



1

# NeuroElectro: A Window to the World's Neuron Electrophysiology Data

Shreejoy J. Tripathy<sup>1,2,3,4,\*</sup>, Judith Savitskaya<sup>1,6</sup>, Shawn D. Burton<sup>1,2</sup>, Nathaniel N. Urban<sup>1,2</sup>, and Richard C. Gerkin<sup>1,2,5,\*</sup>

<sup>1</sup>Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

<sup>2</sup>Center for the Neural Basis of Cognition, Pittsburgh, PA, USA

<sup>3</sup>Present Address: Centre for High-Throughput Biology, University of British Columbia, BC, Canada

<sup>4</sup>Present Address: Department of Psychiatry, University of British Columbia, BC, Canada

<sup>5</sup>Present Address: School of Life Sciences, Arizona State University, Tempe, AZ, USA

<sup>6</sup>Present Address: Graduate Program in Bioengineering, University of California, Berkeley and University of California, San Francisco, CA, USA

Correspondence\*:

Shreejoy J Tripathy

177 Michael Smith Laboratories, 2185 East Mall, University of British Columbia, BC, Canada, V6T 1Z4, stripat3@gmail.com

Richard C Gerkin

School of Life Sciences PO Box 874501 Arizona State University, Tempe, AZ, 85287-4501, rgerkin@asu.edu

## Neuroinformatics Infrastructure

## 2 ABSTRACT

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

**Keywords:** neuroinformatics, electrophysiology, database, text-mining, metadata, API, machine learning, natural language processing

## 1 INTRODUCTION

21 Brains achieve efficient function through implementing a division of labor, in which different types of  
22 neurons serve distinct functional and computational roles. One striking way in which neuron types differ  
23 is in their electrophysiology properties. Though the electrophysiology of many neuron types has been  
24 previously characterized and documented across decades of research, these data exist across thousands of  
25 journal articles, making cross-study neuron-to-neuron comparisons difficult.

26 Neurophysiology lacks a centralized resource where consensus data on basic physiological  
27 measurements from many neuron types and studies are accessible for reference and subsequent meta-  
28 analyses. For example, though it is common for neurophysiologists to measure and report neuronal  
29 measurements such as resting membrane potential and input resistance, there is not a public database  
30 which compiles this information. In other domains of neuroscience such efforts have made more progress.  
31 In the domain of neuroanatomical connectivity, information on connectivity between different brain  
32 regions is being compiled by experts at the Brain Architecture Management System project (BAMS)  
33 across hundreds of publications (**Bota et al.**, 2005). Parallel to this effort is the WhiteText Project,  
34 which addresses a complementary goal by algorithmically mining brain region connectivity statements  
35 from journal abstracts using biomedical natural language processing (bioNLP) methods (**French et al.**,  
36 2009, 2012). Similarly, in the domain of neuroimaging, the NeuroSynth Project has mined fMRI-  
37 based brain activation maps from published x,y,z coordinate data tables from thousands of neuroimaging  
38 publications (**Yarkoni et al.**, 2011). These literature-based methods can be contrasted with projects such  
39 as NeuroMorpho.org (**Parekh and Ascoli**, 2013) and ModelDB (**Migliore et al.**, 2003; **Hines et al.**,  
40 2004), which index neuron morphological reconstructions and computational models for simulating  
41 neuron activity by obtaining this information directly from investigators.

42 Success among these projects can be defined according to different criteria. Such criteria include  
43 completeness and comprehensiveness; for example, what percentage of relevant connectivity studies  
44 are indexed within BAMS? How many different neuron types are contained within the NeuroMorpho  
45 database? Alternatively, success can be defined in terms of the utility of these databases in driving  
46 subsequent research, like the use of BAMS as a resource for discovering relationships between brain  
47 region connectivity and gene expression (**French and Pavlidis**, 2011) or the use of NeuroMorpho to  
48 discover general scaling relationships among the morphologies of neuron types (**Teeter and Stevens**,  
49 2011). Similarly, NeuroSynth is widely used by cognitive scientists as a starting point for designing  
50 functional imaging studies. Thus while these projects are not yet comprehensive and likely contain data  
51 records of varying quality, these resources may nevertheless be employed to draw novel inferences.

52 These projects are logically divided according to their methods for obtaining the source data: through  
53 the use of manual methods like expert curation or user contributions versus automated methods such  
54 as text-mining. Notably, these approaches differ in their scale and accuracy; while algorithmic methods  
55 can “scale-up” and be applied to arbitrary numbers of publications, they typically have a lower accuracy  
56 relative to human-curated content (**French et al.**, 2009). This lower accuracy is often attributed to the  
57 rich lexical complexity of biomedical texts which often require considerable context and background  
58 knowledge to understand and parse (**Dickman**, 2003; **Ambert and Cohen**, 2012). The competing  
59 constraints of scale versus accuracy pose a challenge for large-scale compilation of neuroscientific data.

60 Here, we built a custom infrastructure framework for extracting electrophysiological measurements for  
61 specific neuron types from published neurophysiology articles. These measurements included properties  
62 such as input resistance and resting membrane potential, as well as associated metadata (i.e., article-  
63 specific methodological details). Our methods combine algorithmic literature text-mining, drawing from  
64 the approach used by NeuroSynth (**Yarkoni et al.**, 2011) where neurophysiological measurements are  
65 primarily extracted from data tables, as well as manual curation, leveraging the background knowledge

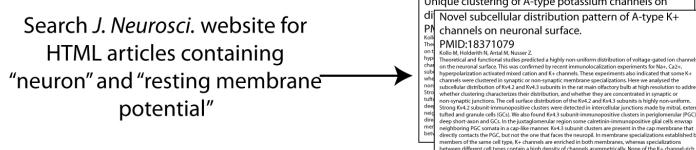
66 of domain experts. The resulting neurophysiology database, named NeuroElectro, can be interactively  
 67 viewed and explored through a public web interface at <http://neuroelectro.org>.

## 2 MATERIALS, METHODS, & RESULTS

### 2.1 OVERVIEW

68 We describe and validate our semi-automated methodology for obtaining neuronal biophysical  
 69 measurements directly from published reports in the literature (summarized in Fig. 1). After obtaining  
 70 full article texts from publishers, we then used text-mining algorithms to identify concepts specific to  
 71 electrophysiology and neuron types, which we then validated manually.

#### 1. Obtain full texts of relevant articles



#### 2. Find articles containing data tables

The diagram shows a search query on the left: "Look for data tables by finding full texts containing html <table> tags". An arrow points from this query to a table on the right. The table has three columns: "RMP (mV)", "RS Cell", and "FS Cell". The data is as follows:

	RS Cell	FS Cell
RMP (mV)	-65 +/- 2	-70 +/- 1
AP threshold (mV)	-45 +/- 1	-50 +/- 1
Tau (ms)	20 +/- 5	45 +/- 9

#### 3. Map concepts and extract values from data table

- 
- The diagram shows a search query on the left: "Look for data tables by finding full texts containing html <table> tags". An arrow points from this query to a list of five steps on the right:
1. Electrophysiology concept mapping  
"RMP (mV)" -> resting membrane potential  
(fuzzy-string matching against electrophysiology property synonym lists)
  2. Neuron concept mapping  
"RS Cell" -> Neocortex pyramidal cell layer  
2-3  
(usually done manually, new neuron types added when necessary)
  3. Data value mapping  
"-65 +/- 2" -> mean: -65  
error: 2
  4. Manual validation of concept mapping and data extraction
  5. Addition of extracted data to Neuro-Electro database

**Figure 1.** Illustration of workflow for obtaining electrophysiological information from the research literature.

### 2.2 ARTICLE IDENTIFICATION

72 We obtained electrophysiological data from 10 neuroscience specific journals (Table 1), which include:  
 73 *Journal of Neuroscience*, *Journal of Neurophysiology*, and *Journal of Physiology* (among others). We  
 74 selected these journals because they often devote a significant fraction of an article's main text, tables,  
 75 and figures to detailed characterizations and summaries of intrinsic neuronal biophysical properties.

76 We obtained tens of thousands of potentially relevant full article texts directly from publisher websites.  
 77 We first identified potential articles that were likely to contain information relevant to neuron biophysics  
 78 using the native search functions provided within the journal websites and only downloaded articles  
 79 containing in their full text any of a specific list of terms including "input resistance" and "resting

**Table 1.** Statistics of journals represented in the NeuroElectro database. Listing of journals and counts of articles downloaded (articles obtained), articles with published data tables containing neurophysiological information which has been manually validated by an expert curator (validated), and articles which likely contain information in a data table which has not yet been manually curated (not validated). Not validated articles are those which have at least 4 algorithmically assigned electrophysiological concepts within data tables.

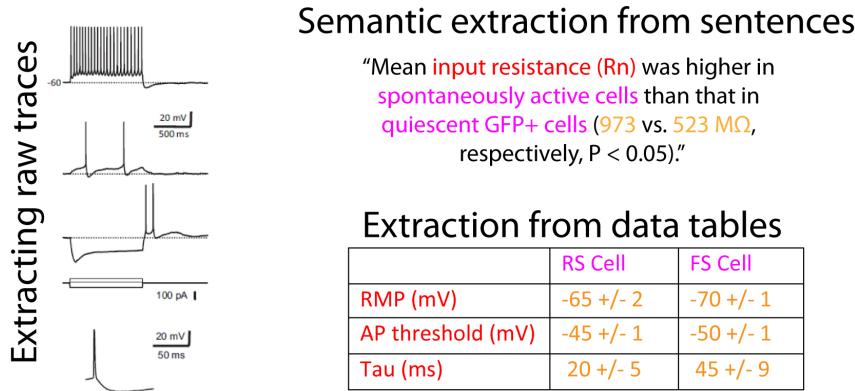
Journal	Articles obtained	Validated	Not validated
J. Neurosci.	19002	104	560
J. Neurophysiol.	12078	94	555
J. Physiol. (Lond.)	10543	44	235
Neuroscience	3035	14	205
Eur. J. Neurosci.	2495	7	117
Brain Res.	3017	7	146
Neuron	1657	4	43
Epilepsia	463	2	23
Neurosci. Lett.	1468	2	34
Hippocampus	208	2	10

80 membrane potential” (Fig. 1). This pre-selection step allowed us to identify and download only articles  
 81 that contained data relevant to our project. Upon identifying candidate articles, we then downloaded the  
 82 full text of each potentially-relevant article as HTML; articles downloaded from the publisher Elsevier  
 83 (e.g., *Neuron* and *Brain Research*) were downloaded as XML using the provided text-mining API and  
 84 subsequently converted to HTML. We chose to work with HTML (as opposed to PDF or XML) because  
 85 HTML provides a machine-readable markup of the article’s content, allowing us easily to identify relevant  
 86 elements within the article – such as data tables and the methods section – using publicly available  
 87 HTML-parsing tools (here we used the Beautiful Soup HTML-processing library implemented in Python:  
 88 <http://www.crummy.com/software/BeautifulSoup/bs4/doc/>). Furthermore, because  
 89 HTML is a single semi-structured standard used across publishers, we could write relatively generic  
 90 HTML-processing algorithms applicable to content published across journals. Our focus on using HTML  
 91 limits us to relatively newer articles - typically those published after 1996 - because before this time  
 92 most publications are only available as scanned PDF files. However, because the rate of publication  
 93 across the field has grown exponentially, this HTML-available subset constitutes the majority of published  
 94 neuroscience articles.

95 We stored the HTML-enhanced full text of each article in our database and associated each article  
 96 with its corresponding PubMed ID (<http://www.ncbi.nlm.nih.gov>). These 8-digit IDs serve as  
 97 publisher-independent unique identifiers for each article, and allow us to use PubMed-specific tools, such  
 98 as a powerful API (i.e., PubMed eutils, <http://www.ncbi.nlm.nih.gov/books/NBK25500/>).  
 99 For example, this API provides the ability to query each article’s MeSH terms (MEdical Subject Headings)  
 100 and returns basic methodological information such as animal species and strain.

## 2.3 ELECTROPHYSIOLOGICAL PROPERTY IDENTIFICATION

101 **2.3.1 Rationale for focusing on electrophysiological property extraction from data tables.** In order to  
 102 algorithmically extract information on neuron electrophysiology from these articles, we needed to first  
 103 specify the data types of interest. Our preference was to obtain as much detailed information about neuron  
 104 electrophysiological properties as possible: ideally, this would include raw data corresponding to recorded  
 105 electrophysiological traces. In mining information from articles, we were presented with multiple options  
 106 (illustrated in Fig. 2), including extraction from: 1) the text of the article including figure captions, 2)



**Figure 2.** Illustration of the sources within an article containing information relevant to neuron electrophysiological properties. Data on neuronal electrophysiological properties are presented within article figures and raw traces, sentences within the article text, and formatted data tables. The raw traces and references are from van Brederode et al. (2011) and the data table is a constructed example. Colored text indicates electrophysiological concepts (red), neuron concepts (pink), or neurophysiological data (yellow).

107 the figures of the article, or 3) data tables presented within the article. In addition to these, authors often  
108 submit supplemental materials and figures which also contain neurophysiological data.

109 Given the challenges in mining raw electrophysiological traces from figure images, we instead focused  
110 on obtaining information about basic neuronal electrophysiological properties, such as input resistances  
111 and resting membrane potentials. Though this information is often presented within the text of the article,  
112 it is usually presented in complex sentence structures that are difficult to accurately parse algorithmically.  
113 Published data tables, on the other hand, present a unique opportunity for electrophysiological data  
114 extraction, since common techniques exist for extracting information from structured tables (Yarkoni  
115 et al., 2011). Moreover, because tables succinctly summarize multiple attributes of a collected dataset, the  
116 effort of an expert curator can be put to best use when validating tables relative to validating content mined  
117 from article sentences or figure panels. While we estimate that only 5-10% of electrophysiology articles  
118 contain data tables, there is sufficient redundancy within the field (i.e., multiple investigators often publish  
119 articles on the same neuron type) that focusing on data tables nevertheless yields substantial coverage of  
120 electrophysiological properties across many major neuron types.

121 **2.3.2 Extracting information on electrophysiological properties.** In extracting electrophysiological  
122 data, we took advantage of the fact that certain measurements are commonly made during intracellular  
123 recordings. For example, such recordings are commonly used to: 1) measure a neuron's resting membrane  
124 potential, 2) apply hyperpolarizing current injections for measurement of input resistance and membrane  
125 time constant, and 3) apply depolarizing current steps to evoke action potentials (spikes) and enable  
126 measurement of characteristics such as spike threshold, width, and amplitude.

127 We developed an electrophysiological lexicon comprising 28 measurements that we found to be  
128 commonly reported in the literature, largely based on previously published definitions (Toledo-Rodriguez  
129 et al., 2004; Ascoli et al., 2008). To account for subtle differences in terminology that authors use to refer  
130 to the same electrophysiological concept (e.g., resting membrane potential is often referred to as "rmp")

131 and “ $V_{rest}$ ”), we also identified a common list of synonyms to map to each concept. Together, these  
132 electrophysiological concepts and their synonyms define a preliminary ontology for electrophysiological  
133 concepts (included in Supplemental Materials). Moreover, this physiological measurement ontology can  
134 serve as a scaffolding for a more in-depth ontology of electrophysiological investigations (e.g., Ontology  
135 for Experimental Neurophysiology, **Bruha et al.** (2013)). The terms in our preliminary ontology are also  
136 indexed and defined within NeuroLex (<http://neurolex.org>, **Larson and Martone**, 2013)).

137 To identify data corresponding to electrophysiological properties reported within a data table,  
138 we developed algorithms to search data table header elements and assess whether these elements  
139 corresponded to any of the electrophysiological concept synonyms in our ontology. We first identified  
140 table header elements by searching for table elements composed primarily of non-numeric characters.  
141 For each putative header element, we then used fuzzy string matching algorithms (implemented using  
142 the fuzzywuzzy library in Python: <https://github.com/seatgeek/fuzzywuzzy>), to assess  
143 the textual match between the header element and each of the electrophysiological synonyms. These  
144 fuzzy matching algorithms combine a number of string match metrics into a single “match value”,  
145 including whether a pair of strings completely match, contain matching substrings, or contain matching  
146 but misordered substrings. If the table header and electrophysiological synonym match value exceeded  
147 a specified threshold, the table header and corresponding row or column of numeric values were  
148 automatically mapped to the electrophysiological concept. Similarly, we mapped whole rows or columns  
149 to specific neuron types recorded during normotypic or “wild-type” conditions.

150 We then manually corrected cases where these algorithms misassigned an electrophysiological concept.  
151 For example, a common algorithmic mis-assignment was the case when an author used the string “EPSP  
152 amplitude” to refer to the electrophysiological concept excitatory post-synaptic potential amplitude. In  
153 these cases, our algorithms incorrectly mapped this string to “spike amplitude” because the former concept  
154 is not in our current ontology. In a test sample of 279 articles that were manually curated, we found that  
155 78% of concept-matchings (901/1152) were identified correctly with no supervision, with the remainder  
156 manually corrected.

157 2.3.3 *Accounting for differences in electrophysiological definitions across investigators.* By focusing  
158 on textually matching the electrophysiological terms in each table to a list of electrophysiological  
159 concepts, we are implicitly assuming that electrophysiological properties are measured in the same way  
160 by investigators across different articles. For example, the most common method that electrophysiologists  
161 use to measure a neuron’s spike properties is to record from the neuron in current-clamp mode and apply  
162 peri-threshold depolarizing currents to evoke 1-2 spikes over several hundred milliseconds or more. The  
163 neuron’s spike amplitude is then commonly measured by calculating the difference between the neuron’s  
164 voltage at spike threshold and spike peak for the first evoked spike (e.g. (**Connors et al.**, 1982; **Toledo-**  
165 **Rodriguez et al.**, 2004)). However, experimental differences exist between how investigators measure and  
166 compute these properties; we divide these differences into roughly 3 categories: *protocol*, *calculation*, and  
167 *condition* differences. For example, investigators can use different experimental protocols to measure the  
168 spike amplitude, like evoking spikes using current steps much greater than rheobase current required to  
169 elicit a single spike (*protocol differences*). Additionally, the spike amplitude itself can be calculated in  
170 different ways, such as using the neuron’s resting membrane potential as the baseline instead of the spike  
171 threshold (*calculation differences*). Furthermore, the value of spike amplitude that an investigator reports  
172 will also be affected by specific experimental conditions such as the animal species or age and recording  
173 solution temperature or contents (*condition differences*).

174 When manually curating the text-mined content for some of the most commonly reported  
175 electrophysiological properties, we accounted for an investigator’s calculation of an electrophysiological  
176 measurement using an inconsistent methodology (e.g., protocol or calculation differences). We did so  
177 by normalizing such measurements to a common reference definition or removing such data when  
178 normalization was not possible. However we note that we could not identify all of these cases (in  
179 particular: spike amplitude, input resistance, and membrane time constant), in part because investigators  
180 did not always explicitly define how these measurements were calculated within their article. We note that

A

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

	+/+	stg/stg	P
$V_{\text{r}}$ , mV	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)	ns
$R_{\text{in}}$ , MΩ	170 ± 25 (20)	170 ± 13 (22)	ns
Time constant, ms	26.9 ± 2.6 (14)	32.1 ± 3.1 (22)	ns
AP overshoot, mV	37.0 ± 3.7	34.1 ± 3.14 [11–58] (23)	ns

B

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

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

**Figure 3.** Example data table illustrating mark-up and annotation of entities. A. Example published data table containing neurophysiological information. Data table from **Pasquale et al.** (1997). B. Same as A, but semantically marked up with algorithmic and manually curated annotations. Markups in red and pink indicate electrophysiological and neuron type concepts and yellow indicates extracted data measurements. Note that here the textual string “+/+” and “stg/stg” refers to the normotypic and manipulated condition, respectively.

181 in cases where we pool measurements which are measured using inconsistent protocols or calculations,  
 182 this will tend to add unexplained variance to our data set. Given these measurement inconsistencies, we  
 183 provide our recommendations for how these electrophysiological properties should be reported in future  
 184 investigations via our electrophysiology ontology (see Supplemental Materials).

## 2.4 NEURON TYPE IDENTIFICATION

185 **2.4.1 Using neuron types defined by NeuroLex.** To extract physiological information specific to  
 186 individual neuron types, we had to identify which neuron types were reported in each article. However,  
 187 in many cases uniquely identifying the neuron type(s) reported in any given study and mapping these to a  
 188 canonical “neuron type” is difficult. This difficulty arises in part because investigators use different criteria  
 189 for classifying neurons, including electrophysiological, morphological, or molecular characteristics  
 190 (**Ascoli et al.**, 2008; **Huang and Zeng**, 2013; **Fishell and Heintz**, 2013).

191 To define canonical neuron types, we chose to use an existing list of approximately 250 neuron types  
 192 and definitions provided by NeuroLex, a community-sourced, expert-defined collection of neuron types  
 193 (<http://neurolex.org>; **Shepherd** (2003); **Hamilton et al.** (2012); **Larson and Martone** (2013)).  
 194 Moreover, we chose to use NeuroLex to keep our database consistent with existing resources and to enable  
 195 future researchers to combine these resources seamlessly. NeuroLex also provides synonyms for each  
 196 neuron type, which we utilized to identify the neuron type(s) in each article. In cases where a neuron type  
 197 was investigated in the literature across multiple articles but not indexed within NeuroLex (e.g. cerebellar  
 198 nucleus neurons), we manually added this neuron type to our database’s listing and provided this neuron  
 199 type to the NeuroLex neuron curators for incorporation (Gordon Shepherd, personal communication).  
 200 Our specific criteria for identifying each of the neuron types reflected in the database are given in the  
 201 Supplemental Materials.

202 2.4.2 *Identifying specific neuron types within an article.* Because of the complexity in unambiguously  
203 identifying neuron types, we used a mixed text-mining and manual approach to map the neuron types  
204 studied in each article to canonical NeuroLex neuron types. First, we used text-mining algorithms to  
205 provide an initial “best guess” of the most likely neuron type. Specifically, we used a bag-of-words  
206 approach (Aldous, 1985) on the full article text. This approach ignores the serial structure of the words  
207 in the document and utilizes only the frequency of occurrence of each word within the document. We  
208 next compared the article’s word-frequency histogram to the listing of neuron synonyms provided by  
209 NeuroLex, ranking all neuron types by their likelihood of being actually studied within that article.  
210 In comparison to articles that we manually curated, we found that this automated approach accurately  
211 identified the neurons studied in each article with an accuracy of 30% (120 of 399 total) and up to 55%  
212 when defining success as the studied neuron appearing as one of the top three neuron types suggested  
213 by the bag-of-words method. Because of the relatively low accuracy of an automated-only approach,  
214 we added a manual curation step where a curator identified the recorded neuron type using HTML drop  
215 down menus enriched by the bag-of-words search (e.g., Fig. 4). As previously described, we mapped  
216 individual data table elements and corresponding rows or columns to specific neuron types recorded  
217 under normotypic conditions. We note that currently we only identify data from normotypic or “control”  
218 neurons represented in tables, but plan to identify data from additional conditions in future work (e.g.  
219 from pharmacologically manipulated or genetically modified animals).

## 2.5 EXTRACTION OF ELECTROPHYSIOLOGICAL DATA VALUES

220 After identifying specific electrophysiological properties and neuron types reported in a data table  
221 (corresponding to row or column table headers), we then algorithmically extracted the data corresponding  
222 to the table intersection of these (Fig. 3). We developed custom string regular expressions (Thompson,  
223 1968) to parse the string corresponding to the numeric data. Specifically, we found that data strings were  
224 often of the form: “XX ± YY (ZZ)”, where XX, YY, and ZZ refer to the mean, error term, and sample  
225 size (i.e. the “n”), respectively. Often the number of replicates or error measurement were not reported  
226 or were reported in alternative ways within the table. Presently, the error term is not resolved as either a  
227 standard deviation or standard error measurement in the current version of NeuroElectro, but could easily  
228 be resolved in future iterations.

229 When designing our processing algorithms, we parsed data strings from right to left: first searching for  
230 data entities contained within parentheses, then for entities contained to the right of the ± term, and finally  
231 the remaining term which we assumed to refer to the mean term. We found that occasionally data were  
232 reported as “XX (LL - HH)” – where LL and HH indicate the lower and upper limits of a data range – and  
233 accounted for these cases similarly. We used regular expressions to identify entities such as digits, decimal  
234 signs, parentheses, and ± signs. We then converted the individual data elements which were encoded as  
235 textual strings of digits to double precision decimal entities before storing these into our database. Our  
236 focus here was primarily on parsing the mean value from a data record (i.e., summarizing the properties  
237 of a number of recorded neurons), but we also extracted and stored the error term and sample size where  
238 possible. Using these methods, we were able to extract 2176 electrophysiological values for 93 distinct  
239 neuron types within 279 articles.

## 2.6 MANUAL VALIDATION OF AUTOMATED DATA EXTRACTION

240 Following these automated concept identification and data extraction steps, we manually validated  
241 associated concepts and corrected incorrect concept mappings as necessary. We developed custom-HTML  
242 and javascript code to allow human curators to graphically interact with downloaded HTML data tables  
243 and “mark-up” entities within the table (Fig. 4). This code allows for textual based elements of the HTML  
244 table to be semantically annotated using drop down menus and text fields. Moreover, because annotation  
245 is implemented via user interfaces composed of interactive web pages and drop down menus, these user  
246 interfaces are simple enough to be utilized by other expert curators with little formal instruction.

**Table 2.** A partial listing of metadata attributes and extraction methodology. Metadata attributes are extracted through combining PubMed Medical Subject Heading terms (MeSH Terms) and custom regular expressions (Regex). Regular expression column (or MeSH Term column) indicates specific regular expressions (or MeSH terms) used for identifying metadata concept entities.

Metadata Concept	Values	Extraction method	Regular Expression	MeSH Term
Species	Rats Mice	MeSH Term only		Rats Mice
Electrode Type	Guinea Pigs Patch-clamp sharp	MeSH Term + Regex	"whole cell" or "patch clamp" "sharp electrode"	Guinea Pigs Patch-Clamp Techniques
Animal Strain	Fischer 344 Long-Evans Sprague-Dawley Wistar C57BL BALB C	MeSH Term only		Rats, Inbred F344 Rats, Long-Evans Rats, Sprague-Dawley Rats, Wistar Mice, Inbred C57BL Mice, Inbred BALB C
Prep Type	in vitro in vivo cell culture model	MeSH Term+Regex	"slice" or "in vitro" "in vivo" "culture" "model"	Cell Culture Techniques Computer Simulation
Jxn Potential	Not Corrected Corrected	Regex	"not junction potential" "junction potential"	
Rec Temperature	Continuous value Room temperature	Regex	"record ... C" or "experiment C" "record room temperature"	
Animal Age	Continuous value	Regex	Find digits near: "P#-#" or "P#-P#"	

## 2.7 METADATA IDENTIFICATION

247 Given the strong relationships between experimental conditions, such as animal species or recording  
 248 temperature, and electrophysiological measurements (e.g., input resistances are known to decrease when  
 249 measured in neurons from older animals (**Zhu**, 2000; **Okaty et al.**, 2009; **Kinnischtzke et al.**, 2012)),  
 250 we also identified information on article-specific experimental conditions by extracting this information  
 251 primarily from each article's methods section. For each article, we found the methods section by  
 252 developing custom HTML tag filters for each journal (e.g., common publisher-defined HTML tags for  
 253 methods sections are "Methods" or "Experimental procedures"). For each metadata entity that we focused  
 254 on (species, animal strain, electrode type, preparation type, liquid junction potential correction, animal  
 255 age, recording temperature), we devised custom automated text searching methods to identify these  
 256 based on combining regular expressions (**Thompson**, 1968) with PubMed MeSH terms (Table 2). In  
 257 other words, rather than taking a machine-learning based approach and training classifiers (**Mccallum**,  
 258 2002), we took a rule-based approach and developed custom rules for identifying metadata entities. For  
 259 example, to identify whether the recording electrode's liquid junction potential was corrected for in the  
 260 study (**Neher**, 1992), we searched for whether the character string "junction potential" was mentioned  
 261 within the methods section and, if so, whether the sentence or phrase containing the term was explicitly  
 262 negated (indicating that the junction potential was not corrected for). Here, we identified and parsed

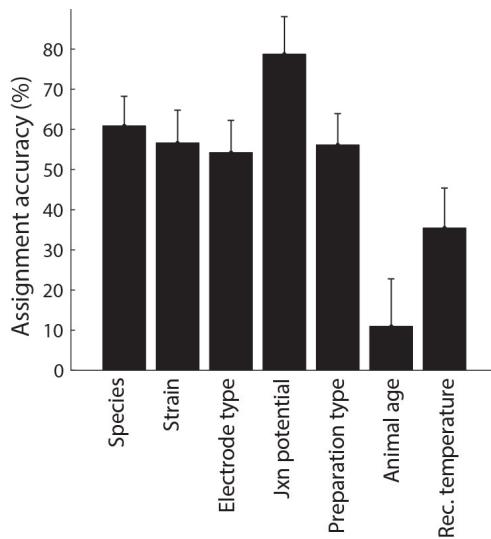
Table 1.

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

	+/ Concept: Neocortex pyramidal cell layer 5-6	stg/stg
V <sub>r</sub> , mV Concept: resting membrane potential	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)
R <sub>in</sub> , MΩ Concept: input resistance		
•Ephys Prop Neuron		
'input resistance' <input type="button" value="Submit"/>	170 ± 25 (20)	170 ± 13 (22)
'None selected' 'cell capacitance'		
'input resistance'	26.9 ± 2.6 (14)	32.1 ± 3.1 (22)
'resting membrane potential'		
'membrane time constant'		
'spike amplitude'		
'spike half-width'		
'spike threshold'		
'rheobase'		
'firing frequency'		
'AHP duration'		
'cell diameter'		
	[12–67] (14)	
		6.9 ± 1.1 (14)
		1.5 ± 0.7 (12)

**Figure 4.** Example of human validation of algorithmically assigned content. All textual elements of a table are enhanced using HTML and javascript to allow for assignment of neuron or electrophysiological concepts using drop down menus. Example data table from Pasquale et al. (1997).

263 distinct sentences within the methods section using tools provided within the Natural Language Tool Kit  
 264 in Python (Bird et al., 2009).

**Figure 5.** Accuracy of metadata assignment using automated methods alone. Error indicate 95% binomial confidence intervals.

265 Following automated identification of article metadata, we then manually checked each article to  
 266 ascertain that algorithmically-tagged metadata was identified correctly and, as before, we corrected  
 267 misidentified content as necessary through the use of custom HTML forms. We found that the mean  
 268 accuracy of algorithmic metadata assignment was approximately 50% (Fig. 5) and was typically lower  
 269 for identifying continuous-valued metadata (e.g. animal age or recording temperature) relative to nominal  
 270 metadata such as species and electrode type.

## 2.8 OBJECT MODELS AND RELATIONAL DATABASE

271 We stored extracted data and metadata using a relational database implemented in MySQL (<http://dev.mysql.com/doc/refman/5.6/en/>) built from a Python Django object model (<https://www.djangoproject.com/>). The object model contains classes for a number of fields, such as full  
272 article texts, electrophysiological properties, neuron types, synonyms, electrophysiological data values,  
273 and experimental metadata (Fig. S1). A useful feature of the relational nature of the database is that it  
274 enables linking between classes (e.g., linking between neuron types and electrophysiological properties  
275 reported by a single investigator across multiple articles). This linking feature facilitates efficient and  
276 arbitrary querying of data; for example, querying for known electrophysiological data on olfactory bulb  
277 mitral cells recorded *in vitro* and published between the dates 2000 and 2004. For example, such a feature  
278 could be used to assess whether measurements of olfactory bulb mitral cells have changed as a function  
279 of time or are dependent upon whether the data are collected *in vitro* or *in vivo*.  
280  
281

## 2.9 WEB APPLICATION

282 The primary results of NeuroElectro are viewable at <http://www.neuroelectro.org> where the  
283 data can be interactively explored.  
284

284 **2.9.1 Human interface** The web interface is organized around neuron types and electrophysiological  
285 properties. For example, each neuron type has its own webpage where extracted data corresponding  
286 to specific electrophysiological properties is graphically and interactively displayed (graphical plot  
287 interactivity implemented using the jqPlot javascript toolbox, <http://www.jqplot.com/>). Users  
288 can thus visualize the mean and variability of electrophysiological values across papers, view references  
289 plus experimental metadata, and easily navigate to primary data from specific papers. Furthermore, users  
290 can view electrophysiological data across all of the neuron types in the database - putting phenotypic  
291 properties of a given neuron type into the larger context of other neuron types located throughout the  
292 nervous system.  
293

293 The web application also contains preliminary features to allow website visitors to contribute to the  
294 NeuroElectro resource. For example, users can suggest articles which contain electrophysiological data  
295 which are not already in the database. We also invite visitors to become “expert curators” for neurons  
296 of interest. In the future, we plan to build functionality that will allow investigators to upload raw and  
297 summary data, such as recorded voltage and current traces. In addition, we plan to continue mining the  
298 literature and adding neurophysiological measurements as they are published.  
299

299 **2.9.2 API** An initial API (application programmer interface) providing public access to the  
300 electrophysiological data is described at <http://neuroelectro.org/api/docs/>. This RESTful  
301 API allows contents of the NeuroElectro database to be dynamically retrieved in JSON or XML format to  
302 utilize this data within external applications. For example, using the current API, a developer could build  
303 an application which dynamically queries NeuroElectro for all data corresponding to layer 2/3 neocortical  
304 pyramidal cells and then uses this data to constrain parameters for a Hodgkin-Huxley type neuron model  
305 (**Hodgkin and Huxley**, 1952). Example use cases of the current API (version 1) include:  
306

- 306 • <http://neuroelectro.org/api/1/n/> : Returns a list of all neurons with electrophysiological  
307 data indexed in NeuroElectro.
- 308 • <http://neuroelectro.org/api/1/nedm/?nlex=sao830368389> : Returns a list of all  
309 indexed data on CA1 pyramidal cells (queried using the NeuroLex identifier for CA1 pyramidal cells,  
310 *sao830368389*).
- 311 • [http://neuroelectro.org/api/1/nes/?e\\_name=Input%20Resistance&n\\_name=Cerebellum%20Purkinje%20Cell](http://neuroelectro.org/api/1/nes/?e_name=Input%20Resistance&n_name=Cerebellum%20Purkinje%20Cell) : Returns a data record composed of the mean, standard  
312

313 deviation, and sample size n, summarizing input resistance measurements from cerebellar Purkinje  
314 cells based on all indexed articles in NeuroElectro database. Here the database query is performed  
315 using the textual strings for the electrophysiological and neuron type concepts.

316 Our future plans are to work with domain ontologists to further develop the existing API into a formal  
317 relational data format (RDF) specification, allowing further querying and extending of NeuroElectro into  
318 additional resources. All code used in the project and the website is available at [http://bitbucket.org/rgerkin/neuroelectro\\_org](http://bitbucket.org/rgerkin/neuroelectro_org).

### 3 DISCUSSION

320 We have developed, applied, and validated a methodology and pipeline for extracting – from existing  
321 literature on cellular neurophysiology – measurements of basic biophysical properties from diverse neuron  
322 types throughout the nervous system. Currently, the NeuroElectro database contains 2344 manually  
323 curated electrophysiological measurements from 98 neuron types from 335 publications. Of these  
324 electrophysiological measurements, 2176 (93%) were obtained from 279 (83%) publications using the  
325 semi-automated approach described here. In addition, we machine-extracted and manually validated 1667  
326 methodological conditions (metadata) from these publications. This represents the single largest collection  
327 of neurophysiological data ever compiled and represents a potentially valuable tool for scientific discovery.

#### 3.1 SPECIFIC BENEFITS PROVIDED BY THE SEMI-AUTOMATED APPROACH

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

335 Given our laboratory's focus on olfactory circuits, we conducted a natural experiment to compare  
336 the efficacy of biophysical property extraction using these semi-automated methods versus traditional  
337 methods which do not make use of algorithmic text-mining as a pre-processing step. In a seven-hour  
338 curation session (evoking the classic American parable of John Henry versus the steam-powered hammer),  
339 a senior graduate student in our laboratory identified 91 electrophysiological measurements (focusing on  
340 resting membrane potential, input resistance, membrane time constant, spike amplitude, spike width, and  
341 spike threshold) from 35 articles for 7 olfactory bulb neuron types using only prior knowledge of which  
342 articles and investigators were likely to have reported such electrophysiological data. In a comparable  
343 seven-hour curation session using our semi-automated methods, a single curator (with similar expertise  
344 to the first curator) identified 551 electrophysiological measurements from 70 articles across 40 neuron  
345 types throughout the nervous system. Moreover, this comparison would likely tilt even more in favor of  
346 the semi-automated methods had the curators been less familiar with the primary literature.

#### 3.2 SCALABILITY OF CURRENT APPROACH

347 We note that multiple steps in our approach require manual intervention by an expert curator in order  
348 for electrophysiological measurements to be extracted with an acceptably low error rate. Namely, an  
349 expert curator needs to confirm which of the machine-identified candidate neuron types are recorded from  
350 in each article and where data from the normotypic or “control” states of these neurons are textually  
351 referenced within a data table. Moreover, given the current accuracy of the unsupervised algorithmic  
352 assignment of electrophysiological concepts and experimental metadata (78% and 50%, respectively),

353 these also need to be manually validated and corrected and normalized as required by an expert. Given the  
354 necessity of these manual steps, the scalability of our current approach is limited by our ability to manually  
355 curate this information or by our ability to improve the error rate of the automated methods. Despite this  
356 limitation, our current pipeline is still much faster than a purely manual one. The methodology could be  
357 further improved by correcting falsely matching entities (such as EPSP amplitude in section 2.3.2). These  
358 could be corrected by simply adding these valid concepts to the electrophysiological ontology. Moreover,  
359 these improvements would facilitate formally computing the sensitivity and specificity of these entity  
360 recognition methods.

### 3.3 PRELIMINARY USE OF NEUROELECTRO IN SCIENTIFIC WORK

361 The NeuroElectro project is intended to facilitate scientific investigation by providing easy access to large  
362 quantities of data about neurons. Because the data is machine-readable, we have already begun to conduct  
363 several analyses that would not be possible without this resource. First, we have begun an investigation  
364 of the relationships between neurons as defined by the similarity of their electrophysiological properties.  
365 This information can be used to make predictions about as yet unmeasured properties. Second, we have  
366 begun to explore the relationship between patterns of gene expression (using both the Allen Brain Atlas  
367 (**Lein et al.**, 2007) and single cell qPCR approaches) and electrophysiological properties of neurons.  
368 Third, we have begun automated testing of quantitative neuron models in concert with SciUnit (**Omar**  
369 **et al.**, 2014), under the reasonable assumption that these models should be constrained by the available  
370 experimental data. These projects are described in manuscripts currently in preparation.

### 3.4 EXTENSIONS AND IMPROVEMENTS TO THE CURRENT SEMI-AUTOMATED ALGORITHMS

371 Currently, neuron type identification is a critical bottleneck in our approach. One potential improvement  
372 would be to replace the nonspecific bag-of-words approach we are currently using in favor of a  
373 bioNLP classifier-based approach (**Mccallum**, 2002). Specifically, we propose adapting the named entity  
374 recognition methodology used by the WhiteText project for tagging brain regions mentioned in literature  
375 (**French et al.**, 2009; **French and Pavlidis**, 2012) and first identifying spans of text likely to pertain to a  
376 neuron type before mapping these textual spans to a individual neuron type within the neuron ontology.

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

383 Given the relatively low accuracy of the automated approach to identifying neuron types, there may be  
384 several avenues through which this process can be improved. For example, we note that the automated  
385 approach was particularly ineffective when the neuron type investigated within an article was not already  
386 described in NeuroLex or when the neuron had an insufficient list of synonyms associated with it.  
387 The current implementation of NeuroElectro also does not consider common neuron type acronyms  
388 (e.g., that olfactory bulb mitral cells are commonly referred to as “MCs”). Adding acronym and  
389 abbreviation identification to future iterations will thus likely improve the automated approach (**Okazaki**  
390 **and Ananiadou**, 2006; **French and Pavlidis**, 2012). Moreover, our current implementation of the bag-  
391 of-words algorithm would likely be enhanced via minor improvements, such as only identifying neurons  
392 using the text of the abstract or results and discarding text from the introduction or discussion. As neuron  
393 identification forms the major bottleneck in the scalability of NeuroElectro due to the requirement for  
394 manual curation, we plan to address this bottleneck in future revisions.

### 3.5 FUTURE METHODS FOR DATA EXTRACTION

395 A more pressing issue with the current approach is its focus on extraction from data tables. We estimate  
396 that only 5-10% of published electrophysiological data is contained within tables, while the remaining  
397 90-95% is presented within article text or figure images. Given our preference to obtain data in their  
398 most raw form, we initially considered extraction of data from figures, e.g. voltage traces of neuronal  
399 activity. However, digitizing article figures (presented by publishers as images) into a form that can  
400 be further analyzed presents multiple challenges. Though techniques and tools exist to digitize figures,  
401 substantial amounts of manual effort are required to employ them correctly, making this figure-based  
402 approach difficult to scale to increasing numbers of articles without also employing a large team of  
403 human curators. While automatically extracting measurements from figure images will likely prove  
404 challenging, our methods can likely be adapted to operate on article text, perhaps by making use of  
405 bioNLP methodologies currently used for relationship extraction in the identification of connected brain  
406 regions (**French et al.**, 2012) or interacting pairs of proteins (**Kim and Wilbur**, 2011).

407 Future developments in machine extraction of data from the scientific literature will be of great benefit.  
408 These should include better semantic understanding of context, ranging from relatively unambiguous  
409 notations such as units, to syntax-parsing of free-form prose that relates objects of study to their reported  
410 properties. Much progress has been made by computer scientists in some of these areas, and more future  
411 engagement with their research should enable vastly more data to be extracted from the literature.

412 We believe that, if successful, the use of NeuroElectro will influence the practices of scientists writing  
413 papers and reporting results. Specifically, we recommend the usage of common standards and definitions  
414 for basic physiological measurements (**Toledo-Rodriguez et al.**, 2004) and neuron types (**Ascoli et al.**,  
415 2008; **Larson and Martone**, 2013). Moreover, we advocate that, where possible, scientists report  
416 more basic physiological data overall and report such data using machine-parsable data tables. These  
417 recommendations could be made informally by journals (in particular, requested by reviewers during  
418 manuscript review) as well as by funding agencies. This change would make it easier for scientists to find  
419 and make use of data collected by others. Such a culture shift has the potential to make science function  
420 more effectively and efficiently to facilitate discovery.

## DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

421 The authors declare that the research was conducted in the absence of any commercial or financial  
422 relationships that could be construed as a potential conflict of interest.

## ACKNOWLEDGEMENT

423 We thank William Cohen, Gordon Shepherd, and Renaud Richardet for discussions and comments on  
424 the manuscript. We are especially grateful to all of the investigators whose collected data are represented  
425 within the NeuroElectro database. We thank the academic journal publishers (in particular, Elsevier and  
426 Wiley and Highwire) for allowing us access to their full-texts for text-mining. We are especially grateful  
427 to all of the investigators whose collected data are represented within the NeuroElectro database.

428 **Funding:** This work was supported by a National Science Foundation Graduate Research Fellowship  
429 and a R. K. Mellon Foundation Fellowship (to S.J.T.), an Achievement Rewards for College Scientists  
430 Foundation Fellowship and NIDCD NRSA F31DC013490 (to S.D.B.), NIDCD award F32DC010535 and  
431 NIMH award R01MH081905 (in support of R.C.G.), and NIDCD award R01DC005798 (to N.N.U.).

## REFERENCES

- 432 Aldous, D. J. (1985), Exchangeability and related topics, in P. L. Hennequin, ed., cole d't de Probabilit's  
433 de Saint-Flour XIII 1983 (Springer Berlin Heidelberg), number 1117 in Lecture Notes in Mathematics,  
434 1–198
- 435 Ambert, K. H. and Cohen, A. M. (2012), Text-mining and neuroscience, *International review of*  
436 *neurobiology*, 103, 109–132, doi:10.1016/B978-0-12-388408-4.00006-X, PMID: 23195123
- 437 Ascoli, G. A., Alonso-Nanclares, L., Anderson, S. A., Barrientos-Piccione, R., Burkhalter, A., et al. (2008), Petilla terminology: nomenclature of features of GABAergic  
438 interneurons of the cerebral cortex, *Nature Reviews Neuroscience*, 9, 7, 557–568, doi:10.1038/nrn2402,  
439 WOS:000256929300015
- 440 Bird, S., Klein, E., and Loper, E. (2009), Natural language processing with Python (O'Reilly, Beijing;  
441 Cambridge [Mass.])
- 442 Bota, M., Dong, H.-W., and Swanson, L. W. (2005), Brain architecture management system,  
443 *Neuroinformatics*, 3, 1, 15–48, doi:10.1385/NI:3:1:015, PMID: 15897615
- 444 Bruha, P., Papez, V., Bandrowski, A., Grewe, J., Mouek, R., Tripathy, S., et al. (2013), The ontology for  
445 experimental neurophysiology: a first step toward semantic annotations of neurophysiology data and  
446 metadata., *Frontiers in Neuroinformatics*, 7, doi:10.3389/conf.fninf.2013.09.00026
- 447 Connors, B. W., Gutnick, M. J., and Prince, D. A. (1982), Electrophysiological properties of neocortical  
448 neurons in vitro, *Journal of neurophysiology*, 48, 6, 1302–1320, PMID: 6296328
- 449 Dickman, S. (2003), Tough mining, *PLoS Biol*, 1, 2, e48, doi:10.1371/journal.pbio.0000048
- 450 Fishell, G. and Heintz, N. (2013), The neuron identity problem: Form meets function, *Neuron*, 80, 3,  
451 602–612, doi:10.1016/j.neuron.2013.10.035
- 452 French, L., Lane, S., Xu, L., and Pavlidis, P. (2009), Automated recognition of brain region mentions in  
453 neuroscience literature, *Frontiers in Neuroinformatics*, 3, 29, doi:10.3389/neuro.11.029.2009
- 454 French, L., Lane, S., Xu, L., Siu, C., Kwok, C., Chen, Y., et al. (2012), Application and evaluation of  
455 automated methods to extract neuroanatomical connectivity statements from free text, *Bioinformatics*,  
456 28, 22, 2963–2970, doi:10.1093/bioinformatics/bts542, PMID: 22954628
- 457 French, L. and Pavlidis, P. (2011), Relationships between gene expression and brain wiring in the adult  
458 rodent brain, *PLoS computational biology*, 7, 1, e1001049, doi:10.1371/journal.pcbi.1001049, PMID:  
459 21253556
- 460 French, L. and Pavlidis, P. (2012), Using text mining to link journal articles to neuroanatomical databases,  
461 *The Journal of Comparative Neurology*, 520, 8, 17721783, doi:10.1002/cne.23012
- 462 Hamilton, D. J., Shepherd, G. M., Martone, M. E., and Ascoli, G. A. (2012), An ontological approach  
463 to describing neurons and their relationships, *Frontiers in Neuroinformatics*, 6, 15, doi:10.3389/fninf.  
464 2012.00015
- 465 Hines, M. L., Morse, T., Migliore, M., Carnevale, N. T., and Shepherd, G. M. (2004), ModelDB: a  
466 database to support computational neuroscience, *Journal of Computational Neuroscience*, 17, 1, 7–11,  
467 doi:10.1023/B:J CNS.0000023869.22017.2e, PMID: 15218350
- 468 Hodgkin, A. L. and Huxley, A. F. (1952), A quantitative description of membrane current and its  
469 application to conduction and excitation in nerve, *The Journal of physiology*, 117, 4, 500–544, PMID:  
470 12991237
- 471 Huang, J. and Zeng, H. (2013), Genetic approaches to neural circuits in the mouse, *Annual Review of*  
472 *Neuroscience*, 36, 1, 183–215, doi:10.1146/annurev-neuro-062012-170307, PMID: 23682658
- 473 Kim, S. and Wilbur, W. J. (2011), Classifying protein-protein interaction articles using word and syntactic  
474 features, *BMC bioinformatics*, 12 Suppl 8, S9, doi:10.1186/1471-2105-12-S8-S9, PMID: 22151252
- 475 Kinnischtzke, A. K., Sewall, A. M., Berkepile, J. M., and Fanselow, E. E. (2012), Postnatal maturation of  
476 somatostatin-expressing inhibitory cells in the somatosensory cortex of GIN mice, *Frontiers in neural*  
477 *circuits*, 6, 33, doi:10.3389/fncir.2012.00033, PMID: 22666189
- 478 Larson, S. D. and Martone, M. E. (2013), NeuroLex.org: an online framework for neuroscience  
479 knowledge, *Frontiers in Neuroinformatics*, 7, 18, doi:10.3389/fninf.2013.00018
- 480

- 481 Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., et al. (2007), Genome-  
482 wide atlas of gene expression in the adult mouse brain, *Nature*, 445, 7124, 168–176, doi:10.1038/  
483 nature05453
- 484 McCallum, A. (2002), MALLET: a machine learning for language toolkit
- 485 Migliore, M., Morse, T. M., Davison, A. P., Marenco, L., Shepherd, G. M., and Hines, M. L. (2003),  
486 ModelDB, *Neuroinformatics*, 1, 1, 135–139, doi:10.1385/NI:1:1:135
- 487 Neher, E. (1992), [6] correction for liquid junction potentials in patch clamp experiments, in Bernardo  
488 Rudy, ed., Methods in Enzymology, volume Volume 207 (Academic Press), 123–131
- 489 Okaty, B. W., Miller, M. N., Sugino, K., Hempel, C. M., and Nelson, S. B. (2009), Transcriptional and  
490 electrophysiological maturation of neocortical fast-spiking GABAergic interneurons, *The Journal of  
491 Neuroscience*, 29, 21, 7040–7052, doi:10.1523/JNEUROSCI.0105-09.2009, PMID: 19474331
- 492 Okazaki, N. and Ananiadou, S. (2006), Building an abbreviation dictionary using a term recognition  
493 approach, *Bioinformatics*, 22, 24, 3089–3095, doi:10.1093/bioinformatics/btl534, PMID: 17050571
- 494 Omar, C., Aldrich, J., and Gerkin, R. (2014), Collaborative infrastructure for test-driven scientific model  
495 validation, *ICSE: New Innovations and Emerging Results*
- 496 Parekh, R. and Ascoli, G. (2013), Neuronal morphology goes digital: A research hub for cellular and  
497 system neuroscience, *Neuron*, 77, 6, 1017–1038, doi:10.1016/j.neuron.2013.03.008
- 498 Pasquale, E. D., Keegan, K. D., and Noebels, J. L. (1997), Increased excitability and inward rectification in  
499 layer V cortical pyramidal neurons in the epileptic mutant mouse stargazer, *Journal of Neurophysiology*,  
500 77, 2, 621–631, PMID: 9065835
- 501 Ramakrishnan, C., Patnia, A., Hovy, E., and Burns, G. A. (2012), Layout-aware text extraction from  
502 full-text PDF of scientific articles, *Source code for biology and medicine*, 7, 1, 7, doi:10.1186/  
503 1751-0473-7-7, PMID: 22640904
- 504 Shepherd, G. M., ed. (2003), The Synaptic Organization of the Brain (Oxford University Press, USA), 5  
505 edition
- 506 Teeter, C. and Stevens, C. (2011), A general principle of neural arbor branch density, *Current Biology*,  
507 doi:10.1016/j.cub.2011.11.013
- 508 Thompson, K. (1968), Programming techniques: Regular expression search algorithm, *Commun. ACM*,  
509 11, 6, 419422, doi:10.1145/363347.363387
- 510 Toledo-Rodriguez, M., Blumenfeld, B., Wu, C., Luo, J., Attali, B., Goodman, P., et al. (2004), Correlation  
511 maps allow neuronal electrical properties to be predicted from single-cell gene expression profiles in  
512 rat neocortex, *Cerebral Cortex*, 14, 12, 1310–1327, doi:10.1093/cercor/bhh092
- 513 van Brederode, J. F. M., Yanagawa, Y., and Berger, A. J. (2011), GAD67-GFP+ neurons in the nucleus  
514 of the raphe: a possible source of inhibitory input to hypoglossal motoneurons. i. morphology and firing  
515 properties, *Journal of neurophysiology*, 105, 1, 235–248, doi:10.1152/jn.00493.2010, PMID: 21047932
- 516 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011), Large-scale  
517 automated synthesis of human functional neuroimaging data, *Nat Meth*, 8, 8, 665–670, doi:10.1038/  
518 nmeth.1635
- 519 Zhu, J. J. (2000), Maturation of layer 5 neocortical pyramidal neurons: amplifying salient layer 1 and  
520 layer 4 inputs by Ca<sup>2+</sup> action potentials in adult rat tuft dendrites, *The Journal of physiology*, 526 Pt 3,  
521 571–587, PMID: 10922009
- 522

## SUPPLEMENTAL DATA

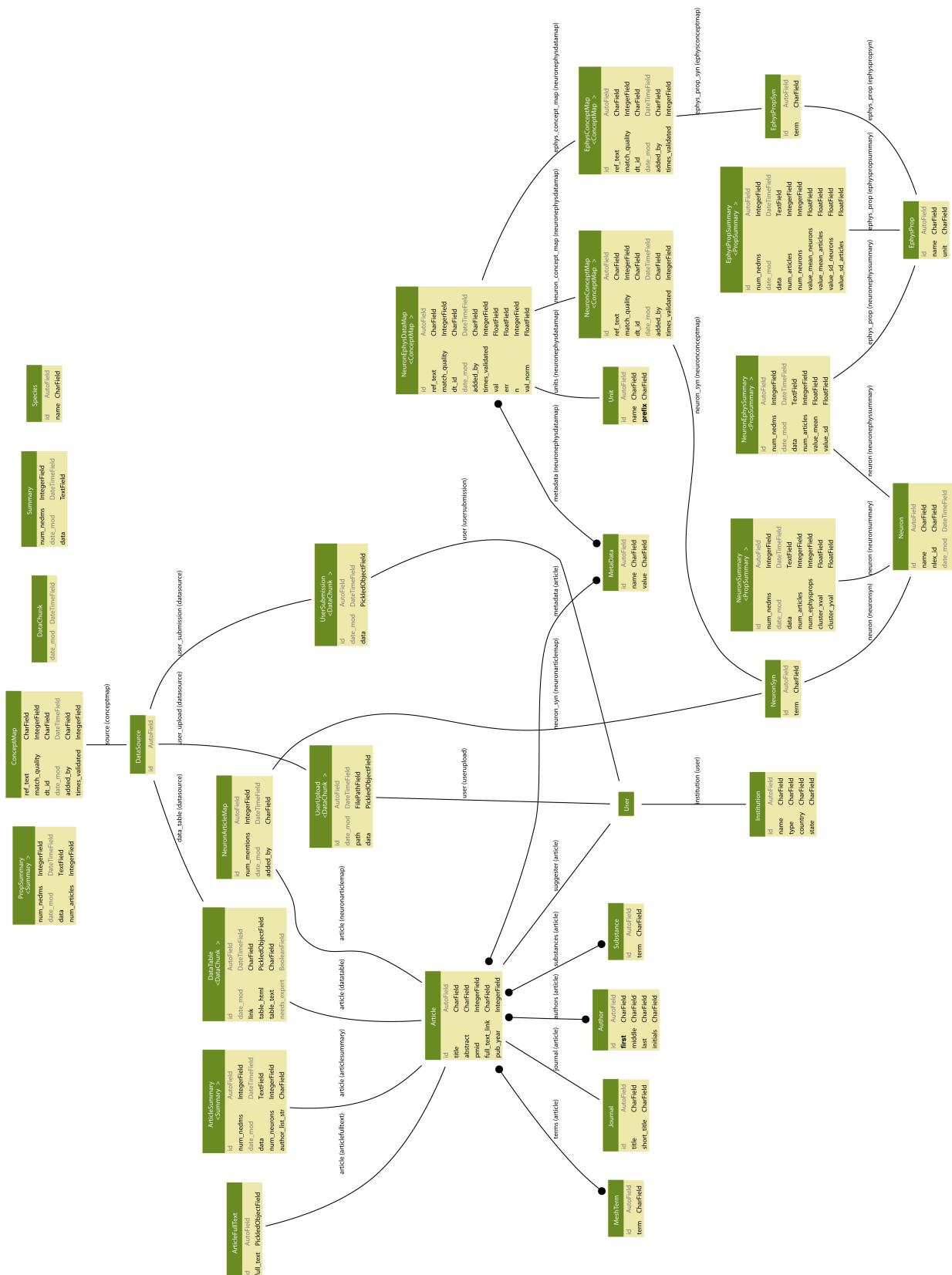


Fig. S1: Illustration of NeuroElectro relational database schema