly, though). The changes in concentration values over time are likely caused by researchers who discover beneficial effects of certain chemicals on the state of neurons during electrophysiological experiments. It is possible that chloride concentrations increased because it was just the counter-ion for another chemical deemed beneficial for ephys recordings. Since NeuroElectro does not have an easy way of tracking new compounds being used in chemical solutions, we would have a difficult time identifying these over-time shifts in the concentrations of uncommonly used compounds. To address the issue in the future, it is possible to extend the NeuroElectro solutions text-mining algorithm to use a dictionary of all known chemical compounds. However, it is important to note that such an approach would introduce many instances of extremely rare compounds and no general conclusions could be drawn from such cases due to sparsity of data.

4.1.3 Implications of modeling study-to-study electrophysiological variability

The field of neuron electrophysiology has been in dire need of methods that enable better comparisons of ephys data between experiments. My analysis of electrophysiological data stored in NeuroElectro has shown that, even within similar types of neurons, ephys measurements cannot always be directly compared between experiments. Strikingly, even the well-defined hippocampal CA1 pyramidal neuron type has a wide range of reported resting membrane potential values: from -55.1 mV (Kim and Connors, 2012) to -80.0 mV (Booth et al., 2014). My analysis has demonstrated that such study-to-study variability can be partially explained by experimental conditions (metadata), specifically solution compositions. Electrophysiology data, collected from hundreds of articles and stored in NeuroElectro, is very heterogeneous when compared to results reported by a single lab. The innate high variability of neuronal ephys properties could be masking the impact of experimental conditions I use to predict them. Additionally, there are other experimental conditions in neuroscientific articles that NeuroElectro does not keep track of, such as pipette properties (resistance, glass type), scientific kit (amplifier), for *in vitro* recordings - time between brain extraction, slicing, incubation and ephys measurements. Ideally, we would like to store all the experimental setup information provided by the authors of every article, but that would require a huge time investment into creating the infrastructure capable of supporting such a task. To accommodate for the middle ground between storing all or none metadata information, NeuroElectro extracts the most commonly and consistently reported types of metadata (Tripathy et al., 2015).

My initial approach of using univariate linear models for predicting electrophysiological data with one compound concentration at a time did not yield many nominally significant results. Additionally, they did not pass the significance threshold after the multiple testing correction adjustment. The major cause for the poor performance of univariate linear models is the fact that very few compound concentrations are distributed evenly over a wide range of values. As discussed, most of the compounds tend to be tightly normally distributed around specific values, or they are only used at several (rarely more than two) set concentrations. Factoring in the innate variability of ephys properties, I found it unsurprising that univariate models were not able to explain much of the ephys variance, even when considering a single neuron type (in the case of hippocampus CA1 pyramidal neurons).

Employing non-linear multivariate models to predict the variance of ephys properties using metadata features proved to be the better approach. Comparing models that use six different sets of metadata features, I confirmed that solutions (as a group) can provide valuable information when predicting input resistance, action potential threshold and amplitude. Neuron types alone cannot predict input resistance values at similar accuracy levels when compared to models using additional basic and solutions features. That could be caused by the fact that the most common neuron type in NeuroElectro is “other”, which is an aggregation of all neuron types that cannot be directly assigned a NeuroLex term. Surprisingly, adding solution information impeded the model when predicting resting membrane potentials, despite the known relationship between RMP and several major ions given by the Goldman-Hodgkin-Katz equation, along with temperature. My explanation is that certain solution components are likely useful when

predicting RMPs, but their predictive power is masked by the noise introduced by the other solution components.

To test that hypothesis, I used random forest feature importance ranking and the corrected Akaike Information Criterion to construct models that use only the top few best features per ephys property (the cut-off chosen using AICc). As expected, neuron type proved to be the most valuable source of information when modeling almost all ephys properties, because neurons are classified into types based on their genetic, morphological and electrophysiological features (among other characteristics). Other commonly chosen basic metadata types include: strain, which is generally more informative than species in NeuroElectro, electrode type, recording temperature and animal age. It is very reassuring that the models often chose to include basic metadata types that have already been shown to significantly correlate with study-to-study ephys variability (Tripathy et al., 2015).

I addressed my previous concern of some solution features masking the impact of predictive ones when predicting ephys properties by further examining the solution features that often get chosen for the best model. Several solution components are consistently included for the resting membrane potential predictions: internal and external potassium, internal sodium and internal EGTA. Internal EGTA is weakly correlated with RMP ($r = 0.12$), but it could be making the difference in distinguishing several otherwise unpredictable RMP values. Internal sodium could be masking the effects of internal chloride on RMP predictions, because their concentration values are also correlated ($r = 0.20$) and internal sodium provides more explanatory power than chloride.

External compound concentrations are tightly distributed around specific values, implying that most experiments use very similar compound concentrations in their ACSF. My models can only detect signal from metadata that possess enough variance to uncover it, thus extracellular solutions cannot provide sufficient explanatory power to the models. I am not claiming that external solution concentrations do not matter when predicting the variability in ephys values, but rather that given the current dataset, I could not detect their relationships. On the other hand, internal major ion concentration values are widely spread, making them good candidates for modeling certain ephys properties (R_{in}, RMP, AP_{thr}, membrane time constant, AHP_{amp}). Additionally, compounds that were rarely seen in solution recipes (BAPTA, EDTA, cesium) were included into the model selection procedure as negative controls.

Comparing the feature-selected models to the previously considered ones, I note their consistent relatively good performance. They do not always achieve the highest R^2 values, but they also never completely fail. The feature selected models cannot magically achieve much better performance, than the several models I created initially, since they often share the same features (Example: neuron type + basic metadata model). The purpose of the new models is to eliminate features that introduce noise, which is the case with most of the solution features. Additionally, these new models are dynamic and with more data added to NeuroElectro they can change, incorporating previously inconsequential features.

As it stands, for those interested in applying my models to shift NeuroElectro data to the baseline defined by their experimental conditions, I recommend using all models except the ones

predicting adaptation ratio, maximum firing frequency and capacitance. There is currently not enough data in NeuroElectro (~300 data rows or less) to even attempt to explain the variance in those ephys properties. In comparison, rheobase also has only ~300 mentions, but it can be robustly modeled, possibly because it is reported more consistently than the other three rare properties. Finally, I adjusted the five most commonly reported ephys properties in NeuroElectro to the experimental conditions used by AIBS. The effects of my models include removing explainable ephys variability from NeuroElectro data, making it more comparable with the AIBS ephys measurements.

4.2 Future directions

The text-mining algorithm currently employed in NeuroElectro can be extended in several ways. First, improving the performance of the chemical compound concentrations text-mining algorithm by enabling it to comprehend common spellings of compounds, for example, we know that calcium chloride implies $CaCl_2$ and not CaCl, but the text-mining algorithm does not currently have access to the ionic valence information. To truly address this task, all chemical compounds need to be identified and their concentrations extracted by the text-mining algorithm. That, in turn, requires access to a comprehensive database of chemical compounds, their formulas, common and uncommon spellings of their names. The backbone code for this project already exists in the NeuroElectro codebase (assign_metadata.py).

The second text-mining algorithm extension option should be less challenging. It is possible to enhance experimental solution type assignment to solution-containing sentences by extracting compound concentrations first, and then assigning the solution type. For example, I have shown that voltage-clamp experiments very consistently use cesium in their internal solution recipes, instead of potassium. In the event of an article listing multiple ephys experiments (voltage- and current-clamp), internal solutions could be automatically assigned to the correct experiment.

The third potential improvement involves enhancing the existing algorithm for text-mining electrophysiology properties from tables and text of neuroscientific papers. Text-mining ephys properties from all articles stored in NeuroElectro would enable many new types of analyses and

make the analysis described here many times more powerful. As discussed, increasing the number of text-mined papers dramatically increases model performance ($r = 0.76$, $p < 0.001$). Approaching the point of diminishing returns would decrease that correlation, but we are not there yet, thus, as it stands, it makes perfect sense to add new articles to the analysis. It is possible that the models that cannot even partially explain ephys variability (maximum firing frequency, adaptation ratio, capacitance) could be improved with more data provided to them.

It is possible to further explore the sources of ephys variability. As discussed earlier, there are many experimental conditions that are not tracked by the existing version of NeuroElectro, which could provide valuable information to this type of analysis. For example, an ongoing project in the Pavlidis lab searches for links between similarities of experimental conditions used by pairs of neurophysiologists and how related they are in terms of training. Specifically, we are trying to link NeuroTree, an online database of neuroscientific genealogies (David and Hayden, 2012), and the experimental metadata stored in NeuroElectro.

Additionally, given that the solutions text-mining algorithm can be extended to extract all solutions components, the effects of counter-ions from the major ions salts can be studied (Examples: $NaHCO_3$, K_2SO_4, $KMeSO_4$). This information could provide compounds dissociation information to the multiple regression models, improving the performance of major ions when predicting electrophysiological properties.

4.3 Conclusion

In conclusion, my integrative meta-analysis approach addresses the neuroscience need for comparing electrophysiological data across different studies. The models proposed here partially adjust for ephys variability that can be explained by experimental conditions, generally enabling improved comparisons of reported electrophysiological properties across experiments. Electrode solution constituents show greater diversity in their concentration values when compared to extracellular recipes, resulting in the more common inclusion of internal solutions components into the models. External solution recipes used by the neurophysiological community are too similar, possibly preventing this type of analysis from detecting their true effects on the variability in ephys properties. I conclude that while experimental solutions possess certain explanatory power when modeling the variability of ephys properties, the larger part of the study-to-study ephys variability remains left to be explained.

Bibliography

Aghajanian, G.K., and Rasmussen, K. (1989). Intracellular studies in the facial nucleus illustrating a simple new method for obtaining viable motoneurons in adult rat brain slices. Synapse *3*, 331–338.

Agmon, A., and Connors, B. (1991). Thalamocortical responses of mouse somatosensory (barrel) cortexin vitro. Neuroscience *41*, 365–379.

Aivar, P., Valero, M., Bellistri, E., and Prida, L.M. de la (2014). Extracellular Calcium Controls the Expression of Two Different Forms of Ripple-Like Hippocampal Oscillations. J. Neurosci. *34*, 2989–3004.

Altman, N.S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am. Stat. *46*, 175–185.

Anderson, D.R., and Burnham, K.P. (2002). Avoiding pitfalls when using information-theoretic methods. J. Wildl. Manag. 912–918.

André, V.M., Cepeda, C., Cummings, D.M., Jocoy, E.L., Fisher, Y.E., William Yang, X., and Levine, M.S. (2010). Dopamine modulation of excitatory currents in the striatum is dictated by the expression of D1 or D2 receptors and modified by endocannabinoids. Eur. J. Neurosci. *31*, 14–28.

Archer, K.J., and Kimes, R.V. (2008). Empirical Characterization of Random Forest Variable Importance Measures. Comput Stat Data Anal *52*, 2249–2260.

Armentia, M.L.D., and Sah, P. (2004). Firing Properties and Connectivity of Neurons in the Rat Lateral Central Nucleus of the Amygdala. J. Neurophysiol. *92*, 1285–1294.

Bird, S. (2006). NLTK: the natural language toolkit. (Association for Computational Linguistics), pp. 69–72.

Boehlen, A., Henneberger, C., Heinemann, U., and Erchova, I. (2013). Contribution of near-threshold currents to intrinsic oscillatory activity in rat medial entorhinal cortex layer II stellate cells. J. Neurophysiol. *109*, 445–463.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilita (Libreria internazionale Seeber).

Booth, C.A., Brown, J.T., and Randall, A.D. (2014). Neurophysiological modification of CA1 pyramidal neurons in a transgenic mouse expressing a truncated form of disrupted-in-schizophrenia 1. Eur. J. Neurosci. *39*, 1074–1090.

Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.

Burdakov, D., Luckman, S.M., and Verkhratsky, A. (2005). Glucose-sensing neurons of the hypothalamus. Philos. Trans. R. Soc. B Biol. Sci. *360*, 2227–2235.

Burnham, K.P., and Anderson, D.R. (2004). Multimodel inference understanding AIC and BIC in model selection. Sociol. Methods Res. *33*, 261–304.

Cembrowski, M.S., Bachman, J.L., Wang, L., Sugino, K., Shields, B.C., and Spruston, N. (2016). Spatial gene-expression gradients underlie prominent heterogeneity of CA1 pyramidal neurons. Neuron *89*, 351–368.

CHANDLER, R.C. (1995). Practical considerations in the use of simultaneous inference for multiple tests. Anim. Behav. *49*, 524–527.

Chen, I.-W., Helmchen, F., and Lütcke, H. (2015). Specific Early and Late Oddball-Evoked Responses in Excitatory and Inhibitory Neurons of Mouse Auditory Cortex. J. Neurosci. *35*, 12560.

Cooper, D.C., Moore, S.J., Staff, N.P., and Spruston, N. (2003). Psychostimulant-induced plasticity of intrinsic neuronal excitability in ventral subiculum. J. Neurosci. Off. J. Soc. Neurosci. *23*, 9937–9946.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

Cui, R.J., Li, X., and Appleyard, S.M. (2011). Ghrelin inhibits visceral afferent activation of catecholamine neurons in the solitary tract nucleus. J. Neurosci. *31*, 3484–3492.

David, S.V., and Hayden, B.Y. (2012). Neurotree: A Collaborative, Graphical Database of the Academic Genealogy of Neuroscience. PLOS ONE *7*, e46608.

Derchansky, M., Jahromi, S.S., Mamani, M., Shin, D.S., Sik, A., and Carlen, P.L. (2008). Transition to seizures in the isolated immature mouse hippocampus: a switch from dominant phasic inhibition to dominant phasic excitation. J. Physiol. *586*, 477–494.

Dunn, T.F., and Goldstein, L.G. (1959). Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. Educ. Psychol. Meas.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. Science *349*, 636–638.

Fleuren, W.W., and Alkema, W. (2015). Application of text mining in the biomedical domain. Methods *74*, 97–106.

Freedman, D.A. (2009). Statistical models: theory and practice (cambridge university press).

Fujiwara-Tsukamoto, Y., Isomura, Y., Kaneda, K., and Takada, M. (2004). Synaptic interactions between pyramidal cells and interneurone subtypes during seizure-like activity in the rat hippocampus. J. Physiol. *557*, 961–979.

Gall, D., Roussel, C., Susa, I., D'Angelo, E., Rossi, P., Bearzatto, B., Galas, M.C., Blum, D., Schurmans, S., and Schiffmann, S.N. (2003). Altered neuronal excitability in cerebellar granule cells of mice lacking calretinin. J. Neurosci. *23*, 9320–9327.

Gittis, A.H., and Lac, S. du (2007). Firing Properties of GABAergic Versus Non-GABAergic Vestibular Nucleus Neurons Conferred by a Differential Balance of Potassium Currents. J. Neurophysiol. *97*, 3986–3996.

Goldfarb, M., Schoorlemmer, J., Williams, A., Diwakar, S., Wang, Q., Huang, X., Giza, J., Tchetchik, D., Kelley, K., and Vega, A. (2007). Fibroblast growth factor homologous factors control neuronal excitability through modulation of voltage-gated sodium channels. Neuron *55*, 449–463.

Golding, N.L., Mickus, T.J., Katz, Y., Kath, W.L., and Spruston, N. (2005). Factors mediating powerful voltage attenuation along CA1 pyramidal neuron dendrites. J. Physiol. *568*, 69–82.

Graves, A.R., Moore, S.J., Bloss, E.B., Mensh, B.D., Kath, W.L., and Spruston, N. (2012). Hippocampal pyramidal neurons comprise two distinct cell types that are countermodulated by metabotropic receptors. Neuron *76*, 776–789.

Hajdu, S.I. (2003). Discovery of the Cerebrospinal Fluid. Ann. Clin. Lab. Sci. *33*, 334–336.

Hall, J.E. (2015). Guyton and Hall textbook of medical physiology (Elsevier Health Sciences).

Hernandez, R.V., Navarro, M.M., Rodriguez, W.A., Martinez, J.L., and LeBaron, R.G. (2005). Differences in the magnitude of long-term potentiation produced by theta burst and high frequency stimulation protocols matched in stimulus number. Brain Res. Brain Res. Protoc. *15*, 6–13.

Hille, B. (1984). Ionic Channels of Excitable Membranes (Sunderland, Mass: Macmillan Education Australia).

Hirschberg, J., and Manning, C.D. (2015). Advances in natural language processing. Science *349*, 261–266.

Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. *117*, 500–544.

Isenberg, G. (1976). Cardiac Purkinje fibers: cesium as a tool to block inward rectifying potassium currents. Pfluegers Arch. *365*, 99–106.

Jaderberg, M., Mnih, V., Czarnecki, W.M., Schaul, T., Leibo, J.Z., Silver, D., and Kavukcuoglu, K. (2016). Reinforcement Learning with Unsupervised Auxiliary Tasks. ArXiv161105397 Cs.

Jorgenson, L.A., Newsome, W.T., Anderson, D.J., Bargmann, C.I., Brown, E.N., Deisseroth, K., Donoghue, J.P., Hudson, K.L., Ling, G.S.F., MacLeish, P.R., et al. (2015). The BRAIN

Initiative: developing technology to catalyse neuroscience discovery. Phil Trans R Soc B *370*, 20140164.

Kandel, E.R., Markram, H., Matthews, P.M., Yuste, R., and Koch, C. (2013). Neuroscience thinks big (and collaboratively). Nat. Rev. Neurosci. *14*, 659–664.

Kim, J., and Connors, B. (2012). High temperatures alter physiological properties of pyramidal cells and inhibitory interneurons in hippocampus. Front. Cell. Neurosci. *6*, 27.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. pp. 1137–1145.

Koyama, S., and Appel, S.B. (2006). Characterization of M-Current in Ventral Tegmental Area Dopamine Neurons. J. Neurophysiol. *96*, 535–543.

Lamsa, K., Irvine, E.E., Giese, K.P., and Kullmann, D.M. (2007). NMDA receptor-dependent long-term potentiation in mouse hippocampal interneurons shows a unique dependence on Ca(2+)/calmodulin-dependent kinases. J. Physiol. *584*, 885–894.

Larson, S.D., and Martone, M. (2013). NeuroLex. org: an online framework for neuroscience knowledge. Front. Neuroinformatics *7*, 18.

Le Cessie, S., and Van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. J. R. Stat. Soc. Ser. C Appl. Stat. *41*, 191–201.

Lee, J.C.F., Callaway, J.C., and Foehring, R.C. (2005). Effects of Temperature on Calcium Transients and Ca2+-Dependent Afterhyperpolarizations in Neocortical Pyramidal Neurons. J. Neurophysiol. *93*, 2012–2020.

Liaw, A., and Wiener, M. (2012). Random Forest: Breiman and Cutler's Random Forests for Classification and Regression. R Package Version 4.6-7.

Linoff, G.S., and Berry, M.J.A. (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (John Wiley & Sons).

Lübke, J., Frotscher, M., and Spruston, N. (1998). Specialized electrophysiological properties of anatomically identified neurons in the hilar region of the rat fascia dentata. J. Neurophysiol. *79*, 1518–1534.

MacGregor, D.G., Chesler, M., and Rice, M.E. (2001). HEPES prevents edema in rat brain slices. Neurosci. Lett. *303*, 141–144.

Miller, J.F. (1981). Assessing language production in children: Experimental procedures (Univ Park Press).

Moyer, J.R., and Brown, T.H. (1998). Methods for whole-cell recording from visually preselected neurons of perirhinal cortex in brain slices from young and aging rats. J. Neurosci. Methods *86*, 35–54.

Nassar, M., Simonnet, J., Lofredi, R., Cohen, I., Savary, E., Yanagawa, Y., Miles, R., and Fricker, D. (2015). Diversity and overlap of Parvalbumin and Somatostatin expressing interneurons in mouse presubiculum. Front. Neural Circuits *9*.

Novkovic, T., Shchyglo, O., Gold, R., and Manahan-Vaughan, D. (2015). Hippocampal function is compromised in an animal model of multiple sclerosis. Neuroscience *309*, 100–112.

Perkowski, J.J., and Murphy, G.G. (2011). Deletion of the mouse homolog of KCNAB2, a gene linked to monosomy 1p36, results in associative memory impairments and amygdala hyperexcitability. J. Neurosci. *31*, 46–54.

Pilarski, J.Q., Wakefield, H.E., Fuglevand, A.J., Levine, R.B., and Fregosi, R.F. (2011). Developmental Nicotine Exposure Alters Neurotransmission and Excitability in Hypoglossal Motoneurons. J. Neurophysiol. *105*, 423–433.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., and Jensen, L.J. (2015). DISEASES: Text mining and data integration of disease–gene associations. Methods *74*, 83–89.

Prestori, F., Rossi, P., Bearzatto, B., Lainé, J., Necchi, D., Diwakar, S., Schiffmann, S.N., Axelrad, H., and D'Angelo, E. (2008). Altered neuron excitability and synaptic plasticity in the cerebellar granular layer of juvenile prion protein knock-out mice with impaired motor control. J. Neurosci. *28*, 7091–7103.

Rang, H.P., and Dale, M.M. (2003). Pharmacology (Churchill Livingstone).

Richerson, G.B., and Messer, C. (1995). Effect of composition of experimental solutions on neuronal survival during rat brain slicing. Exp. Neurol. *131*, 133–143.

Routh, V.H., Song, Z., and Liu, X. (2004). The role of glucosensing neurons in the detection of hypoglycemia. Diabetes Technol. Ther. *6*, 413–421.

Savin, N.E. (1984). Multiple hypothesis testing. Handb. Econom. *2*, 827–879.

Scorza, C.A., Araujo, B.H.S., Leite, L.A., Torres, L.B., Otalora, L.F.P., Oliveira, M.S., Garrido-Sanabria, E.R., and Cavalheiro, E.A. (2011). Morphological and electrophysiological properties of pyramidal-like neurons in the stratum oriens of Cornu ammonis 1 and Cornu ammonis 2 area of Proechimys. Neuroscience *177*, 252–268.

Staff, N.P., Jung, H.-Y., Thiagarajan, T., Yao, M., and Spruston, N. (2000). Resting and active properties of pyramidal neurons in subiculum and CA1 of rat hippocampus. J. Neurophysiol. *84*, 2398–2408.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. BMC Bioinformatics *9*, 307.

Suter, B.A., Migliore, M., and Shepherd, G.M.G. (2013). Intrinsic electrophysiology of mouse corticospinal neurons: a class-specific triad of spike-related properties. Cereb. Cortex N. Y. N 1991 *23*, 1965–1977.

Tanaka, Y., Tanaka, Y., Furuta, T., Yanagawa, Y., and Kaneko, T. (2008). The effects of cutting solutions on the viability of GABAergic interneurons in cerebral cortical slices of adult mice. J. Neurosci. Methods *171*, 118–125.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol. *58*, 267–288.

Tripathy, S.J., Savitskaya, J., Burton, S.D., Urban, N.N., and Gerkin, R.C. (2014). NeuroElectro: a window to the world's neuron electrophysiology data. Front. Neuroinformatics *8*, 40.

Tripathy, S.J., Burton, S.D., Geramita, M., Gerkin, R.C., and Urban, N.N. (2015). Brain-wide analysis of electrophysiological diversity yields novel categorization of mammalian neuron types. J. Neurophysiol. jn.00237.2015.

Uzuner, Ö., and Stubbs, A. (2015). Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. J. Biomed. Inform. *58*, S1–S5.

Van Der Walt, S., Colbert, S.C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. *13*, 22–30.

Van Rijsbergen, C. (1979). Information retrieval. dept. of computer science, university of glasgow. URL Citeseer Ist Psu Eduvanrijsbergen79information Html.

Wu, W.W., Chan, C.S., and Disterhoft, J.F. (2004). Slow Afterhyperpolarization Governs the Development of NMDA Receptor–Dependent Afterdepolarization in CA1 Pyramidal Neurons During Synaptic Stimulation. J. Neurophysiol. *92*, 2346.

Yan, X. (2009). Linear regression analysis: theory and computing (World Scientific).

Yang, W., Carrasquillo, Y., Hooks, B.M., Nerbonne, J.M., and Burkhalter, A. (2013). Distinct Balance of Excitation and Inhibition in an Interareal Feedforward and Feedback Circuit of Mouse Visual Cortex. J. Neurosci. *33*, 17373–17384.

Zaitsev, A.V., Povysheva, N.V., Gonzalez-Burgos, G., Rotaru, D., Fish, K.N., Krimer, L.S., and Lewis, D.A. (2009). Interneuron diversity in layers 2-3 of monkey prefrontal cortex. Cereb. Cortex N. Y. N 1991 *19*, 1597–1615.

Zhou, F.-W., Fortin, J.M., Chen, H.-X., Martinez-Diaz, H., Chang, L.-J., Reynolds, B.A., and Roper, S.N. (2015). Functional Integration of Human Neural Precursor Cells in Mouse Cortex. PLOS ONE *10*, e0120281.

Appendices

Appendix A

Figure 17: Principal component analysis of patch-clamp internal experimental solution components (5 major ions). Arrows represent the original ion concentrations on the PC1-PC2 space.

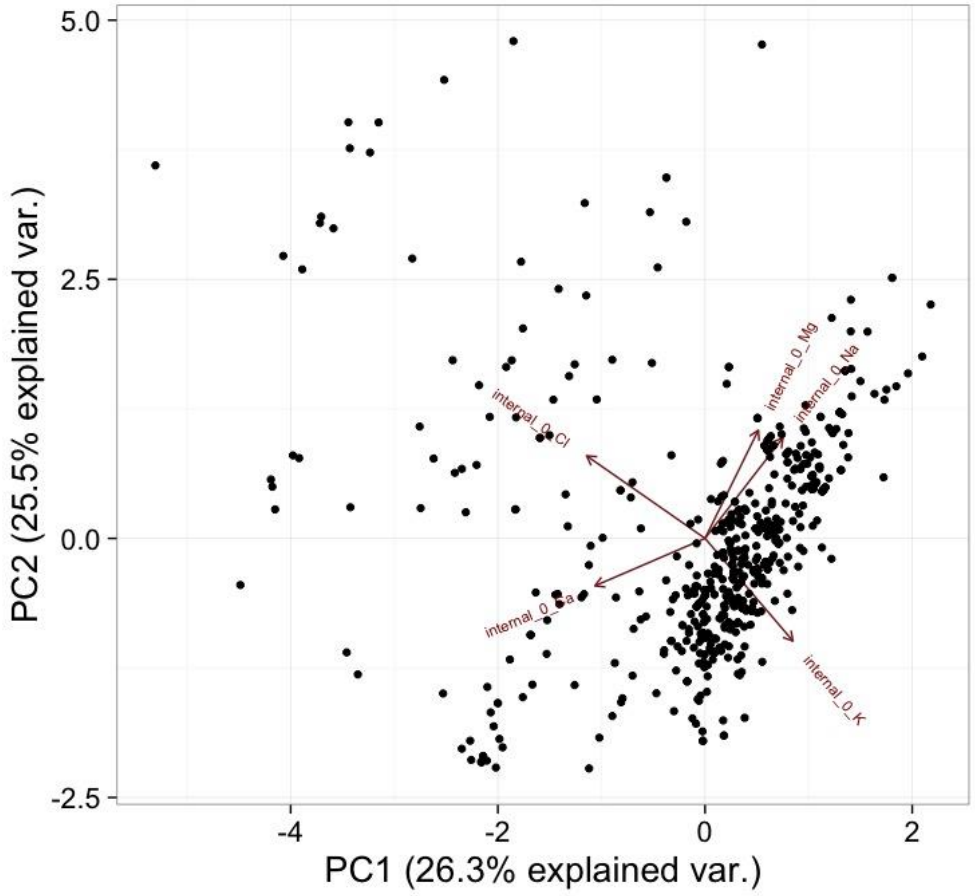


Figure 18: Principal component analysis of patch-clamp extracellular experimental solution components (5 major ions). Arrows represent the original ion concentrations on the PC1-PC2 space.

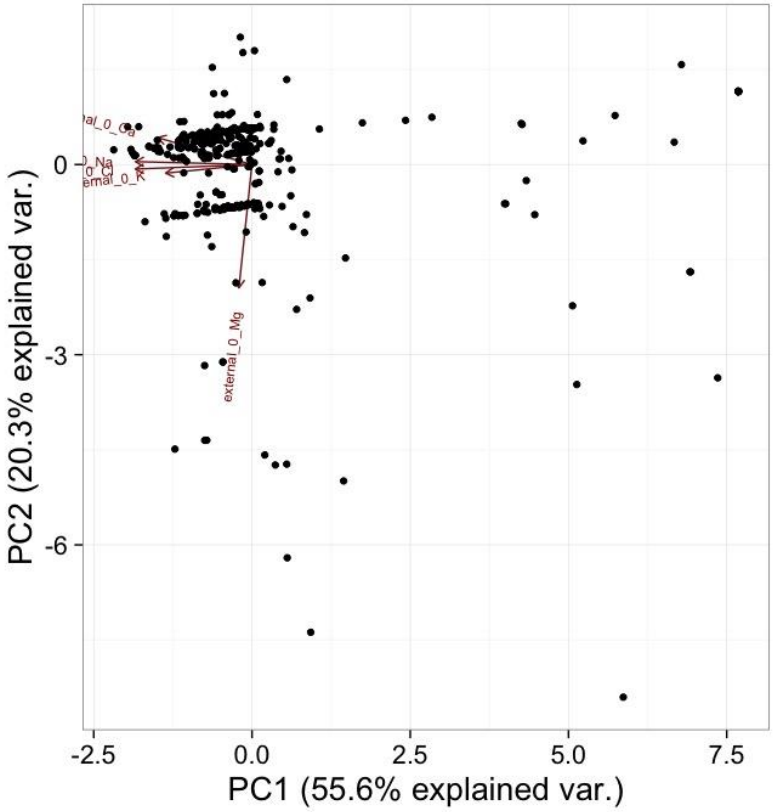
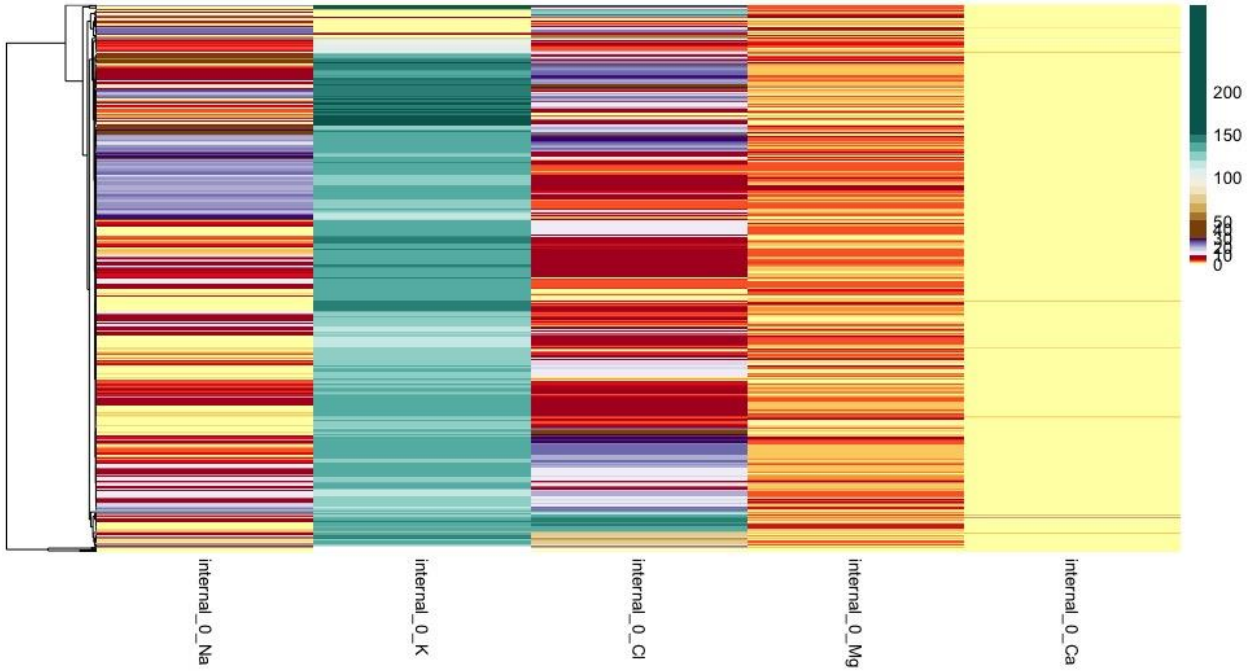


Figure 19: Heatmaps with hierarchical clustering, manual color breaks for concentration differentiations (in mM).



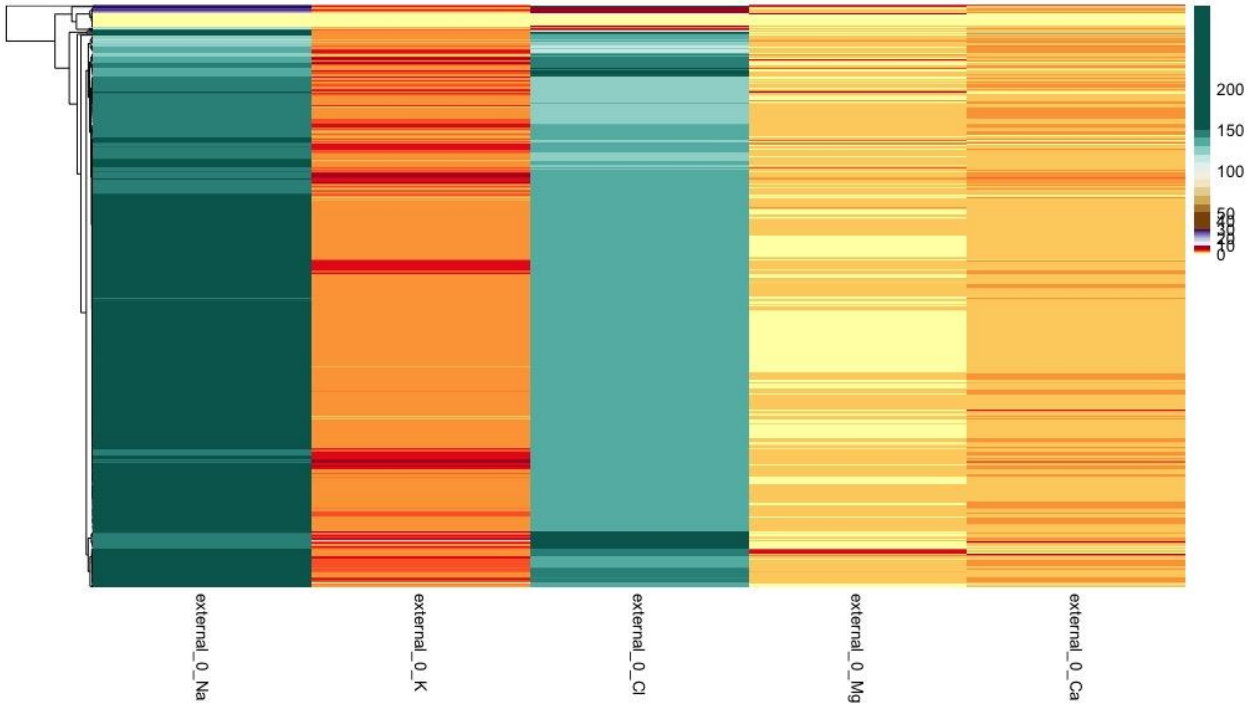
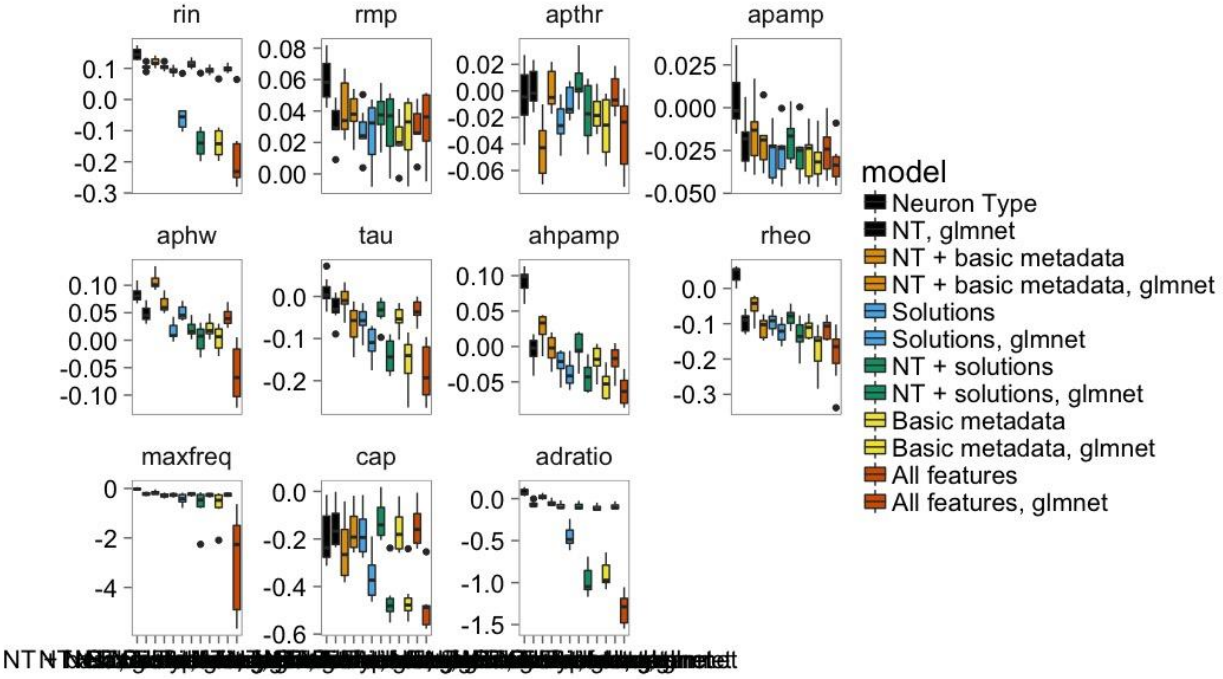
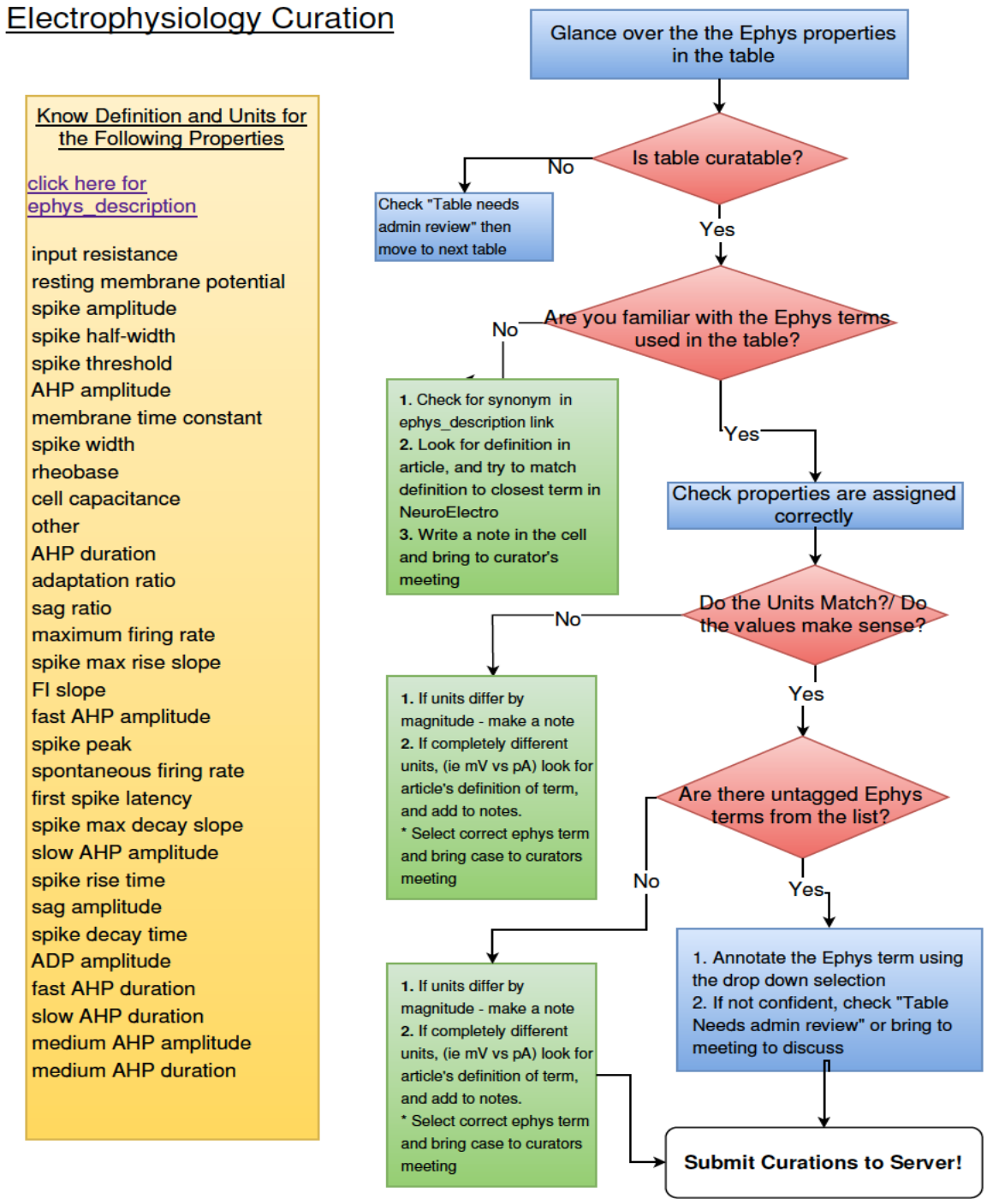


Figure 20: Comparing the performances of SVM and RandomForest models. For each pair of models (colors) – RF is on the left, SVM (glmnet package, version 2.0-5) is on the right. RF's performance is consistently more stable than SVM's.



Appendix B

Figure 21: Detailed NeuroElectro curation protocol



Cell type Curation

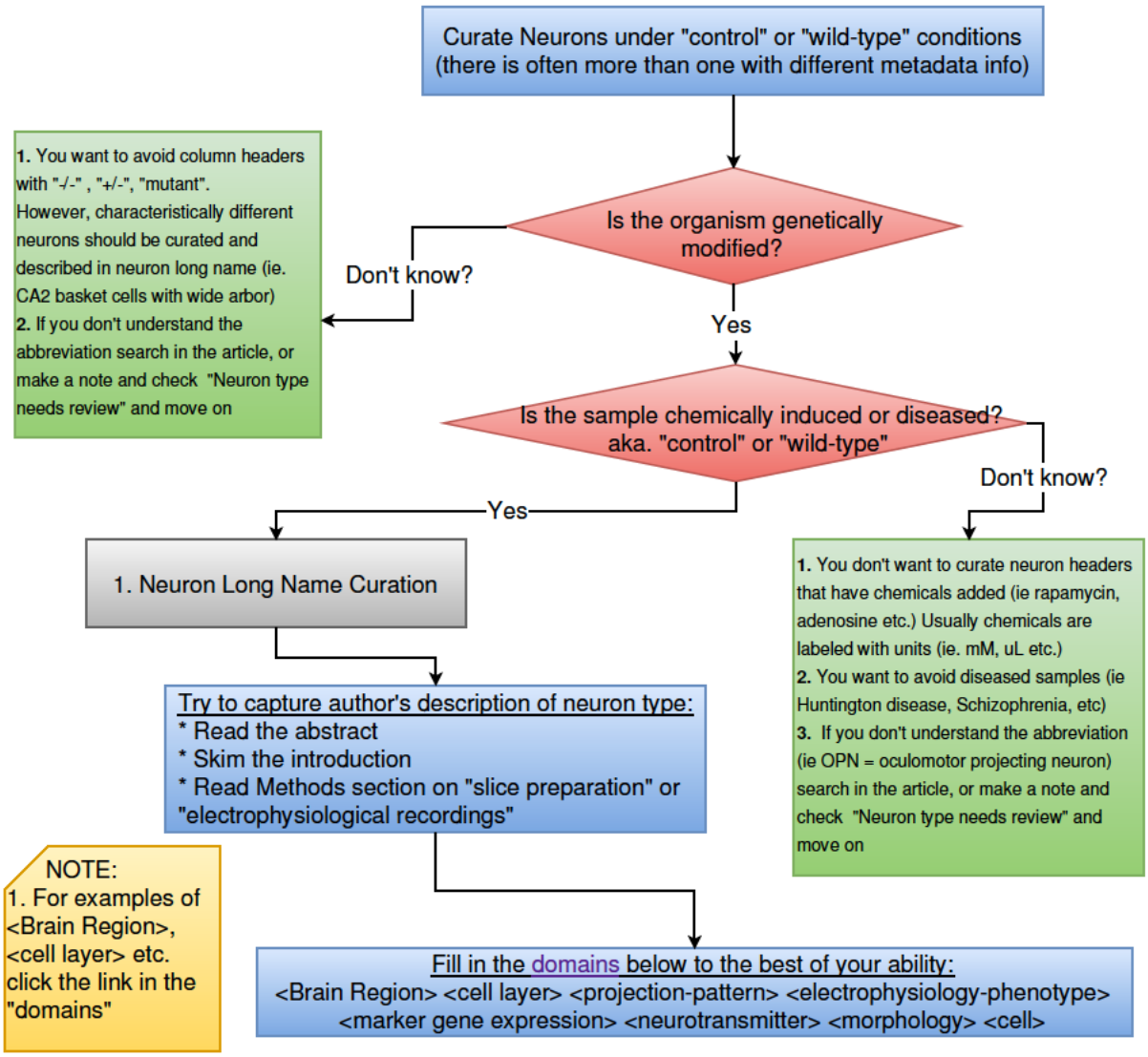
Overall approach

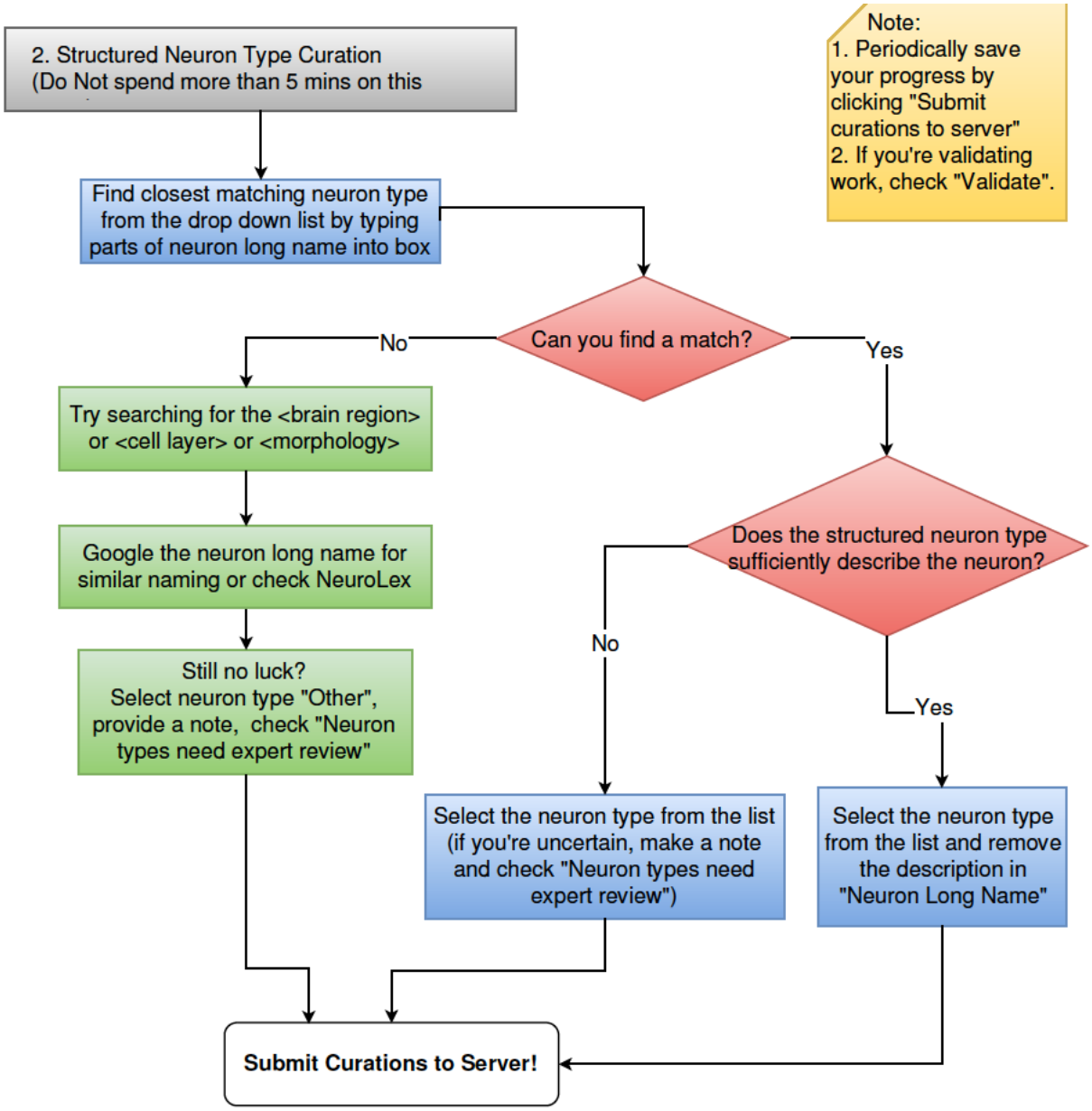
NeuroElectro employs 2 methods for defining neuron types. The first approach is a 'free text' approach, where curators are asked to recapitulate the neuron types described in the manuscript using free text descriptors. The second is a 'structured' approach, where we normalize mentions of neuron types to a predefined list of neuron types provided by NeuroLex.org.

Both methods rely on the idea that to define a neuron type, you need to specify a few things about it, like:

- *what brain region and layer is the neuron's cell body located in?*
- *what is the cell's morphology (shape)?*
- *what is the cell's neurotransmitter?*

Please read examples of neuron type descriptions here




2. Structured Neuron Type Curation
(Do Not spend more than 5 mins on this
Note:
1. Periodically save your progress by clicking "Submit curations to server"
2. If you're validating work, check "Validate".
Find closest matching neuron type from the drop down list by typing parts of neuron long name into box
Can you find a match?
No
Yes
Try searching for the <brain region> or <cell layer> or <morphology>
Google the neuron long name for similar naming or check NeuroLex
Still no luck?
Select neuron type "Other", provide a note, check "Neuron types need expert review"
Does the structured neuron type sufficiently describe the neuron?
No
Yes
Select the neuron type from the list (if you're uncertain, make a note and check "Neuron types need expert review")
Select the neuron type from the list and remove the description in "Neuron Long Name"
Submit Curations to Server!

Metadata Curation

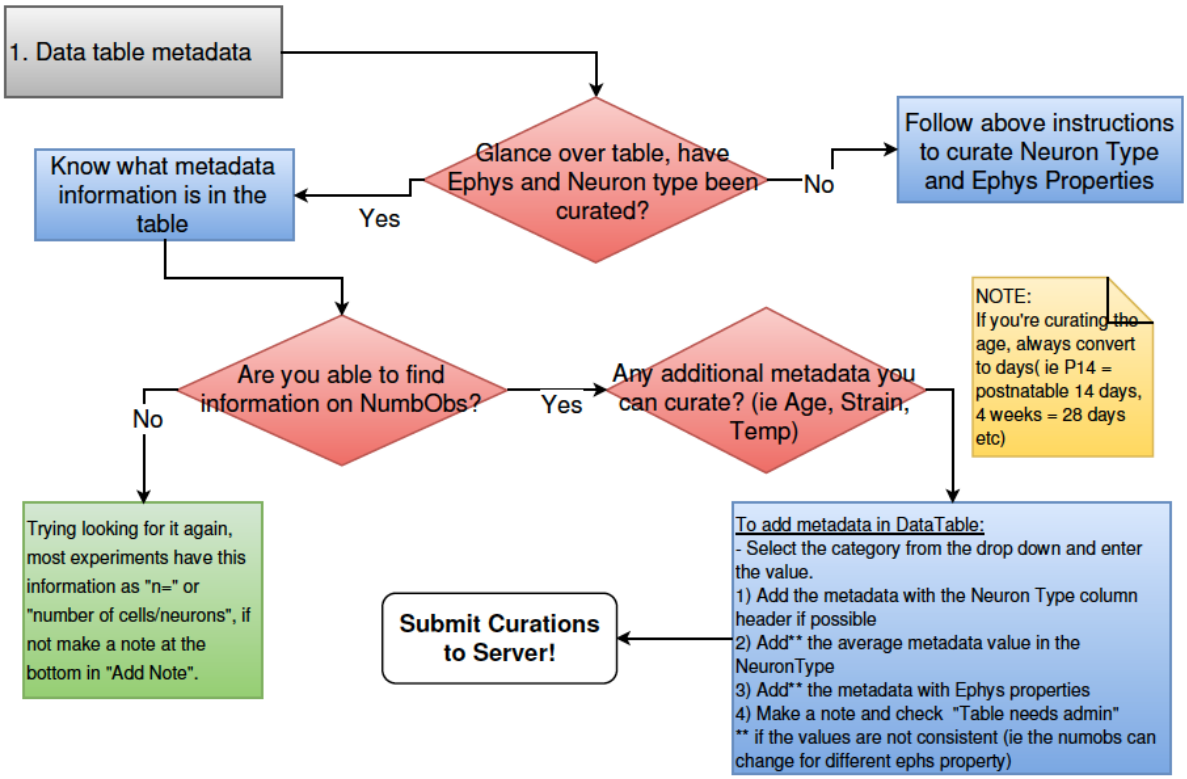
Overall approach

NeuroElectro uses 2 levels of curation to capture the metadata information. The first is in the DataTable and the second is in the 'Article metadata'. The DataTable will have experimental metadata specific to individual columns or rows in the table, whereas, the 'Article metadata' will have information relevant to the entire electrophysiology experiment reported in the table. When the same information can be recorded in both the DataTable and the Article metadata, know that the curation in the DataTable supersedes what's in the Article metadata.

Together, the 2 levels of curation try to capture the following metadata about the experiment:

- *Species*
- *Strain*
- *Electrode Type (patch- clamp, sharp, perforated patch-clamp, extracellular electrode)*
- *Prep Type (in vitro, cell culture, in vivo, model)*
- *NumbObs (number of recorded cells)*
- *Animal Age*
- *Animal Weight*
- *Recording temperature*
- *Junction Potential*
- *Junction Potential Offset*
- *External solution*
- *Internal solution*

Read the full instructions here before you start



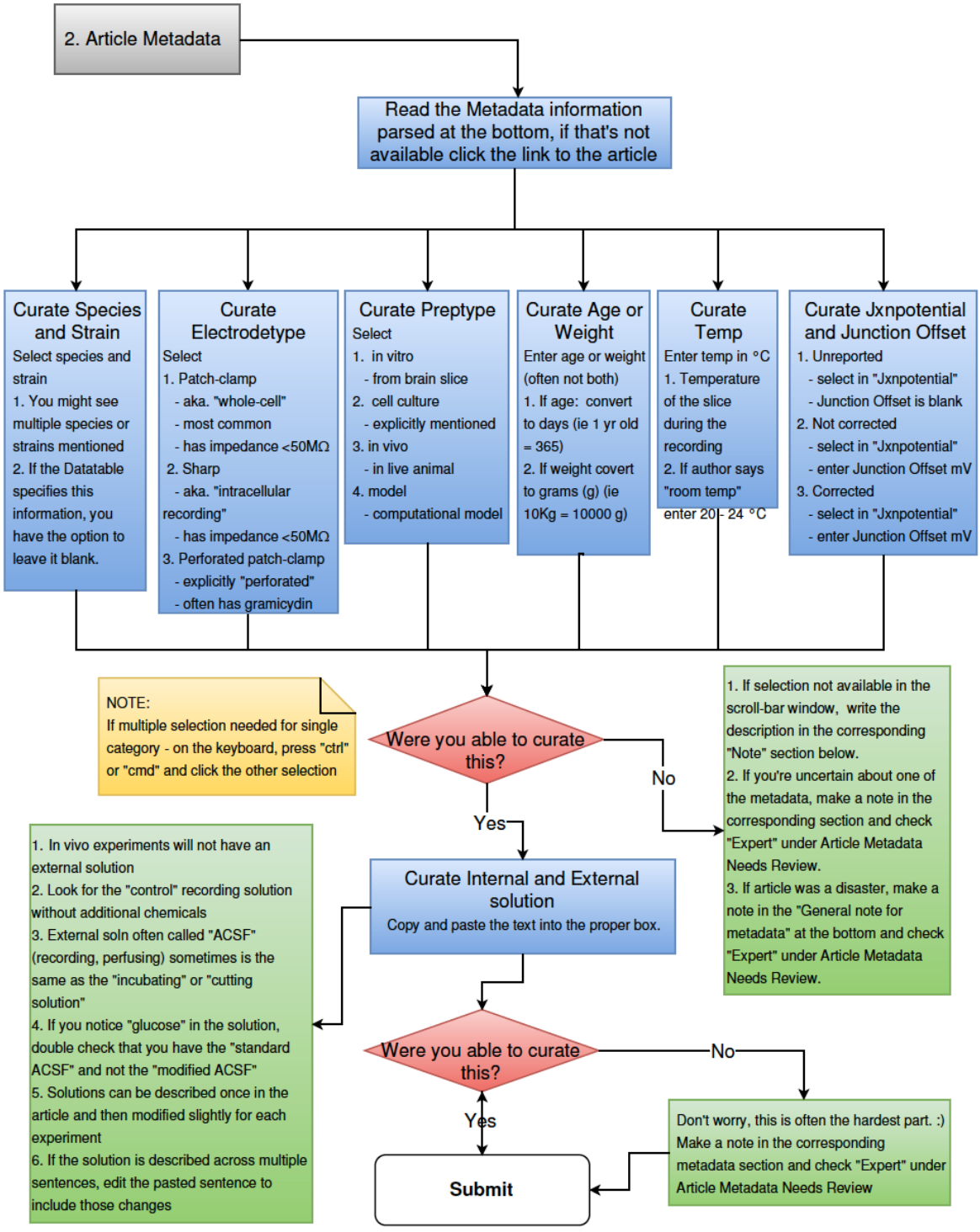

2. Article Metadata
Read the Metadata information parsed at the bottom, if that's not available click the link to the article
Curate Species and Strain
Select species and strain
1. You might see multiple species or strains mentioned
2. If the Datatable specifies this information, you have the option to leave it blank.
Curate Electrodetype
Select
1. Patch-clamp
- aka. "whole-cell"
- most common
- has impedance <50MΩ
2. Sharp
- aka. "intracellular recording"
- has impedance <50MΩ
3. Perforated patch-clamp
- explicitly "perforated"
- often has gramicydin
Curate Preptype
Select
1. in vitro
- from brain slice
2. cell culture
- explicitly mentioned
3. in vivo
- in live animal
4. model
- computational model
Curate Age or Weight
Enter age or weight (often not both)
1. If age: convert to days (ie 1 yr old = 365)
2. If weight covert to grams (g) (ie 10Kg = 10000 g)
Curate Temp
Enter temp in °C
1. Temperature of the slice during the recording
2. If author says "room temp" enter 20 - 24 °C
Curate Jxnpotential and Junction Offset
1. Unreported
- select in "Jxnpotential"
- Junction Offset is blank
2. Not corrected
- select in "Jxnpotential"
- enter Junction Offset mV
3. Corrected
- select in "Jxnpotential"
- enter Junction Offset mV
NOTE:
If multiple selection needed for single category - on the keyboard, press "ctrl" or "cmd" and click the other selection
Were you able to curate this?
No
1. If selection not available in the scroll-bar window, write the description in the corresponding "Note" section below.
2. If you're uncertain about one of the metadata, make a note in the corresponding section and check "Expert" under Article Metadata Needs Review.
3. If article was a disaster, make a note in the "General note for metadata" at the bottom and check "Expert" under Article Metadata Needs Review.
Yes
Curate Internal and External solution
Copy and paste the text into the proper box.
1. In vivo experiments will not have an external solution
2. Look for the "control" recording solution without additional chemicals
3. External soln often called "ACSF" (recording, perfusing) sometimes is the same as the "incubating" or "cutting solution"
4. If you notice "glucose" in the solution, double check that you have the "standard ACSF" and not the "modified ACSF"
5. Solutions can be described once in the article and then modified slightly for each experiment
6. If the solution is described across multiple sentences, edit the pasted sentence to include those changes
Were you able to curate this?
No
Don't worry, this is often the hardest part. :) Make a note in the corresponding metadata section and check "Expert" under Article Metadata Needs Review
Yes
Submit

Figure 22: The new curation interface with staged RMP measurement for 2 neuron types. Input resistance annotation scheduled for deletion (example only, the annotation is right).

Article: *Transition to seizures in the isolated immature mouse hippocampus: a switch from dominant phasic inhibition to dominant phasic excitation.*

Full Text (publisher's website) ; Article Metadata ; Article Data (extracted) ; Full Text (on NeuroElectro)
Derchansky M; Jahromi SS; Mamani M; Shin DS; Sik A; Carlen PL
J. Physiol. (Lond.), 2008

Below is the rendering of the article data table as stored on our server.

Table 1. Electrophysiological characterization of pyramidal, and fast-spiking (FS) and non-FS interneurons

Cell type	Pyramidal Staged concept: Hippocampus CA1 pyramidal cell Staged note: New annotation! NumObs: 53.0	Non-FS O-LM Staged concept: Hippocampus CA1 oriens lacunosum moleculare neuron Staged neuron long name: Hippocampus CA1 oriens lacunosum moleculare non-fast-spiking neuron NumObs: 66.0	FS	
			Trilaminar Concept: Other Neuron long name: Hippocampus CA1 trilaminar fast spiking interneuron Note: Recording recurrent seizures in pyramidal neurons and interneurons in the CA1 region of the intact, isolated mouse hippocampus NumObs: 21.0	Basket Concept: Hippocampus CA1 basket cell Neuron long name: Hippocampus CA1 fast spiking basket cell NumObs: 20.0
n	53 (4)	66 (6)	21 (5)	20 (5)
RMP (mV) Staged concept: resting membrane potential	−68.0 ± 4.0	−62.0 ± 3.0	−62.0 ± 6.0	−63.0 ± 4.0
Input resistance (MΩ) Concept: input resistance Undo annotation removal	260.0 ± 20.0	267.0 ± 20.0*	145.0 ± 10.0	110.0 ± 15.0
Membrane time constant (ms) Concept: membrane	40.0 ± 3.0	32.0 ± 3.0*	18.0 ± 3.0	12.0 ± 4.0