Facilitating NeuroMorpho.Org Curation via Neuronal and Glial Metadata Analysis

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Yasmeen Zoubi
Bachelor of Science
McGill University, 2018

Director: Giorgio A. Ascoli, Professor
Department of Biology

Summer Semester 2021
George Mason University
Fairfax, VA

ACKNOWLEDGEMENTS

I would like to extend my gratitude to Dr. Giorgio A. Ascoli, Dr. Ancha Baranova, Dr. Saleet Jafri, and Kayvan Bijari for their guidance, commitment, and support with preparing and writing this thesis. I have garnered an appreciation of various neuroscience concepts and applications in my journey with them. Additionally, they lent a hand in the expansion of my scientific writing skills and expertise. They have sparked my interest in biological and neurological research, which I seek to attain a potential career in. My experience with them has made me enthusiastic about the ever-expanding progression of scientific research. While this progression is not without challenges and a lengthy series of trial and error, it renders a guaranteed sentiment of reward and self-motivation to contribute to the advancement of neurobiological discoveries.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

ABSTRACT

FACILITATING NEUROMORPHO.ORG CURATION VIA NEURONAL AND GLIAL METADATA ANALYSIS

Yasmeen Zoubi, M.S.

George Mason University, 2021

Thesis Director: Dr. Giorgio A. Ascoli

NeuroMorpho.Org is a scientific database of digital reconstructions of neurons and glia. It serves as a large-scale repository of a wide range of morphological information that can be accessed all over the world, thus encouraging data sharing and communication amongst the international neuroscience community. The curation of such data is very helpful in mining and understanding the relationships between dendritic and axonal branching, glial processes, brain connectivity, and synaptic signaling. Metadata refers to information about the data. NeuroMorpho.Org specifically provides metadata for each curated cell, including details on the animal subject, brain region, cell type, and experimental protocol. This information is extracted from the corresponding peer reviewed publications that describe the reconstructed neurons or glia.

The manual process of metadata extraction and annotation can be labor intensive, time consuming, and error prone. In this regard, machine learning can be employed to overcome such challenges by facilitating and eventually automating the identification of relevant information. To ensure efficacy, machine learning tools must be trained with a corpus of

existing annotations. Here we deployed a two-pronged approach for analyzing NeuroMorpho.Org metadata to provide a useful training set to aid the ongoing development of semi-automated annotation. First, we investigated our records of metadata in order to deduce any systematic patterns that may underlie neurobiological rules or statistical trends and could be expressed into artificial intelligence heuristics. Specifically, we used a frequency-based data mining algorithm known as "Apriori", which makes use of association rules to compute frequent itemsets consisting of neuronal and glial metadata. Second, we utilized machine learning tools in extracting key metadata via an approach known as "named entity recognition", or NER, such that metadata acquisition can be automated. In this case, it is necessary to perform several rounds of manual annotations that the algorithm can learn from, thus making automated annotation as precise as possible. Altogether, our investigation can potentially aid technologies in training algorithms for robust metadata annotation, which can lead to the expansion and enhancement of NeuroMorpho.Org.

SPECIFIC AIMS

Neuroscience is a dynamic field seeking to study the nervous system. In recent years, computational neuroscience has taken form as a neuroscience sub-discipline that integrates principles of anatomy, physiology, and medicine with those of mathematics and computer science (Trappenberg, 2010). Within computational neuroscience, neuronal morphology is used to study neuron networking, connectivity, plasticity, and information processing. NeuroMorpho.Org is the largest existing web-accessible repository of neuronal and glial digital reconstructions and their respective metadata. The public availability of this database and related services advances neuroscience research at both the conceptual and technological level (Halavi et al., 2008).

The main goal of this dissertation is to facilitate the extraction of neuronal reconstruction metadata for NeuroMorpho.Org curation. We seek to accomplish this mission through the following specific aims:

Specific Aim I: To recognize systematic correlations in NeuroMorpho.Org metadata in order to assess possible neurobiological rules or detect key statistical trends. These patterns can be leveraged to facilitate future annotations. In particular, we will:

(a) Evaluate the relative frequency of entries in each metadata dimension. For example, rat, mouse, and Drosophila melanogaster are most frequent amongst species and neocortex and hippocampus are most frequent amongst brain regions.

(b) Assess the probability of cooccurring metadata entry pairs. These cooccurrences can reflect both biological rules (for example, "if the cell type is Purkinje cell, the brain region must be cerebellum") and simple statistical associations (for example, "if the brain region is the neocortex, the cell type is most likely pyramidal").

(c) Explore the possibility of identifying negation patterns involving pairs of metadata entries which never appear together. This may correspond to biological rules (for example, "if the brain region is the mushroom body, the species cannot be mouse") or statistical trends (for example, "if the brain region is retina, the cell type is unlikely GABAergic").

Specific Aim II: NeuroMorpho.Org requires integration with machine learning in order to render data curation more robust and less error prone, hence facilitating the further expansion of the database. An approach known as "named entity recognition" (NER) can, in principle, extract key neuronal metadata from texts (e.g. peer-reviewed papers). Such techniques make use of training sets consisting of biological rules and existing metadata annotations. We will compile an annotated dataset for NER training and explore the possibility of using the correlations identified under the Specific Aim I in NER application. A useful outcome of this work will be to enable thorough error detection of NeuroMorpho.Org metadata in the event where we encounter existing annotations conflicting with identified biological rules. The long-term goal of this thesis is to pave the way for the application of machine learning techniques. Learning models will be trained to perform metadata annotation in order to bypass the challenges of time consumption and human error, hence making data readily available.

BACKGROUND

This section will cover the basics of NeuroMorpho.Org, metadata analysis, the Apriori algorithm, and machine learning via named entity recognition.

2.1 NeuroMorpho.Org and Metadata Analysis

NeuroMorpho.Org is an online inventory of over 130,000 digitally reconstructed neuronal and glial morphologies (as of v.8.0 released June 2020) described in peer reviewed publications by more than 650 independent laboratories. The repository grants users the ability to browse datasets by cell type (e.g. basket, granule, etc.), brain region (e.g. neocortex, cerebellum, etc.), animal species (e.g. mouse, chimpanzee, etc.), or lab of origin (Ascoli et al., 2007). Due to the wide representation of different neuron types across the brain, researchers are able to garner an appreciation of the nervous system complexity (Chu et al., 2015).

The acceleration of digital reconstruction collection and accessibility heavily depends on the organized contribution from multiple researchers in neuroscience. This joint effort is required to enable curation, annotation, storage, and distribution of digital reconstructions. Users can analyze existing data initially collected from many studies, which can result in several new morphological discoveries. Hence NeuroMorpho.Org encourages data sharing and scientific communication by enabling labs to freely post their neuronal reconstructions (Parekh & Ascoli, 2013).

Metadata is a diverse collection of information that is created, stored, and shared with the purpose of describing an item or set of items. This information increases the functionality of item repositories by enabling users to recognize, record, and share key data trends. The availability and pervasiveness of metadata hence encourages access, analysis, and feedback from multiple parties (Riley & National Information Standards Organization (U.S.), 2017). Different metadata types exist: descriptive, administrative, and structural (Halavi et al., 2008). NeuroMorpho.Org provides descriptive metadata about each neuron reconstruction, adding to its scientific value and permitting effective user searches for specific data of interest (e.g. by experimental protocol). This metadata can be typically extracted from the peer-reviewed publications describing the analyzed neuron (Bijari et al., 2020).

NeuroMorpho.Org reconstructions are annotated with metadata under four categories: animal, experiment, anatomy, and source. The animal category provides details of the subject, such as the species, strain, gender, weight, and age. The experiment category provides metadata which includes the protocol (e.g. in vivo, culture), condition (e.g. control, knock-out), label or stain used to visualize the cell, thickness of tissue section, slicing orientation, objective type and magnification, shrinkage, and reconstruction software. The metadata of the anatomy category include the brain regions and sub-regions, as well as cell type and sub-type. Finally, the source category identifies the contributing laboratory, reference publication, names of the individual neuron files, original digital file formats, and the dates of receipt and website upload. When specific metadata cannot be obtained from the literature or the authors, they are listed as “not reported” (Parekh et al., 2015).

2.2 Apriori Algorithm and Pattern Analysis

Data mining refers to the process of extracting interesting and valuable relationships and structures within a given dataset (Han & Kamber, 2012) and is thus considered a key step in knowledge discovery from databases (Hand, 2007). Data mining solutions can also assist in making the process of information extraction more efficient. The data mining algorithm we employed to analyze neuronal metadata is Apriori, which functions in discovering all significant association rules between items in a large dataset (Agrawal, 1994). In our case the algorithm counts the occurrences of certain items in our metadata, such as a given brain region, reconstruction software, or experimental condition. It then generates frequent itemsets through a series of iterations. It does so by generating the candidate itemsets, computing their "support" and pruning the candidate itemsets to the frequent itemset in each iteration (Ye & Chiang, 2006). From this, we can assess the frequency of each itemset and ascertain biological rules or statistical trends. For example, if the algorithm presents a certain cell type (e.g. Purkinje) being frequently associated with a certain brain region (e.g. cerebellum), then we can stipulate that if the cell type is a Purkinje cell then the corresponding brain region is cerebellum.

2.3 Learning from the Data and Named Entity Recognition

Machine learning is a subfield of artificial intelligence which relies on the use and analysis of large sets of data with the purpose of finding patterns and relationships amongst different entries. Recently the amount of neuroscience information has increased exponentially,

making machine learning a useful or even essential tool for data curation (Ozoh et al., 2020).

Data extraction from the text of scientific articles is usually a challenging, time-consuming, and error-prone task. Therefore, the quality of the extracted data might not be optimal. Additional problems such as the lack of adherence to standardized terminologies can also be problematic. In order to overcome these challenges, natural language processing tools, such as "named entity recognition", can be used to assist in the process of metadata extraction and annotation (Zhang et al., 2004).

Named entity recognition (NER) is a text-mining tool used for learning, identifying, and labeling key elements (e.g. words, concepts, sentences, and named entities) within texts and publications (Bijari et al., 2020). Learning models can thus be trained in principle to extract named entities corresponding to neuroscience metadata of interest, such as neuron types, brain regions, and experimental protocols, from peer reviewed publications and spreadsheets so as to facilitate data annotation (Shardlow et al., 2019). This long-term goal is an essential step in reducing the burden (and human error rate) of manual sifting through continuously increasing numbers of publications (Bijari et al., 2020).

The quality of the results of machine learning algorithms highly depends on the quality of the training data. The biological rules that we extract and develop from data mining become useful and necessary in future training of machine learning models. With enough training, our ultimate goal is to enable machines to independently annotate future metadata.

METHODS

This section will describe how the Apriori algorithm works, the process of rule evaluation, and the process of machine learning via named entity recognition.

3.1 The Mechanism of Apriori Algorithm and the Process of Itemset Creation

A common approach to assess the relationship between two items involves the use of association rules (Nengsih, 2015). Apriori algorithm efficiently mines association rules among the frequently appearing substructures of a given dataset (Inokuchi et al., 2000). The number of frequent itemsets the algorithm extracts depends on a user-defined minimum support threshold. This threshold can be small (e.g. 0.2%) or large (e.g. 55%), with small thresholds yielding larger quantities of candidate itemsets (Hossain et al., 2019). Apriori uses a method referred to as "level-wise" search, where k-itemsets are used to explore (k+1)-itemsets. It follows the principle that every subset of a frequent itemset must also be frequent. The algorithm scans the dataset in search of the first frequent 1-itemsets which satisfy the support. Moreover, frequent 2-itemsets are found by using the frequent 1-itemsets, continuing until frequent k-itemsets can be found (Han & Kamber, 2012). To further explain, consider the set [c, d, e] with subsets {c}, {d}, {e}, {c, d}, {c, e}, and {d, e}. The Apriori algorithm assumes that if set [c, d, e] is frequent, then its subsets must also be frequent. On the other hand, given an infrequent set [a, b], its supersets (i.e. {a, b, e}, {a, b, c, d}), must also be infrequent (Tan et al., 2006).

Initially every item in a given dataset is a candidate 1-itemset. The ones that do not meet our support threshold (i.e. {a}, {b}) are discarded from the analysis. Another round of iterations is performed, where candidate 2-itemsets are generated using only the frequent 1-itemsets (i.e. {c, d}, {c, e}). The 2-itemsets that do not meet the support threshold are in turn removed from the analysis. The remaining sets will be used in generating 3-itemsets, and the process is repeated until we have sufficient number of sets that meet the support threshold (Tan et al., 2006).

Our itemsets are collections of metadata entries from NeuroMorpho.Org extracted in the form of a spreadsheet. Each neuron is listed by row with their corresponding specifications (e.g. species, cell type, reconstruction software, experimental conditions, etc.) being organized by column (Table 1). Thousands of neurons can be saved in a single spreadsheet.

Table 1. Spreadsheet organization of neurons and their metadata specifications.

neuron	**species**	**brain region**	**cell type**	**…**	**reconstruction software**
1	mouse	hippocampus	principal	…	Amira
2	mouse	cerebellum	Purkinje	…	Neurolucida
…	…	…	…	…	…
n	monkey	neocortex	interneuron	…	Amira

After Apriori identifies all frequent sets from our original ('input') metadata spreadsheet, several binary itemsets can be exported into a different ('output') spreadsheet to critically evaluate as possible association rules. Association rules follow the general form: "If A occurs, there is a high probability of B occurring", typically written in the simplified style

"A → B"; A is referred to as the "antecedent" and B is referred to as the "consequent". In our output spreadsheet, antecedents and consequents are listed in adjacent columns (Table 2). The metadata fields of reference (e.g. cell type, brain region, stain, protocol, etc.) are represented in brackets for both antecedents and consequents.

Table 2. Sample spreadsheet representing binary itemsets in potential association rules of the form "antecedent → consequent".

antecedent	**consequent**
Multidendritic-dendritic arborization (DA) (cell type)	larval (developmental stage)
Abdominal (region)	Drosophila melanogaster (species)
Han Wistar (strain)	Neurolucida (reconstructions software)
Canton S G1 x w1118 (strain)	in vitro (protocol)
water or oil (objective type)	biocytin (stain)
stratum granulosum (region)	interneuron (cell type)
ventral striatum (region)	nucleus accumbens (region)
protocerebrum (region)	Drosophila melanogaster (species)
Rapid Golgi (stain)	in vitro (protocol)

For example, the pair "Multidendritic-dendritic arborization (DA) (cell type) → larval (developmental stage)" means "if the cell type is Multidendritic-dendritic arborization (DA), the developmental stage is likely larval".

3.2 Rule Evaluation

For better interpretation of itemset associations, it is useful to distinguish whether they reflect scientifically meaningful facts/mechanisms or coincidental statistical trends. Examples of the former are as follows:

- "If the neuron type is adult-born [and the species is human], the anatomical region is the hippocampus".

- “If the stain is Rapid Golgi, the protocol is in vitro”.
- “If the neuron type is Multidendritic-dendritic arborization (DA), the anatomical region is the peripheral nervous system”.

The first example reflects the biological knowledge that, in human brains, neurogenesis only occurs in the dentate gyrus, a portion of the hippocampal formation. In the second case, the association corresponds to the necessity of excising the tissue prior to silver impregnation, which makes this labeling technique impossible for in vivo applications. The third example reflects the fact that Multidendritic-dendritic arborization (DA) cells are specialized neurons of the invertebrate peripheral mechanosensory system.

In many cases, however, the antecedent and consequent have a high likelihood of cooccurring due to relative abundances of data often determined by the comparative popularity of certain scientific models. For example:

- "If the neuron type is Kenyon cell, the species is most likely Drosophila melanogaster”.
- “If the anatomical region is the retinal ganglion layer, the neuron type is most likely ganglion”.
- “If the mouse strain is C57BL/6”, the age class is most likely adult”.

Kenyon cells are the intrinsic neurons of the mushroom body, a structure of the nervous system that is present in several invertebrates, including all insects and some annelids (e.g. earthworms) (Tomer et al., 2010). However, when cells in this structure are searched for in NeuroMorpho.Org, a great percentage of the corresponding neurons come from Drosophila melanogaster, with only a few coming from silkworm, due to the tremendous

penetration of the fruit fly model in neuroscience. In the second case, while ganglion cells are the most abundant neurons in the retinal ganglion layer, displaced amacrine cells are also found in the same region. Even more extremely, the C57BL/6 mouse strain can obviously be of any age group. However, the vast majority of NeuroMorpho.Org neurons corresponding to this strain happen to come from adult animals. Note that although associations in this second category do not reflect necessary biological mechanisms, they may still be useful to predict the correct metadata annotation given existing trends in the literature.

In addition to co-occurrences, we will also evaluate possible 'negation rules', which also can reflect scientific necessities or statistical trends. Self-explanatory examples of the former are:

- "If the cell type is pyramidal neuron, the region cannot be peripheral nervous system".
- "If the stain used is GFP, the objective type cannot be electron microscopy".
- "If the developmental stage is larval, the species cannot be mouse".

Instances of negation statistical trend include:

- "If the cell type is dopaminergic, the brain region is unlikely hypothalamus".
- "If the reconstruction software is Eyewire, the protocol is unlikely in vivo".
- "If the anatomical region is frontal cortex, the species is unlikely lemur".

According to NeuroMorpho.Org, many different brain regions possess dopaminergic neurons, yet very few entries happen to come from the hypothalamus. Eyewire is a software tool that is especially popular among labs studying fixed tissue and therefore seldom co-

occurs with live imaging. The frontal lobe is found in many mammal species besides the lemur and most neuron in this region are traced from humans and mice.

Quantitative metrics from set theory can also aid us in the evaluation of an itemset. Here, the symbol "⋂" denotes intersection items. For example, let A be mouse and B be pyramidal cell. If A → B (i.e. mouse → pyramidal), the following can be considered:

- A ⋂ B (# of times mouse and pyramidal cell cooccur in the dataset)
- A ⋂ -B (# of times mouse cooccurs with a different cell type – e.g. granule cell)
- -A ⋂ B (# of times a different animal species cooccurs with pyramidal cell – e.g. rat)
- -A ⋂ -B (# of times a different animal species cooccurs with a different cell type – e.g. chimpanzee and Purkinje)

Here, it can be inferred that the greater the value of "A ⋂ B" amongst the total data of various animal species and various cell types, the stronger the correlation is. The application of intersection metrics, which were previously calculated, are applied to our data (Table 3). The "total" was calculated by summing the values of the four intersection metrics. Because mouse → principal cell has the greatest percentage value for A ⋂ B, it can be inferred that this correlation is the strongest.

Table 3. Sample data as seen in our spreadsheet representing the calculated intersections of "mouse" associated with various cell types.

antecedent	consequent	A∩B	A∩-B	- A∩B	- A∩-B	total
mouse	principal cell	25336	11407	59202	26159	122104
mouse	pyramidal	11777	24966	18007	67354	122104
mouse	interneuron	6783	29960	15904	69457	122104
mouse	granule	5827	30916	1035	84326	122104
mouse	glia	3849	32894	5626	79735	122104
mouse	microglia	2775	33968	4321	81040	122104
mouse	adult-born	2741	34002	445	84916	122104
mouse	iba1-positive	2160	34583	3705	81656	122104
mouse	ganglion	2133	34610	1027	84334	122104
mouse	medium spiny	1662	35081	859	84502	122104

Next, it is useful to consider two additional metrics: the number of archive (independent labs or datasets) and of article (peer-reviewed publications or references) supporting a particular association. The larger these numbers are, the more solid the association is. From our collected data, the highest counts for archive and article are 315 and 509 respectively (Table 4).

Table 4. Comparing different numbers of archives and articles corresponding to different correlations.

antecedent	consequent	archive number	article number	relative count
principal (class 1)	pyramidal (class 2)	254	398	high
oil (objective type)	in vitro (protocol)	176	276	high
neocortex (region 1)	somatosensory (region 2)	74	134	moderate
rat (species)	Wistar (strain)	66	115	moderate
occipital (region 2)	primary visual (region 3)	27	34	low
mouse (species)	glia (class 1)	22	24	low

Additional set theory metrics can help interpret and evaluate the quality of our association rules. These include abundance/support, confidence, and lift.

Support and confidence are used to measure the "strength" of an association rule. The support determines how often a rule is applicable to a given dataset (Tan et al., 2006) by calculating the fraction of entries A and B occurring together amongst all collected neuron records:

- Abundance or support = $\frac{\#(A \cap B)}{\#(A \cap B + A \cap -B + -A \cap B + -A \cap -B)}$

Confidence determines how frequently a given consequent B appears to be associated with a given antecedent A. It measures the reliability of the implication made by an association rule. That is to say, the higher the confidence, the more likely it is for A and B to be observed together:

- Confidence = $\frac{\#(A \cap B)}{\#(A \cap -B)}$

In some cases, high confidence rules can be misleading due to the fact that the confidence metric ignores the support of the consequent. This issue is addressed by applying the "lift" metric, which divides the confidence by the number of instances the consequent cooccurs with an antecedent different from the antecedent of interest, or -A ⋂ B (Tan et al., 2006):

- Lift = $\frac{\text{Confidence}}{\#(-A \cap B)} = \frac{\#(A \cap B)}{\#(A \cap -B) \times \#(-A \cap B)}$

Examples of computed support, confidence, and lift values for a sample of itemset are illustrated in Table 5.

Table 5. Confidence, lift, and abundance (support) values for each itemset as seen in our database.

antecedent	consequent	confidence	lift	abundance
Multidendritic-dendritic arborization (DA) (cell type)	larval (age class)	1	18.9309	0.0179
abdominal (region)	Drosophila melanogaster (species)	1	3.8464	0.0143
Han Wistar (strain)	Neurolucida (reconstruction software)	1	2.0864	0.0117
prelimbic (region)	pyramidal (cell type)	0.6528	1.2.6761	0.0141
sensory (cell type)	in vivo (protocol)	0.6078	1.9091	0.0418
Neurolucida (reconstruction software)	neocortex (region)	0.6003	1.6092	0.4793
peripheral nervous system (region)	green fluorescent protein (stain)	0.3811	1.4912	0.0368
Knossos (reconstruction software)	retina (region)	0.3634	9.0155	0.0366
glia (cell type)	M (gender)	0.2713	0.8199	0.0776

3.3 Machine Learning and Name Entity Recognition

The process of machine learning heavily relies on the quality of the data extracted from electronic databases and documents. Many tools have been developed to assist in this process by means of automatic summarization and information retrieval. The learning model recognizes key pieces of information that are usually represented by words, sentences, and concepts. Additionally, specific patterns such as word positioning and frequency are detected in this automated analysis (Zhang et al., 2004).

Algorithms can be trained to recognize key named entities that have been previously annotated by neuroscientists who skim through a corpus of neuroscience-based documents

and texts. With NER, we initially establish the categories of named entities corresponding to neuronal metadata. We must annotate a corpus of documents and publications that contain the named entities of interest. Following that, the NER learning model can be trained to detect named entities in new documents and publications with the goal of reaching a performance above or similar to that of prior annotations (Shardlow et al., 2019). A data annotation interface known as DataTurks was utilized in labelling key neural named entities corresponding to, for instance, brain regions and cell types. It specifically enables data annotation for different machine learning projects. On this interface, we are provided with sentences and paragraphs containing neural named entities. These sentences and paragraphs are extracted from a given publication describing a NeuroMorpho.Org dataset. Plural forms, descriptive adjectives, abbreviations, hyphenations, and suffix-changes corresponding to these entities are all highlighted. Likewise, we label experimental condition, species, strains, and other common metadata entries, in a color-coded fashion (Figure 1, 2). If uncertain whether an entity should be labeled or not, the metadata search on NeuroMorpho.Org can be used for confirmation.

Figure 1. Color-coding scheme of key neural metadata entries as seen on the DataTurks annotation interface.

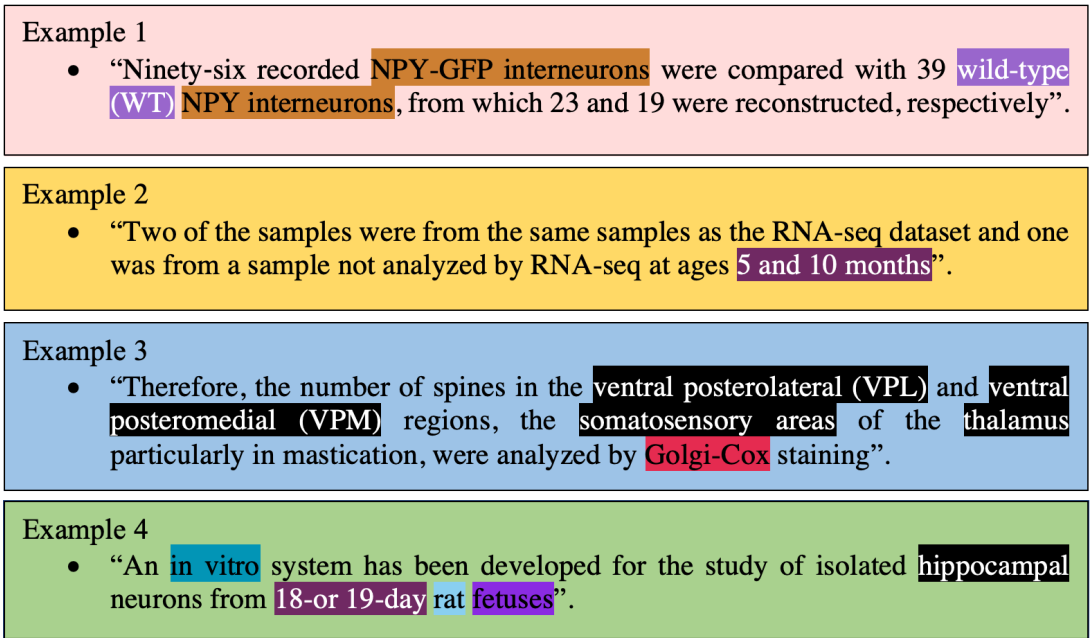

Figure 2. Examples of color-coded labeled entries performed by us on the DataTurks annotation interface.

In Example 1, the NeuroMorpho.Org metadata search portal was used to confirm the presence of "NPY-GFP" and "NPY interneurons" as these cell types are not commonly mentioned in publications. In this case, the cell classes are in plural form and are highlighted in order to train the algorithm to annotate both singular and plural entries. Additionally, the experimental condition "wildtype" can be written in several different ways, such as "wildtype", "wild-type", "wild type", and "WT". All distinct writings will be annotated such that the algorithm can recognize the alternate forms when performing independent tagging.

In Example 2, "5 and 10 months" is labelled as an age. While NeuroMorpho.Org identifies developmental stages by phase (e.g. embryonic, neonatal, young, adult, old) as opposed to

a week, month or year timeframe, it is useful to tag all relevant information in our entries to allow the machine to learn by inference. Hence, the additional tag “age” is created to be distinguished from the “developmental stage” tag.

In Example 3, adjectives describing the position of specific regions, such as “ventral posterolateral (VPL)” and “ventral posteromedial (VPM)” are highlighted as regions. In many cases, certain positions and directions of a given brain region will be mentioned in an entry, and it is necessary for the machine to recognize them.

In Example 4, “hippocampal” is selected as a brain region as it is a derivative of “hippocampus”. Annotating various derivative words of neural entries will prevent the learning model from missing important entries and help the model to generalize on different word-forms.

Once several entries are labeled, the learning model will have a sufficient training corpus to learn how to detect these named entities independently. Association rules such as those described above can be applied in machine training to reduce the proportion of incorrectly annotated entries. Specifically, identifying relations between extracted entries, as well as the context they are used in, can be used in correctly labeling entries and filtering unwanted entries (Shardlow et al., 2019). Examples of mistake the learning model might make are labeling antibodies as species (e.g. rabbit anti-mouse) and labelling a brain region as a cell type (e.g. retinal vs retinal ganglion cell).

RESULTS

This section will explain the results obtained from evaluating our rules and performing annotation tagging. Negation rules and machine learning results will also be discussed.

4.1 Evaluation Metrics

Upon analysis, the total number of cooccurrence entries present in our spreadsheet is 2816. The quality of these potential rules was assessed based on the metrics explained in section 3.2 and briefly summarized below:

(a) A∩B and Support – measures the ratio of a given association over the total record number (e.g. rat → Sprague-Dawley being A∩B, where rat is a species and Sprague-Dawley is the strain, and the support being A∩B divided by the total instances of a species being associated with a strain).

(b) Confidence – measures the ratio of A∩B over A∩-B (e.g. rat → Sprague-Dawley is A∩B and rat → a strain other than Sprague-Dawley is A∩-B).

(c) Archive count – the total number of instances a given association is observed in selected independent labs or datasets

(d) Article count – the total number of instances a given association is observed in selected peer-reviewed publications or references

A∩B and Support: Amongst our data, the lowest and highest A∩B observed were 379 and 44192, respectively. The lowest and highest support values were 0.0031 and 0.6923, respectively.

Confidence: In our data, the confidence ranges from 0.0045 to 1. The values were grouped into 3 sets: low, moderate, and high. Upon inspecting the cooccurrences, it was assessed that a confidence value of [0.0045 to 0.5[is low, from [0.5 to 0.75[is moderate and from [0.75 to 1] is high. Out of all the 2816 correlations, 1618 of those had low confidence values, 352 had moderate confidence values and 846 had high confidence values (Figure 3). In the initial part of our analysis, the confidence threshold of 0.5 inclusive has been set, such that rules with moderate and high confidence values are evaluated.

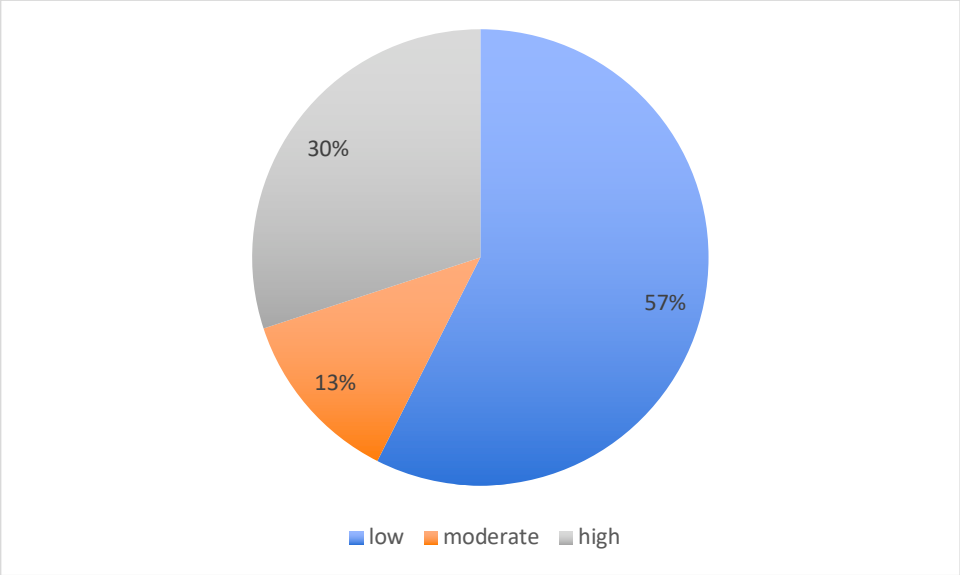

Figure 3. Proportion the ratios of low, medium, and high confidence correlations amongst our data. Note that over half of our entries have a low confidence value.

Archive and Article Count: Based on our data, the ranges for archive and article counts are 1 to 315 and 1 to 509 respectively. Several rules had archive and/or article counts of 1, meaning that the source and reliability is very limited. To eliminate low counts while

preventing the issue of data limitation, a threshold of 10 counts inclusive was set for both categories. This combined with the confidence threshold of 0.5 resulted in 406 entries ready for analysis. A sample of the resulting data is illustrated below (Table 6).

Table 6. Sample data as seen in our spreadsheet with thresholds applied to confidence, A⋂B/support, archive count, and article count values.

antecedent	**consequent**	**A⋂B**	**confidence**	**abundance**	**archive count**	**article count**
ganglion (class 2)	ganglion layer (region 2)	1883	0.5959	0.0259	21	23
embryonic (age class)	culture (protocol)	4633	0.9717	0.039	40	41
primary visual (region 3)	M/F (gender)	3755	0.7638	0.0403	14	14
rat (species)	neocortex (region 1)	24318	0.6616	0.301	68	115
Neurolucida (reconstruction software)	principal cell (class 1)	44192	0.7551	0.4793	307	497
biocytin (stain)	in vitro (protocol)	7338	0.9529	0.0631	136	255

When analyzing all our values in a comprehensive manner, it is logical to evaluate those of higher values as they are more likely to be strong rules. While numerical analysis is useful in understanding associations, the application of existing biological knowledge is essential in the evaluation process. Additionally, the analysis of these rules would be useful for the prediction and machine learning phase in the long run. Here, evaluation was based on the 4 categories described in 3.2:

(a) Category 1: The rule possesses an established biological significance. Example: “If the secondary cell class is medium spiny, the primary region is basal ganglia”.

(b) Category 2: The rule presents a positive statistical trend, where the antecedent and consequent have a high chance of being seen together. Example: “If the species is a zebrafish, the age class is most likely larval”.

(c) Category 3: The negation rule is true from a biological standpoint. Example: “If the strain name is Wistar, the species cannot be Drosophila melanogaster”. Because our rules are based on extracted entries where each antecedent and consequent have been observed together in the past, no negation rules will be observed in our spreadsheet at this time.

(d) Category 4: The rule presents a negative statistical trend, where the antecedent and consequent are observed together at a low frequency. Example: “If the primary region is hippocampus, the secondary cell class is unlikely fast-spiking”.

All 406 associations have been categorized 1-4 and have been evaluated. The results are as follows (Table 7):

- 7 associations fell under category 1.
- 301 fell under category 2.
- 71 fell under category 4.

Table 7. Evaluated association samples as seen in our spreadsheet. Reasoning section was included to explain how each rule was categorized.

antecedent	consequent	category	reasoning
medium spiny (class 2)	basal ganglia (region 1)	1	medium spiny cells are known to exist in the basal ganglia
dry (objective type)	in vitro (protocol)	2	most NMO protocol entries for the dry objective type were in vitro
biocytin (stain)	neocortex (region 1)	4	most NMO region 1 entries for the biocytin stain were hippocampus
adult-born (class 3)	dentate gyrus (region 2)	1	adult-born cells are known to exist in the dentate gyrus
main olfactory bulb (region 1)	adult (age class)	2	Most NMO age class entries for the main olfactory bulb region were adult
Drosophila melanogaster (species)	female (gender)	4	Most NMO gender entries for the Drosophila melanogaster species was not reported

Additionally, there were 27 remaining associations that could not be evaluated:

- 6 of these associations involved either region 2 → region 3, class 1 → class 2 or 3, and class 2 → class 3 associations. These are not suitable, and the algorithm failed to disregard them from its analysis when analyzing the list of associations (Table 7).
- 21 of these associations were vague due to the poor specificity of certain brain regions. Some examples include "dorsal" or "layer 5" as tertiary regions. However, the nervous system consists of several dorsal and layer 5 regions, hence making it impossible to evaluate these associations effectively (Table 8).

Results of categorization are summarized below in Figure 4.

Table 8. Summary of the associations that could not be evaluated. The first 6 could not be evaluated because sub-regions and subclass associations are not feasible and should have been eliminated from the algorithm analysis, the last 21 because it is unclear where in the brain these tertiary regions are located as many layers 2, 3 and 5 and dorsal regions exist throughout the nervous system. Only 6 of the 21 examples are provided below.

antecedent	consequent	association type
glia (class 1)	microglia (class 2)	class 1 → class 2
glia (class 1)	Iba1-positive (class 3)	class 1 → class 3
microglia (class 2)	Iba1-positive (class 3)	class 2 → class 3
spinal cord (region 1)	lumbar (region 2)	region 1 → region 2
dentate gyrus (region 2)	granule layer (region 3)	region 2 → region 3
occipital (region 2)	primary visual (region 3)	region 2 → region 3
dorsal (region 3)	pyramidal (class 2)	region 3 → class 2
layer 2-3 (region 3)	pyramidal (class 2)	region 3 → class 2
layer 3 (region 3)	pyramidal (class 2)	region 3 → class 2
layer 5 (region 3)	pyramidal (class 2)	region 3 → class 2
right (region 3)	principal cell (class 1)	region 3 → class 1
left (region 3)	in vitro (protocol)	region 3 → protocol

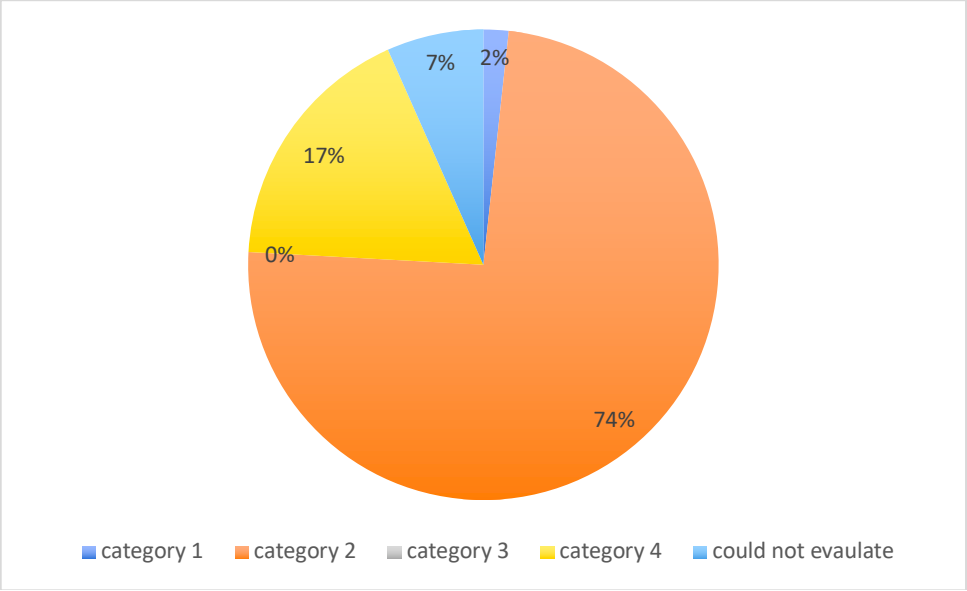

Figure 4. Percentage of each evaluated category amongst the 406 associations. Most of these associations were noted to be positive statistical trends (category 2).

In the following scenario, the significance of setting the confidence threshold value is assessed. It is expected that with a lower confidence threshold, there will be a smaller percentage of category 2 rules and a higher percentage of category 4 rules, given that the higher the confidence, the more likely antecedent A and consequent B are seen together. The confidence threshold was reduced from 0.5 to 0.3 inclusive. All support values were included in the analysis, and the archive and article count thresholds were kept the same (e.g. 10 inclusive). The new threshold resulted in 244 new associations ready to be evaluated. The categorization result was as follows:

- 0 associations fell under category 1.
- 172 fell under category 2.
- 65 fell under category 4.
- 7 could not be evaluated.

Results of the new categorization are illustrated below in Figure 5.

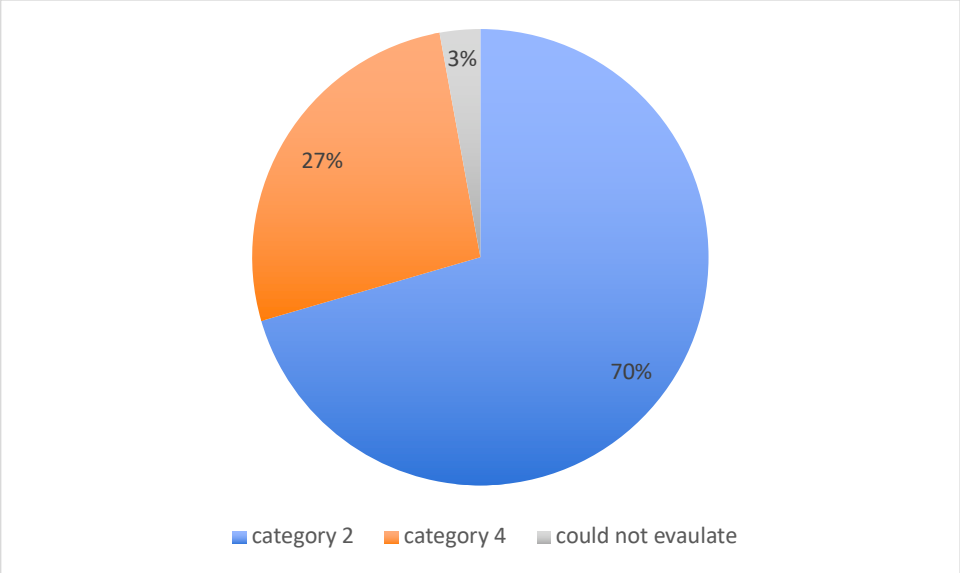

Figure 5. Percentage of the newly evaluated categories, with the total count being 244. Note that the majority of these associations fell under category 2 (e.g. positive statistical trends). However, this is less than what was observed in Figure 2 percentage-wise (74%).

4.2 Conclusion on Rule Analysis

Because these new associations are of a smaller confidence range $[0.3, 0.5[$, it was expected that they would be weaker than the ones evaluated in section 4. As previously mentioned, confidence refers to the frequency of antecedent A and consequent B being observed together. Hence a high confidence value indicates that A and B are seem more often and represent a positive statistical value (category 2), and vice versa. In the associations from section 4, 74% of the total evaluated entries fell under category 2, while the percentage is lower amongst the newly evaluated entries (70%). On the other hand, 7 associations from section 4 fell under category 1 (biologically true), while no associations from this section did. Additionally, amongst the total associations section 4, 17% of them fell under category 4 (weak statistical trends), while it was raised to 27% amongst the new set of associations.

Hence, this proves that lowering the confidence value will yield weaker associations more often than those with higher confidence values.

Certain patterns in associations were detected within each evaluated category. In category 1, biological rules were observed. These were evaluated based on already existing biological knowledge and NeuroMorpho.Org search. A few examples detected in the data are as follows:

- "If the secondary cell class is the Multidendritic-dendritic arborization (DA), the age class is larval".
- "If the secondary cell class is medium spiny, the primary region is the basal ganglia".
- "If the tertiary cell class is adult-born, the primary region is the hippocampus".

Below are a few examples of positive statistical trends observed in the data analysis. As mentioned, these are associations that are not necessarily biologically true, but are observed together at a high frequency when searching on NeurMorpho.Org.

- "If the primary cell class is glia, the primary region is most likely hippocampus".
- "If the age class is adult, the species is most likely Drosophila melanogaster".
- "If the strain is Wistar, the protocol is most likely in vitro".
- "If the primary region is neocortex, the primary cell class is most likely principal".

Glia are a cell type that do not produce electrical impulses, but rather serve a protective function within the central and peripheral nervous systems. Theoretically they are found in several regions. However, most hits on NeuroMorpho.Org reveal that glial cells are mostly associated with the hippocampus as a primary region. Additionally, several different

species can be adults, yet most hits were associated with Drosophila melanogaster. Several stains can be used with different original formats, where Neurolucida.dat was observed frequently with immunostaining. There are several ways neural tissue from the rat Wistar strain can be prepared (e.g. in vivo, ex vivo, culture, etc.), most corresponding NeuroMorpho.Org protocol entries were in vitro. The neocortex is a complex structure which possesses several cell types, such as glial, interneurons, and sensory, with most entries coming from principal.

Conversely, a handful of associations were observed to fall under category 4 (weak statistical trends). Here, when searching via NeuroMorpho.Org, the associations were not frequently observed together. In other words, antecedent A was often seen with a different consequent that consequent B. A few examples are the following:

- "If the protocol is culture, the secondary cell class is unlikely pyramidal".
- "If the secondary region is somatosensory, the species is unlikely mouse".
- "If the age class is neonatal, the species is unlikely rat".
- "If the protocol is culture, the objective type is unlikely oil".

When using the NeuroMorpho.Org search tool, the secondary cell class commonly observed with the culture protocol was not reported. Additionally, the species often seen with the somatosensory region was rat. Several mammals exhibit a neonatal, or newly born, developmental stage. When searching for species corresponding to the neonatal age class in NeuroMorpho.Org, few entries came from rat. NeuroMorpho.Org also presented a small number of entries possessing oil as an objective type when searching for the "culture' protocol.

Finally, a handful of rules could not be evaluated. As previously mentioned, these were either cell classes or regions associated with their subclasses and subregions respectively. These cannot be evaluated as they are too broad and generic. Metadata search on NeuroMorpho.Org illustrates that a given region or cell class will have several subregions or classes, making these associations unfavorable and impractical for our analysis. Additionally, several brain regions consist of layers 2, 2-3 and 5, and dorsal regions so the primary and secondary regions associated with these cannot be determined. As a result, these associations could not be properly assessed.

4.3 Negation Rule Formation

It is possible to generate negation rules based on our previously evaluated association rules. To reiterate, a negation rule involves associations that cannot exist due to biological limitations. A few examples are as follows:

- "If the primary region is neocortex, the species cannot be cricket".
- "If the strain name is Thr-Gal4, the secondary cell class cannot be pyramidal".
- "If the objective is electron microscopy, the stain cannot be red fluorescent protein".

In the first case, the neocortex is a mammalian structure and therefore cannot be found in crickets. Likewise, pyramidal cells are found in mammals and cannot be seen in Drosophila strains such as Thr-Gal4. Finally, only a light microscopy objective can be utilized in visualizing ref fluorescent protein, unlike electron microscopy which cannot detect fluorescence.

In the process of negation rule formation, we initially created an Excel sheet which contains each metadata entry found on NeuroMorpho.Org along with the metadata category (e.g. region, cell class, strain, species, objective type, etc…) and their observed frequency. This is illustrated in Table 9 below.

Table 9. Sample of metadata entries with their corresponding observed frequencies in NeuroMorpho.Org. There was a total of 4148 terms in the sheet. Note that the frequency was calculated by dividing the number of times each term is observed in NeuroMorpho.Org by the total number of digital reconstructions (~147K).

term	**category**	**frequency**
monkey	species	0.0416165726440395
CD1	strain name	0.0106353463423656
protocerebrum	region 1	0.06558904
infralimbic	region 3	0.000343502
microglia	class 2	0.070114016
in vivo	protocol	0.30863643
tangential	slicing direction	0.00249038855346078
horseradish peroxidase	stain	0.00175053837312229
transgenic	experimental condition	0.00130134362077393

Next, we color coded each term by metadata category such that they can be individually grouped. We then organized each term by descending frequency. Here, the term with the highest frequency was control (experimental condition) at 0.704119380111241; the term with the lowest frequency was DYNC1I1 knockdown+BRAF (experimental condition) at $6.60580518159358 \times 10^{-6}$. When forming negation rules, the most useful rules are those containing entries which are more frequently observed. As a result, a threshold of 1% (0.01) inclusive was set to include terms that are observed at an equal or greater frequency. With

this threshold applied, terms with smaller observed frequencies were discarded and this resulted in a total of 179 terms ready to be utilized in the negation rule formation process. To continue the process, combinations of different antecedents and consequents were formed in a different Excel sheet. Here we paired different terms with one another and calculated their expected frequencies by multiplying the two frequencies together. This is illustrated in Table 10 below. From this, we ended up with a total of 15931 non-repeated pairs.

Table 10. Sample of term pairs and calculated expected frequencies for each pair. For example, the expected frequency of control and principal cell was calculated by multiplying the control frequency (0.70411938) by the principal cell frequency (0.69870262), giving us 0.49197006. This was conducted on all 15931 pairs.

	control					
principal cell	0.49197006	**principal cell**				
in vitro	0.36032501	0.35755304	**in vitro**			
adult	0.34117571	0.33855106	0.24795902	**adult**		
Neurolucida	0.32816609	0.32564152	0.23850391	0.22582874	**Neurolucida**	
oil	0.26418779	0.26215541	0.19200589	0.18180183	0.17486941	**oil**
neocortex	0.25455035	0.25259211	0.18500161	0.17516979	0.16849026	0.1356419

Once all the pairs were formed and their expected frequencies were calculated, candidate negation pairs were extracted. Here, we sifted through the pairs and determined which ones cannot exist from a biological standpoint. From this, we ended up with 1157 potential negation rules. A few examples are shown below in Table 11.

Table 11. Sample of potential negation rules and their expected frequencies extracted from the total term pairs

antecedent	antecedent type	consequent	consequent type	expected frequency
neocortex	region 1	Drosophila melanogaster	species	0.079301756
larval	age class	Cynomolgus	strain name	0.002351279
microglia	class 2	retina	region 1	0.002683546
CA3	region 2	day5 born	class 3	0.000201429
electron microscopy	objective type	lucifer yellow	stain	0.005659078
in vivo	protocol	coronal	slicing direction	0.095499062

In addition to the expected frequency, it is necessary to determine the observed frequency of each negation rule pair such that both frequencies can be compared. This was done by conducting a NeuroMorpho.Org search on each evaluated rule in order to obtain the number of instances they are seen together in the database. Using the keyword search feature, they were searched for in the style of "A & B", where "A" is the antecedent and "B" is the consequent. From this, the number of hits from the current criteria were identified. In theory, if the number of hits is zero, the association is likely a negation rule. As noted in our Excel sheet, many negation rules did not have an expected frequency of zero. Therefore, this implies that we expect to see the antecedent and consequent at a given frequency on NeuroMorpho.Org even if it is impossible for them to be seen together. This emphasizes the importance of determining the observed frequency as it confirms that our potential negation rules are indeed negation rules if they have a zero number of hits.

Rules that did not result in zero number of hits were removed from the analysis. In the end, the entire process resulted in a total of 1006 non-reciprocal negation rules. A sample of these rules is illustrated in Table 12. More examples of our negation rules will be discussed below.

Table 12. Sample of the negation rules generated from the term pairs. Each negation rule association has been searched for on NeuroMorpho.Org such that each had zero number of hits.

antecedent	**consequent**
embryonic (age class)	mushroom body (region 2)
day7 born (class 3)	cerebellum (region 1)
electron microscopy (objective type)	biocytin (stain)
neocortex (region 1)	Trh-Gal4 (strain)
Multidendritic-dendritic arborization (DA) (class 2)	ganglion layer (region 2)
hypothalamus (region 1)	mitral (class 2)
zebrafish (species)	primary somatosensory (region 3)
Han Wistar (strain)	adult central complex (region 1)

A few examples of negation rules and their explanations are described below. With the term Sprague-Dawley (strain), we generated the following negation rules:

- "If the strain name is Sprague-Dawley, the age class cannot be larval".
- "If the strain name is Sprague-Dawley, the primary region cannot be optic lobe".
- "If the strain name is Sprague-Dawley, the tertiary cell class cannot be day1 born".

Sprague-Dawley is a rat strain and therefore does not go through a larval stage. Moreover, the optic lobe is a structure found in insects, not mammals. Finally, day1 born cells are only found in Drosophila melanogaster.

Moreover, with the term protocerebrum (region 1), we produced the following negation rules:

- "If the primary region is protocerebrum, the species cannot be monkey".
- "If the primary region is protocerebrum, the strain name cannot be C57BL-6J".

- "If the primary region is protocerebrum, the secondary cell class cannot be medium spiny".

The protocerebrum is a region typically found in insects. Therefore, it is never found in monkeys or in mouse strain C57BL-6J. Additionally, it does not possess any medium spiny cells as they are found in the basal ganglia.

We created the following negation rules using human (species) as the antecedent:

- "If the species is human, the secondary cell class cannot be Multidendritic-dendritic arborization (DA)".
- "If the species is human, the secondary region cannot be right medulla".

Multidendritic-dendritic arborization (DA) cells and the right medulla are found in insects, and therefore cannot exist in humans.

While the majority of our negation rules involved either species, strain, cell classes and regions, we were able to generate a few rules containing other metadata categories such as objective type and stain. The following rules were produced using electron microscopy (objective type) as the antecedent:

- "If the objective type is electron microscopy, the stain cannot be green fluorescence protein".
- "If the objective type is electron microscopy, the stain cannot be lucifer yellow".
- "If the objective type is electron microscopy, the stain cannot be biocytin".

Electron microscopy relies on viewing electron dense (dark) and electron lucent (light) cellular structures. As a result, the structures will be viewed in greyscale, and fluorescently

colored stains such as green fluorescence protein, lucifer yellow, and biocytin cannot be used as they will not be detectable in view.

4.4 Machine Learning and Preliminary NER Results

An NER model named BERT (Bidirectional Encoder Representations from Transformers) was utilized in the machine learning process. BERT is a well-known language modeling algorithm recently proposed and supported by Google. It benefits in its application of bidirectional training for natural language processing application as opposed to traditional algorithms that rely on examining at a text from a solely left-to-right or a right-to-left sequencing. This enables the algorithm to understand, with greater clarity, the context of a named entity recognition given the words surrounding it (Alammar, 2018).

A set of training data must be prepared for the BERT language model. About 10k sentences need to be manually tagged for this task. While 10k may sound extensive, it is necessary to supply the algorithm with a large quantity of tagged annotations as it has a larger corpus it can learn from, thus making independent annotations more successful. The manual tagging process is performed based on an active learning method, where a human annotator will tag certain neuronal named entities (e.g. brain region, cell class, reconstruction software) based on neuroscience knowledge. These annotations are then provided back and forth as a feedback to the machine learning algorithm until we have a final set. A few examples of tagged sentences are illustrated in section 3.3. Additional examples are represented below in Figure 6.

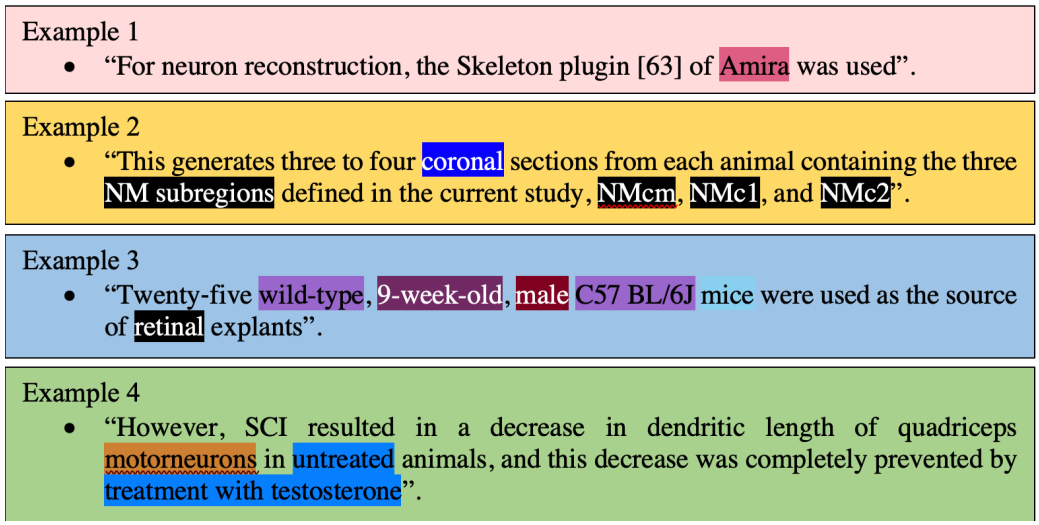

Figure 6. Additional examples of color-coded labeled entries we annotated. Note that the color-coding scheme is identical to the one observed in section 3.3.

The sentences which are ready for annotation are extracted from 1000+ manually curated peer-reviewed articles and publications available on NeuroMorpho.Org. After the training phase, the learning model, BERT, is required to extract different neuronal metadata fields. Based on the frequency of named entities and where they occur within the article a certain probability is assigned to them and they can be chosen as the target named entity based on that probability. Additionally, the model makes use of the association rules in its final analysis. It does so by computing the application ratio of different rules being observed within the entries of interest. Hence, a set of extracted named entities and their associated probability compared to the ones extracted by a human annotator can be produced (Table 13).

Table 13. Probabilities of extracted named entities from a specific article by the machine learning model compared to the target named entities extracted by a human annotator. The numbers next to each metadata term correspond to the frequency at which they are seen in peer-reviewed article texts.

metadata category	human annotator	machine learning algorithm
species	mouse	mouse 0.94. human 0.05
strain	BALB/c	BALB/c 0.75, C57BL 0.10, C57BL/6 0.04, …
gender	female	female 0.66, male 0.33
developmental stage	adult	day 7 0.33, 6 and 14 week of age 0.16, …
cell type	glia, microglia, Iba1-positive	microglia 0.71, t cell 0.12, microglium cell 0.047, large 0.01, …
brain region	neocortex	Cn 0.60, bone marrow 0.12, cortex 0.11, central 0.05, …
staining method	immunostaining	immunostaining 0.6, multiple 0.15, mantel-cox 0.09, …
protocol	in vitro	in vitro 0.4, in vivo 0.4, culture 0.2
reconstruction software	Imaris	Imaris 1.0

The key neural metadata terms the machine learning algorithm detects while sifting through peer reviewed publications are saved in an Excel sheet, along with the article PMID of the article. The named entity cases that have been evaluated as of now are objective (dry), gender (male, female, male/female), and protocol (in vivo, in vitro, and ex vivo).

While the machine learning model does have a considerable success rate at detecting named entities of interest, it does have its limitations. Here, it cannot understand the contextual background of a given sentence or text, hence making it more difficult to identify metadata terms of interest within publications. For example, within objective usage, it is assumed that a dry objective is used in microscopic analysis unless stated

otherwise (i.e. usage of water or oil immersion objectives). As a result, the objective type is not explicitly mentioned in the article. In this case, a human reading the publication will be able to recognize that a dry objective is being used, but the machine was unable to do so in its independent analysis. Moreover, in some rare instances, a "dry" objective is referred to an "air" objective. Because the model has not been trained to recognize the term "air", it failed to identify the term when performing independent named entity recognition. Additional instances of limitations were observed in the recognition of gender named entities. The model has been trained to recognize the terms "male", "female", or both "male" and "female". However, there are many other terms and phrases a publication will use when referring to the gender or sex of the animals used in the study, such as "gender", "sex", "animals of either sex", "cross/crosses/crossed/crossing", "mate/mates/mated/mating", and "breed/breeds/bred/breeding". As a result, the model was unable to recognize them because it has not yet been trained to do so.

Limitations were also observed in the analysis of protocol named entities. While sifting through the articles, the model was only able to identify the terms "in vivo", "in vitro", and "ex vivo". However, peer reviewed publications rarely utilize these terms when describing their experimental protocols. In other words, they use conceptual phrases and sentences to describe which protocol is being used. For example, terms and phrases describing an in vivo protocol include "conditioning" (e.g. fear), "dye injection", "electrode/electrodes", "implants", "live", "maze" (e.g. water), "object recognition test", and "viral injection". On the other hand, terms and phrases describing an in vitro or ex vivo protocol include "fixed", "immunohistochemistry", "mounted", "section/sections/sectioned", "slice/slices/sliced",

"slides", "tissue preparation", and "tissue/tissues". Furthermore, the terms "dish", "Petri dish", and "plate" are used in describing a culture protocol. A human sifting through these articles can identify the protocol being used by making note of these contextual words and phrases, Because the model has not been trained to identify these conceptual terms and phrases, it failed to recognize them in its independent annotating process. Some additional keywords are represented in Table 14 below.

Table 14. Classification of protocol keywords as observed in the publications. They are sorted according to protocol type (e.g. in vivo, in vitro, ex vivo, culture, and all).

in vivo	**ex vivo**	**in vitro**	**culture**	**all**
conditioning (e.g. fear)	fixed	fixed	dish	prepared
dye injection	immunohistochemistry	immunohistochemistry	Petri dish	protocol
electrode(s)	mounted	mounted	plate	-
implant(s)	section(ed)	section(ed)	-	-
live	slice(s)	slice(s)	-	-
maze (e.g. water)	slide(s)	slide(s)	-	-
object recognition test	tissue preparation	tissue preparation	-	-
viral injection	tissue(s)	tissue(s)	-	-

We analyzed and sifted through the articles on NeuroMorpho.Org to assess the integrity of the machine learning model. Then, we searched for keywords which represent the type of metadata (e.g. objective, gender of animal, and protocol) being used in the experimental procedure. The sentence or sentences describing the metadata terms of interest were

recorded as well. These have been saved in the form of Excel sheets and a sample of this is illustrated in Table 15.

Table 15. Sample of evaluated metadata cases as seen on our Excel sheets. The specific metadata case illustrated here is gender. Note that sentences containing or describing the metadata term of interest have been recorded from every evaluated article. Sentence examples will be included below.

PMID	machine learning model prediction	reported in the article	our findings
14978208	none	“male”	none
17382886	“male”, “female”	“male/female”	“male/female”
21653851	none	“male/female”	“male/female”
25906337	none	“male”	“male”
27651000	both	“male/female”	“male/female”

As mentioned, sentences containing keywords of interest were recorded in our Excel sheet. Below are some examples of sentences we found when searching for the dry/air objective metadata in peer reviewed publications:

- “Imaging was performed either in a Zeiss Observer Z.1 microscope equipped with a Plan-Apochromat 20× air objective (0.8 numerical aperture); an AxioCam HRm camera and Zen Blue 2011 software”.
- “Confocal images were obtained using a Leica TCS SPE confocal microscope with a 10× air objective (numerical aperture: 0.3)”,
- “Cells were imaged at room temperature using an inverted Nikon Eclipse T microscope using a 0.75 NA40x air objective, Andor Xyla camera, and Micro-Manager software (Edelsteinetal., 2010)”.

Some examples of sentences we recorded which describe gender metadata of interest are illustrated below:

- "Acute brain slices were prepared from C57Bl6N mice of either sex except for population Ca2+ imaging experiments with GCaMP".
- "We used C57BL/6 mice of both genders at the age of 3 months".
- "Both male and female mice were used for all experiments".
- "The queens of the colonies used for these experiments were inseminated with the semen of a single drone (different for each colony) in order to reduce genetic variation among daughter workers (honey bee queens naturally mate with 10–20 males) in the colony." Note that the keyword here is "queens", which refers to female bees.
- "Briefly, horizontal brain slices (300 μM) were prepared from male, Wistar rats >30 days old under protocols approved by Rutgers-NJMS, Newark, NJ, IACUC".

Finally, sentences which state or describe protocol metadata entries of interest include the following:

- "Transverse hippocampal slices (300 μm thickness) were cut from brains of 18- to 25-day-old mice using a vibratome (DTK-1000, Dosaka)."
- "For immunohistochemistry and biolistic transfection, the eyes were then transferred to oxygenated mouse artificial cerebrospinal fluid (mACSF, pH 7.4) containing (in μm) 119 NaCl, 2.5 KCl, 2.5 CaCl2, 1.3 MgCl2, 1 NaH2PO4, 11 glucose and 20 HEPES at room temperature".

- “Spinal cords and brains were isolated, post-fixed in 4% PFA, and dehydrated in 30% sucrose”.
- “Confocal imaging was performed on live, morphologically normal animals expressing EGFP or one of the Kv3.3-EGFP fusion proteins”.
- “General assessment of neurology, testing in motor coordination tests, prepulse inhibition (PPI), fear conditioning (FC), 8-arm radial maze (8ARM), open field (OF), tail suspension test (TST), forced swim test (FST), bright open field (BOF), light and dark box (LDB), elevated plus maze (EPM) and quantification of neurotransmitters by high-pressure liquid chromatography (HPLC) were performed in parallel on age-matched male and female littermates”.
- “Mutant monosynaptic rabies virus Rb-B19-ΔG-GFP (purchased from Salk Institute) was bilaterally injected into the intermediate gray matter (laminae V – VIII) of the L2 spinal cord (0.75µl/side, 10∧9TU/ml) using a stereotaxic apparatus”.

CONCLUSION

In this thesis, we investigated ways to facilitate the process of metadata extraction and annotation for NeuroMorpho.Org. We determined the best methods on how to automate this process by preparing data and analyzing the statistical trends of current metadata records in the neural repository. In this regard, we initially described the process of evaluating the statistical trend of extracted metadata rules from the NeuroMorpho.Org's repository (the chances of rules occurring in high or low frequencies). Additionally, we generated negation rules (trends that are less likely to occur in the future records of the data) based on biological knowledge and searching for the number of hits on NeuroMorpho.Org. Afterwards, we paved the way for and explained the incorporation of a machine learning model in the process of independent data annotation. This was performed by providing tagged sentences containing named entities of interest for the machine learning model.

In the process of rule evaluation, the majority of the rules we evaluated were likely statistical trends. However, when we lowered the confidence threshold, we noticed that the percentage of likely statistical trends observed has decreased. Negation rules were generated from our evaluated rules, such that the rules cannot exist biologically or cannot be found via NeuroMorpho.Org search. Finally, we performed named entity recognition to search for key metadata terms of interest in peer-reviewed publications. We then compared our annotations to the model's annotations. This proved that while the model's responses are quite promising at independent annotation, it has limitations as it cannot detect

contextual phrases which describe the metadata of interest. As a result, it will have to learn from a larger corpus of sentences containing terms of interest with more contextual information such that it can conduct a more successful independent annotation process in the future.

.

REFERENCES

Agrawal, R. (1994). Fast Algorithms for Mining Association Rules. Proceedings of the 20th International Conference on Very Large Data Bases, 487–499.

Alammar, J. (2018). The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning).

Ascoli, G. A., Donohue, D. E., & Halavi, M. (2007). NeuroMorpho.Org: A Central Resource for Neuronal Morphologies. Journal of Neuroscience, 27(35), 9247–9251. https://doi.org/10.1523/JNEUROSCI.2055-07.2007

Bijari, K., Akram, M. A., & Ascoli, G. A. (2020). An open-source framework for neuroscience metadata management applied to digital reconstructions of neuronal morphology. Brain Informatics, 7(1), 2. https://doi.org/10.1186/s40708-020-00103-3

Chu, P., Peck, J., & Brumberg, J. C. (2015). Exercises in Anatomy, Connectivity, and Morphology using Neuromorpho.org and the Allen Brain Atlas. Journal of Undergraduate Neuroscience Education: JUNE: A Publication of FUN, Faculty for Undergraduate Neuroscience, 13(2), A95–A100.

Halavi, M., Polavaram, S., Donohue, D. E., Hamilton, G., Hoyt, J., Smith, K. P., & Ascoli, G. A. (2008). NeuroMorpho.Org Implementation of Digital Neuroscience: Dense Coverage and Integration with the NIF. Neuroinformatics, 6(3), 241–252. https://doi.org/10.1007/s12021-008-9030-1

Han, J., & Kamber, M. (2012). Data mining: Concepts and techniques (3rd ed). Elsevier.

Hand, D. J. (2007). Principles of Data Mining: Drug Safety, 30(7), 621–622. https://doi.org/10.2165/00002018-200730070-00010

Helmstaedter, M. (2015). The Mutual Inspirations of Machine Learning and Neuroscience. Neuron, 86(1), 25–28. https://doi.org/10.1016/j.neuron.2015.03.031

Hossain, M., Sattar, A. H. M. S., & Paul, M. K. (2019). Market Basket Analysis Using Apriori and FP Growth Algorithm. 2019 22nd International Conference on Computer and Information Technology (ICCIT), 1–6. https://doi.org/10.1109/ICCIT48885.2019.9038197

Inokuchi, A., Washio, T., & Motoda, H. (2000). An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In D. A. Zighed, J. Komorowski, & J.

Żytkow (Eds.), Principles of Data Mining and Knowledge Discovery (Vol. 1910, pp. 13–23). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45372-5_2

Jun, S. (2014). A Technology Forecasting Method using Text Mining and Visual Apriori Algorithm. Applied Mathematics & Information Sciences, 8(1L), 35–40. https://doi.org/10.12785/amis/081L05

Martínez-Romero, M., O'Connor, M. J., Egyedi, A. L., Willrett, D., Hardi, J., Graybeal, J., & Musen, M. A. (2019). Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases. Database, 2019, baz059. https://doi.org/10.1093/database/baz059

Nengsih, W. (2015). A comparative study on market basket analysis and apriori association technique. 2015 3rd International Conference on Information and Communication Technology (ICoICT), 461–464. https://doi.org/10.1109/ICoICT.2015.7231468

NeuroMorpho.Org. (n.d.).

Ozoh, P. A., Adigun, A.A., & Omotosho, L. O. (2020). A Comparative Analysis of Machine Learning Techniques. International Journal of Research and Innovation in Applied Science (IJRIAS), V(IV), 146–152.

Parekh, R., Armañanzas, R., & Ascoli, G. A. (2015). The importance of metadata to assess information content in digital reconstructions of neuronal morphology. Cell and Tissue Research, 360(1), 121–127. https://doi.org/10.1007/s00441-014-2103-6

Parekh, R., & Ascoli, G. A. (2013). Neuronal Morphology Goes Digital: A Research Hub for Cellular and System Neuroscience. Neuron, 77(6), 1017–1038. https://doi.org/10.1016/j.neuron.2013.03.008

Riley, J., & National Information Standards Organization (U.S.). (2017). Understanding metadata: What is metadata, and what is it for? http://www.niso.org/publications/understanding-metadata-riley

Shardlow, M., Ju, M., Li, M., O'Reilly, C., Iavarone, E., McNaught, J., & Ananiadou, S. (2019). A Text Mining Pipeline Using Active and Deep Learning Aimed at Curating Information in Computational Neuroscience. Neuroinformatics, 17(3), 391–406. https://doi.org/10.1007/s12021-018-9404-y

Squire, L. R. (Ed.). (2013). Fundamental neuroscience (4th ed). Elsevier/Academic Press.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining (1st ed). Pearson Addison Wesley.

Tomer, R., Denes, A. S., Tessmar-Raible, K., & Arendt, D. (2010). Profiling by Image Registration Reveals Common Origin of Annelid Mushroom Bodies and Vertebrate Pallium. Cell, 142(5), 800–809. https://doi.org/10.1016/j.cell.2010.07.043

Trappenberg, T. P. (2010). Fundamentals of computational neuroscience (2nd ed). Oxford University Press.

Yanbin Ye, & Chia-Chu Chiang. (2006). A Parallel Apriori Algorithm for Frequent Itemsets Mining. Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06), 87–94. https://doi.org/10.1109/SERA.2006.6

Zhang, L., Pan, Y., & Zhang, T. (2004). Focused named entity recognition using machine learning. Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval - SIGIR '04, 281. https://doi.org/10.1145/1008992.1009042

BIOGRAPHY

Yasmeen Zoubi graduated from McGill University with a Bachelor of Science in Anatomy and Cell Biology in 2018. She continued her education and obtained a Master of Science in Biology at George Mason University in 2021. She wishes to pursue a career in academic research, specifically in the clinical and health care applications of neuroscience.