Neuroinformatics
https://doi.org/10.1007/s12021-022-09566-7

The Neuron Phenotype Ontology: A FAIR Approach to Proposing and Classifying Neuronal Types

Thomas H. Gillespie[1] · Shreejoy J. Tripathy[2,3,4] · Mohameth François Sy[5] · Maryann E. Martone[1] · Sean L. Hill[2,3,4,5]

Accepted: 19 January 2022
© The Author(s) 2022

Abstract

The challenge of defining and cataloging the building blocks of the brain requires a standardized approach to naming neurons and organizing knowledge about their properties. The US Brain Initiative Cell Census Network, Human Cell Atlas, Blue Brain Project, and others are generating vast amounts of data and characterizing large numbers of neurons throughout the nervous system. The neuroscientific literature contains many neuron names (e.g. parvalbumin-positive interneuron or layer 5 pyramidal cell) that are commonly used and generally accepted. However, it is often unclear how such common usage types relate to many evidence-based types that are proposed based on the results of new techniques. Further, comparing different types across labs remains a significant challenge. Here, we propose an interoperable knowledge representation, the Neuron Phenotype Ontology (NPO), that provides a standardized and automatable approach for naming cell types and normalizing their constituent phenotypes using identifiers from community ontologies as a common language. The NPO provides a framework for systematically organizing knowledge about cellular properties and enables interoperability with existing neuron naming schemes. We evaluate the NPO by populating a knowledge base with three independent cortical neuron classifications derived from published data sets that describe neurons according to molecular, morphological, electrophysiological, and synaptic properties. Competency queries to this knowledge base demonstrate that the NPO knowledge model enables interoperability between the three test cases and neuron names commonly used in the literature.

Keywords Neurons · Cell types · Ontology · Knowledge base · Interoperability · Knowledge integration · FAIR principles

Introduction

The modern description and classification of neurons and the diversity of their properties began with the work of Santiago Ramon y Cajal over 100 years ago. Cajal benefitted from a newly discovered technique, the Golgi stain, to reveal neurons as individual entities of remarkably different shapes, which he described as the "butterflies of the soul". Our knowledge of neuron types (as with cell types) has continued to evolve as new experimental techniques emerge. For this reason, a centerpiece of the US Brain Initiative is to re-examine what constitutes a cell type in light of new ways of probing the nervous system. Through the BRAIN Initiative Cell Census Network (BICCN) researchers are generating large pools of data using cutting edge methods that are being integrated across data types through the use of standards such as common spatial and semantic mappings (Ecker et al., 2017). The BICCN joins several other large initiatives such as the Blue Brain Project (Markram, 2006), Human Cell Atlas (Regev et al., 2017), and SPARC (https://sparc.science/) which also seek to provide foundational knowledge on the types of cells that make up the nervous system. As these data are analyzed and synthesized, new ways to distinguish

✉ Sean L. Hill
sean.hill@epfl.ch

1 Department of Neuroscience, University of California, San Diego, CA, USA

2 Department of Psychiatry, University of Toronto, Toronto, ON, Canada

3 Department of Physiology, University of Toronto, Toronto, ON, Canada

4 Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, ON, Canada

5 Blue Brain Project, École Polytechnique Fédérale de Lausanne (EPFL), Campus Biotech, 1202 Geneva, Switzerland

Published online: 10 March 2022

among different classes of neurons are being proposed and published.

One of the end goals of these large projects is to integrate and analyze large quantities of cellular data to derive new taxonomic classification of neurons across neural structures and to arrive at a new understanding of what constitutes a cell type in the nervous system. To manage this process, some have called for a consistent naming scheme for neurons, so that as new types are discovered, their findings can be reported and compared in an organized way (DeFelipe et al., 2013; Hamilton et al., 2012; Shepherd et al., 2019). Biology has a long history of successfully developing and deploying taxonomies and naming conventions for new entities, e.g., species, enzymes. The process usually involves the commissioning of an authoritative body that comes up with a regularized method and vocabulary for distinguishing among different types and applying an appropriate nomenclature. This approach has been attempted for neuron types. For example, the Petilla terminology proposed a set of criteria and controlled terminology for naming cortical interneurons based on traditional electrophysiological and morphological measurements (Petilla Interneuron Nomenclature Group et al., 2008). However, developing taxonomies and naming conventions pre-supposes that we understand the key dimensions across which neurons should be classified and the foundations of what constitutes a cell type. If the methodological foundations for the classification have not yet reached something universally agreed upon as foundational, such as a nucleotide or amino acid sequence, then the classification remains technique dependent. Thus, as new technologies enable further characterization of additional dimensions,including some that may be foundational, our concept of cell types is likely to evolve. While we know that existing techniques for determining cell type have not yet been able to measure something as foundational as a nucleotide sequence, recent large integrative data gathering exercises have tended to refine our current concepts rather than replace them (Osumi-Sutherland, 2017). A single cell transcriptomic analysis of retinal bipolar cells, (Shekhar et al., 2016), detected 17 different types of RBC, 15 of which had been previously described. The challenge remains to define a knowledge representation that can readily adapt to and integrate results from new data-driven taxonomic efforts but which still supports references to classical naming schemes to ensure integration with the large amount of historical published knowledge. Further, even when foundational techniques can be routinely deployed at scale, not all experiments and certainly not all clinical use cases will be able to employ those techniques directly. Thus, our knowledge management systems need to explicitly account for the techniques that are required to perform such classification so that mappings to other techniques can be developed.

Most proposed schemes, to date, comprise a hierarchical method based on various phenotypic properties for their foundation, i.e., key molecular, physiological, and connectivity signatures that distinguish a neuron type. Phenotypic properties are typically properties of a neuron which are consistent across a variety of measurements, although many phenotypic properties can only be consistently reproduced with a specific experimental technique or protocol. Given the multiple dimensions across which neurons can be differentiated, a phenotype-based approach for classification could effectively generate an almost infinite number of ways to categorize neurons, depending on the granularity at which the distinctions are expressed. A single taxonomy that effectively organizes neurons across these dimensions is unlikely. The recent proposal for naming cortical neurons by (Shepherd et al., 2019) shows how quickly the number of phenotypes can explode, particularly when trying to address the results of dense phenotypic sampling such as array expression. Thus for neuronal cell types, given the complexity and variety of potentially distinguishing features and the likely evolution of these over time, any system for communicating and comparing across phenotypes will require a firm computational foundation.

Traditionally, such proposed classifications are communicated through the research paper, where any taxonomy proposed is presented in the form a table, dendrogram or some other figure (e.g., Paul et al., 2017, Table S7; Markram et al., 2015, Table 1). The problem with our traditional way of constructing and communicating these taxonomies is that they require a human being to understand, compare, and reconcile them (Petilla Interneuron Nomenclature Group et al., 2008). Anyone who has attempted to read through multiple articles, each with their own proposal for classifying cell types within a region understands the difficulties in trying to reconcile the different schemes, even when they are based on limited numbers of data dimensions. The multiplicity of papers proposing classification schemes just for cortical interneurons illustrates this point (Cauli et al., 1997). With the BICCN and other large scale consortia tasked to map the cellular landscape of the brain and body, the potential number of these taxonomies is likely to explode beyond the current already unmanageable number, as researchers apply new types of analytics to understand the data. For neuroscience to move beyond paper-based forums for discussion and integration, we need to treat taxonomies and names as computable artifacts that comply with the FAIR data principles, FAIR = Findable, Accessible, Interoperable and Reusable; (Wilkinson et al., 2016).

Towards that end, we have developed an ontology-based data model, the Neuron Phenotype Ontology (NPO). The NPO aims to provide an interoperable representation of cell

Table 1 The current Phenotypic Dimensions of the NPO and the associated ontologies/vocabularies used to populate the data model. When NIFSTD appears in this table the terms were nearly always added to support the NPO. Examples are drawn from Fig. 2

Phenotypic dimension	Definition	Vocabularies/ontologies
Taxonomic Example: Species	The species or taxon rank in which the phenotype inheres	NCBI taxonomy[1]
Anatomical Example: Brain Region	The regions of the nervous system containing parts of the neuron. Primary location is indicated by the location of the cell soma, but anatomical location may be assigned to any cell part through a series of predicates	UBERON; various brain atlases via NIFSTD parcellation[2]
Morphological	Distinguishing morphological characteristics	NIFSTD[3]
Molecular Example: Expression	Distinguishing molecular constituents	NCBI Gene[4], CHEBI[5], Protein Ontology[6]
Physiological	Expresses a relationship between a neuron type and an electrophysiological phenotype concept. This should be used when a neuron type is described using a high level electrophysiological concept class, e.g., bursting	NIFSTD Petilla Conventions (Petilla Interneuron Nomenclature Group, 2008)
Connection	Indicates a synaptic relationship between cell types. Further elaborated into connectivity determined by different techniques, e.g., physiology, electron microscopy	Gene Ontology[7]
Circuit role Example: Projection	Indicates whether the neuron is an Intrinsic neuron (local circuit neuron), projection neuron, or sensory neuron	NIFSTD (Bug et al., 2008)
Projection targets Example: Projection	Expresses a relationship between a neuron type and a brain region to which it sends axons. Synaptic relationships are represented through the connection relationship	UBERON (Mungall et al., 2012)/various atlases/NIF Gross Anatomy (Bug et al., 2008)

[1] https://www.ncbi.nlm.nih.gov/taxonomy

[2] https://github.com/SciCrunch/NIF-Ontology/blob/master/docs/brain-regions.org

[3] https://github.com/SciCrunch/NIF-Ontology

[4] https://www.ncbi.nlm.nih.gov/gene

[5] https://www.ebi.ac.uk/chebi/

[6] https://proconsortium.org/

[7] http://geneontology.org/

types that can evolve as our phenotypic knowledge evolves, from initial data gathering to modeling and synthesis (Fig. 1). The NPO provides a computable representation of cell types defined by collections of phenotypic properties, designed to enable interoperability between neuronal taxonomies. It is designed to enable scientists to discover which cell types (or potential cell types) share similar properties and to help scientists understand when the cell types they observe are the same or similar to other cell types described in the literature or from other laboratories. Here, we show how the NPO can be used to express taxonomies proposed by different research groups using modern techniques, enable comparisons between them, and enable queries with commonly used neuron types from the literature.

Methods

Overview of NPO

The NPO as well as all data and code referenced below are available for reuse under open licenses (see Data and Code availability statement).

The NPO is composed of two parts. A set of core ontology files that define a data model for neuron types, and the NPOKB, the collection of neuron types defined using the NPO core ontology data model. See supplemental methods for details.

The NPO provides a data model for modeling a neuron type as a "bag of key phenotypes", that is, neurons are represented as a collection of phenotypic properties (Fig. 2) formalized as Web Ontology Language (OWL) classes. These properties can then be used to communicate about and compare phenotypes across laboratories, species, and experimental techniques. This approach has been demonstrated previously in the context of text-based queries of neuron type mentions (Richardet et al., 2015). The original set of object properties for the ontology were sourced from the existing NeuroLex (RRID:SCR_005402) model for neurons (Larson & Martone, 2013). As we developed the CUTs and EBTs we added new properties as needed based on the phenotypes that were measured in particular experiments.

Each of these dimensions is linked to a formal vocabulary or ontology, which is used to provide the descriptors for qualitative phenotypic attributes (Table 1). When possible,

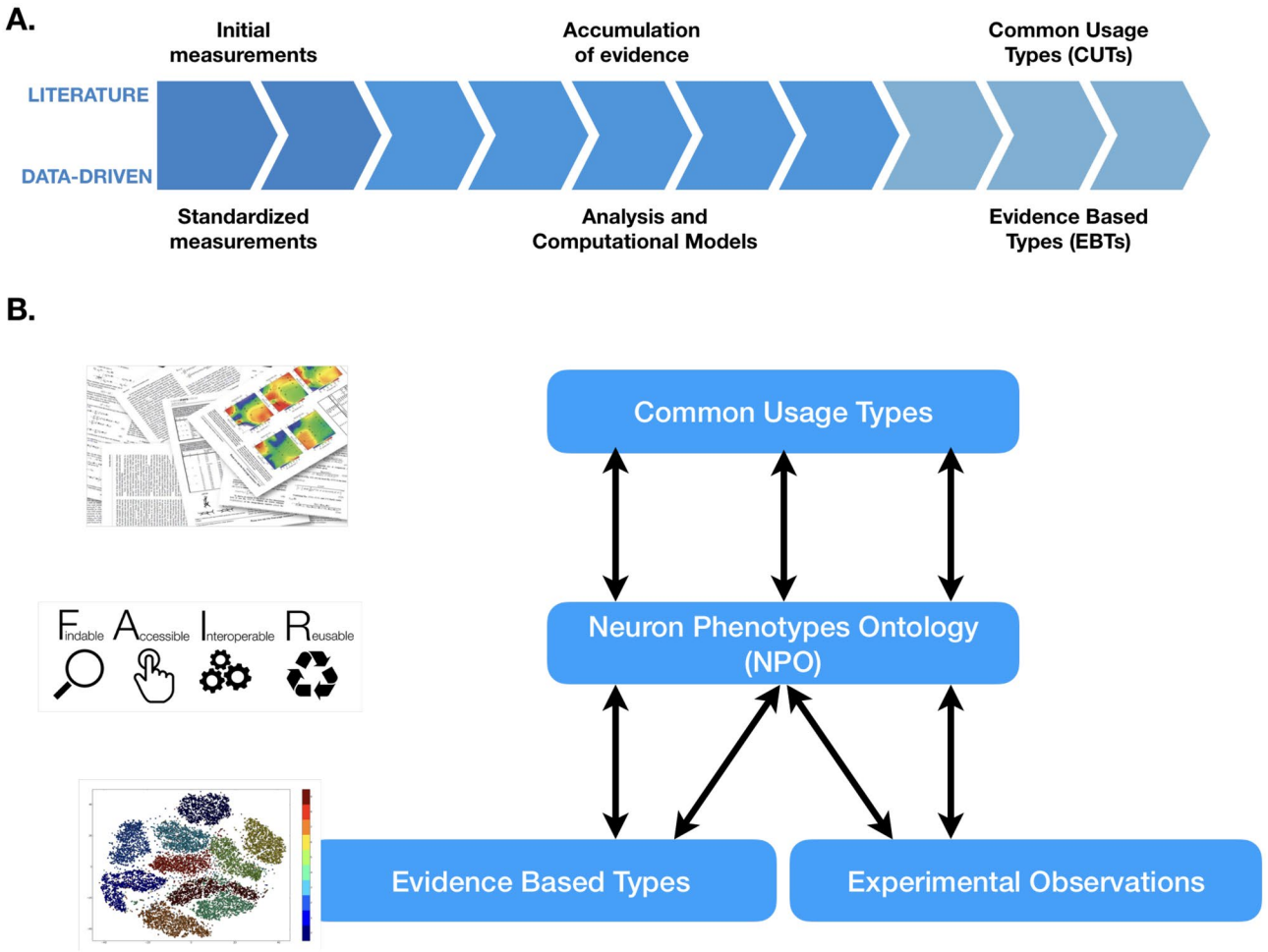

Fig. 1 Evolution of neuron knowledge **A** Common usage types (CUTs) emerge in the literature as evidence accumulated for generally accepted neuron types with implicitly known properties. Data-driven studies generate evidence-based types (EBTs) based on explicitly measured standardized properties **B** The Neuron Phenotype Ontology (NPO) provides interoperability between the CUTs from the literature, the EBTs from data-driven studies, and new experimental observations from individual laboratories

the vocabularies are drawn from community ontologies/vocabularies in broad use across biomedicine to aid in interoperability. Those dimensions that were not covered by specific community ontologies were added as classes to the appropriate branches of the NIFSTD ontology. NIFSTD is a harmonized set of neuroscience relevant ontologies developed and maintained by the Neuroscience Information Framework (Bug et al., 2008). These dimensions are further elaborated in a set of predicates that capture more granular aspects of phenotypes. For example, *hasMolecularPhenotype* can be further divided into *hasNeurotransmitterPhenotype*, *hasEpigeneticPhenotype*, and *hasExpressionPhenotype* (Fig. 3). *hasExpressionPhenotype* is further broken down into a set of predicates that captures the methodology used to reveal the phenotype. In the current version (v1) of the NPO, we have not made use of the full set of relationships to simplify the reasoning. Relationships that have not been used in the current version of the NPO or that are not in the NPO core but are planned for inclusion in the future are grayed out in Fig. 3.

For negative phenotypes, that is, where the lack of a particular phenotype is considered to be a distinguishing feature between neuron types, we use negation in OWL semantics, e.g., a parvalbumin negative neuron would be modeled as “not (*hasExpressionPhenotype* some 'parvalbumin alpha')”.

We have also included disjointness axioms[1] in cases where the strength of the assertions from the EBTs were not as definitive as full negation.

For evaluation purposes, we have used the NPO data model to construct a knowledge base of neuronal phenotypes

[1] For an introduction to disjointness axioms in ontologies see Disjointness Between Classes in an Ontology (Stevens & Sattler, 2012) http://ontogenesis.knowledgeblog.org/1260/.

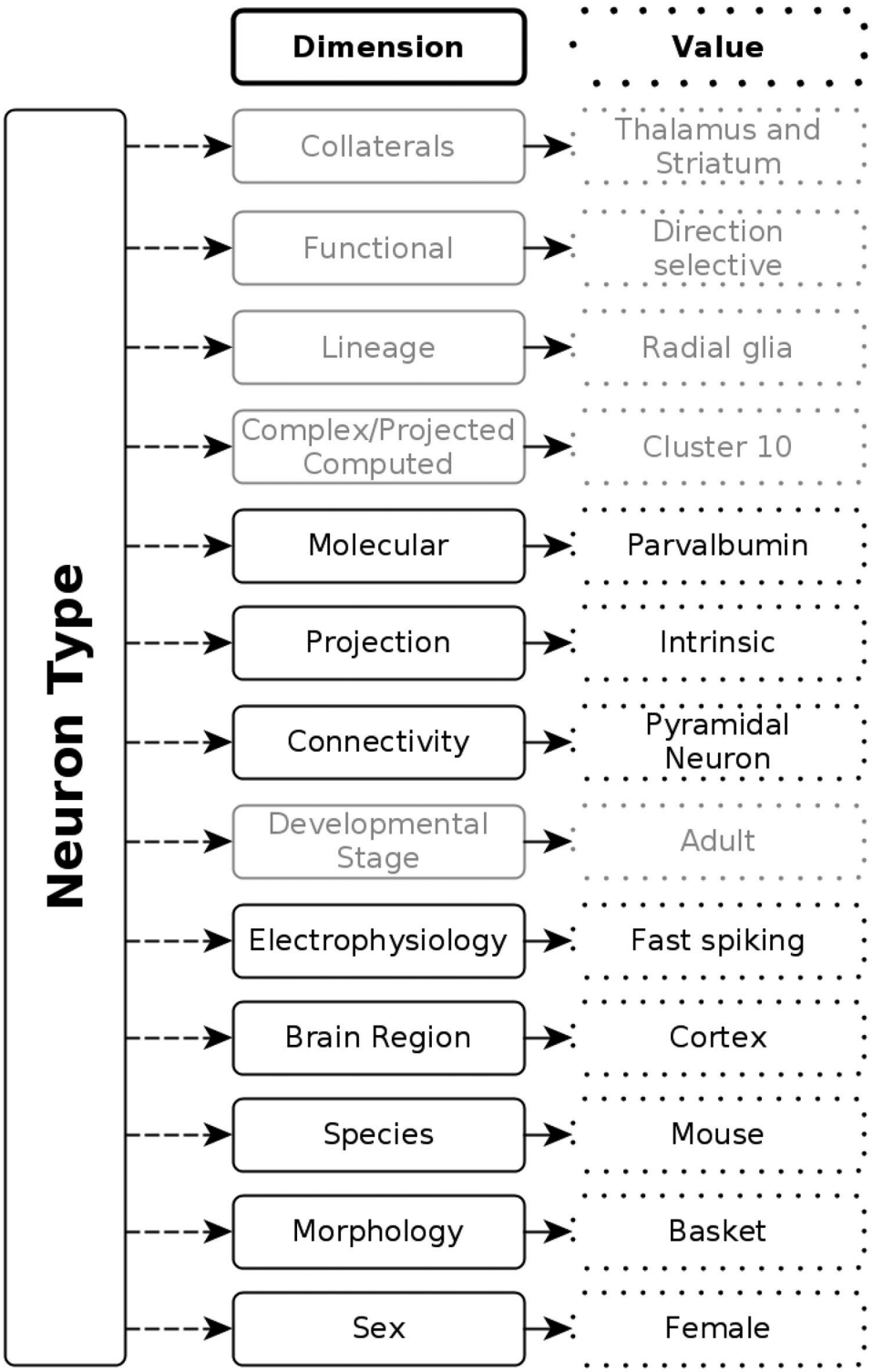

Fig. 2 High level data model for neuron phenotypes. The Neuron Phenotype Ontology characterizes neuron types as bundles of normalized phenotypic properties. Dimensions that have not been used in the current version of the NPO or are planned for the future are grayed out

comprising two branches: 1. Phenotypic representations of common usage types (CUTs) from classical morphological and physiological studies over the past 100 years; 2. Classification models arising from newer experimental techniques tied to individual projects,laboratories or initiatives, termed evidence based types (EBTs). The data model is supported by computational tools that enable individual researchers to compose the complex phenotype of a neuron out of any number of individual phenotypes that are tightly linked to individual data sets and analyses (Fig. 4). We have created

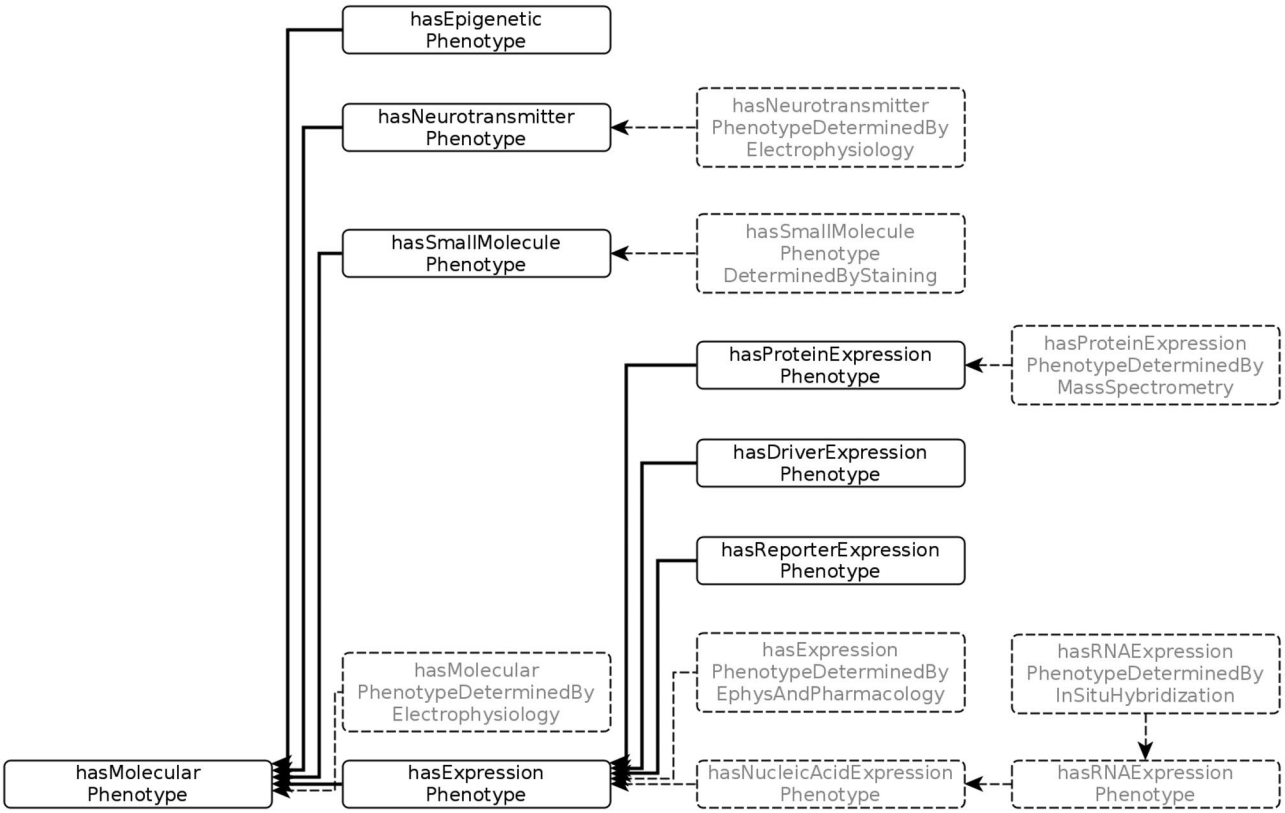

Fig. 3 The set of predicates employed to define molecular phenotypes. Relationships that have not been used in the current version of the NPO or are planned for the future are grayed out

a python library called neurondm that implements Neuron Lang, a domain specific language (DSL) for specifying neuron types. Neuron Lang was created as part of this project to provide a compact representation that expands into more verbose OWL2. The neurondm library provides tools for generating human readable neuron names based on these OWL semantics as well as tools for mapping to and from collections of local names for phenotypes by using ontology identifiers as the common language underlying all local naming. The tools allow us to automatically generate names for neurons in a regular and consistent way using a set of rules operating on the neurons' constituent phenotypes. Neuron types created using neurondm can be exported to Python or to any serialization supported by rdflib, however deterministic turtle[2] (ttl) is preferred. When neurondm generates an OWL ontology it tracks provenance by inserting the exact path and git commit hash for the source python file in the owl:Ontology section via the prov:wasGeneratedBy predicate.

Modeling Decisions

Neuron Class Names

Each neuron in the NPO is identified by a full uniform resource identifier (URI) and a compact identifier for ease of reference. The compact identifier has the prefix npokb and the ontology is registered in BioPortal[3] (RRID:SCR_002713) using the NPOKB prefix as NPO prefix was taken. Each class has multiple human readable labels assigned as annotation properties. Neurons are named according to the phenotypic properties they display. These labels are generated automatically based on the collection of phenotypic properties reported for each cell type using the neurondm Python library. Phenotypes are expressed as OWL2.0 restrictions, and neuron types as equivalent to the intersection of those restrictions (Fig. 4). NPO provides two versions of these names. *Local label* records molecular properties in the native form in which they were measured, e.g., genes, proteins, transgenes, while the *rdfs:label* contains a normalized view where molecules are assigned a common

[2] https://github.com/tgbugs/pyontutils/blob/cc538d9c790d607cbc8c2af8a3c25f1bfa3bfc0b/ttlser/docs/ttlser.md

[3] https://bioportal.bioontology.org/

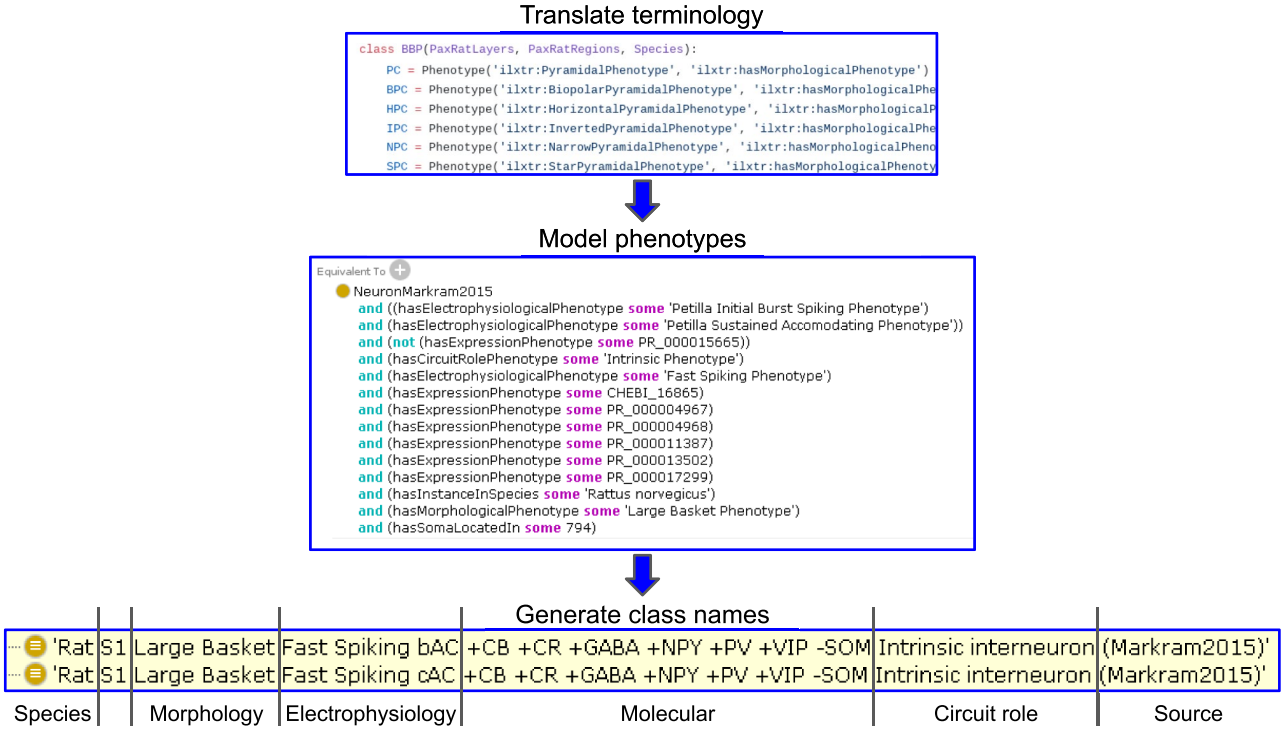

Fig. 4 Process used to translate local terminology into ontology-based representations and machine-generated names. Using neurondm, phenotypes are first mapped by a user into ontology identifiers (top panel). Neuron types are constructed and neurondm automatically translates these mappings into OWL equivalence statements (middle panel). From the same internal representation of these restrictions neurondm generates a set of human readable labels (bottom panel)

molecular abbreviation regardless of the form in which it was measured (see below). For ease of reference we also preserve the common name for the CUT and the original name assigned by the investigator for EBTs if it was provided. These can be found under *origLabel*, and they also appear as *skos:prefLabel* when they are present, otherwise *skos:prefLabel* is populated from *rdfs:label* so there are no neurons missing a preferred label.

For the NPOKB, we generally follow the ordering recommended by Hamilton et al. (2012) and Ecker et al. (2017). In both papers, the recommendation was to create an ordered taxonomy based on key phenotypic features, arranged roughly hierarchically, starting from the highest level, species, followed by anatomical regions, then a set of standardized names for morphological, physiological, molecular or connectional phenotypes (Fig. 4, lower panel). In this way, as proposed originally by Hamilton et al., (2012), it is easy to generate a human readable list of neurons from a given species or brain region and to compare across complex phenotypes. In addition, while we are still sorting out what constitutes a cell type, we define the local environment in which the neuron resides.

Molecular Indicators

For EBTs, NPO preserves the means by which molecular phenotypes are determined. If gene expression is measured, we use the identifier for the gene; if the expression of a transgene is measured, we include the transgene; if the protein is measured, we include the protein. For CUTs, we only use the protein, peptide or small molecules that are thought to define the class. In order to tie together these different measurements, we created a class called phenotype indicator (PhenotypeIndicator) that groups together the different forms of molecular entities, e.g., a somatostatin indicator is equivalent to Sst, SST, SOM, Sst-IRES-Cre, Sst-IRES-FLpO. A somatostatin neuron is then defined as equivalent to any neuron that has some somatostatin indicator as a molecular phenotype. In this way, we simplify the reasoning required to retrieve all somatostatin neurons, but we also clearly preserve the statements made by investigators in their instances or model assertions as preserved in the *localLabel*. In addition, to translate all of the different representations of a particular molecular entity into a consistent human readable label, we have assembled a set of short names that

represent each class based largely on common conventions or the names used in NCBI for mouse genes. These short names are used in the *skos:hiddenLabel* for each class and are suffixed with " (indicator)" to create the *rdfs:label*. For example, when generating a label, phenotype indicators for parvalbumin are shown as PV. These labels are available through the "hidden label" annotation property under the *ilxtr:PhenotypeIndicator* class.

Knowledge Base Construction

The basic process for constructing CUTs and EBTs from external sources has four steps where the first three can be done in any order. First we identify the names for the cell types in the source material. For example, in the Paul et al. (2017) paper the header of Table S7 contains the names of 6 neuron types PVBC, CHC, CCK, MNC, ISC, and LPC. Second we identify the phenotype values that are associated with those cell types. For example, in the Markram et al. (2015) paper these include values such as CB, PV, CR, NPY, VIP, SOM, bNAC, cAC, dNAC, bAC, cIR, fast spiking, non-accommodating, non-adapting, late spiking, etc. Third we identify the local names for the phenotypic dimensions that are being used and map those dimensions to and existing object property or create a new one if we determine that the dimensions is determined to be new and not captured by an existing object property, or is a specialization of some more general dimension. For example, for the Allen cell types these are sex_full_name, transgenic_line type,structure, hemisphere, cell_soma_location, and dendrite type. Finally we convert all phenotype values to ontology identifiers, match the values to the dimensions, construct the owl restrictions, and bag them into neuron types with a local name that matches the one provided in the original source.

For EBTs terms are selected for use as a phenotype value as follows. If a term exists and in one of the community ontologies listed in Table 1 we use it, but in some cases (e.g. during development) it is easier to create "new" terms (mint new identifiers) that match the local nomenclature used in the paper to simplify matching the EBTs to the original source. Those "new" terms are then replaced or mapped e.g. with NCBIGene identifiers. For anatomical phenotypes we use terms from species specific anatomical atlases (e.g. the Allen Mouse Reference Atlas Ontology (RRID:SCR_021000)) whenever the original source data mentions them as a reference. Sometimes papers use specific terminology that is not contained in a community ontology in which case we created a new term. There are three areas where new terms were created specifically for the NPO: phenotype indicators which are used to subsume multiple different types of evidence for a phenotype, neuron morphology (which has been upstreamed to PATO), and Petilla electrophysiological classification terms.

For CUTs phenotype indicators are used for molecular phenotypes and Uberon identifiers are used for anatomical phenotypes. Other phenotypic dimensions such as morphology do not have values with the same diversity of indicators and techniques, and thus phenotype classes are chosen in the same way as described for EBTs.

Each set of EBTs was constructed from the original source using a different approach. For the Allen cell types we retrieve the input data in a computationally accessible form from the Allen REST API. Since the original source is computationally accessible there is no issue validating reproducibility of an individual conversion. For the Markram and Huang models the original sources are opaque and computationally inaccessible. As a result we created computationally accessible representations of the original figures and tables. For Markram we converted a figure into a text representation that could be parsed and then manually checked that the accessible version was consistent with the opaque version. For Huang we did not attempt to convert the original source into a visually similar representation and instead encoded the information direction in the Python source file because the underlying sources were rasterized images and a table in a pdf. The effort needed to write custom code for parsing and converting directly from the underlying source could not be justified. See the supplemental methods for details.

When modeling CUTs curators try to the best of their ability to follow the consensus in the literature if there is one. Validation that a CUT is correct is derived from whether the neuron types that are inferred to be subclasses of the CUT include EBTs that should classify under the CUT, and similarly EBTs that should not classify as a CUT are excluded.

While developing the NPO we routinely checked for unexpected classifications, and the competency queries have been developed in part to detect such cases. In principle careful construction of disjointness axioms could be used to cause reasoning errors if an EBT does not classify under the expected CUT, however this has not been implemented.

It is not currently possible to run the code to regenerate the common usage types from the archive reference in this paper because neurondm is configured to pull data from the google sheets API v4. Even if we were to make a copy of the sheet available publicly users would need to configure API access to google sheets which is a significant stumbling block. The neurondm code could be updated to transparently switch between google sheets and an archival source, however this has not been done at this time.

Data and Code Availability

A docker image that captures the environment and code for this paper is available at https://hub.docker.com/layers/156844166/tgbugs/musl/npo-1.0-neurondm-build/images/sha256-c64fef99a0315184b604d20a571e6881de17c4da201edd74830b6169ee0d276a and an archive of that image has been archived on Zenodo at https://doi.org/10.5281/zenodo.5033493. The docker files specifying the image are part of https://github.com/tgbugs/dockerfiles/blob/d942371dc399510914d039022d2b4f92303bc120/source.org#npo-10-neurondm-build and the archive is on Zenodo at http://doi.org/10.5281/zenodo.5068491. See supplemental methods for how to use the image.

The NPO can be viewed by loading the.ttl file available at https://raw.githubusercontent.com/SciCrunch/NIF-Ontology/npo-1.0/ttl/npo.ttl into the Protégé Ontology Tool (RRID:SCR_003299) v5.5.0 or higher. Note that WebProtégé is not capable of running the reasoners required by the NPO. As described in the supplemental methods, npo.ttl is the "light" version of the full ontology that makes it less reliant on the full import chain. Additional information about the structure of the NPO and working with the NPO can be found in the supplemental methods. The NPO is distributed under a CC-BY 4.0 Attribution license, but it imports community ontologies that may be covered under different licenses.

The work here describes v1.0 of the NPO which can be accessed at https://raw.githubusercontent.com/SciCrunch/NIF-Ontology/npo-1.0/ttl/npo.ttl. In the import closure of npo.ttl there are no external imports except for http://purl.obolibrary.org/obo/bfo.owl which had versionIri http://purl.obolibrary.org/obo/bfo/2019-08-26/bfo.owl at the time npo 1.0 was released. All other ontology iris resolve to the neurons branch of the NIF-Ontology except for http://ontology.neuinfo.org/NIF/ttl/generated/parcellation-artifacts.ttl. As a result, importing npo.ttl directly in Protégé will result in the newest version of the imports on the neurons branch being used, which may lead to some small differences in the results compared to what are presented here. However, it is possible to use the NIF-Ontology catalog file to load an exact view of version 1.0 of npo.ttl by cloning the git repository and checking out the npo-1.0 tag. In the event that the npo-1.0 tag is somehow lost at some point in the future, it names the sha1 commit hash 7bb15aa5fda9391809032a6765419dfb2486b2fa which is a merge commit with parents d6615f8 and cdffa6e. The NIF-Ontology repository can identified by root commit hash sha1 ba8482cfccd934b45591e6bbfd6378ef165d0e31 and/or 4f3e0493d926a2c42459b8622dda4de148cf2c5d.

The NPOKB is available on BioPortal at https://bioportal.bioontology.org/ontologies/NPOKB. A loaded graph that can be used with SciGraph, a neo4J-based database for serving ontologies, is available at https://github.com/SciCrunch/NIF-Ontology/releases/tag/npo-1.0.

The content of the NPO is also accessible via the UCSD SciCrunch SciGraph API at https://scicrunch.org/api/1/sparc-scigraph/. Documentation for access can be found at http://ontology.neuinfo.org/docs/NIF-Ontology/README.html#using-nifstd.

The neurondm git repo is https://github.com/tgbugs/pyontutils/tree/master/neurondm. The pyontutils repository can be identified by the root commit hash sha1 6d96945e85d4e949215910f13f3e620495b5e165.

All python code bears an MIT license and is available on the Python Package Index (PyPI) https://pypi.org/project/neurondm/. It can be installed via `pip install neurondm`. Additional instructions are available in the README.[4]

An archive of the code corresponding to this publication is also available on Zenodo at https://doi.org/10.5281/zenodo.4005727. Additional release artifacts are also available on the GitHub release page https://github.com/tgbugs/pyontutils/releases/tag/neurondm-0.1.3.

The full list of CUTs is available at: https://github.com/tgbugs/pyontutils/releases/download/neurondm-0.1.3/data-bundle-2020-08-28.zip.

The full datasets produced for the competency queries (see Results) are available at: Gillespie et al. (2020) https://zenodo.org/record/4007065#.X03TD2dKiAZ.

Results

Common Usage Types

Common usage types represent neuron types that have been reliably identified over many years by multiple groups using multiple techniques. The criteria we used to identify CUTs is provided in Supplementary Table S1. Any type that meets these criteria can and (given sufficient resources) will ultimately be included as a CUT. A master spreadsheet was created in Google Spreadsheets and populated with a list of neuron "stubs" that were created automatically by taking the list of major brain regions in the UBERON ontology and creating two classes per region: Region X projection neuron and Region X intrinsic neuron. These anatomical regions were at a fairly coarse level and comprised the major brain and spinal cord regions, but generally not subregions, for example, cerebral cortex and not motor

[4] https://github.com/tgbugs/pyontutils/blob/master/neurondm/README.md

cortex. Individual brain regions were then augmented with the list of neuron types extracted from online knowledge bases. We started with the list of approximately 300 mammalian neurons from Neurolex Wiki (RRID:SCR_005402) (Larson & Martone, 2013) that had been compiled through expert input via the Neuron Registry Task Force of the INCF (Hamilton et al., 2012), as well as by community contributions. This list was then cross referenced to NeuroElectro (RRID:SCR_006274), BAMS Cells (RRID:SCR_003531), Hippocampome.org (RRID:SCR_009023), NeuroMorpho.org (RRID:SCR_002145) and Blue Brain Project (RRID:SCR_002994). All of these sources were accessed via the Neuroscience Information Framework (RRID:SCR_002894) project to find a set of cells that were referenced in multiple databases. As NeuroElectro maps their nomenclature to the Neurolex names, we used this database to examine representation of these cell types in the neurophysiology literature. We selected all neurons that were referenced in more than one paper.

This procedure resulted in a working list of ~ 350 neurons (for full list see Data Availability Statement). From this list, we then selected ~ 100 neurons for which we had basic morphological and molecular properties available. We also included the neurotransmitter for the majority. We elected to focus in v1.0 primarily on molecular and morphological phenotypes, rather than the full complexity available in the NPO (Fig. 2), as these are the most well known for CUTs and are the most frequent types encountered in the EBTs (Zeng & Sanes, 2017). We also elected in the modeling to take a minimalist approach, that is, our representation is meant not to represent an exhaustive list of every molecule that has been identified within a neuron, but the minimum set of molecules and morphological features that are characteristic for that type. This decision allowed us to construct OWL equivalence statements for each CUT that defined the necessary and sufficient conditions that would allow EBTs to classify under these CUTs. Additional phenotypes were still recorded but added through the Subclassof axiom. Subclassof represents a weaker form of restriction, representing a necessary but not sufficient condition for membership in a class. In order to avoid logical inconsistencies that would interfere with classification, we only included positive phenotypes in necessary and sufficient conditions for CUTs. If distinguishing negative phenotypes were present, they were modeled as entailments rather than OWL restrictions.

Following (Larson et al., 2007), the primary anatomical location of a neuron is assigned based on the brain region in which the soma is located, e.g., cerebellar neuron is equivalent to a neuron with a cell soma in any part of the cerebellum.

Evidence-based Types

EBTs represent cell types and taxonomies proposed by a single group based on an analysis of experimental evidence. In an ideal world the experimental types for every paper ever published and every database involving neurons would be part of the NPO. For this version of the NPO and for the purposes of evaluating our phenotype model, we focused on 3 projects that have generated cortical classifications based on large amounts of experimental data:

A. Cortical cell types proposed by the Blue Brain Project (Markram et al., 2015), as elaborated in the text and Table 1. In this study, 56 total types across 9 morphological types are identified and physiologically characterized from cells in cortical area S1 of rats ranging from P11-P15 from which they recorded physiological properties. Cell-specific molecular markers were confirmed by immunohistochemistry and RT-PCR. (Markram et al., 2015) utilize a nomenclature aligned to the Petilla conventions (Petilla Interneuron Nomenclature Group et al., 2008) to annotate their physiological properties. For NPO V1.0, we included the molecular, morphological and electrophysiological phenotypic dimensions.
B. The classification of proposed cortical GABAergic cell types from Josh Huang and colleagues as summarized in Table S7 of Paul et al. (2017) supplemented with additional information from Fig. 1. The latter was used primarily to create disjointness axioms (see Fig. 1b). For NPO v1.0, we concentrated primarily on the gene expression phenotypes presented in this table, supplemented with information from the rest of the paper, e.g., disjointness axioms based on Fig. 1b. Synaptic and physiological phenotypes will be included in a later version.
C. The ~ 800 cell classes contained in the Allen Cell Types database (RRID:SCR_014806), a database of experimental electrophysiological, morphological and transcriptomic data derived from single cell data. In the Cell Types database, no classification scheme was proposed; rather the records represent statistical summaries of properties measured from these classes of cells identified in transgenic lines. We therefore include this as an EBT. For this version, we focused on molecular measurements from mouse cortex.

Competency Queries

The NPO was designed to classify neurons according to phenotype dimensions, regardless of whether they represent EBTs or CUTs. To test the integrity of the knowledge base and the structure of the ontology, we developed a set of competency queries (CQ):

Table 2 Examples of EBT and CUT neurons returned from Competency query CQ1: Find all examples of parvalbumin containing neurons. The form of the parvalbumin indicator is highlighted in red. Only one example is provided from the Allen EBT (total 59). Full results are available in Gillespie et al. (2020). The compact identifier for each class is prefixed (in bold) to the localLabel for ease of reference. The local label preserves the form in which the molecule was measured. The Common/original name represents the common name from the superclass for all of the physiological subtypes for the Markram cells. However, for the local label we provide a subtype as the superclass does not include the full molecular profile in the name

Type	#	Common/original name	NPO localLabel
CUT	6	**nifext:56**: Neocortex basket cell	**nifext:56**: Mammalia neocortex L2/3 Basket + **PV** + GABA intrinsic neuron
EBT Markram	16	**npokb:112**: Nest basket cell	**npokb:112**: Rattus norvegicus S1 Nest basket (intersectionOf AC b) Fast spiking + GABA + calbindin + CR + NPY + **PV** + VIP -SST intrinsic neuron (Markram2015)
EBT Huang	2	**npokb:43**: PVBC cortical neuron	**npokb:43**: Mus musculus neocortex Basket + GABA + **PV-cre** intrinsic neuron (Huang2017)
EBT Allen	59	none	**npokb:434**: Mus musculus female left cerebral hemisphere VISrl2_3 -Apical Dendrite -Spiny + **Pvalb-T2A-FlpO** + Vipr2-IRES2-Cre + Ai65(RCFL-tdT) neuron (AllenCT)

1. Find all parvalbumin + neurons
   Description Logic (DL) Query: hasPhenotype some 'parvalbumin (indicator)'.
2. Find all cortical neurons that contain somatostatin
   DL Query: hasPhenotype some 'somatostatin (indicator)' and hasSomaLocatedIn some (neocortex or 'part of' some neocortex).
3. How do basket cells described in Paul et al. (2017) and Markram et al. (2015) compare on key dimensions?
   DL Query: (NeuronHuang2017 or NeuronMarkram2015) and hasPhenotype some 'Basket phenotype'.
4. What EBTs are related to the Martinotti cell?
   Determine which neurons classify under the CUT Neocortex Martinotti cell
   DL Query: NeuronEBM and hasPhenotype some 'Martinotti phenotype'

All of the results presented below were produced by issuing OWL DL queries as specified above in Protégé v5.5.0 on a MacBook Pro using the ELK 0.4.3 reasoner unless otherwise noted. More information on loading the ontology into Protégé can be found in the Supplemental Methods.

CQ1: Find All Examples of Parvalbumin Neurons

This query should return all neurons that have a phenotype associated with parvalbumin, regardless of exactly what molecule was measured (DNA, RNA, protein) or how it was measured. In this version of the NPO, we achieve this by creating phenotype indicators without specifying the relationships between these measures through the npokb:parvalbumin (indicator) class. The results of this query are summarized in Table 2. A total of 86 neurons are returned, including EBTs (Huang, N = 2, Markram, N = 16 and Allen; N = 59) and CUTs (N = 9). To aid in comparison across these classes, we illustrate with one example each from the Markram EBTs and Allen data. The complete list of neurons is provided in Gillespie et al., (2020). The original label is provided for each EBT and the common name for the CUT. These are followed by the *localLabel* names that preserve the form of molecule upon which the classifications were based to illustrate how the NPO can be used to compare across different assertions about molecular identity (Markram2015, Huang2017, AllenCT). Related phenotypic values are color coded to aid in comparison. In this case, we use the *localLabel* that preserves the original type of molecule upon which the classifications were based. For a complete list of abbreviations, see Table S2.

Three of the neuron classes indicate that the parvalbumin cells are basket cells, while the Allen data does not specify morphology beyond noting that these cells lack an apical dendrite and dendritic spines.

CQ2: Find All Cortical Neurons That Contain Somatostatin

This query should return all cortical neurons that contain somatostatin regardless of cortical subregion or atlas brain region. Details about how atlas brain regions are handled are provided in the supplemental methods. This query returns a total of 100 neurons, including the neocortex Martinotti cell from the CUT and EBTs from the three classification schemes (Table 3). For Markram, we show only one subtype from each of the 3 main types. For Allen, we selected a few representative examples. Note that Allen neurons are returned from retrosplenial cortex (RSPd2/3) and two areas of primary visual cortex (VISal6a, VISl5) while Markram is returned for primary somatosensory cortex (S1). Both Huang and Allen cells use the same transgenic line for Sst expression, however it is extremely difficult to tell by looking at the laboratory nomenclatures (as demonstrated by Table 3) because they are called SST by Huang and Sst-IRES-FlpO by Allen. Thus, while the local labels preserve the nomenclature

Table 3 Results for CQ2: Find all cortical neurons containing somatostatin. Full results are available in Gillespie et al. (2020). The compact identifier for each class is prefixed (in bold) to the local label for ease of reference. The local label preserves the form in which the molecule was measured. The Common/original name represents the common name from the superclass for all of the physiological subtypes for the Markram cells. However, for the local label we provide a subtype as the superclass does not include the full molecular profile in the name. Similar entities across cell types are color coded. Brain region = blue; somatostatin indicator = red

Type	#	Common/original name	NPO localLabel
CUT	1	**nifext:55**: Neocortex Martinotti cell	**nifext:55**: Mammalia **neocortex** (unionOf EGL L3 L5) (with-axon-in cortical layer I) Martinotti + **Sst** + GABAR + GluR + GABA intrinsic neuron'
EBT Markram	31	• **npokb:114:** Small basket neuron • **npokb:111**: Martinotti neuron • **npokb:109**: Double bouquet neuron	• **npokb:75**: Rattus norvegicus **S1** Small basket (intersectionOf NAC d) Fast spiking + GABA + calbindin + NPY + **SST** + VIP -CR -PV intrinsic neuron (Markram2015) • **npokb:89**: Rattus norvegicus **S1** Martinotti (intersectionOf AC b) Regular spiking non pyramidal + GABA + calbindin + NPY + **SST** -CR -PV -VIP intrinsic neuron (Markram2015) • **npokb:87**: Rattus norvegicus **S1** Double bouquet (intersectionOf IR c) Regular spiking non pyramidal + GABA + calbindin + CR + **SST** + VIP -NPY -PV intrinsic neuron (Markram2015)
EBT Huang	4	• **npokb:42**: MNC neuron • **npokb:45**: LPC neuron	• **npokb:42**: Mouse **Neocortex** Martinotti + GABA (intersectionOf + Adcy2 + Calb2 + Grin3a + Inhbb + Nppc + Pde2a + Rgs6 + Rgs7 + Sst + Zip1 + Znt3) + CR + **SST** interneuron (Huang2017) • **npokb:45**: Mouse **Neocortex** + GABA (intersectionOf + Calca + Chrm2 + Cort + Gpr88 + Gucy1a3 + Gucy1b3 + Hcrtr1 + Kcnmb4 + Nos1 + Opn3 + Oxtr + Pde1a + Penk + Prkg2 + Ptn + Rln1 + Slc7a3 + **Sst** + Syt4 + Syt5 + Syt6 + Tacr1 + Trpc6 + Unc5d + Wnt2) + **SST** + NOS1 projection (Huang, 2017)
EBT Allen	64	none	• **npokb:296**: Mus musculus female right cerebral hemisphere **RSPd2_3** -Apical Dendrite (intersectionOf Spiny sparse) + **Sst-IRES-FlpO** + Nos1-CreERT2 + Ai65(RCFL-tdT) neuron (AllenCT) • **npokb:415**: Mus musculus female left cerebral hemisphere **VISl5** -Apical Dendrite -Spiny + **Sst-IRES-Cre** + Ai14(RCL-tdT) neuron (AllenCT) • **npokb:412**: Mus musculus female right cerebral hemisphere **VISp6a** -Apical Dendrite (intersectionOf Spiny sparse) + **Sst-IRES-Cre** + Ai14(RCL-tdT) neuron (AllenCT)

used in the source (Paul et al., 2017 and Allen Cell Types Database respectively), they are difficult or impossible to use for alignment. The NPO resolves this issue by mapping to identifier systems wherever possible by reviewing the source to see what the local nomenclature actually means. The default labels for neurons (not shown in Table 3) are generated from the underlying identifier which makes it possible to see that Huang and Allen use the same transgenic line (JAX:028579) developed by the Huang lab, regardless of the different local nomenclature. In the NPO, if a transgene is involved, and it was derived from a transgenic mouse line, we use the Jackson lab stock number to represent transgenic phenotype when it is available.

CQ3:How do Basket Cells Described in Paul et al. (2017) and Markram et al. (2015) Compare on Key Dimensions?

This query returned EBT cells from the two groups that were assigned the morphological phenotype "basket". A total of 22 neurons were returned, 20 from Markram and two from Huang. A subset are illustrated in Table 4 and related phenotypes are color coded across the different types for ease of comparison. For the Markram cells, we only show one subtype for each main class.

Two classes of basket neurons are returned for Huang, while three are returned for Markram. Each of the three Markram classes are distinguished by distinct basket morphologies: small basket phenotype, large basket phenotype,

Table 4 Neurons that have a basket phenotype. Similar entities across the cell are color coded to aid in comparison. The full results list is available in Gillespie et al (2020). Similar entities are color coded across cell types: blue = brain region; green = morphology; purple = neurotransmitter; dark red = parvalbumin indicator; red = somatostatin indicator

Original name	NPO ID	NPO Label
PVBC Neuron (Huang2017)	npokb:43	Mus musculus **neocortex Basket + GABA** (intersectionOf + Adm + Cckbr + PV + ilxtr:Kv3 + Rspo2 + Adcy8 + Cox6c + Gabra1 + Gabra4 + Gabrd + Gria1 + Gria4 + Mef2c + Pparg + Ppargc1a + Rgs4 + Slit2 + Slit3 + Tac1 + Arhgef10 + Esrrg + Nefh + Adcy1 + Rasl11b) + **PV** intrinsic neuron (Huang2017)
CCKC Neuron (Huang2017)	npokb:40	Mus musculus **neocortex Basket + GABA** (intersectionOf + Crh + Cck + Cck + Cnr1 + Edn3 + Htr3a + Igf1 + VIP + VIP + Vipr1 + Adcy9 + Chrm3 + Cplx2 + Htr2c + Pnoc + Npy1r + Tac2 + Cplx3 + Pde7b + Prok2 + Hs6st3 + Syt10 + Rgs12) + Cck + **VIP** intrinsic neuron (Huang2017)
Large basket cell (Markram2015): subtype	npokb:59	'Rattus norvegicus **S1 Large Basket** (intersectionOf AC b) Fast Spiking + **GABA** + Calb + Calb2 + Npy + **PV + VIP** -Sst interneuron (Markram2015)'
Nest basket cell (Markram2015): subtype	npokb:65	'Rattus norvegicus **S1 Nest Basket** (intersectionOf AC b) Fast Spiking + **GABA** + Calb + Calb2 + Npy + **PV + VIP** -Sst interneuron (Markram2015)'
Small basket cell (Markram2015): subtype	npokb:73	'Rattus norvegicus **S1 Small Basket** (intersectionOf AC c) Fast Spiking + **GABA** + Calb + Npy + Sst + **VIP** -Calb2 **-PV** interneuron (Markram2015)'

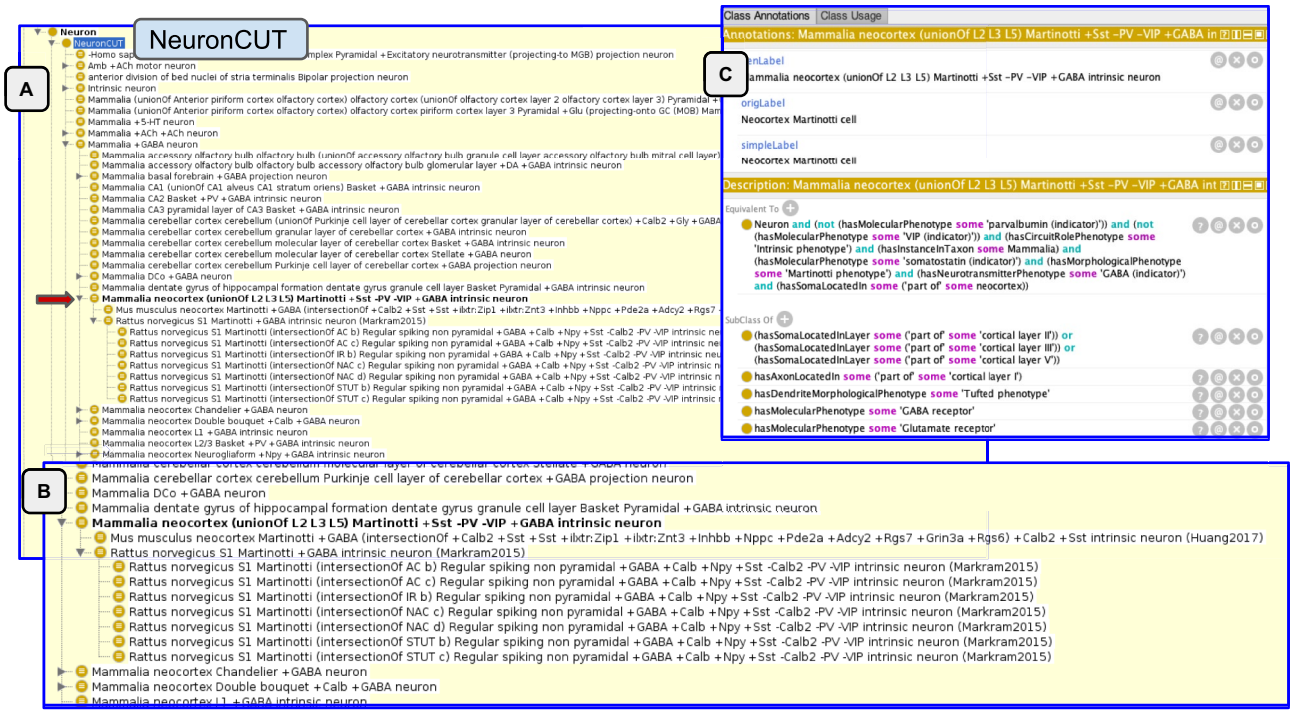

Fig. 5 Inferred hierarchy after reasoning over the ontology for the Martinotti cell. Panel **A** shows the hierarchy generated under the NeuronCUT class. The position of the Marinotti CUT is indicated by the lower red arrow. An enlargement of the Martinotti classification is shown in panel **B**. Panel **C** shows the OWL representation of the Martinotti CUT

and nest basket phenotype. These morphologies are modeled as subtypes of BasketPhenotype.

For these types of comparisons, the NPO facilitates comparison across diverse experimental techniques and anatomical nomenclatures and can help to generate testable hypotheses regarding phenotypes. In this example, it is difficult to tell from the information provided whether there is a 1:1 correspondence between any of the Huang and Markram cells. The only molecules mentioned by all 5 cells are GABA, PV and VIP. The Huang PVBC neuron is PV + while the CCKC neuron is VIP + . Two Markram neurons are positive for both PV and VIP, while the small basket cell is asserted to be PV + and VIP-. No negative phenotypes were recorded for the Huang neurons, as we based the equivalence classes on the information available in Table S7 which only included positive phenotypes. In the NPO, we operate under an open world assumption, that is, unless there is an explicit statement that a molecule is lacking, we do not assume that it is absent. We do provide additional information in the form of disjointness axioms based on Fig. 1b of Paul et al. (2017) that the PV-containing and the VIP-containing cells are non-overlapping. This approach dovetails with EBTs making assertions about disjointness of cell types within a species which can be true even if there is not a universal axiom about molecular constituents. Disjointness therefore doesn't mean that there is no expression, but an inspection of the data provided in Fig. 1e indicates that expression of PV in the CCKC neuron is very low. Inspecting the data therefore suggests that the CCKC neuron is VIP + and PV-, consistent with the small basket cell of Markram.

This example illustrates some of the difficulties involved in comparing across phenotypes, particularly when the different phenotypes are measured across experiments. It also illustrates the importance of tying EBTs to experimental data, so that predictions generated from these comparisons can be explored. In this case, Paul et al. (2017) provided expression data for several key molecules in Fig. 1e. This figure shows that while the CCKC neuron expresses little to no PV, consistent with the small basket cell, it also expresses little to no Sst and detectable Calb2, in contrast to the small basket cell. However, as is easily seen in the labels, the Huang and Markram cells come from mouse and rat respectively and how complex molecular phenotypes compare across species is unknown (Yuste et al., 2020).

CQ4: What EBTs are Related to the Martinotti Cell?

To address this competency query, we reasoned over the ontology to determine which neurons would classify under the Neocortex Martinotti neuron CUT. For a neuron to be classified as a type of Martinotti cell, it has to share necessary and sufficient conditions of that class as coded in

Table 5 This rubric (Hodson et al., 2018) organizes the 15 FAIR principles (Applicable principles) into a hierarchical table according to how easy they are to achieve, starting from a basic core (Summary) and rates data according to level of compliance, from 1 to 4 * (Rating). We provide an evaluation of the NPO/NPOKB against these principles in column 4

Rating	Summary	Applicable principles	NPO/NPOKB
*	The basic core: metadata, PID & access	F2. data are described with rich metadata F1. (meta)data are assigned a globally unique and persistent identifier A1. (meta)data are retrievable by their identifier using a standardized communications protocol	• F2: Full descriptive metadata for the ontology are included in the.ttl file. Metadata for the datasets and code are included in Pypi from setup.py, Zenodo, MIRO; The NPOKB includes complete authoring metadata • F1: All datasets referenced in this paper have been assigned DOIs • F1. The NPOKB is assigned a unique identifier (RRID) RRID:SCR_017403 • A1. RRIDs are resolvable through identifiers.org: https://identifiers.org/RRID:SCR_017403 and through the SciCrunch Registry resolver service: https://scicrunch.org/resolver/RRID:SCR_017403 by the Neuroscience Information Framework and dkNET
**	Enhanced access: catalogues for discovery, standard (controlled) access & licences	F4:. (meta)data are registered or indexed in a searchable resource A1.1. the protocol is free, open and universally implementable A1.2. the protocol allows for an authentication and authorization procedure, where necessary R1.1. (meta)data are released with a clear and accessible data usage license	• F4: All python code is available via pypi. ebuilds for Gentoo are available from tgbugs-overlay • F4. The NPO is registered in BioPortal and in the SciCrunch Registry (RRID:SCR_017403) • A1.2 API access is provided via Bioportal and also via SciGraph maintained by the Neuroscience Information Framework and dkNET • R1.1 The NPO is covered under a CC-BY 4.0 license
***	Use of standards: for metadata and data	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation R1.3. (meta)data meet domain relevant community standards F3: metadata clearly and explicitly include the identifier of the data it describes	• I1: The ontology is built in OWL2, a recognized standard for ontologies • R1.3: The phenotype bags are built out of terms from community standard ontologies • F3: All terms are defined by a URI as well as a compact identifier
****	Rich, FAIR metadata	R1. (meta)data are richly described with a plurality of accurate and relevant attributes I2. (meta)data uses vocabularies that follow FAIR principles	• R1: The ontology has complete metadata associated with it • I2: The NPO has been designed in accordance with the FAIR principles. Documentation • I2: The NPO/NPOKB imports relevant community vocabularies (see Table 1) that adhere to the FAIR principles
*****	Provenance and additional context	R1.2 (meta)data are associated with data provenance I3. (meta)data include qualified references to other (meta)data A2. metadata are accessible, even when the data are no longer available	• R1.2: References that support assertions are included in the annotations although unfortunately OWL does not provide an easy way to annotate specific triples •!3: I2: The NPO/NPO-KB imports relevant community vocabularies (see Table 1) that adhere to the FAIR principles • A2: The NPO and associated tools have been registered with the SciCrunch Registry, which maintains metadata pages for similar resources. They ensure that their metadata is accessible even if the resource is no longer available

the equivalence statements. As discussed in the methods, we deliberately chose to model a minimum of properties as necessary and sufficient due to the large variability in the number of phenotypes recorded for the EBTs. Additional properties are included (Fig. 5C) but not in the form of OWL restrictions, so they do not factor into the reasoning. We also only represent the major classes of CUTs and do not include subtypes, as these are less well agreed upon. In OWL, if we were to require that a Martinotti neuron must have calretinin,then if a given EBT did not state that calretinin was a defining characteristic, the neurons would not classify. In fact, according to Rudy et al. (2011), Martinotti cells contain two subclasses, one that contains calretinin and one that does not. In the NPO, the NeuronHuang2017 EBT notes the presence of calretinin (+Calb2), while the NeuronMarkram2015 EBT says it is absent (-Calb2), perhaps representing these two subclasses.

As Fig. 5 shows, the Allen EBTs do not classify under the Martinotti CUT. In v1.0 of the NPO, we only model morphological phenotypes at a coarse level, e.g., Martinotti phenotype, which is assigned to the level of the entire cell. In contrast, NeuronACT provided morphological information only for the dendrites of each cell. For the cortical somatostatin containing cells, it was noted that they lack an apical dendrite and dendritic spines, but no assertion was made about a Martinotti phenotype, unlike in the other two classifications. In the future, the NPO will include additional defining features of a Martinotti phenotype.

FAIR Properties of the NPO

The NPO was designed to be consistent with the FAIR principles. In Table 5, we show how the NPO achieves FAIR using the rubric in Hodson et al. (2018). The key features are machine readability, the use of identifiers (FAIR vocabularies), common knowledge representation languages and community standards. We provide a comparison with other cellular ontologies in Table S1.

Discussion and Conclusion

The NPO provides a semantically-enriched, FAIR data model for representing the complex cellular phenotypes being generated by neuroscientists involved in individual and large scale brain initiatives. It allows the creation of machine generated taxonomies, and provides a consistent naming convention that is machine configurable. Using the NPO, we showed that we could take cellular data arising from high throughput activities, e.g., the Allen Cell Atlas, large projects like the Blue Brain Project, and from individual investigators to cross between different techniques to show areas of agreement and non-alignment. This exercise is not trivial, as the multiplicity of techniques, the incomplete sampling, and the complex nomenclature present challenges. However, the NPO helps to mitigate these by allowing translation of custom lab nomenclature and experimental results into a common, semantic, and computable representation using community ontologies. The names themselves can be customized to conform to any nomenclature standard that might emerge for human consumption (e.g., Shepherd et al., 2019), but this process is managed as a formal specification rather than through agreed upon naming conventions.

We have focused our efforts on addressing the problem of cell classification vs the issue of determining neuronal types by providing a means to compare our current knowledge about cell types (our common usage types) with the many different classifications being generated by data driven methods and other experimental techniques. The distinction between a neuron type vs a neuron class is not entirely clear, and the terms are often used interchangeably. We use class here to refer to a set of neurons that satisfy a set of criteria, e.g., GABAergic neurons = all neurons that use GABA as a neurotransmitter. The number of potential classes given the number of phenotypic dimensions measured is therefore very large. Types, however, refer to neurons that are sufficiently distinct that the presence of a given set of features will reliably predict the presence of additional features that have not been measured. For example, when a cerebellar Purkinje cell is identified by a Nissl stain based on its size, shape, and location, we can reliably infer that it contains parvalbumin and calbindin, has dendrites densely covered in dendritic spines, and uses GABA as a neurotransmitter whether or not we explicitly measure them. This definition is similar to that proposed by Zeng and Sanes (2017) who propose that types represent discrete groups which notionally serve a specific function while classes represent aggregates of types that share common features. Types are also the categories of cells that must be accounted for when building circuit diagrams of the nervous system (Luo et al., 2008).

The NPO allows us to communicate about and compare measured neuronal phenotypes in a way that reflects human understanding but that can also be fully managed using modern computational methods. Genomics benefitted enormously from a community ontology for annotation of experimental results that allowed them to be communicated in a consistent and machine-processable manner. The issue of neuron typology will also benefit from a consistent annotation framework. Although there are challenges, phenotypes lend themselves to a consistent annotation framework, e.g. genes and morphological features. However, the issue of cell type itself is more fluid. Thus the NPO implements a model that distinguishes between observations in single cells (instances), proposals about cell types derived from computational analyses (EBTs), and cell types that have been recognized by one or more criteria across multiple

labs and techniques (CUTs). None of these categorizations represent ground truth. Nevertheless, transcriptomics combined with data driven approaches have shown promise as a unifying technique that may allow stable cell populations to be described within a probabilistic framework (Yuste et al., 2020). Such abstractions will still likely reference entities such as brain regions, marker genes, morphology, and connections. Likewise many of these abstractions will map onto well-known cell types (Yuste et al., 2020). Disagreements are still likely to arise about the nature of these populations, particularly at finer levels of granularity. The NPO and the associated knowledge environment provide a bridge between classifications generated using high throughput and integrative techniques and our accumulated knowledge over the past 100 years on cell types in the nervous system.

Looking to the future, extension of the NPO beyond the contents described in this paper is already underway. We have started to create new evidence based types for the peripheral nervous system as part of the NIH SPARC consortium (Osanlouy et al., 2021). Application to the peripheral nervous system is an extension along the location dimension. Extensions along other dimensions are also possible. The taxonomic dimension is an obvious candidate. The inclusion of invertebrate and avian neuron types would significantly broaden the generality of the content of the NPO and further test the flexibility of the approach. To truly understand the nervous system we will likely need to study it in all its variation across a menagerie of clades and dimensions. We designed the NPO to have a flexible data model so that it could not only accommodate such diversity, but also be enhanced by it. The ongoing initiatives to exhaustively catalog neuron types for Drosophila melanogaster seem like they could provide a tractable testing ground for applying the NPO at scale and for the infrastructure that will be needed to manage the flood of vertebrate data that will be collected over the coming years.

The work reported here should be considered a proof-of-concept; in order for the NPO to be used at the scale we envision significant additional tooling would be required. Currently, the Python code can be used by a researcher to translate their phenotypes into NPO and they can compare their neurons locally to the NPOKB using Protégé. To gain traction, increase ease of use, and populate the knowledge base, we envision a set of on-line tools that would assist researchers in translating their phenotypes into the NPO, along with a web-accessible growing knowledge base with visualization and analysis tools for researchers to compare their neurons to what is known. Yuste and colleagues (2020) also envision an online community knowledge base where information on cell types is accumulated and linked. In addition, the NPO currently only provides the skeleton of discrete types on top of which the continuous nature of measurements needs to be integrated. Nonetheless, the goals of the BRAIN initiative and other large scale data projects are to transform our understanding of the brain using new technologies and data science and understanding the "parts list" of the nervous system is a key objective (Zeng & Sanes, 2017). If we accept the premise that no single project or group can do it alone, then neuroscientists must produce data and knowledge artifacts like atlases and taxonomies in a way that is amenable to computation. The FAIR data principles outline some of the basic ways to do that (Table 5). Integral to FAIR is the use of community standards that make the process of searching, aggregating, and reusing data more tractable. The proposed methods do not require that we all think alike, rather, they ensure that we can employ computational methods to compare and contrast across different classification schemes. Although the proposed approaches would require a significant investment by funders and researchers alike to develop and adopt these methods, we have to measure this against the time we currently spend trying to reconcile computationally opaque and un-FAIR neuroscience data. In an ideal world, we would focus our resources on grappling with the innate complexity of the issue of cell types in the brain, rather than having to focus on reconciling the myriad number of ways we can refer to common entities in neuroscience.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12021-022-09566-7.

Acknowledgements This work was supported by NIH Brain Initiative award 5U24MH114827-04, a Canadian Institute for Health Research post-doctoral fellowship and National Institutes of Health grant MH111099, the Krembil Foundation, and by funding to the Blue Brain Project, a research center of the École polytechnique fédérale de Lausanne (EPFL), from the Swiss government's ETH Board of the Swiss Federal Institutes of Technology. Meetings supporting this work were facilitated by the International Neuroinformatics Coordinating Facility (INCF) through the Neuroinformatics for Cell Types Special Interest Group. The authors thank Felix Schürmann and Karin Holm for helpful comments.

Funding Open access funding provided by EPFL Lausanne.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Bug, W. J., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A. R., Larson, S. D., et al. (2008). The NIFSTD and BIRNLex Vocabularies: Building Comprehensive Ontologies for Neuroscience. *Neuroinformatics, 6*(3), 175–194.

Cauli, B., Audinat, E., Lambolez, B., Angulo, M. C., Ropert, N., Tsuzuki, K., Hestrin, S., & Rossier, J. (1997). Molecular and Physiological Diversity of Cortical Nonpyramidal Cells. *The Journal of Neuroscience: THe Official Journal of the Society for Neuroscience, 17*(10), 3894–3906.

DeFelipe, J., López-Cruz, P. L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., Burkhalter, A., et al. (2013). New Insights into the Classification and Nomenclature of Cortical GABAergic Interneurons. *Nature Reviews Neuroscience, 14*(3), 202–216.

Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., He, Y., et al. (2016). The Cell Ontology 2016: Enhanced Content Modularization and Ontology Interoperability. *Journal of Biomedical Semantics, 7*(1), 44.

Ecker, J. R., Geschwind, D. H., Kriegstein, A. R., Ngai, J., Osten, P., Polioudakis, D., Regev, A., Sestan, N., Wickersham, I. R., & Zeng, H. (2017). The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron, 96*(3), 542–557.

Franklin, K. B. J., & Paxinos, G. (2008). *The Mouse Brain in Stereotaxic Coordinates* (3rd ed.). Academic Press.

Gillespie, T. H., Martone, M. E., & Hill, S. L. (2020). Results of Neuron Phenotype Ontology Competency Queries. https://doi.org/10.5281/zenodo.4007065

Hamilton, D. J., Shepherd, G. M., Martone, M. E., & Ascoli, G. A. (2012). An Ontological Approach to Describing Neurons and Their Relationships. *Frontiers in Neuroinformatics, 6*, 15.

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., et al. (2019). Conserved Cell Types with Divergent Features in Human versus Mouse Cortex. *Nature, 573*(7772), 61–68.

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., & Wittenburg, P. (2018). Turning FAIR Data into Reality: Interim Report from the European Commission Expert Group on FAIR Data. https://doi.org/10.5281/zenodo.1285272

Larson, S. D., Fong, L. L., Gupta, A., Condit, C., Bug, W. J., & Martone, M. E. (2007). A Formal Ontology of Subcellular Neuroanatomy. *Frontiers in Neuroinformatics, 1*, 3.

Larson, S. D., & Martone, M. E. (2013). NeuroLex.org: An Online Framework for Neuroscience Knowledge. *Frontiers in Neuroinformatics*, *7*, 18

Luo, L., Callaway, E. M., & Svoboda, K. (2008). Genetic Dissection of Neural Circuits. *Neuron, 57*(5), 634–660.

Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience, 7*(2), 153–160.

Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., Ailamaki, A., et al. (2015). Reconstruction and Simulation of Neocortical Microcircuitry. *Cell, 163*(2), 456–492.

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., & Haendel, M. A. (2012). Uberon an Integrative Multi-Species Anatomy Ontology. *Genome Biology, 13*(1), R5.

Osanlouy, M., Bandrowski, A., de Bono, B., Brooks, D., Cassarà, A. M., Christie, R., Ebrahimi, N., Gillespie, T., Grethe, J. S., Guercio, L. A., Heal, M., Lin, M., Kuster, N., Martone, M. E., Neufeld, E., Nickerson, D. P., Soltani, E. G., Tappan, S., Wagenaar, J. B., … Hunter, P. J. (2021). The SPARC DRC: Building a Resource for the Autonomic Nervous System Community. *Frontiers in Physiology, 12*(929), 693735. https://doi.org/10.3389/fphys.2021.693735

Osumi-Sutherland, D. (2017). Cell Ontology in an Age of Data-Driven Cell Classification. *BMC Bioinformatics, 18*(Suppl 17), 558.

Paul, A., Crow, M., Raudales, R., He, M., Gillis, J., & Huang, Z. J. (2017). Transcriptional Architecture of Synaptic Communication Delineates GABAergic Neuron Identity. *Cell, 171*(3), 522–539.

Petilla Interneuron Nomenclature Group, Ascoli, G. A., Alonso-Nanclares, L., Anderson, S. A., Barrionuevo, G., Benavides-Piccione, R., Burkhalter, A., et al. (2008). Petilla Terminology: Nomenclature of Features of GABAergic Interneurons of the Cerebral Cortex. *Nature Reviews Neuroscience, 9*(7), 557–568.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., et al. (2017). The Human Cell Atlas. *eLife, 6*, e27041. https://doi.org/10.7554/eLife.27041

Richardet, R., Chappelier, J. C., Tripathy, S., & Hill, S. (2015, November). Agile text mining with Sherlok. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1479–1484). IEEE. https://doi.org/10.1109/BigData.2015.7363910

Rudy, B., Fishell, G., Lee, S., & Hjerling-Leffler, J. (2011). Three Groups of Interneurons Account for Nearly 100% of Neocortical GABAergic Neurons. *Developmental Neurobiology, 71*(1), 45–61.

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., et al. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell, 166*(5), 1308–1323.

Shepherd, G. M., Marenco, L., Hines, M. L., Migliore, M., McDougal, R. A., Carnevale, N. T., Newton, A. J. H., Surles-Zeigler, M., & Ascoli, G. A. (2019). Neuron Names: A Gene- and Property-Based Name Format, With Special Reference to Cortical Neurons. *Frontiers in Neuroanatomy, 13*, 25.

Stevens, R., & Sattler, U. (2012). Disjointness Between Classes in an Ontology. *Ontogenesis*

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3, 160018

Yuste, R., Hawrylycz, M., Aalling, N., Aguilar-Valles, A., Arendt, D., Armañanzas, R., ... & Lein, E. (2020). A community-based transcriptomics classification and nomenclature of neocortical cell types. *Nature Neuroscience*, *23*(12), 1456–1468. https://doi.org/10.1038/s41593-020-0685-8

Zeng, H., & Sanes, J. R. (2017). Neuronal Cell-Type Classification: Challenges Opportunities and the Path Forward. *Nature Reviews Neuroscience, 18*(9), 530–546.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.