**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

© Dmitry Tebaykin, 2016

Abstract

This is the Abstract.

The abstract is a concise and accurate summary of the research contained in the thesis. It states the problem, the methods of investigation, and the general conclusions, and should not contain tables, graphs or illustrations. It **must not** exceed 350 words, and should contain relevant keywords that will make your thesis more likely to be found in an electronic search. **Do not** put a separate list of keywords. There must be a single abstract for the entire thesis.

Preface

This is the Preface. As of May 2013, **all** theses and dissertations are **required** to have a preface, which appears immediately after the Abstract and before the Table of Contents. The preface **must** include a statement indicating your contribution to the following:

* Identification and design of the research program,
* Performance of various parts of the research, and
* Analysis of the research data.

In addition, if any part of your research was partly or wholly published in articles, was part of a collaboration, or required the approval of one or more of UBC’s Research Ethics Boards, you must include that information in your preface as well. This information would be included as follows:

* A statement indicating the relative contributions of all collaborators and co-authors of publications (if any), emphasizing details of your contribution and stating the proportion of research and writing conducted by you
* A list of any publications arising from work presented in the thesis, and the chapter(s) in which the work is located. **Do not** include the actual publications at any point in your thesis.
* The name of the particular UBC Research Ethics Board, and the Certificate Number(s) of the Ethics Certificate(s) obtained, if ethics approval was required for the research. **Do not** include the actual Ethics Certificate(s) at any point in your thesis. If you are not sure if your research requires ethics approval, please contact the [Office of Research Services](http://www.ors.ubc.ca/home) as early in your research process as possible.

Examples of statements regarding collaborations or lists of publications arising from the research:

Chapter 2 is based on work conducted in UBC’s Maple Syrup Laboratory by Dr. A. Apple, Professor B. Boat, and [student’s name]. I was responsible for tapping the trees in forests X and Z, conducted and/or supervised all boiling operations, and performed frequent quality control tests on the product.

or

A version of chapter 3 has been published. [Student’s name], Apple, A. and Boat. B. (2010) Frequency of Quality Testing in Syrup Creation. Maple Science J. 255:139-144. I conducted all the testing and wrote most of the manuscript. The section on “Testing Implements” was originally drafted by Boat, B.

If your research did not require ethics approval, was conducted as part of a collaboration, was not previously published in whole or in part, and was designed, carried out, and analyzed by you alone, your preface may be very short. The contents of the preface must be verified by your supervisor. Please note that preface requirements may change over time; please check the [Faculty of Graduate Studies website](http://www.grad.ubc.ca/current-students/dissertation-thesis-preparation) to ensure that you are following the most current requirements.

Table of Contents

Abstract ii

Preface iii

Table of Contents v

List of Tables viii

List of Figures ix

Acknowledgements xii

Dedication xiii

Chapter 1: Introduction 14

1.1 Intro section 1 14

1.2 Intro section 2 14

Chapter 2: Extracting solution components from papers via automated text-mining and manual curation 19

1.3 Methods 19

1.1.1 Text-mining and curating electrophysiology-relevant chemical solutions 20

2.1.1 Manual curation 25

3.1.1 Data and code availability 28

4.1.1 Statistical analysis 29

1.1.1.1 Data preprocessing 29

1.1.1.2 Exploration of common solution recipes 31

1.4 Results 33

1.1.1.3 Analysis of experimental solution compositions used by neurophysiologists 37

Chapter 3: Modeling variability of electrophysiological properties with experimental conditions information 45

1.5 Methods 45

Modeling the effects of experimental conditions on the variability in electrophysiological properties 45

Constructing univariate linear models 46

Multiple regression approach 46

Custom model creation 49

Validating custom models 52

1.6 Results 54

1.1.1.1 Data overview 54

1.1.1.2 Assessing within neuron type electrophysiological variability 58

1.1.1.3 Modeling ephys properties with experimental metadata 61

Univariate linear models approach 61

Multiple regression approach 63

1.1.1.4 Custom models for predicting common electrophysiological properties using important experimental conditions 70

Selecting the most predictive experimental conditions per ephys property 70

Validating custom models with NeuroElectro and AIBS data 74

Chapter 4: Discussion and conclusion 78

1.7 Discussion 79

1.1.1.5 Solutions text-mining and curation 79

1.1.1.6 Experimental solution recipe trends 82

1.1.1.7 Modeling electrophysiological variability 84

5.1.1 Future directions 89

6.1.1 Conclusion 91

7.1.1 Closing remarks 91

Chapter 5: Bibliography 92

Chapter 6: Appendices 94

List of Tables

Table 1: Summary of data stored in NeuroElectro database. Color highlights: green – top 11 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 41

List of Figures

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. 24

Figure 2: 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. 47

Figure 3: 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. 55

1. Figure 4: Internal sodium concentration increases with time. 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). 59

2. Figure 5: 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) Resting membrane potential linear fit is driven by 3 outliers in the 7-7.5 mM range of external magnesium concentration. 60

3. Figure 6: 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. 62

Figure 7: 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. 65

Figure 8: 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 R2 values of 10 runs for that model. The number of data rows per property decreases from left to right. 66

4. Figure 9: Multivariate model performance improves with N. R2 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. 68

5. Figure 10: Model comparison with AICc score. One run of 10-fold cross-validation, each line is an AICc curve calculated by adding top X (from 1 to 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 performance (X-axis). <Numbers will be replaced with feature names, 12 features that represent synaptic and channel blockers will be removed> 70

Figure 11: Feature importance, based on the frequency of inclusion into the custom models. Ephys properties and metadata features are listed vertically and horizontally, respectively. Color represents number of times the feature has been chosen for the ephys property’s model. The heatmap shows how often a metadata feature gets chosen for the best model per EP, from 0 to 100 times. 72

6. Figure 12: Comparison of custom models to basic models. <Better labels, bigger text, highlight best model in the legend on the right. This data is on the server> 74

7. Figure 13: Adjusting NeuroElectro data to AIBS conditions. Violin plots of NeuroElectro data (blue), Allen Institute for Brain Science data (red), adjusted NeuroElectro data using the custom models designed above (green). Each dot is a mean ephys property value reported by an experiment. The model correction tends to squeeze NeuroElectro data around the mean and bring it closer to AIBS value. <Zoom in on APhw plot, add units to each EP> 76

8. Figure 14: Reported action potential amplitudes of CA1 pyramidal cells vary with time. Each point is a population mean APamp value for Hippocampus CA1 pyramidal cells, reported by an article published in the corresponding year. Violins outline the distributions of values for each year. 87

Acknowledgements

This is the Acknowledgements section.

The following are usually mentioned in the Acknowledgements:

* Supervisor and committee
* Grant support
* Helpful fellow students, lab mates, etc.
* Family support

Acknowledgements may extend for more than one page, but should be no longer than two pages.

Example:

I offer my enduring gratitude to the faculty, staff and my fellow students at UBC, who have inspired me to continue my work in this field. I owe particular thanks to Dr. J. M. Schneider, whose penetrating questions taught me to question more deeply.

I thank Dr. O. Meyer for enlarging my vision of science and providing coherent answers to my endless questions.

Special thanks are owed to my parents, whose have supported me throughout my years of education, both morally and financially.

Dedication

This is the Dedication.

The dedication must be listed in the Table of Contents. You may decide you do not want to have the word “Dedication” appear on this page. In Microsoft Word, to get “Dedication” to show up in the Table of Contents without appearing here, change the colour of the word “Dedication” to white. You will then need to change the font colour of “Dedication” back to black in the Table of Contents.

# 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. For example, the first goal of the US NIH BRAIN project is to generate a “census of cell types”, involving neuron comparison using genetic, morphological and electrophysiological characteristics. However, the common practice among neurophysiologists is to only analyze data that they have collected themselves. This is largely because it is colloquially accepted that subtle variation in experimental conditions introduces certain variation into the corresponding measurements. However, this approach imposes sample size restrictions, since a single scientist or lab can only collect and analyze a limited amount of data on their own. The goal of my project is to enable the comparison and aggregation of electrophysiological data across different experiments.

## Intro section 1

## Intro section 2

To provide some context for the methodology used in intracellular electrophysiology, a typical 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) solution at a constant temperature. Finally, a recording electrode is inserted inside 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.

Combining and comparing electrophysiology data across labs directly and on a large scale is, perhaps, questionable because such data is often collected under different experimental conditions. 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 Ca2+ concentrations (Aivar et al. 2014) result in changes in electrophysiological properties of neurons. Therefore, comparing data across differently designed experiments without accounting for variability introduced by experimental conditions could lead to incorrect or inconsistent results.

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 ions inside relative to the outside. This causes sodium (ENa) / potassium (EK) reversal potentials, calculated using the Nernst equation (Schmidt-Nielsen, pp. 478-480), 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. During the neuron action potential (AP), sodium ion permeability (GNa) across the cell membrane increases dramatically, allowing Na+ ions to flood inside the neuron due to the driving force of sodium. As the AP approaches its peak, both sodium driving force and permeability decrease, but potassium permeability (GK) and driving force rise, causing it to flood outside of the neuron, eventually restoring membrane potential to its resting state. Thus, neuron electrophysiological properties depend heavily on the precise ion concentrations inside and outside of the cell.

Many neurophysiologists address the task of exploring the effects of experimental conditions on neuron ephys properties using experimental electrophysiology techniques (Kim et al. 2012, Armentia et al. 2004, Lee et al. 2004). 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.

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. Using a large-scale meta-analysis method, he showed that animal age, recording temperature, electrode type choices significantly correlate with the study-to-study variance in reported ephys values (Tripathy et al. 2015)

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. Once the data was collected, 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 multivariate approach. Additionally, I explored the external and internal solution recipes commonly used by electrophysiologists.

I found the effect of solution compositions on the variance in electrophysiological properties to be relatively small, likely because different labs use similar solutions, thus their explanatory power is limited. Additionally, using experimental conditions (neuron type, recording temperature, animal age, species, solution compositions, etc.) I created custom models for ephys properties commonly reported by neurophysiologists. My models can be used to remove a portion of the ephys variance when comparing results from different experiments, making such comparisons more reliable. Further applications of the models include normalization of ephys values and adjusting them from one set of experimental conditions to another. To validate and showcase the last scenario, 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.

# Extracting solution components from papers via automated text-mining and manual curation

## 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 et al. 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 custom 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.

## Text-mining and curating electrophysiology-relevant chemical solutions

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).



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 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.

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; CaCl2, 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 “CaCl2, 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 CaCl2” would be parsed as 2 mM of Ca2+ and 4 mM of Cl-), even if the compound is fully spelled out (disodium sulfate instead of Na2SO4 or sodium creatine instead of Na2-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 F1-score metrics (Rijsbergen, 1979).

<CITE libraries, re-write paragraph>

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), RE (regular expressions), Numpy (scientific computing methods), FuzzyWuzzy (partial String matching). 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).

### Manual curation

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.

### 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>.

### Statistical analysis

#### 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 different solutions that are in contact with each other. Unfortunately, I discovered that only 46% of articles report junction potential value and 24% correct for it. Thus, I decided that it would be simpler 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.

#### 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) (Orna & Stock, 1989). Sodium example:

(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 (Appendix A). 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.

## Results

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

**Extracting experimental solutions from methods sections of neurophysiology articles stored in NeuroElectro via text-mining and curation**

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 consists of two parts: first it identifies sentences of methods sections from neurophysiology articles stored in NeuroElectro that contain extracellular and intracellular solutions. The entire sentences get stored in the NeuroElectro database for curation by trained NeuroElectro curators. Secondly, the algorithm extracts concentration values of major ions (Na, K, Cl, Ca, Mg) and other common compounds (HEPES, glucose, EGTA, EDTA, BAPTA, ATP, GTP, Cs). 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 et al. 1991). It is important to note that curation protocol involves verification of external and internal solution sentences, but not the correctness of concentration extraction.

For the 100 articles mentioned above, I evaluated each component in the solutions data acquisition pipeline. I considered: 1) accuracy of identifying sentences corresponding to external and internal solutions; 2) accuracy of trained NeuroElectro curators; 3) accuracy of major ions and other compounds concentration extraction.

First, I evaluated the accuracy of correct identification of solution-containing sentences that were used in the ephys recordings (**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. 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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Precision** | **Recall** | **F1 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 2: 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.

had their type assigned as “external”. I used F1 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.

I have 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 sentences that mention multiple solutions at the same time: “Electrophysiology Patch electrodes were … filled with two internal solutions consisting of the following (in mM): 1) 140 KMeSO4, 10 KCl, 10 HEPES, 4 Mg2ATP, and 0.4 Na3GTP or 2) 130 KMeSO4, 10 KCl, 10 HEPES, 10 BAPTA, 4 Mg2ATP, and 0.4 Na3GTP.” (Wu et al. 2004). It is extremely difficult for the algorithm to separate the two internal solutions listed in one sentence, generally the sentence gets parsed as a single internal solution, effectively doubling several chemical concentration values. Only the first solution should be identified as internal for this article, because it was the one used to record ephys properties under control conditions.

The **Compound concentration extraction** algorithm had difficulties correctly parsing inconsistently listed solutions. 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 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.

#### Analysis of experimental solution compositions used by neurophysiologists

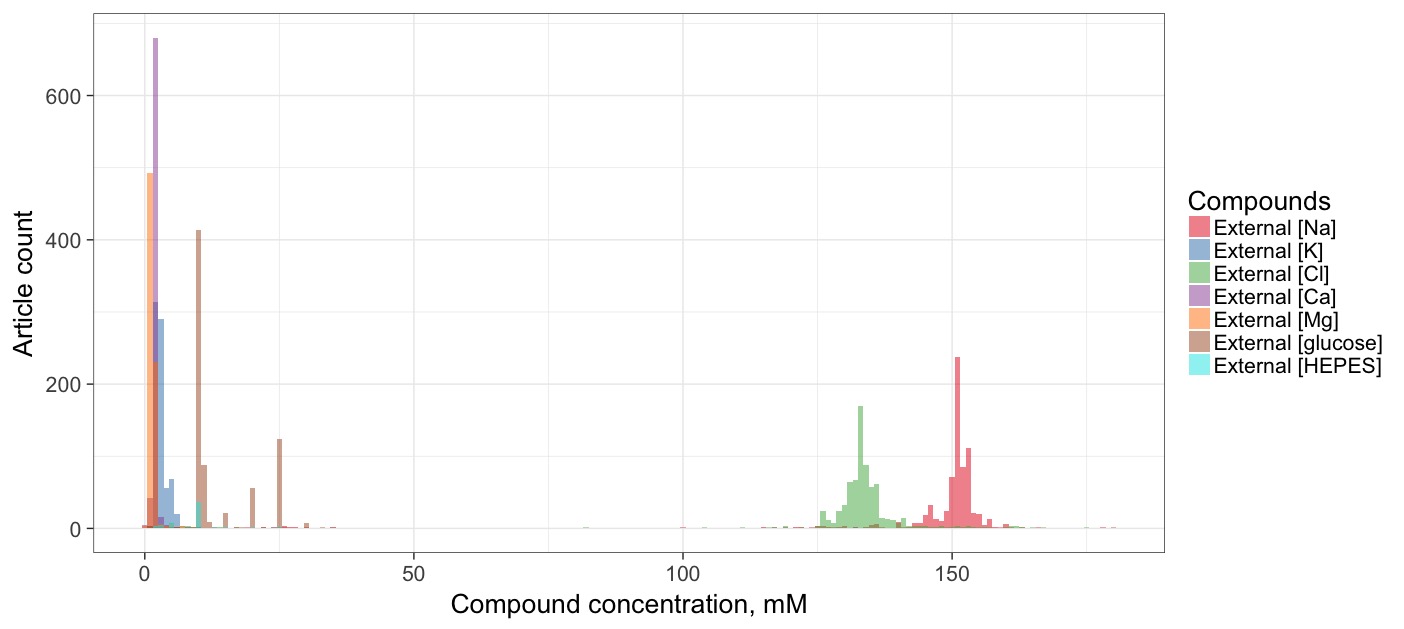
The first step to understanding the effect of solutions on electrophysiological variance is determining the magnitude of variance within solutions themselves. If most labs use very similarly designed artificial cerebrospinal fluids (ACSFs) and pipette solutions, it would be difficult to explain the existing ephys variance with solutions differences. Initially, my approach was to extract the solution constituents that are consistently present and are known to contribute to electrophysiological processes within neurons [CITE textbook, H&H]. 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 millimoles and 141 millimoles of chloride are extracted from the following recording solution: “130 mm NaCl, 3 mm KCl, 1.25 mm NaH2PO4, 26 mm NaHCO3, 2 mm MgCl2, 2 mm CaCl2 and 10 mm glucose oxygenated with 95% O2/5% CO2, pH 7.2–7.4, 290–310 mOsm.” (André et al., 2010). Figure 4 shows the distributions of concentrations of these major ions and other common compounds from curated articles stored in NeuroElectro.

Because electrophysiologists generally mimic their extracellular solutions after cerebrospinal fluid, similar major ion and common compounds concentrations tend to be used throughout the neuron electrophysiology community. Additionally, external solution recipes are similar between patch-clamp and sharp electrodes. 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 (observational data, I am not extracting acetate or methylsulfate concentration values).

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 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 10% of the recipes it was present at a concentration of 10 mM.

Surprisingly, concentration values of intracellular solutions do not follow normal distributions. 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, though they are not used by the same recipes. Current-clamp ephys recordings use potassium-based solutions, while voltage-clamp recordings utilize cesium-based recipes. 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.



B

A

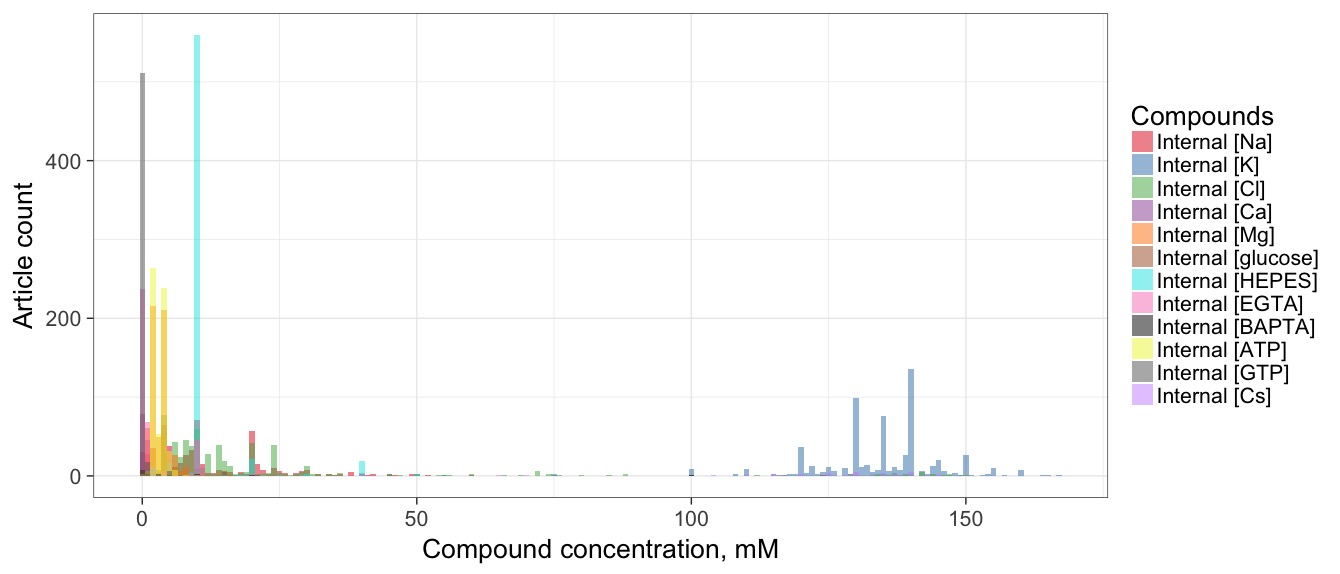


Figure 4: Chemical compositions of experimental solutions. Data from 731 curated Patch-clamp solutions. Histograms of compounds that are commonly found in: A) External (extracellular, ACSF) and B) internal (pipette, electrode) solutions. The ion concentrations were calculated by summing concentrations 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.

I then quantified the frequency of identical solution recipes usage by different articles. Out of the 731 curated Patch-clamp electrode articles, when looking only at the five major ions: 47 recipes for extracellular solutions were used 2 times, 14 – 3 times, 12 – 4 times, 7 – 5 times and certain specific recipes were used 6 or more times. 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 NaHCO3, 1.25 mM NaH2PO4, 1 mM MgCl2, 2 mM CaCl2, and 25 mM dextrose” (Lübke et al. 1998, Staff et al. 2000, Cooper et al. 2003, Golding et al. 2005, Graves et al. 2012, Cembrowski et al. 2016).

In contrast, intracellular solutions are much more diverse: 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). The most common recipes for both solutions are *in vivo* experiments (no extracellular solution) that used 1 M of K (8 articles), 2 M of K (7 articles) and 3 M of K (4 articles).

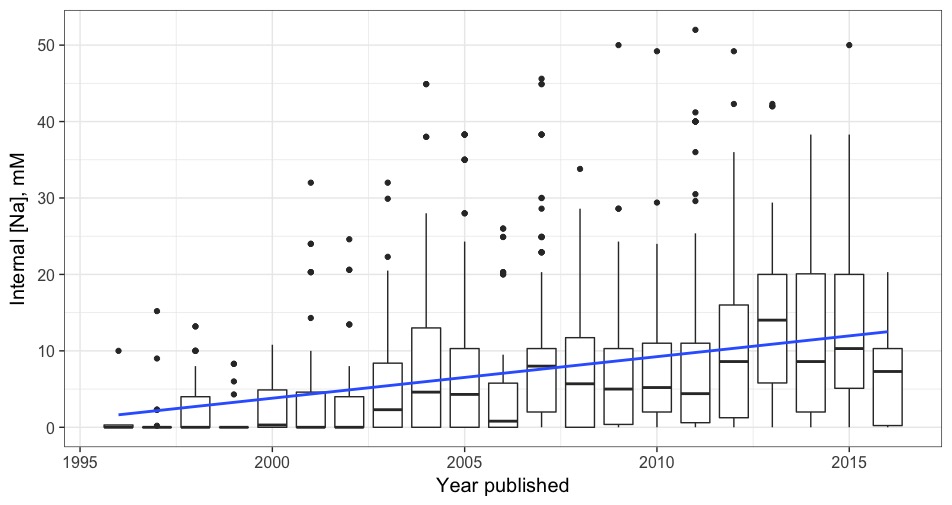
To summarize, identical solution recipes tend not to be re-used from paper to paper, percentage of unique solutions per electrode and solution type: Patch-clamp external – 49%, Patch-clamp internal – 66%, Sharp external – 63% and Sharp internal – 48%. I used principal component analysis supplemented by hierarchical clustering to identify trends in recipe creation for recording and pipette solutions. No obviously distinct clusters presented themselves meaning that electrophysiologists use similar recipes with slight variations, within biologically reasonable concentration values. However, two trends were identifiable among the Patch-clamp solution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chemical** | **Electrode Solution, mM** | | **Recording 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 3: 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 specific solution composition.

recipes: internal solutions could be separated into those with low Na, Cl concentrations and high Na, Cl (Table 3); 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 due to the nature of electrophysiology properties it stores.

An interesting variation in Patch-clamp pipette solution recipes is internal sodium concentrations increase throughout the years (Figure 6). It seems to be caused by the introduction of 10-20 millimoles of Na2-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.



r = 0.29

p < 0.0001

Figure 5: Internal sodium concentration increases with time. 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).

After exploring the distributions of solution components and verifying that neurophysiologists tend to use 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.

# Modeling variability of electrophysiological properties with experimental conditions information

## Methods

### 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/](http://neuroelectro.org/ephys_prop/index/).

#### 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; Xin, 2009). The built-in R function for linear models were 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 (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, 1959 and 1961).

#### 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).

(2)

where Vm denotes resting membrane potential and P denotes ion permeability across the cell membrane (Junge, 1981). Ionic permeabilities were approximated with the default text-book values: (Hille, 2001).

Before delving into multiple regression model selection, I established the training and testing datasets, because testing a model on 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.

Next, I examined the benefits and drawbacks of several commonly used supervised multiple regression models: K-nearest neighbor, Neural Networks, Support Vector Machines, Random Forest. 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. Neural Networks, specifically deep learning, is a powerful regression algorithm. It is probably one of the most talked about regression algorithms, it is being used by Google for image/voice recognition and interpreting streams of sensorimotor data (Jaderberg, 2016). Unfortunately, neural networks require huge amounts of data when compared to the number of features (Linoff & Berry, 2011). Since that is not the case with my dataset, I must 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 & Vapnik, 1995). I compared the performance of SVMs to my models of choice and showed their relative instabilities for some folds of the 10-fold cross-validation testing (Appendix A, Figure 3). 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, 2012). I use 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 has 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.

#### Custom model creation

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. 2009). However, cforest performs slightly worse when used for regression modeling than randomForest, because that task is not fully optimized yet (Hothorn, 2016).

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).

(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 (Burnham & Anderson, 2002). This is particularly important for less popular ephys properties. AICc adds a correction term for limited datasets (4) to the Akaike Information Criterion.

(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 (Burnham & Anderson, 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 custom 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).

(5)

where 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).

(6)

Models with the lowest AICc should have the optimal performance when predicting the ephys property. The optimal model selection procedure was performed 10 times using 10-fold cross-validation (100 total runs) to ensure the robustness of the results. Finally, features that were chosen at least 90% of the time were included into the custom models for each ephys property.

### Validating custom models

To validate the performance of custom models, I compared them to the previously created models that use distinct feature sets: neuron type only, basic metadata, solutions, and their combinations. As before, 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 (unpublished data). The experimental conditions baseline shifting formula was developed by Shreejoy Tripathy (7).

(7)

where NEadj denotes the shifted NeuroElectro ephys values, NEobs – default ephys values stored in NeuroElectro, NEpred – ephys values from NeuroElectro, predicted with the custom models, trained using 10-fold cross-validation without seeing the values of the articles they are trying to predict in this step, and AIBSshift is the ephys values predicted with NeuroElectro custom models, using AIBS experimental conditions. The observed and adjusted ephys values from NeuroElectro were then compared to the reported AIBS ephys values.

## 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 custom 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.

### Data 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 (Rin, 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 |
| CellCapacitance (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 – top 11 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. Each data entry in NeuroElectro is annotated with one of 120 neuron types (Table 1). Neuron types are defined by the NeuroElectro extended dictionary of neuron types that was 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: Rin, 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 (Perkowski et al. 2011, Cui et al. 2011) or AP threshold (Novkovic et al. 2015, Boehlen et al. 2013). 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 (Lamsa et al. 2007, Derchansky et al. 2008). 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).

### Assessing within neuron type 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 [**ADD P-VALUE**]. 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 conditions must be false. These inter-study differences must be partially due to differences in experimental procedures.

../../Neuroelectro%20documents/Plots/exampleRMP.pdf

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.

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 potential explanations for inter-study variance.

### Modeling ephys properties with experimental metadata

#### Univariate linear models approach

B

A

../../Neuroelectro%20documents/Plots/rinVsNaInt.pdf../../Neuroelectro%20documents/Plots/rmpVsMgExt.pdf

r = 0.35

p < 0.001

r = -0.16

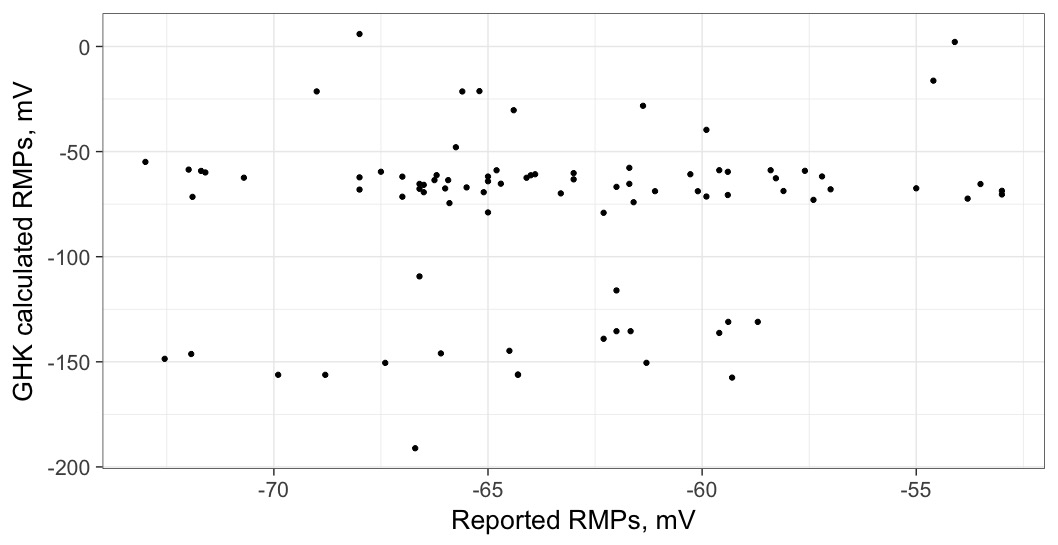
p > 0.05

Figure 8: 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) Resting membrane potential linear fit is driven by 3 outliers in the 7-7.5 mM range of external magnesium concentration.

My initial approach was to identify the effect of major ion concentrations on electrophysiological properties one at a time, reasoning that strong correlations between specific solution components and ephys properties would reveal themselves using this approach. Figure 7 illustrates two examples of this univariate approach, when applied to hippocampus CA1 pyramidal neuron type: input resistance significantly correlates with internal sodium concentration, resting membrane potential relationship with external magnesium is driven by outliers. The univariate approach allows me to detect significant relationships between ephys properties and single ion or compound concentrations on per cell type basis. However, I am interested in general effects of solutions metadata on ephys properties that transcend cell types.

Extending univariate linear models to incorporate all neuron types, and after applying multiple testing correction, I did not find significant relationships between ephys data and individual ion or compound concentrations. 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.

#### Multiple regression approach



R2 = -0.06

Figure 9: 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.

First, I evaluated the performance of the GHK equation for modeling membrane potentials of neurons at rest by comparing its predictions to the observed ephys values. Strikingly, the R2 value was essentially 0, implying that the GHK equation cannot reliably predict the reported RMPs of hippocampal CA1 pyramidal neurons. A possible explanation is that using generic membrane permeability values did not provide a reasonable substitution for the actual Na, K and Cl permeability values of CA1 neurons. Therefore, I optimized these membrane permeability values with respect to the R2 value. However, it is unlikely that slightly adjusted permeability values can correct the differences between the predicted and the observed RMP values.

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. The models were designed to capture the relative impact of various combinations of: neuron names, solution composition information and basic metadata like species, age, temperature, electrode type, etc. (Table 1).

Briefly, I chose Random Forest over the previously used linear regression approach because it is a non-linear model that empirically better handles statistical overfitting (i.e., the fit model does not generalize well to unseen test data) when using datasets with many features relative to sample size (CITE RF textbook). All models were trained and tested using 10-fold cross-validation, where models were trained on a randomly chosen 90% of the articles that have reported the ephys property and then tested on the remaining unseen 10%, the procedure was repeated 10 times. From the predictions made by the model using the unseen testing set of articles and their respective reported ephys values, I calculated an R2 value, which serves as a measure of the model’s performance. As a point of reference, an R2 value of 1 means that the model was 100% correct in all predictions. An R2 value of 0 means that the predictions are as accurate as using the average value of the observed ephys properties for the prediction. A negative R2 means that the model performs worse than the mean because of overfitting to the training data.

To compare the effect of solutions metadata to basic metadata when modeling the variability in ephys properties, I designed several initial 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 Random Forest algorithm to my data, I used the initial models that related metadata features to input resistance (Figure 9). The Random Forest algorithm was given several hundred articles to learn the relationships between experimental conditions (features) and resulting ephys measurement (Rin), thus creating a model. Then, the predicted values are the model’s best estimates of what the observed values should be given the features from each observed 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 9B, refer to color scheme in table 1 for metadata groups). 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 data from the same 10 folds, there is no reshuffling of data between different models. Judging by the model performances in Figure 9B, solution features help to predict input resistance. All features model and the neuron type + solution features perform very similarly and better than neuron type + basic metadata. Interestingly, solutions on their own perform worse

B

A



R2, model performance

Figure 10: 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.

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 solutions features.

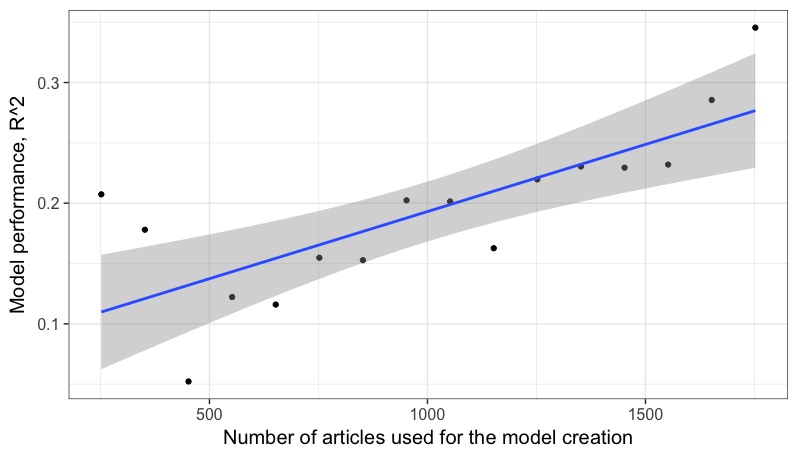
Expanding input resistance modeling to 11 commonly reported ephys properties, I evaluated the effectiveness of each model type in predicting them (Figure 10). The baseline is generated by



Figure 11: 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 R2 values of 10 runs for that model. The number of data rows per property decreases from left to right.

randomly shuffling ephys values, so the model should not be able to learn anything useful. It serves as a lower bound for the worst predictions that could be made when the model is essentially predicting noise. The rest of Figure 10 compares six types of models and their predictive power of the common ephys properties. Only for the properties with a lot of data (input resistance and resting membrane potential) models can explain some of the variance, however, in most cases the models are only slightly better 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 train the multivariate models. I observe a general increase of 0.2-0.5 in the predictive power of my models when comparing the reshuffled resultswith the R2 values reported by our models. On average, solutions contribute less to the overall model predictive power than neuron name. However, adding solutions information to other metadata in some cases increases model’s performance (Rin, APthr, APamp), thus their contributions include different information than neuron name or basic metadata.

Next, 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 R2 values of a model that predicts it (Figure 11). 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.



r = 0.76

p < 0.001

Figure 12: Multivariate model performance improves with N. R2 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.

### Custom models for predicting common electrophysiological properties using important experimental conditions

##### Selecting the most predictive experimental conditions per ephys property

The final step of my project was to create custom models out of only the best features, possibly 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 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. 2009) 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.

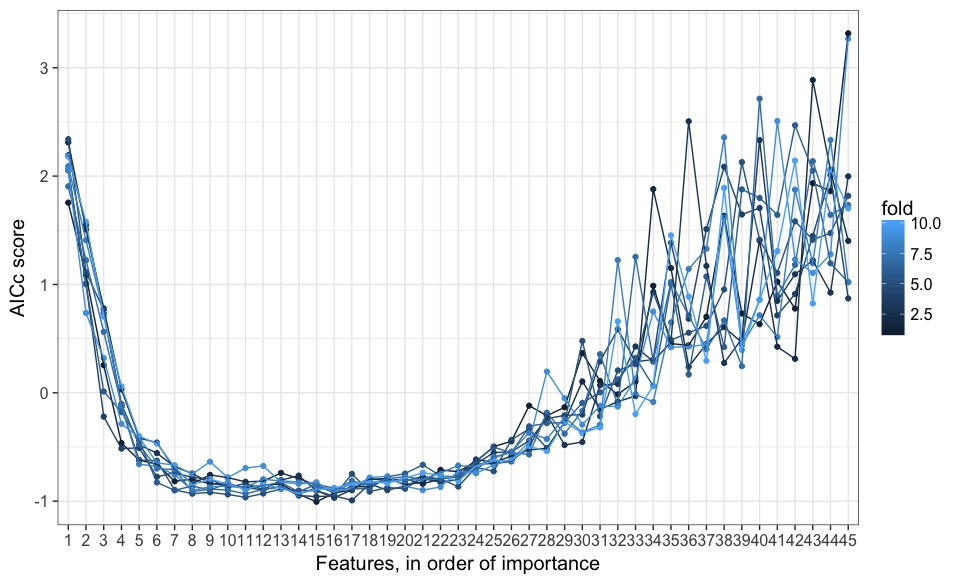


Figure 13: Model comparison with AICc score. One run of 10-fold cross-validation, each line is an AICc curve calculated by adding top X (from 1 to 33) features to the model that predicts input resistance. Model with the lowest AICc score is the best performing one. Metadata features are ordered from high to low based on their performance (X-axis). <Numbers will be replaced with feature names, 12 features that represent synaptic and channel blockers will be removed>

To choose the best model for input resistance I split up the data into 90% and 10% portions. The larger portion of the data was used to rank all features by their importance for predicting input resistance. 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 Rin. 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 (Figure 12). 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. Here, I use it to evaluate input resistance models with regards to each other. The lower AICc is, the better the model. Adding 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 the 23d would hurt the best 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 10 features (23d through 33d) illustrate the amount of instability and noise bad features can add to a model (the effect of overfitting).

To generalize the above Random Forest variable importance ranking complimented by AICc approach for model selection, I applied the same algorithm to the 11 of the most abundant electrophysiological properties. 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.

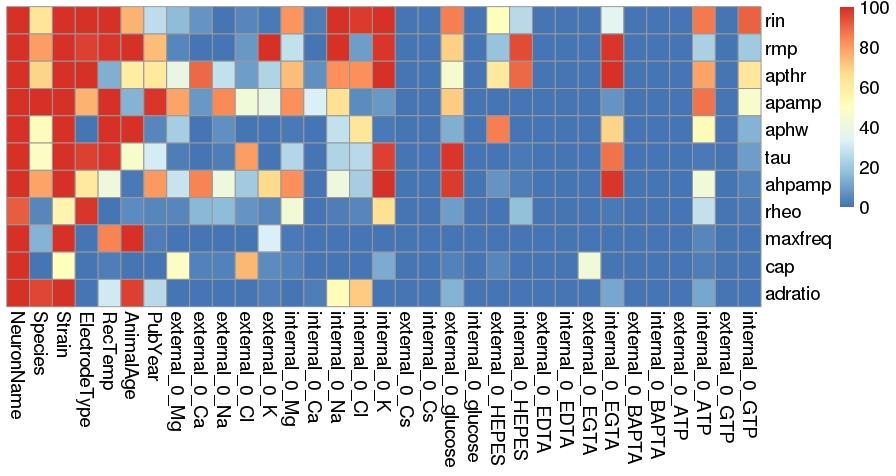


Figure 14: Feature importance, based on the frequency of inclusion into the custom models. Ephys properties and metadata features are listed vertically and horizontally, respectively. Color represents number of times the feature has been chosen for the ephys property’s model. The heatmap shows how often a metadata feature gets chosen for the best model per EP, from 0 to 100 times.

Neuron name and species are almost always chosen for the best model of each ephys property. A few solution features get consistently included for at least 1 ephys property as well: external K for RMP, internal Na for Rin and RMP, etc. It reinforces the earlier observation that certain solution components are very helpful when trying to predict specific electrophysiological properties. 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. 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 custom models were created using features that are included in >90% of the best models to minimize overfitting.

##### Validating custom models with NeuroElectro and AIBS data

Now that I have created custom models for all ephys properties, I need to compare their performance to the models used previously (Figure 10). To achieve that, I once again employ 10-fold cross-validation and calculate R2 values for each model (Figure 14). The best model (per AICc) for each property is shown in green color. It does achieve the highest performance levels (or on par with other models) when predicting Rin, APthr, APamp. However, it does fall short of “no solutions” model when predicting RMP and APhw. The differences in model performances are small and could be explained by randomness of splitting the data into 10 folds.

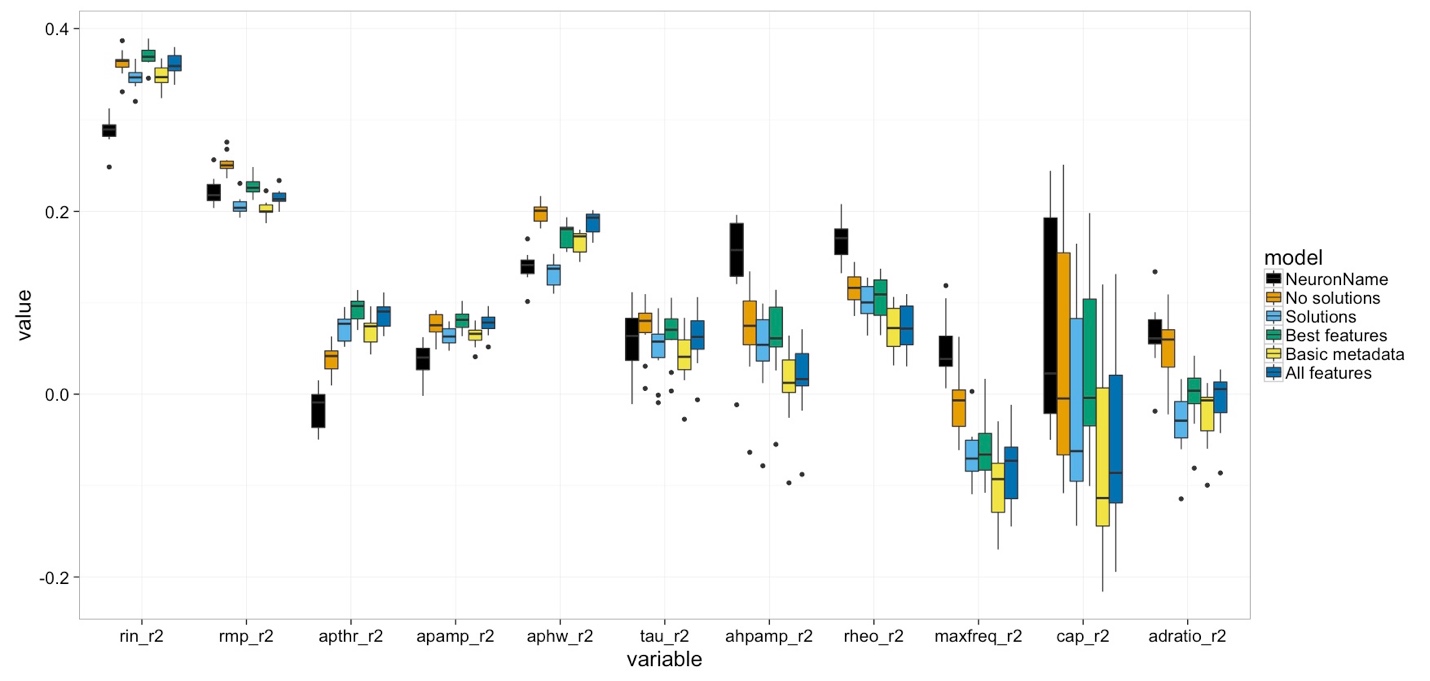


Figure 15: Comparison of custom models to basic models. <Better labels, bigger text, highlight best model in the legend on the right. This data is on the server>

After comparing the custom models to basic ones, I decided to apply these models to never-before-seen data. Fortunately, Allen Institute for Brain Science provided my colleague with their electrophysiological data spreadsheet, supplemented by very detailed experimental conditions that covered all metadata types NeuroElectro keeps track of. AIBS neuron types do not always have a direct NeuroElectro analogue, as such, AIBS neuron types that could not be definitively assigned a NeuroLex cell type were put into the “other” category. This strategy is consistent with the way NeuroElectro treats uncommon neuron types. Since AIBS data was produced during a set of experiments in a single lab – each ephys property measurements were aggregated into a single average value per neuron type, because NeuroElectro stores reported means, not individual measurements.

To evaluate the custom models performance, I used them to ‘shift’ NeuroElectro ephys data (NE) to AIBS experimental conditions baseline (See Methods for details about this step). Briefly, I first predicted and removed ephys variance from NE data that could be explained by my custom models, then I added ephys variance introduced by AIBS experimental conditions. If the models work, the adjusted NE data should be more closely distributed around AIBS electrophysiological measurements. Figure 15 supports that claim. Green violin plots correspond to NE ephys values that have been adjusted to AIBS metadata. Generally, they have less variance, which is hopefully a result of removing explained variance from each article, and their means are closer to the red dots (AIBS means) than raw NeuroElectro data. The only neuron type that does not follow these trends closely is “other”, which is an aggregate of several cell types both in AIBS and NeuroElectro datasets.

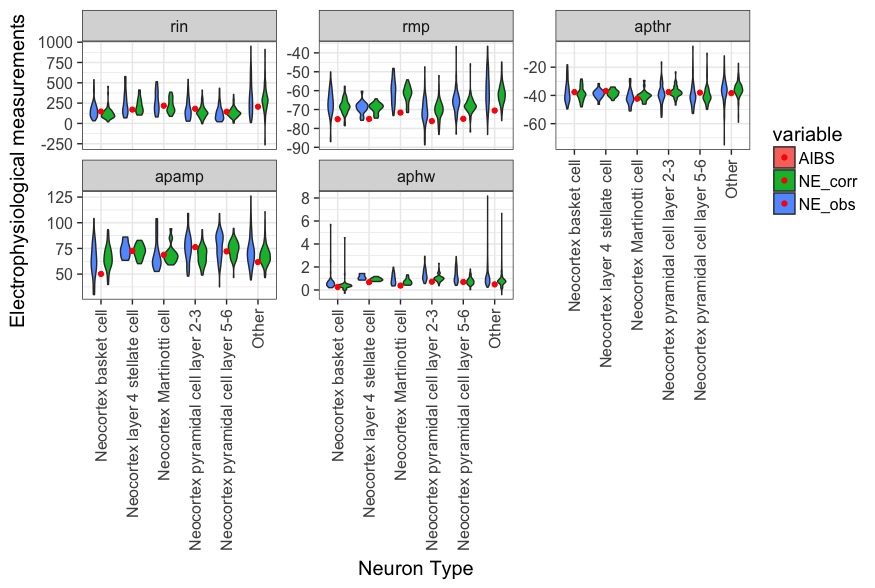


Figure 16: Adjusting NeuroElectro data to AIBS conditions. Violin plots of NeuroElectro data (blue), Allen Institute for Brain Science data (red), adjusted NeuroElectro data using the custom models designed above (green). Each dot is a mean ephys property value reported by an experiment. The model correction tends to squeeze NeuroElectro data around the mean and bring it closer to AIBS value. <Zoom in on APhw plot, add units to each EP>

# Discussion and conclusion

Here, I have researched potential sources of study-to-study variability in electrophysiological properties of neurons. 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 custom 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, yet generally distinct recipes and 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.

## Discussion

#### Solutions text-mining and curation

Extracting targeted words, phrases and sentences from texts written in natural language often proves to be a challenging task [CITE some text-mining papers]. Exceptionally so, when the task is text-mining data from Biomedical literature [Cite people trying to text-mine PubMed]. 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 (Examples: effects of a few compounds on gene expression [CITE], the species of mice that scientists use in behavioural experiments [CITE]). 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 in close proximity to 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 a challenge. Distinguishing sentences that refer to extracellular, electrode and other (cutting, storage, incubation) solutions in the cases of a single electrophysiological experiment per article was a trivial task that was efficiently addressed by creating a dictionary of terms that refer to each solution type. However, articles that described several ephys experiments posed a significant challenge, for example simultaneously reported patch-clamp and sharp electrode recordings or current-clamp and voltage-clamp recordings. 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. However, my analysis of solution recipes revealed several strict rules in internal solution designs: voltage-clamp patch-clamp experiments use cesium instead of potassium to block inwardly rectifying potassium channels (Rang et al. 2003); current-clamp patch-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). It should be possible to utilize internal solution compositions for the task of assigning solution sentences to the correct experiments. Although, 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 X). Its detailed description, as well as curation speed and quality improvement metrics will be included into the future NeuroElectro paper. NeuroElectro primarily stores data ephys tables that was reported under normal control conditions, meaning that 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 1) to the columns of ephys data tables. Thus, the curation interface not only assisted in increasing the number of neurophysiology articles stored in the NeuroElectro database, but it also enabled us to gather more data from certain papers.

One of the most important rules the NeuroElectro curation team had to follow was the “15-minute rule”. It stated 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. Additionally, the curation quality tends to decrease with the increase in article’s complexity (observational point, no data provided). 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 (F1 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.

#### Experimental solution recipe trends

Major ion (Na, K, Cl, Ca, Mg) concentrations used in extracellular solution recipes resemble normal distributions, which could mean that different labs measured these ionic concentrations in cerebrospinal fluids of animal brains (most of NeuroElectro data comes from mice and rats). 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. 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 (Routh et al. 2004, Burdakov et al. 2005). 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 titrating small amounts of a strong base or acid into ACSF [CITE a few papers].

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 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-2000’s electrophysiologists started to consistently add phosphocreatine to their internal solutions [CITE a few] and, since Na2-phosphocreatine is a relatively inexpensive way to fulfill that goal when compared to K2-phosphocreatine, 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 papers that discover beneficial effects of certain chemicals on the state of neurons during electrophysiological experiments. It is possible that chloride concentrations increased because it was paired up with some other chemical, that was 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 conclusions could be drawn from such cases due to sparsity of data.

#### Modeling electrophysiological variability

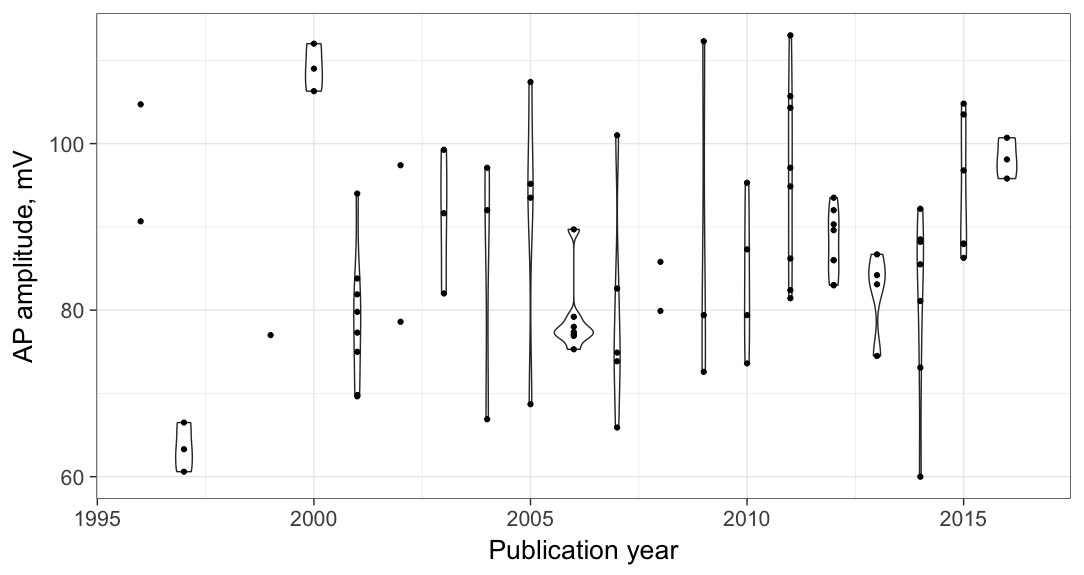
The field of neuron electrophysiology has been in dire need of methods that enable better comparisons of ephys data between experiments. My analysis of electrophysiological data stored in NeuroElectro has shown that, even within similar types of neurons, ephys measurements cannot always be directly compared 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 & Connors, 2012) to -80.0 mV (Booth et al. 2014). However, my analysis has demonstrated that such study-to-study variability can be partially explained by experimental conditions (metadata), specifically solution compositions. However, data collected from hundreds of articles 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, type of scientific kit used, 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 significant results. Additionally, the significant correlations did not pass 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 as well as in conjunction with 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. This is strange since the Goldman-Hodgkin-Katz equation models RMP as a function of the recording temperature and the external and internal concentrations of sodium, potassium and chloride. 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.

I used random forest feature importance ranking and Akaike Information Criterion, corrected for datasets of limited size, to construct custom 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). Surprisingly, publication year was important for AP amplitude (Figure 16). I hypothesize that this effect can be explained by changes in the AP amplitude calculation or measurement protocols.



1. Figure 17: Reported action potential amplitudes of CA1 pyramidal cells vary with time. Each point is a population mean APamp value for Hippocampus CA1 pyramidal cells, reported by an article published in the corresponding year. Violins outline the distributions of values for each year.

I addressed my previous concern of underperforming solution features masking the impact of impactful ones when predicting ephys properties by further examining the solution features that often get chosen for the best model. I would like to stress the fact that several solution components, including major ions, get chosen every time for predicting resting membrane potential: internal and external potassium, internal sodium, and for some reason, internal EGTA. Internal EGTA does not have a significant univariate relationship with resting membrane potential, but it could be making the difference in distinguishing several otherwise unpredictable RMP values. Several major ion features (internal and external chloride and external sodium) were excluded from the custom RMP model. Internal chloride is positively correlated with internal sodium; thus, it is possible that internal sodium is a better predictor for RMP and it masks the effect of internal chloride. External sodium and chloride are both essentially normally distributed with relatively small standard deviations. My models can only detect signal from features that possess enough variance to uncover it. As previously discussed, external concentrations of major ions are tightly normally distributed around a specific value, meaning that the clear majority of experiments use very similar major ion concentrations in their ACSF. On the other hand, internal major ion concentration values are widely spread, making them good candidates for modeling certain ephys properties (Rin, RMP, APthr, membrane time constant, AHPamp). Compounds that were rarely seen in solution recipes (BAPTA, EDTA, cesium) were included into the model selection procedure as negative controls.

Comparing the custom models to the previously considered ones, I note their consistently high performance. They do not always achieve the highest R2 values, but they also never completely fail. The custom 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 custom models is to eliminate features that introduce noise, which is the case with most of the solution features. Additionally, the custom 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 custom 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 adjust the five most commonly reported ephys properties in NeuroElectro to the experimental conditions used by AIBS. The effects of custom models include removing explainable ephys variability from NeuroElectro data, making it more comparable with the AIBS ephys measurements.

## 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 CaCl2 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 (check assign\_metadata.py file).

The second text-mining algorithm extension option is 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 avenue involves enhancing the existing algorithm for text-mining electrophysiology properties from tables and text of neuroscientific papers. The current implementation was developed by my colleague, Shreejoy Tripathy, thus any follow up efforts would need to gain his approval first. 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 custom models that currently 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. Essentially, we are trying to link NeuroTree, an online database of genealogies that keeps track of academic mentorship in neuroscience (David & Hayden, 2012), and the experimental metadata stored in NeuroElectro.

## Conclusion

In conclusion, my integrative meta-analysis approach addresses the neuroscience need for comparing electrophysiological data across different studies. The custom models proposed here partially remove ephys variability that can be explained by experimental conditions, enabling better comparisons of reported electrophysiological properties across experiments. Electrode solution constituents are sometimes included in these models, but extracellular recipes used by the neurophysiological community are too similar, thus they cannot provide meaningful information when modeling the variability of ephys properties.

## Closing remarks

If the field of electrophysiology intends to progress towards large-scale analyses, protocols for measuring and reporting ephys protocols need to be standardized. The first step towards that goal is carefully reporting both calculated and arbitrary decisions that lead to reported ephys values. Make electrophysiology great again, together.

# Bibliography

# Appendices