**Exploring sources of variability in electrophysiology data of mammalian neurons**

**Introduction**

Electrophysiological (ephys) recordings are widely used for characterizing neuron function. The 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, and their interactions throughout the brain. For example, the first goal of the US NIH BRAIN project is to generate a “census of cell types”, a task that involves neuron comparison using genetic, morphological and electrophysiological characteristics. In the case of neuron electrophysiology, a common practice among neurophysiologists is to only analyze data that they have themselves collected. However, this approach imposes sample size restrictions, because a single scientist, lab or even institution can 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.

Though it would be magnificent if electrophysiology data could be combined and compared across labs, doing so 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, or varying extracellular Ca2+ concentrations result in changes in electrophysiological properties of neurons (Suter et al. 2013, Aivar et al. 2014). Thus, comparing data across differently designed experiments without accounting for variability introduced by experimental conditions could lead to incorrect or inconsistent results.

Many neurophysiologists address the task of exploring the effects of experimental conditions on neuron ephys properties by using experimental electrophysiology techniques (Kim et al. 2012, Armentia et al. 2004, Lee et al. 2004). However, the experimental approach is limited to varying a single condition and studying one or several neuron types at a time. Additionally, 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 remained fixed throughout each experiment. Although, the results obtained in these individual experiments are highly reliable and could be used in a meta-analysis.

A typical ephys experiment involves: extracting and cutting the brain of the anesthetized animal at a specified age; letting the slices recover in a bath of a carefully designed solution (sometimes the slices also get incubated); transfer them one by one to the 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 measuring electrode is inserted inside (or is kept right next to the neuron, depending on electrode type and technique) and electrophysiological properties of the neuron can be measured. The electrode also contains the internal solution (intracellular, pipette), which ends up inside the cell.

Previously, it has been shown that animal age, temperature and electrode type choices significantly correlate with the variance in reported ephys values (Tripathy et al. 2015). From the original electrophysiological experiments performed by Hodgkin and Huxley we know that neuronal action potentials rely on the electrochemical gradient across the cell membrane (Hodgkin & Huxley, 1952). Since a typical ephys experiment utilizes custom solutions inside and outside the measured neurons, I hypothesize that between-lab ephys variability can 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 methods to extract experimental solutions used in published neurophysiology articles. Additionally, I integrated my solution extraction algorithms into the existing NeuroElectro database, which contains curated electrophysiological measurements, neuron type and other experimental conditions information from control samples of neuroscientific articles. Once the data was collected, I applied univariate linear models to uncover the effects of solutions on 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 that the effect of solution compositions on the variance in electrophysiological properties is relatively small, mostly because different labs use similar solutions, thus their explanatory power is limited. Additionally, I created custom models for ephys properties commonly reported by neurophysiologists. However, they can only partially predict reported ephys properties, thus experimental parameters that are causing ephys variability might be among the metadata types that are not tracked by NeuroElectro or that are not commonly reported in neuroscientific articles. The innate high variability of neuronal ephys properties could be masking the impact of experimental conditions I use to predict them. Nevertheless, 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 my models include normalization of ephys data by metadata and adjusting values of common ephys properties from one set of experimental conditions to another. To validate my models 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.

<Do I need an Aims: 1), 2), 3) bullet list somewhere in the intro?>

**Methods**

We used the existing NeuroElectro (www.neuroelectro.org) database as a starting point for determining metadata factors that could have an effect on the variability in reported electrophysiological measurements (Tripathy et al. 205). NeuroElectro stores and data-mines thousands of Neuroscience articles that may contain electrophysiology data. Those articles are downloaded in HTML format, text-mined for the most common electrophysiological properties, for example: resting membrane potential, input resistance, action potential spike half-width and amplitude. The extracted data is then curated by a team of trained undergraduate students using a carefully designed curation interface. Neuron types are assigned 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”.

Text-mining tables of HTML articles for ephys properties and Methods sections for experimental conditions is done in C-Python using HTML-parsing tools and libraries. The process of text-mining for ephys properties and previously analyzed metadata (animal age, weight, species, strain; recording temperature, electrode type, junction potential) has not been significantly adjusted since the previous NeuroElectro paper (Tripathy et. al, 2015). However, here we have implemented a robust text-mining approach that is able to handle complicated metadata extraction such as solution compositions used during electrophysiological experiments. We use a Python library Natural Language Toolkit for converting an imported HTML article’s Methods section into a collection of sentences and parsing them. We assume that all the relevant metadata information can be found in the Methods section.

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

Recording chamber (external) and pipette (internal) solutions used during electrophysiological experiments were considered to be the most promising experimental conditions that were not yet included into NeuroElectro. We define the solution extraction task as a 3 step process: 1) ranking each sentence in the Methods section on how likely it is to contain a solution; 2) identifying the solution type for each solution-containing sentence and 3) extracting individual compound concentrations from external and internal solutions (e.g., Na+, K+, HEPES, etc).



**Figure 2.** **Solution text-mining.** Steps in the process of identifying solution-containing sentences and extracting compounds concentrations. Colors represent different processing steps and link to the targeted text. Sentence extracted from Derchansky et al. 2008

The first 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 mentions of ions of interest (Ca, Mg, Na, K, Cl) and has a general solution-describing structure. A solution-containing sentence typically mentions a concentration unit (mM or μM), contains a series of chemical compounds separated by commas or other delimiters and might end with a pH measurement.

To automatically find the most likely solutions used during the recording, we use the existing HTML articles in the NeuroElectro database as well as the implemented electrophysiology data and metadata text-mining infrastructure. The difficulty of this task lies in the fact that ephys recording are often performed in slices, as a result each paper reports more than just the two solutions used in the actual measurements. The most common *other* solution types include: cutting, storage and incubation. While it is true that sometimes the same solution will be used for multiple steps of an experiment and we default the extracellular solution to the last reported *other* solution type, those cases are not common enough to use as a general rule for all articles.

Once a sentence has been identified as solution-containing, the algorithm uses more 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 expanded one sentence at a time by up to 3 sentences before and 1 after the solution-containing sentence. Based on an empirical analysis, internal solutions are identified very consistently, but external solutions can be simply referred to as “the same as storage solution” or “ACSF used for dissecting the brain” <Do I need citations here?>. To solve this, we assume that missing external solutions imply that the last mentioned storage/cutting solution was also used for electrophysiological recordings.

Finally, we extract solution concentration values by identifying the location of each compound of interest in solution sentences using regular expressions. We account for element valence (2 mM CaCl2 would be parsed as 2 mM of Ca2+ and 4 mM of Cl-) even when it is a part of the compound name (disodium sulfate instead of Na2SO4 or sodium creatine instead of Na2-creatine). This step of the algorithm suffers from inconsistencies in the way solutions are reported: using semicolon as compound separator in all cases but one where a comma is used, spelling Cl with an “i” instead of an “l” and other typos in compound names (phosphocreatinine instead of phosphocreatine).

Our algorithm achieves 89% accuracy based on a manually curated set of 60 randomly chosen articles. The most common errors include incorrect solution type identification and a solution that is described over multiple sentences that cannot be parsed correctly. The former refers to an incorrect solution being chosen as external or internal ephys solution used in the article and could be dealt with by a better solution classification algorithm, the latter is more difficult to address since it is problematic to find the adjustments and identify parts of solutions that should be fixed.

**Manual curation**

Briefly, we follow up text-mining with two rounds of curation: 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, 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 and assign a NeuroElectro neuron type to each author-defined neuron. Both rounds of curations check the text-mining output. In our analysis we use only data that has been put through both rounds of curations. The curation pipeline is explained in greater detail in the NeuroElectro database paper.

**Statistical analysis**

**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 table and all metadata we could gather from the Methods section of the article. Each ephys property is stored exactly as it is reported in each article: mean +/- standard deviation or standard error and number of measurements.

Little processing is performed on the metadata entries since some of them are pre-defined categorical variables (species, strain, electrode type, preparation type and junction potential correction status). Continuous variables are checked against possible value thresholds (age and weight cannot be negative, recording temperature should be within a reasonable range). In the text-mining and curation stage we record the sentences containing external and internal solutions, but little analysis can be done on solutions described in text. We define ‘compounds of interest’ that are very commonly used in electrophysiological solutions (Ions: Ca, Mg, Na, K, Cl, Cs; Other: glucose, ATP, GTP, HEPES, EGTA, EDTA, BAPTA and pH of each solution). We programmatically extract the reported concentrations of these compounds and disregard the rest of the solution text. Importantly, we also disregard each individual compound’s dissociation constants since we do not possess that kind of information for every chemical used in electrophysiological experiments.

**Post-processing of the data spreadsheet in RStudio and reasons for each assumption / correction.**

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

Here, we 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 sparsity. Our statistical pipeline does not depend on the number of ephys properties analyzed, but it does require a reasonable amount of articles reporting them, otherwise the resulting models would have insignificant explanatory power. With more articles being added to NeuroElectro, our analysis can be applied to more ephys properties. For a full and most up-to-date list of ephys properties visit [neuroelectro.org/ephys\_prop/index/](http://neuroelectro.org/ephys_prop/index/).

The most important data preparation steps include: assigning a default compound concentration and reversing appropriate ephys values correction for junction potential. The first step involves assigning an arbitrary small default concentration of 10-6 millimoles (mM) for each ion concentration that was not mentioned in the solution sentence. We calculate reversal potentials in our analysis and the data would otherwise be too sparse. Ideally, the second preparation step would be done in the other direction – correcting resting membrane potential and action potential threshold values for junction potentials in the articles that did not do so themselves. Liquid junction potential is the voltage difference between two different solutions that are in contact with each other. Unfortunately, we discovered that only 46% of articles report junction potential value and 24% correct for it, thus we decided that it would be easier to reverse the corrections (since the articles that correct for junction potential are likely to report its value) than attempt to impute unreported junction potentials based on solutions used. It was important to address the junction potential problem since we have to standardize important properties like resting membrane potential and AP threshold.

**Model creation**

To predict electrophysiological property quantitative values from experimental metadata attributes, we built models for predicting each of the 11 most abundant ephys properties using RandomForest package in R. It implements a supervised machine learning approach that is based on creating decision trees from samples of our dataset. 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.

To evaluate model performance, we examined the predictive power of each of the 30 metadata features using 10-fold cross-validation utilizing a leave-one-feature-out and use-one-feature approaches. For each of the 10 folds we ensured that unique articles were assigned to each fold, otherwise the metadata would be predicting PubMed ID instead of ephys properties.

<Too much information, also cforest does not have this problem>

Random Forest 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. Our 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.

We have also tested several thematic models, examples include only using chemical compound concentrations, neuron types, other metadata or combinations of them. Based on these results we propose a model for each of the commonly reported ephys properties that allows to reduce the effect of experimental conditions on the ephys value.

“This will be a good approximation even at quite small sample sizes if the sample elements are (nearly) independent and the underlying distribution is unimodal, and neither badly skewed nor heavy tailed.” (Burnham & Anderson, 2002)

**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 located in <https://github.com/dtebaykin/neuronephys>.

**Results**

My main goal was to identify how the use of different experimental conditions (specifically, experimental solution recipes) affect the results of electrophysiological experiments. I performed a meta-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, 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, enabling comparisons across articles.

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 a major neuron type, thus my full dataset contains 1588 rows of data (Table 1). Neuron types are defined by an extended dictionary that was originally provided by NeuroLex.org. The full list of NeuroElectro neuron types can be found here: <http://neuroelectro.org/neuron/index/>.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
| Cell Capacitance (cap) | 258 | **External [ATP]** | 4 |
| Maximum AP Frequency (maxfreq) | 229 | **External [GTP]** | 4 |
|  | | **Internal [Na]** | 1119 |
| Neuron Type | 1588 | **Internal [K]** | 1466 |
|  | | **Internal [Cl]** | 1340 |
| Basic metadata: | | **Internal [Mg]** | 1217 |
| Species | 943 | **Internal [Ca]** | 244 |
| Strain | 887 | **Internal [Cs]** | 60 |
| Electrode Type | 943 | **Internal [glucose]** | 46 |
| Preparation Type | 943 | **Internal [HEPES]** | 1241 |
| Recording Temperature | 1511 | **Internal [EGTA]** | 797 |
| Animal Age | 1388 | **Internal [EDTA]** | 3 |
| Animal Weight | 272 | **Internal [BAPTA]** | 30 |
| Junction Potential | 1588 | **Internal [ATP]** | 1193 |
| Junction Offset | 551 | **Internal [GTP]** | 1083 |

**Table 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

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/>.

A major challenge in analyzing ephys data across studies is inconsistent standards. For example, ephys properties can be measured from different baselines: 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); adaptation ratio (ratio of durations between early and late APs inter-spike intervals in an AP train) is less standardized and can be reported as a ratio first / last ISI (Nassar et al. 2015, Scorza et al. 2011), ratio of last / first ISI (Novkovic et al. 2015, Zhou et al. 2015), a percentage (Fujiwara-Tsukamoto et al. 2004, Zaitsev et al. 2009), 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 consistently curate and standardize. 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 are known to have relatively large standard deviations, which could conceal the effects of experimental conditions on the results (Tripathy et al. 2015). I first considered whether between-experimental variance in the same neuron types is greater than within-study variance. That is, whether the observed high variance in ephys properties could stem from consistent differences in experimental conditions or their innate high variability. If the latter was true, the meta-analysis approach would likely yield inconclusive results, because: 1) finding significant correlations would be highly unlikely; 2) I would not be able to definitively distinguish the true correlations from artifacts of the dataset.

In the context of a single experiment, the scientist measuring RMPs of hippocampus CA1 pyramidal neurons expects a set of measurements that is normally distributed around the true mean. If experimental conditions do not introduce significant variance when measuring ephys properties, then multiple electrophysiology experiments 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. 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 to -48 mV with a mean SEM of 2.6 mV) and medium spiny neurons (-95 to -61 mV, mean SEM of 2.8 mV), still with relatively small standard errors. The data does 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). Figure 3 shows resting membrane potential only, however, other ephys properties behave very similarly. Thus, the earlier assumption of ephys measurements not being affected by experimental conditions must be false.

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

**Figure 3: 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). Here, I further explore the variance in experimental conditions. I am particularly interested in extracellular (external) and pipette (internal) solution compositions and their effect on the ephys measurements relative to experimental conditions that are already known to be significant.

**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 tested the solutions data extraction pipeline on a 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 validated the correctness of each component in the solutions data acquisition pipeline. Text-mining validation was a multistep process: 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 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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.

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.

<To discussion:>

We know that calcium chloride implies CaCl2 and not CaCl, but it would be time-consuming to design an algorithm that could account for ion valence when necessary.

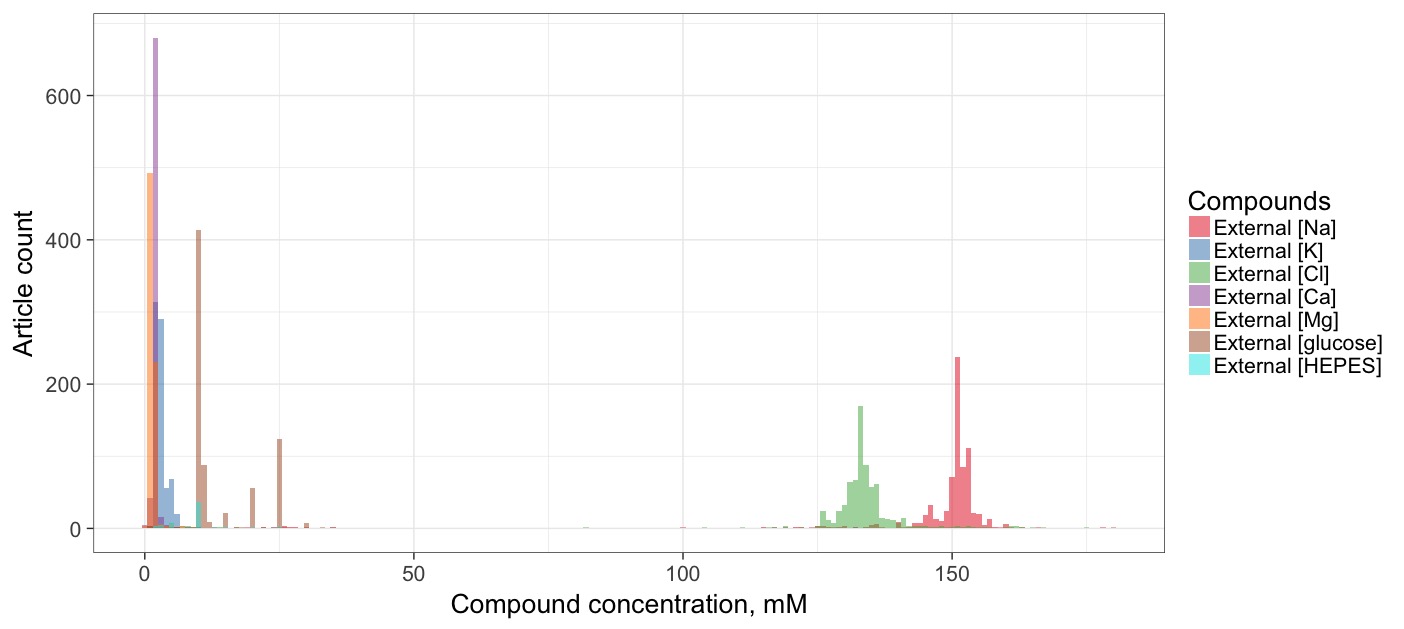
**Exploring solution compositions**

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. 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 8). 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 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 small amounts of chloride (5-20 mM), magnesium (1-8 mM) and sodium (0-50 mM); while the latter tends to have several moles of potassium, typically paired up with acetate, methylsulfate or chloride (observational data, I am not extracting acetate or methylsulfate concentration values).

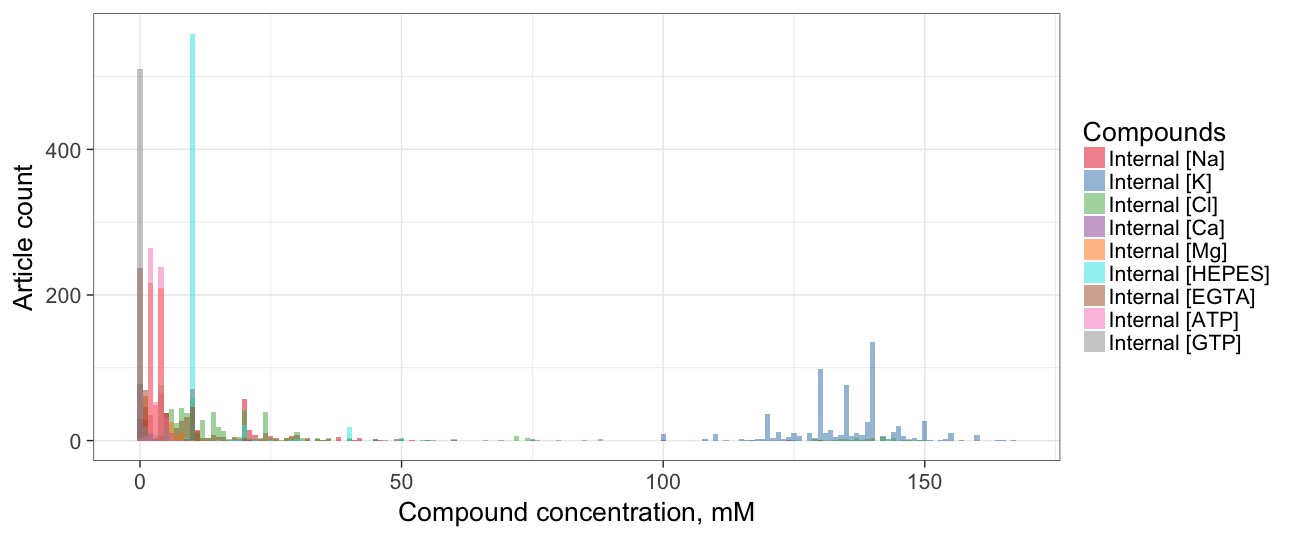
The distributions of chemical components used in experimental solutions vary. External Na, K, Cl, Ca and both internal and external solutions follow almost normal distributions, with some bias towards even

<A paragraph describing observations from solutions histograms – how widely spread are the concentrations, mention outliers – some are errors and some are real. Examples for real ones>



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.

I then quantified how often identical solution recipes were used 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).

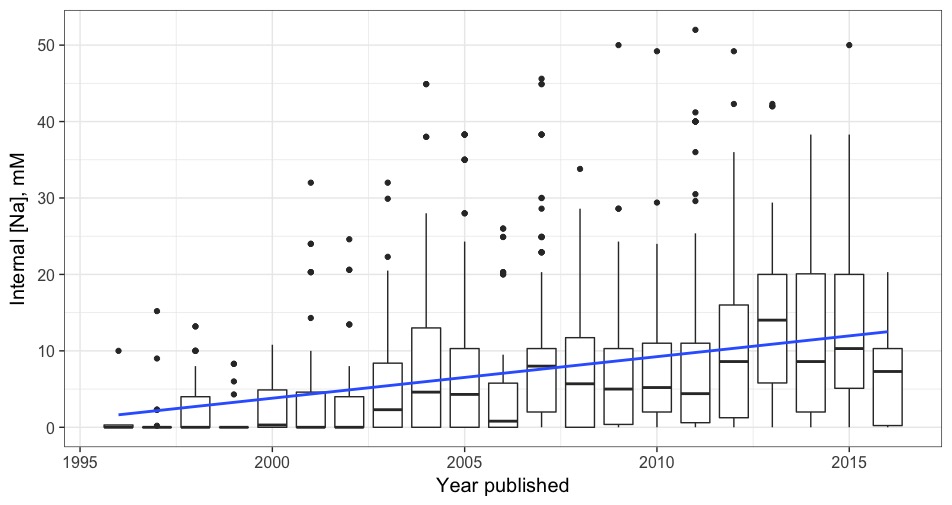
To summarize, identical solution recipes tend not to be re-used from paper to paper, percentage of unique solutions per electrode and solution type: Patch-clamp external – 49%, Patch-clamp internal – 66%, Sharp external – 63% and Sharp internal – 48%. I used principal component analysis supplemented 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 - 3 | 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 < 0.001

**Figure 6: 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 differences help to further explain variability in the commonly reported ephys properties.

**Univariate approach for modeling electrophysiological properties**

B

A

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

r = -0.16

p > 0.05

r = 0.35

p < 0.001

**Figure 7: Univariate relationships between electrophysiological properties and solution concentrations**. Each point is a mean ephys value reported by an article for Hippocampus CA1 pyramidal neurons. Blue line is the best univariate linear fit for the data, grey area shows 95% confidence interval for the linear fit. A) Input resistance increases with internal sodium concentration (r = 0.35, p < 0.001). B) 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, after 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 multivariate approach that incorporates the influence of multiple experimental parameters on the same ephys property simultaneously.

**Multivariate approach for modeling electrophysiological properties**



**Figure 8: Multivariate approach to modeling EPs. <Placeholder> Perhaps make a figure that shows metadata on the input side and EPs on the output side with some Random Forest in the middle.**

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 (Figure 8). 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 methods: Furthermore, I enforced unique articles into each fold to avoid models learning to predict the article PubMed ID rather than reported ephys values.>

<Describe data going into each model and the purpose of having each one>

Applying Random Forest algorithm to my data, I created 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, predicted values are more ‘squished’ than the observed values because “all models are wrong, but some are useful” (Box et al. 1987). In my case, it means that models can capture only a portion of the true variance in the variable they are trying to predict.

B

A



R2, model performance

**Figure 9: Multivariate regression models predict input resistance.** 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.

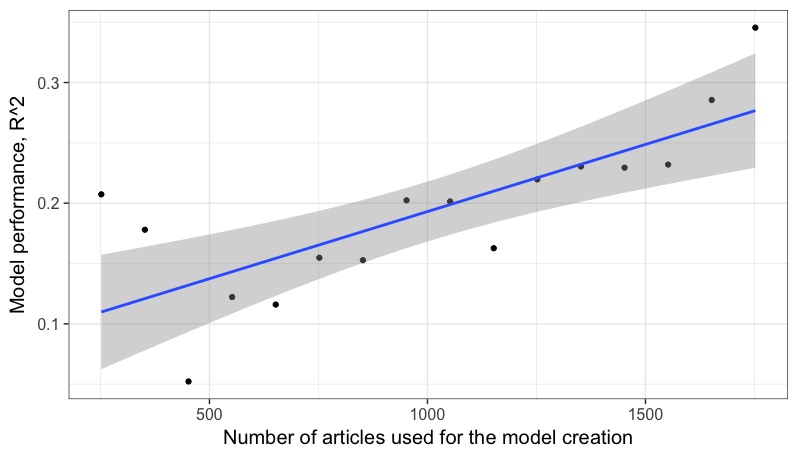
The next step was to evaluate the relative contributions of neuron type, basic metadata and solutions when predicting ephys properties, starting with input resistance (Figure 9B, refer to color scheme in table 1 for metadata groups). Since the folds are assigned to articles in a random fashion, the performance 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 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 10: R2 values for 6 different models predicting 11 common ephys properties.** Baseline is the lower bound for performance of Random Forest. 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.



r = 0.76

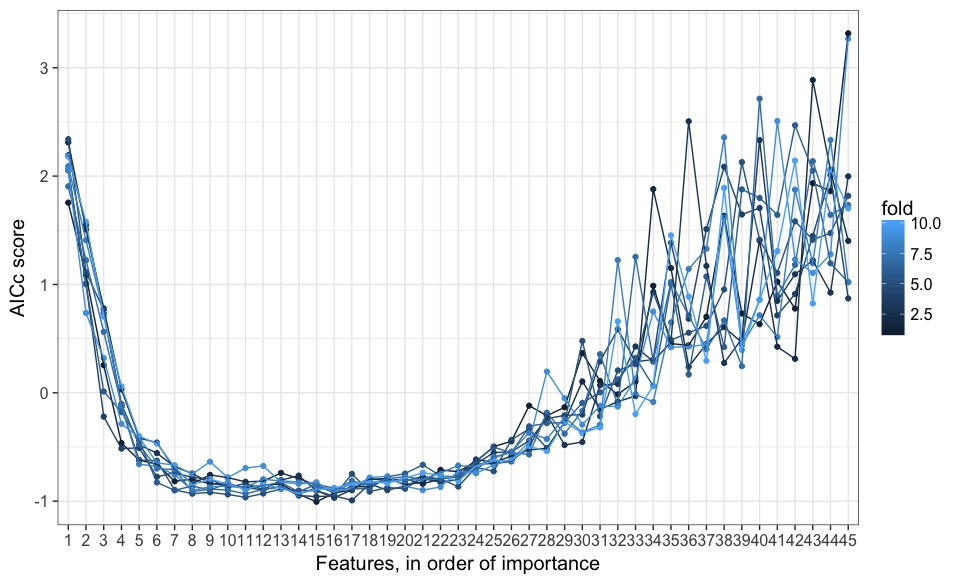
p < 0.001

**Figure 11: 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.

I next 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.

**Designing custom models for common electrophysiological properties**

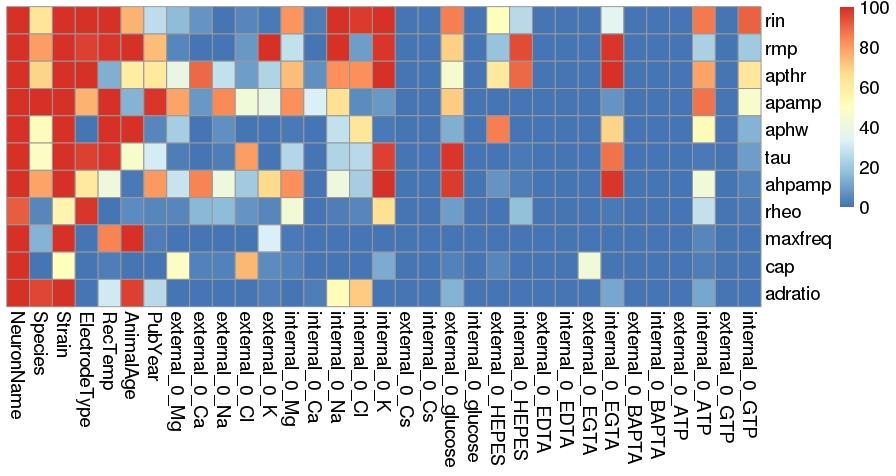
The final step of my project was to create custom models out of only the best features, possibly different for each ephys property. For example, solution features represent an aggregate of 22 different features: 5 major ions and 6 commonly used compounds per internal and external solutions. What if some 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 12: AICc scores for input resistance.** 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 all 11 of the most abundant electrophysiological properties. I have also performed this 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. The heatmap shows how often a metadata feature gets chosen for the best model per EP, from 0 to 100 times.

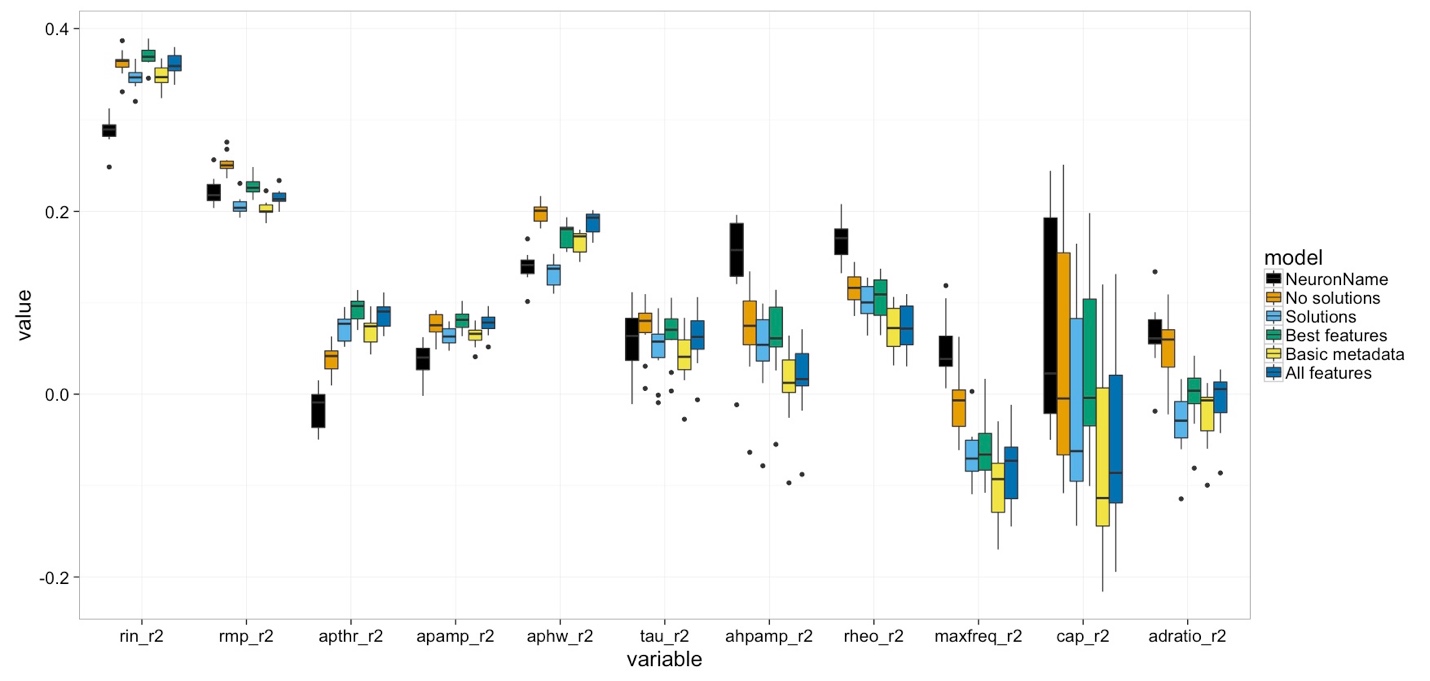


**Figure 13: Feature selection.** Ephys properties and metadata features are listed vertically and horizontally, respectively. Color represents number of times the feature has been chosen for the ephys property’s model.

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**

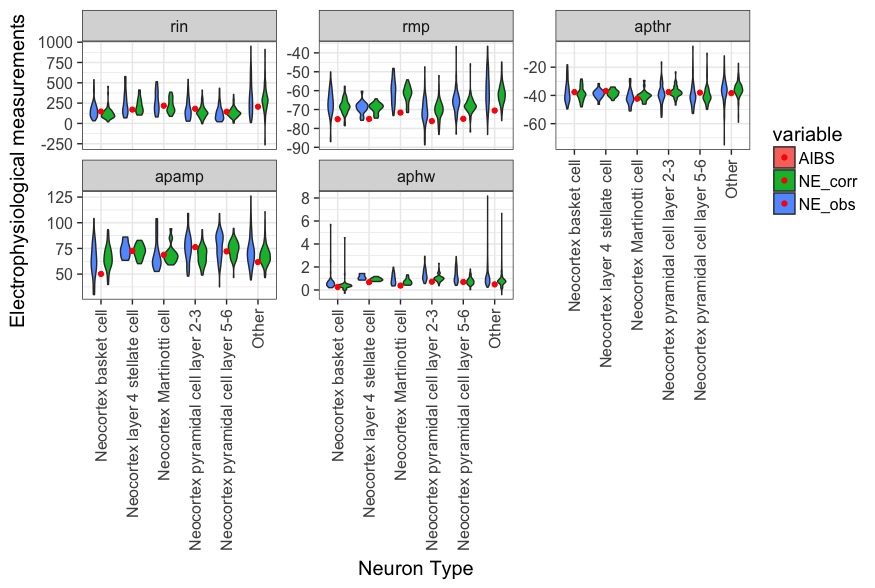
Now that I have created custom models for all ephys properties, I need to compare their performance to the models used previously (Figure 10). To achieve that, I once again employ 10-fold cross-validation and calculate R2 values for each model (Figure 14). The best model (according to AICc) for each 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 fairly small though and could be explained by randomness of splitting the data into 10 folds.



**Figure 14: Comparison of best model 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 15: 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>

<Urban lab data needs to have a similar figure>

**Discussion**

Here, we have researched potential sources of between-lab variability in the reported electrophysiology values via text-mining and curation of published neuroscientific literature. We are using an expanded version of the NeuroElectro database (882 articles) compared to the previously published research article which was using ~300 curated papers (Tripathy et al. 2014). Subsequently, we process more experimental setup information, namely chemical solutions used in recording chambers and inside electrodes/pipettes of electrophysiological experiments. We found that while the metadata we process is not able to fully explain lab-to-lab variability, our models are able to explain some of that variance and they could make reported electrophysiological properties from different labs more comparable.

An argument can be made that Martinotti cells or medium spiny neurons are not referring to the same respective cell types in various labs. However, the hippocampus CA1 pyramidal neuron is a well-established neuron type and as such, we can safely assume different labs talk about measurements from very similar neurons.

**Solutions text-mining and curation**

<The precision of text-mined internal solutions is lower than anticipated, likely due to non-triviality of the task and unfamiliarity of curators with it.> Curated solutions have high accuracy.

Even though my text-mining algorithm was good enough to be applied on the entire corpus of articles, ephys data from them was not available.

Furthermore, given that some papers will employ multiple experimental solutions (e.g., different solutions used for voltage- versus current-clamp, making it difficult to match experimental metadata to corresponding electrophysiological data.

Why wasn’t concentration value extraction part of the curation process? Time and effort constraint. Bang-for-your-buck argument. It was worth it to get the right solution sentence curated, but not each compound concentration individually.

**Experimental setup trends**

, which could be a result of specific experimental needs or simply several original recipes being passed down from instructor to student through the years.

**Schools of thought analysis**

First, in mid-2000’s electrophysiologists started to consistently add phosphocreatine to their internal solutions and, since Na2-phosphocreatine is a relatively inexpensive way to fulfill that goal, internal sodium concentrations increased from 0-5 mM to 20-30 mM. The latter trend is more elusive, extracellular calcium concentrations are impressively similar at 2-2.5 mM across the 703 patch-clamp articles included in this analysis. However, magnesium is often present at 1 mM or 2 mM extracellularly, this split is very clearly defined (with very few articles that use between 1 mM and 2 mM of [Mg]).

<Still searching for a reason for this pattern of Mg concs>

**Predicting ephys data with metadata**

Neurons are classified into types based on their genetic, morphological and electrophysiological features (among other characteristics), as such I expected the neuron type to be the most valuable source of information when modeling ephys properties.

Unfortunately, electrophysiology is a field with low experimental reproducibility. NeuroElectro helps to outline the problem: results for resting membrane potential of hippocampal CA1 pyramidal neurons across 102 articles range from -55.1 mV (Kim & Connors, 2012) to -80.0 mV (Booth et al., 2014). This problem is made worse by the fact that most articles have relatively small sample sizes (number of cells or animals used). We attempt to deal with this problem by using reported experimental conditions as features for predicting 11 commonly measured electrophysiology properties. Consequently, we have validated our initial assumption that neuron type is the most important piece of information one needs in order to compare ephys properties across experiments. The other metadata can be helpful, neutral, or introduce noise into the comparison. Thus we propose models for removing this metadata-explainable bias. Our models are able to explain up to 40% of the variability in ephys values when compared to models trained on randomly shuffled data, however, comparing our models to simply using average values for each ephys property as predictors we observed only a modest increase of 10-20% in their predictive power.

<Can I come up with an upper bound? Perhaps use sampling from SD interval around the mean as predictions. If time permits>

The relatively low influence of experimental metadata on electrophysiology values could be explained by intrinsic high variance in those values even within the same experiments. Alternatively, there are other experimental conditions in neuroscientific articles that we do not target at this time, such as pipette properties, type of scientific kit used, for *in vitro* recordings - time between brain extraction, slicing, incubation and ephys measurements, <neuron types not necessarily same>. It is also possible that a good portion of intrinsic ephys variability arises from unreported procedural steps.

Data collected from hundreds of articles is very heterogeneous when compared to results reported by a single lab. The standardization task is too complicated to rely on automatic approaches; it requires the work of trained curators. The results and implications of this process are discussed in more detail in the NeuroElectro database update paper <Coming soon, citation>.

One of the more common issues is that there is no exact definition of how to measure each electrophysiological property. As a result, neuroscientists can report similar properties, but measured differently. For example, spike amplitude can be reported from resting membrane potential, spike threshold or weirder approaches results from which cannot be converted to our standardized values. Another example is measuring inter-spike interval, which can be reported as: a ratio between first and last spike intervals, a ratio between last and first intervals, one minus the above values or a percentage of the them. Our solution is dealing with these problems during the curation stage since it is hard to quantify the extent to which scientific data reporting styles differ and that, in turn, makes it difficult to resolve the issue programmatically.

<Paragraph on model creation>

The 100% threshold was not chosen because due to random draw 1 run of cross-validation may have had very unfavourable folds for a feature that otherwise should be included (this is also overfitting, but from the other side – it causes a useful feature to be excluded due to poor luck of the draw).

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 reported in ephys articles that cannot be directly assigned a NeuroLex term.

Compounds that were almost non-existent in the data were still used for the models, they served as negative controls, I expected them to not ever be picked as valuable features.

We believe that our integrative meta-analysis approach addresses the neuroscience need for comparing electrophysiological data across labs and experiments. Specifically, we created a large corpus of neuroscientific literature, then we text-mined and curated it for ephys data and corresponding experimental conditions. We created models for removing metadata-induced bias from electrophysiology values, thus making them more comparable between experiments. However, we have encountered many obstacles while collecting the data and the models we propose are a stepping stone on the path to understanding electrophysiological properties behavioural patterns. Nevertheless, our quantitative meta-analysis approach of analyzing electrophysiology data from hundreds (and soon to be thousands) of experiments drastically increases the influence of each individual electrophysiological article, enables neuroscientists to validate their results against their colleagues and inspires exploration of electrophysiological trends across and within neuron types.

<Discussion point: what are different ways of calculating the upper bound? What are the performance expectations for the upper bound?>

The AICc analysis may yield different models in the future, once more data is available. Models with few features are likely to get bigger

Add discussion point: how does a certain feature relate to a certain EP? Example, in the model selection consistency heatmap, why is external Na almost never chosen for the model and internal Na, Cl and K are? Point: we can only detect signal where there is enough variance to uncover it. External Na concentrations are surprisingly consistent across articles. External Cl varies, but it does not significantly correlate with Eps, on the other hand internal K, Na and Cl seem to be robustly important for prediction of some EPs.

