



1

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

Shreejoy J. Tripathy^{1,2,3,4,*}, Judith Savitskaya^{1,6}, Shawn D. Burton^{1,2}, Nathaniel N. Urban^{1,2}, and Richard C. Gerkin^{1,2,5,*}

¹Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

²Center for the Neural Basis of Cognition, Pittsburgh, PA, USA

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

⁴Present Address: Department of Psychiatry, University of British Columbia, BC, Canada

⁵Present Address: School of Life Sciences, Arizona State University, Tempe, AZ, USA

⁶Present 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 looking to supplement their own data, for computational modelers looking to constrain their model parameters, and for theoreticians searching for undiscovered relationships among neurons and their properties.

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 morphology of neuron types (**Teeter and Stevens**, 2011).
49 Similarly, NeuroSynth is widely used by cognitive scientists as a starting point for designing functional
50 imaging studies. Thus while these projects are not yet comprehensive and likely contain data records of
51 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

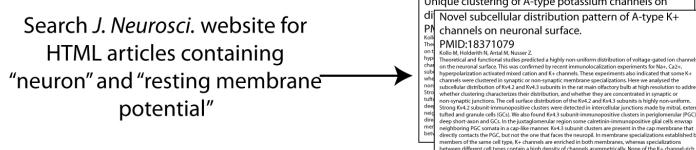
66 of domain experts. The resulting neurophysiology database, named NeuroElectro, can be interactively
 67 viewed and explored through a public web interface at <http://neuroelectro.org>.

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

A screenshot of a search results page from the *J. Neurosci.* website. The search query is "neuron" and "resting membrane potential". The results list several articles, with one article highlighted in yellow. The highlighted article is titled "Unique clustering of A-type potassium channels on..." and has a PMID of 18371079.

	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

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 a specific list of terms including "input resistance" and "resting membrane potential" (Fig. 1).

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 a data table.

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

93
 94 We stored the HTML-enhanced full text of each article in our database and associated each article
 95 with its corresponding PubMed ID (<http://www.ncbi.nlm.nih.gov>). These 8-digit IDs serve as
 96 publisher-independent unique identifiers for each article, and allow us to use PubMed-specific tools, such
 97 as a powerful API (i.e., PubMed eutils, <http://www.ncbi.nlm.nih.gov/books/NBK25500/>).
 98 For example, this API provides the ability to query each article's MeSH terms (MEDical Subject Headings)
 99 and returns basic methodological information such as animal species and strain.

2.3 ELECTROPHYSIOLOGICAL PROPERTY IDENTIFICATION

100 **2.3.1 Rationale for focusing on electrophysiological property extraction from data tables.** In order to
 101 algorithmically extract information on neuron electrophysiology from these articles, we needed to first
 102 specify the data types of interest. Our preference was to obtain as much detailed information about neuron
 103 electrophysiological properties as possible: ideally, this would include raw data corresponding to recorded
 104 electrophysiological traces. In mining information from articles, we were presented with multiple options
 105 (illustrated in Fig. 2), including extraction from: 1) the text of the article including figure captions, 2)
 106 the figures of the article, or 3) data tables presented within the article. In addition to these, authors often
 107 submit supplemental materials and figures which also contain neurophysiological data.

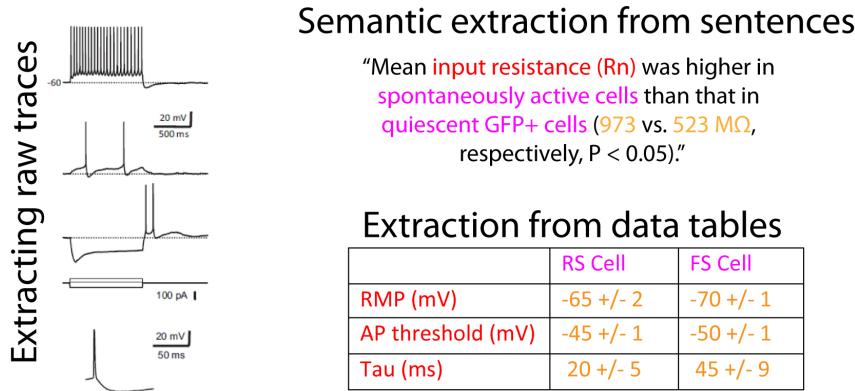


Figure 2. Illustration of the sources within an article containing information relevant to neuron electrophysiological properties. Data on neuronal electrophysiological properties are presented within article figures and raw traces, sentences within the article text, and formatted data tables. The raw 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).

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

120 2.3.2 *Extracting information on electrophysiological properties.* In extracting electrophysiological
 121 data, we took advantage of the fact that certain measurements are commonly made during intracellular
 122 recordings. For example, such recordings are commonly used to: 1) measure a neuron's resting
 123 membrane potential, 2) to apply hyperpolarizing current injections for measurement of input resistance
 124 and membrane time constant, and 3) to apply depolarizing current steps to evoke action potentials (spikes)
 125 and enable measurement of characteristics such as spike threshold, width, and amplitude.

126 We developed an electrophysiological lexicon comprising 28 measurements that we found to be
 127 commonly reported in the literature, largely based on previously published definitions (Toledo-Rodriguez
 128 et al., 2004; Ascoli et al., 2008). To account for subtle differences in terminology that authors use to refer
 129 to the same electrophysiological concept (e.g., resting membrane potential is often referred to as "rmp"
 130 and " V_{rest} "), we also identified a common list of synonyms to map to each concept. Together, these
 131 electrophysiological concepts and their synonyms define a preliminary ontology for electrophysiological
 132 concepts (included in Supplemental Materials). Moreover, this physiological measurement ontology can

A

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

	+/+	stg/stg	P
V_r , mV	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)	ns
R_{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

	+/ <i>Concept: Neocortex pyramidal cell layer 5-6</i>	stg/stg	P
V_r , mV <i>Concept: resting membrane potential</i>	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)	ns
R_{in} , MΩ <i>Concept: input resistance</i>	170 ± 25 (20)	170 ± 13 (22)	ns
Time constant, ms <i>Concept: membrane time constant</i>	26.9 ± 2.6 (14)	32.1 ± 3.1 (22)	ns
AP overshoot, mV <i>Concept: spike overshoot</i>	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.

133 serve as a scaffolding for a more in-depth ontology of electrophysiological investigations (e.g., Ontology
 134 for Experimental Neurophysiology, Bruha et al. (2013)). The terms in our preliminary ontology are also
 135 indexed and defined within NeuroLex (<http://neurolex.org>, (Larson and Martone, 2013)).

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

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

156 2.3.3 *Accounting for differences in electrophysiological definitions across investigators.* By focusing
157 on textually matching the electrophysiological terms in each table to a list of electrophysiological
158 concepts, we are implicitly assuming that electrophysiological properties are measured in the same way
159 by investigators across different articles. For example, the most common method that electrophysiologists
160 use to measure a neuron's spike properties is to record from the neuron in current-clamp mode and apply
161 peri-threshold depolarizing currents to evoke 1-2 spikes over several hundred milliseconds or more. The
162 neuron's spike amplitude is then commonly measured by calculating the difference between the neuron's
163 voltage at spike threshold and spike peak for the first evoked spike (e.g. (Connors et al., 1982; Toledo-
164 Rodriguez et al., 2004)). However, experimental differences exist between how investigators measure and
165 compute these properties; we divide these differences into roughly 3 categories: *protocol*, *calculation*, and
166 *condition* differences. For example, investigators can use different experimental protocols to measure the
167 spike amplitude, like evoking spikes using current steps much greater than rheobase current required to
168 elicit a single spike (*protocol differences*). Additionally, the spike amplitude itself can be calculated in
169 different ways, such as using the neuron's resting membrane potential as the baseline instead of the spike
170 threshold (*calculation differences*). Furthermore, the value of spike amplitude that an investigator reports
171 will also be affected by specific experimental conditions such as the animal species or age and recording
172 solution temperature or contents (*condition differences*).

173 When manually curating the text-mined content for the most commonly reported electrophysiological
174 properties (resting membrane potential, input resistance, membrane time constant, spike half-width,
175 spike amplitude, and spike threshold), we took care to account for where the investigator had calculated
176 an electrophysiological measurement using an inconsistent methodology (e.g., protocol or calculation
177 differences) by normalizing such measurements to a common reference definition or removing such data
178 when 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
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 data record mean value (i.e., summarizing the properties of a
237 number of recorded neurons), but also extracted and stored the error term and sample size where possible.
238 Using these methods, we were able to extract 2176 electrophysiological values for 93 distinct neuron
239 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 fixed 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

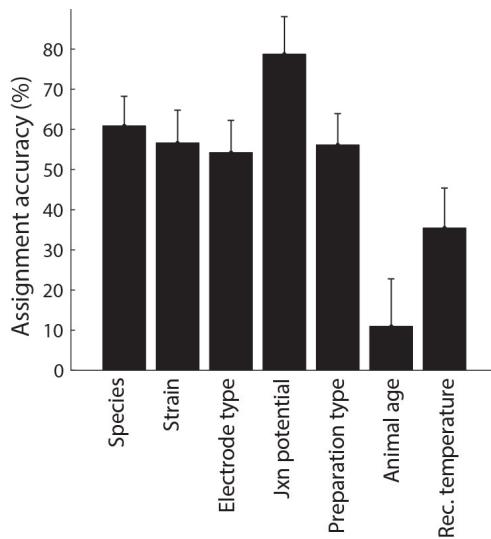
Table 1.

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

	+/ Concept: Neocortex pyramidal cell layer 5-6	stg/stg
V _r , mV Concept: resting membrane potential	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)
R _{in} , 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. 6). 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 the properties of olfactory bulb mitral cells have changed as a function of
279 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 : Returns a data record composed of the
312

313 mean, standard deviation, and n, summarizing input resistance measurements from cerebellar
314 Purkinje cells based on all indexed articles in NeuroElectro database. Here the database query is
315 performed 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.

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-
338 hour curation session, a senior graduate student in our laboratory identified 91 electrophysiological
339 measurements (focusing on resting membrane potential, input resistance, membrane time constant, spike
340 amplitude, spike width, and spike threshold) from 35 articles for 7 olfactory bulb neuron types using only
341 prior knowledge of which articles and investigators were likely to have reported such electrophysiological
342 data. In a comparable seven-hour curation session using our semi-automated methods, a single curator
343 (with similar expertise to the first curator) identified 551 electrophysiological measurements from 70
344 articles across 40 neuron types throughout the nervous system. Moreover, this comparison would likely
345 tilt even more in favor of the semi-automated methods had the curators been less familiar with the primary
346 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 for
348 electrophysiological measurements to be extracted with an acceptably low error rate. Namely, an expert
349 curator needs to specify which neuron types are recorded from in each article and where data from the
350 normotypic or “control” states of these neurons are textually referenced within a data table. Moreover,
351 given the current accuracy of the unsupervised algorithmic assignment of electrophysiological concepts
352 and experimental metadata (78% and 50%, respectively), these also need to be manually validated and

353 corrected and normalized as required by an expert. Given the necessity of these manual steps, the
354 scalability of our current approach is limited by our ability to manually curate this information or by
355 our ability to improve the error rate of the automated methods. Despite this limitation, our current
356 pipeline is still much faster than a purely manual one. The methodology could be further improved by
357 addressing falsely matching entities (such as EPSP amplitude in section 2.3.2) by either auto-suggesting
358 alternative matches returned by string-similarity algorithms or by simply adding these concepts to the
359 electrophysiological ontology. Moreover, these improvements would facilitate formally computing the
360 sensitivity and specificity of these entity 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 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 and single cell qPCR approaches) and electrophysiological properties of neurons. Third, we have begun
368 automated testing of quantitative neuron models in concert with SciUnit (**Omar et al.**, 2014), under the
369 reasonable assumption that these models should be constrained by the available experimental data. These
370 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 Moreover, such an approach could easily incorporate common neuron type acronyms and abbreviations
378 (**Okazaki and Ananiadou**, 2006).

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

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

3.5 FUTURE METHODS FOR DATA EXTRACTION

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

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

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

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

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

ACKNOWLEDGEMENT

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

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

REFERENCES

- 436 Aldous, D. J. (1985), Exchangeability and related topics, in P. L. Hennequin, ed., cole d't de Probabilit's
437 de Saint-Flour XIII 1983 (Springer Berlin Heidelberg), number 1117 in Lecture Notes in Mathematics,
438 1–198
- 439 Ambert, K. H. and Cohen, A. M. (2012), Text-mining and neuroscience, *International review of*
440 *neurobiology*, 103, 109–132, doi:10.1016/B978-0-12-388408-4.00006-X, PMID: 23195123
- 441 Ascoli, G. A., Alonso-Nanclares, L., Anderson, S. A., Barrios, G., Benavides-Piccione, R., Burkhalter, A., et al. (2008), Petilla terminology: nomenclature of features of GABAergic
442 interneurons of the cerebral cortex, *Nature Reviews Neuroscience*, 9, 7, 557–568, doi:10.1038/nrn2402,
443 WOS:000256929300015
- 444 Bird, S., Klein, E., and Loper, E. (2009), Natural language processing with Python (O'Reilly, Beijing;
445 Cambridge [Mass.])
- 446 Bota, M., Dong, H.-W., and Swanson, L. W. (2005), Brain architecture management system,
447 *Neuroinformatics*, 3, 1, 15–48, doi:10.1385/NI:3:1:015, PMID: 15897615
- 448 Bruha, P., Papez, V., Bandrowski, A., Grewe, J., Mouek, R., Tripathy, S., et al. (2013), The ontology for
449 experimental neurophysiology: a first step toward semantic annotations of neurophysiology data and
450 metadata., *Frontiers in Neuroinformatics*, 7, doi:10.3389/conf.fninf.2013.09.00026
- 451 Connors, B. W., Gutnick, M. J., and Prince, D. A. (1982), Electrophysiological properties of neocortical
452 neurons in vitro, *Journal of neurophysiology*, 48, 6, 1302–1320, PMID: 6296328
- 453 Dickman, S. (2003), Tough mining, *PLoS Biol*, 1, 2, e48, doi:10.1371/journal.pbio.0000048
- 454 Fishell, G. and Heintz, N. (2013), The neuron identity problem: Form meets function, *Neuron*, 80, 3,
455 602–612, doi:10.1016/j.neuron.2013.10.035
- 456 French, L., Lane, S., Xu, L., and Pavlidis, P. (2009), Automated recognition of brain region mentions in
457 neuroscience literature, *Frontiers in Neuroinformatics*, 3, 29, doi:10.3389/neuro.11.029.2009
- 458 French, L., Lane, S., Xu, L., Siu, C., Kwok, C., Chen, Y., et al. (2012), Application and evaluation of
459 automated methods to extract neuroanatomical connectivity statements from free text, *Bioinformatics*,
460 28, 22, 2963–2970, doi:10.1093/bioinformatics/bts542, PMID: 22954628
- 461 French, L. and Pavlidis, P. (2011), Relationships between gene expression and brain wiring in the adult
462 rodent brain, *PLoS computational biology*, 7, 1, e1001049, doi:10.1371/journal.pcbi.1001049, PMID:
463 21253556
- 464 French, L. and Pavlidis, P. (2012), Using text mining to link journal articles to neuroanatomical databases,
465 *The Journal of Comparative Neurology*, 520, 8, 17721783, doi:10.1002/cne.23012
- 466 Hamilton, D. J., Shepherd, G. M., Martone, M. E., and Ascoli, G. A. (2012), An ontological approach
467 to describing neurons and their relationships, *Frontiers in Neuroinformatics*, 6, 15, doi:10.3389/fninf.
468 2012.00015
- 469 Hines, M. L., Morse, T., Migliore, M., Carnevale, N. T., and Shepherd, G. M. (2004), ModelDB: a
470 database to support computational neuroscience, *Journal of Computational Neuroscience*, 17, 1, 7–11,
471 doi:10.1023/B:JCN.0000023869.22017.2e, PMID: 15218350
- 472 Hodgkin, A. L. and Huxley, A. F. (1952), A quantitative description of membrane current and its
473 application to conduction and excitation in nerve, *The Journal of physiology*, 117, 4, 500–544, PMID:
474 12991237
- 475 Huang, J. and Zeng, H. (2013), Genetic approaches to neural circuits in the mouse, *Annual Review of*
476 *Neuroscience*, 36, 1, 183–215, doi:10.1146/annurev-neuro-062012-170307, PMID: 23682658
- 477 Kim, S. and Wilbur, W. J. (2011), Classifying protein-protein interaction articles using word and syntactic
478 features, *BMC bioinformatics*, 12 Suppl 8, S9, doi:10.1186/1471-2105-12-S8-S9, PMID: 22151252
- 479 Kinnischtzke, A. K., Sewall, A. M., Berkepile, J. M., and Fanselow, E. E. (2012), Postnatal maturation of
480 somatostatin-expressing inhibitory cells in the somatosensory cortex of GIN mice, *Frontiers in neural*
481 *circuits*, 6, 33, doi:10.3389/fncir.2012.00033, PMID: 22666189
- 482 Larson, S. D. and Martone, M. E. (2013), NeuroLex.org: an online framework for neuroscience
483 knowledge, *Frontiers in Neuroinformatics*, 7, 18, doi:10.3389/fninf.2013.00018
- 484 McCallum, A. (2002), MALLET: a machine learning for language toolkit

- 486 Migliore, M., Morse, T. M., Davison, A. P., Marenco, L., Shepherd, G. M., and Hines, M. L. (2003),
487 ModelDB, *Neuroinformatics*, 1, 1, 135–139, doi:10.1385/NI:1:1:135
488 Neher, E. (1992), [6] correction for liquid junction potentials in patch clamp experiments, in Bernardo
489 Rudy, ed., Methods in Enzymology, volume Volume 207 (Academic Press), 123–131
490 Okaty, B. W., Miller, M. N., Sugino, K., Hempel, C. M., and Nelson, S. B. (2009), Transcriptional and
491 electrophysiological maturation of neocortical fast-spiking GABAergic interneurons, *The Journal of
492 Neuroscience*, 29, 21, 7040–7052, doi:10.1523/JNEUROSCI.0105-09.2009, PMID: 19474331
493 Okazaki, N. and Ananiadou, S. (2006), Building an abbreviation dictionary using a term recognition
494 approach, *Bioinformatics*, 22, 24, 3089–3095, doi:10.1093/bioinformatics/btl534, PMID: 17050571
495 Omar, C., Aldrich, J., and Gerkin, R. (2014), Collaborative infrastructure for test-driven scientific model
496 validation, *ICSE: New Innovations and Emerging Results*
497 Parekh, R. and Ascoli, G. (2013), Neuronal morphology goes digital: A research hub for cellular and
498 system neuroscience, *Neuron*, 77, 6, 1017–1038, doi:10.1016/j.neuron.2013.03.008
499 Pasquale, E. D., Keegan, K. D., and Noebels, J. L. (1997), Increased excitability and inward rectification in
500 layer v cortical pyramidal neurons in the epileptic mutant mouse stargazer, *Journal of Neurophysiology*,
501 77, 2, 621–631, PMID: 9065835
502 Ramakrishnan, C., Patnia, A., Hovy, E., and Burns, G. A. (2012), Layout-aware text extraction from
503 full-text PDF of scientific articles, *Source code for biology and medicine*, 7, 1, 7, doi:10.1186/
504 1751-0473-7-7, PMID: 22640904
505 Shepherd, G. M., ed. (2003), The Synaptic Organization of the Brain (Oxford University Press, USA), 5
506 edition
507 Teeter, C. and Stevens, C. (2011), A general principle of neural arbor branch density, *Current Biology*,
508 doi:10.1016/j.cub.2011.11.013
509 Thompson, K. (1968), Programming techniques: Regular expression search algorithm, *Commun. ACM*,
510 11, 6, 419422, doi:10.1145/363347.363387
511 Toledo-Rodriguez, M., Blumenfeld, B., Wu, C., Luo, J., Attali, B., Goodman, P., et al. (2004), Correlation
512 maps allow neuronal electrical properties to be predicted from single-cell gene expression profiles in
513 rat neocortex, *Cerebral Cortex*, 14, 12, 1310 –1327, doi:10.1093/cercor/bhh092
514 van Brederode, J. F. M., Yanagawa, Y., and Berger, A. J. (2011), GAD67-GFP+ neurons in the nucleus
515 of roller: a possible source of inhibitory input to hypoglossal motoneurons. i. morphology and firing
516 properties, *Journal of neurophysiology*, 105, 1, 235–248, doi:10.1152/jn.00493.2010, PMID: 21047932
517 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011), Large-scale
518 automated synthesis of human functional neuroimaging data, *Nat Meth*, 8, 8, 665–670, doi:10.1038/
519 nmeth.1635
520 Zhu, J. J. (2000), Maturation of layer 5 neocortical pyramidal neurons: amplifying salient layer 1 and
521 layer 4 inputs by ca2+ action potentials in adult rat tuft dendrites, *The Journal of physiology*, 526 Pt 3,
522 571–587, PMID: 10922009
523

SUPPLEMENTAL DATA

