ARTICLE
Enhancing Student Research Experiences with Open Data from the Allen Brain Map

Kaitlyn Casimo
Allen Institute, Seattle, WA, 98109.
https://doi.org/10.59390/NWXO4043

The Allen Brain Map is the main data repository for the Allen Institute for Brain Science, containing big, open datasets commonly used in neuroscience research (Allen Institute for Brain Science, 2022). Open data from the Allen Brain Map can be used to teach core concepts in neuroscience, data analysis methods, and other critical skills and knowledge to neuroscience students. These datasets can be used as the main data source for completely online lab experiences, or analyzed in combination with data students collect themselves. Applications may range in scope and format from a short worksheet used in a single class session to a coding tutorial to a guided independent research project. While open online data cannot fully replace lab experiences for learning techniques, they can be used to expose students to analysis of big data, introduce resources widely used in the field, and teach skills like statistics and coding. This article reviews potential assignment formats where big and open data can be applied, introduces selected popular resources and sample use cases for each, and discusses benefits and limitations of open online data for lab experiences. Some specific applications in the context of distance learning are also detailed.

Key words: open data; big data; data analysis; student independent research; online labs.

The Allen Institute is a nonprofit biomedical research institute located in Seattle, Washington. One of the core principles of the Allen Institute is open science: sharing all data, tools, and research findings freely for anyone to use. The Allen Institute for Brain Science, one of the research divisions of the Allen Institute, creates large-scale, open datasets that address fundamental questions about the brain's components and functions. These datasets and other tools form the Allen Brain Map and are publicly available online at https://brain-map.org. As of fall 2020, the Allen Institute for Brain Science has generated over 14 petabytes of data, all of which is made publicly available after quality control and processing.

These large, open resources were developed with research uses in mind. However, they have obvious utility for student independent and course-based research experiences, either in combination with data collected by the student or as the primary object of analysis. Traditional research applications of the open resources, usually in combination with other data collected by the user, include identifying typical expression patterns of genes of interest (including those associated with neurological conditions), mapping the gene expression patterns of a neural circuit, and identifying cell types associated with a neurological condition (Bernier et al., 2014; Weed et al., 2019; Zeisel et al., 2018).

Using these resources in classes gives students exposure to some large, open datasets that are commonly used in modern neuroscience. These resources also provide access to data types that are not able to be generated in a teaching lab at scale, and for certain methods or data sources (e.g., calcium imaging, living human tissue data), may not be available in teaching labs at all. In the context of a course, applications of these open resources may take such disparate forms as a brief worksheet-guided exploration intended to introduce a new concept or type of data, a guided coding exercise, or a substantial, potentially publishable experimental effort driven by independent student-generated research questions.

Open data for learning neuroscience has various benefits and applications whether students are present in a laboratory setting, in a class that does not have a lab practical, or learning online. Most obviously, the resources are free to use and download. They illustrate some major core concepts in neuroscience such as gene expression in the brain, neuroanatomy, projections and connectivity, and neural coding, which can help students place these concepts into context. The datasets are also big and broad, so a wide array of potential analyses is possible and statistical analysis will be robust with large sample sizes. Some analysis can be done in interactive online viewers, or downloaded data can be analyzed with common, free software such as Python and R. The open resources contain large volumes of multiple types of data with well-documented methods, many of which are impractical for teaching and for many research labs to collect even at smaller quantities (e.g., whole-brain human gene expression data). Some of the resources are best analyzed programmatically and have sample code available, providing opportunities to learn Python and R, which has significant career benefits both in neuroscience and beyond (Juavinett, 2020a). Finally, students learn to analyze and interpret data from resources commonly used in the neuroscience field that will be potentially useful over the course of their career.

This article will review potential assignments that leverage open data to provide experimental experiences that students may not be able to access otherwise, introduce

selected data resources that are especially relevant for educational users, and discuss advantages and limitations of using open online data with undergraduate students for course-based research experiences and other assignments. This introduction expands on prior work that introduced model assignments (Ramos et al., 2007; Wiertelak and Ramirez, 2008; Jenks, 2009; Chu et al., 2015; Grisham et al., 2017) and educational applications for the Allen Mouse Brain Atlas (Gilbert, 2018) by expanding the scope to more of the Allen Brain Map's data resources, as well as introducing a general template for potential research-based assignments.

POTENTIAL ASSIGNMENTS

As of fall 2020, the Allen Brain Map contains over 2 petabytes of processed, quality-controlled open data resources, data analysis tools and code, and laboratory tools. Classroom applications can be broadly divided into two categories: those that exclusively use the open online resources and those that combine the open resources with data students collect themselves. Within these categories, data collection/analysis, reporting, and other assignments can be tailored to the needs of the class.

Introductory Lesson Plans

Lesson plans by Allen Institute staff (Allen Institute, 2020; Gilbert, 2018) and by others (Ramos et al., 2007; Jenks, 2009; Chu et al., 2015; Juavinett, 2020a, 2020b; Ryan and Casimo, 2021; Ho et al., 2022) can provide a starting point for educators looking to incorporate open data into their classes. Some of these lessons are driven by short worksheets that introduce core concepts in neuroscience through representative data. Others include more complex experiments and advanced data analysis, including programmatic analysis. In both models, the lessons can also be used to prepare students to design their own experiments after using the prepared lesson as an introduction to the data resource. The Allen Institute does not require permission to use its open resources for education or research purposes, so educators can incorporate them into lessons quickly without needing to wait for approval, and if students produce notable findings, they can publish them online or in a journal.

Independent Research

The data resources were developed with open-ended research applications in mind and are not curated for a specific application or novice users. They are therefore well suited to a course-based research experience that asks students to generate and investigate a research question over the course of several weeks. This question may use data the student collected themselves in conjunction with one or more of the Allen Brain Map resources, or may use only the open resources. For example, one faculty member assigned students to complete an independently designed project combining data from at least two Allen Brain Map resources.

Because the data resources are expansive and comprehensive, many research questions are options for students. Potential research questions are expanded further if they are able to combine the open data with data they collected themselves in a course-based lab or a faculty research lab. Some sample research applications of the resources are detailed in the next section. In addition to pursuing scientific research questions, students could also use the resources as the basis for projects focused on developing analysis tools and techniques, such as automated image processing.

Whether students are using the data as the primary object of analysis or in combination with their own data, they will need to know methodological details and have past work to build on. The Allen Institute documents the methods used to generate the data via explanatory white papers linked on the Documentation tab of each open resource's web page. Institute scientists also release their own research findings using the open resources in journals (e.g., Gouwens et al., 2020; Harris et al., 2019; Oh et al., 2014; Tasic et al., 2018; complete list available at https://alleninstitute.org/what-we-do/brain-science/research/scientific-publications/). Both types of document can be used as reading assignments for students completing short lessons using the data and provide necessary background for deeper independent research efforts. Because the journal articles use the publicly available data, a potential assignment could also entail a limited replication of the described analysis.

Skills Development

The analysis of large datasets provides a natural opportunity to introduce life sciences students to coding and statistics. Coding skills are increasingly important both within the field of neuroscience and beyond, but few neuroscience programs require coding coursework despite its widespread use in the field (Akil et al., 2016; Grisham et al., 2016, Juavinett, 2021). Sample code is available through the Allen Institute's GitHub (https://github.com/AllenInstitute), which students can use as a starting point for developing their own code. The large and highly quality controlled nature of the datasets also provides students with opportunities to practice robust statistical analysis including more advanced techniques such as clustering analysis.

The Allen Institute publishes its own research on the open data resources, and they are also extensively used by researchers across the world in their work. Students can read a paper describing Institute research on the open data, or outside research combining the open data with other data, as a reading assignment.

Because the resources are well suited to supporting independent student research, such assignments also lend themselves well to practicing standard forms of disseminating results: creating a conference-style poster and presenting it for their classmates/other faculty or writing a short journal-style article. Students who incorporate programmatic analysis into their work may also take the opportunity to document and present a GitHub repository documenting their work.

SAMPLE USE CASES FOR OPEN DATA FROM THE ALLEN BRAIN MAP

Some of the Allen Brain Map resources are particularly popular for educational uses because they demonstrate core neuroscience concepts, use common techniques, or

provide data that students cannot practically collect themselves. Some key properties of the datasets with regard to their educational applications/use by students are summarized in Table 1. A video tutorial introducing these and other open resources from the Allen Brain Map is available at https://youtu.be/b_UvVjWydfo.

For questions about accessing and analyzing data, students and educators are encouraged to explore video tutorials available for many resources and to read and post in the Allen Brain Map Community Forum at https://community.brain-map.org.

These potential research questions are only a starting point, meant to help educators and students brainstorm their own project ideas. Many more questions are also possible, especially in combination with data collected in a course-based lab or research lab, or by combining multiple datasets from the Allen Brain Map. Students with an interest in data science, statistics, or machine learning may also wish to use the data to support projects geared more towards creating analysis tools than basic science questions.

Whole-Brain Gene Expression Atlases

The Allen Brain Map includes several whole-brain, whole-genome, gene-expression datasets across adult and developing mouse and human brains. For the mouse brain, gene expression is estimated using *in situ* hybridization (ISH). For the human brain, whole-brain whole-genome data is available using an RNA microarray; a limited number of genes and brain regions are evaluated with ISH. These resources are relevant to learning about gene expression and central dogma in the context of the brain, methods of measuring gene expression, neuroanatomy, and development. For a detailed overview of using the Allen Mouse Brain Atlas in classrooms, see Gilbert (2018), and for an example lesson using these atlases, see Ryan and Casimo (2021).

In any one of these atlases, the most obvious experimental application of the data is for students to look up the expression pattern of a specific gene or genes: for example, which genes are highly expressed in a region of interest, or in which regions (or networks) are a gene of interest highly expressed. Students may then infer what, if anything, the expression pattern of a gene indicates about the functions of the brain region(s) where it is highly expressed. Students can also investigate colocalization of genes; the AGEA search tool facilitates searching for genes whose expression colocalizes with a selected gene of interest or brain regions with covarying expression patterns. The brains contained in the atlas are all from wild-type adults (mouse) or postmortem donations from healthy individuals (human); students may identify the typical expression patterns for a gene that has been associated with a genetic disorder in both the adult and developing brain.

These datasets can also be combined to conduct cross-species analyses. For example, students who have conducted an experiment related to gene expression in the fruit fly can look up the expression patterns of the homologous gene in the human or mouse brain. This may also include self-contained comparisons between the Allen Mouse Brain Atlas and Human Brain Atlas. Similarly, students can examine gene expression across stages of development or compare data from the adult and developing brain, either within the resources provided by the Allen Brain Map or in combination with data they collect themselves.

Allen Cell Types Database

This collection of resources interrogates properties of single cells with the ultimate goal of classifying them into well-characterized cell types. The database includes electrophysiological characterization of cells (patch clamp), 3D morphological reconstructions of cells, and single-cell, whole-genome transcriptomics (RNA-seq). These profiles, particularly the transcriptomics, are then used to define and characterize individual cell types. Additional datasets combine all three types of data from a limited number of cells (patch-seq dataset) and investigate how cells interact (synaptic physiology dataset). This dataset can be used to introduce students to these major fundamental properties of cells and how they vary.

Institute scientists have previously grouped cells into types based on the expression of ~20 marker genes (Tasic et al., 2016). Students can use these classifications for research questions such as how the expression of other, non-marker genes of interest vary according to cell type or identifying which cell types are likely to be affected in a disease with a known genetic component. The data from the Allen Cell Types Database can also be powerfully combined with the whole-brain gene expression atlases to place projects into the context of broader neuroanatomy. Unlike the whole-brain atlases, gene expression is assessed at the level of individual cells (single-cell RNAseq for mouse brain and single-nucleus RNAseq for human brain). Another potential line of research for students is to examine co-expression patterns of selected genes between the whole-brain atlas and the single-cell data.

The Allen Cell Types Database is also well suited to introducing students to using programming for analysis, especially for the electrophysiology and transcriptomic data. The Allen Institute makes its code available (primarily in Python for electrophysiology data and R for transcriptomic data). For example code at the level of a student with no computer science experience, please see Juavinett (2020) and Ho et al. (2021). For a non-coding lesson plan introducing the electrophysiology and transcriptomic data, see the Allen Institute's Building Blocks of the Brain lesson (Allen Institute, 2020).

Allen Mouse Brain Connectivity Atlas

This resource uses anterograde fluorescent viral tracers to map the projections from selected brain regions. The injection visualizes the targets of all neurons whose cell bodies are located at the injection site (called the source) or, for Cre line experiments, of all neurons of the selected cell type. The Allen Cell Types Database described above includes morphological data on individual cells but does not evaluate their projection patterns, so this resource can help students place neuronal projections into the context of anatomy, as well as demonstrating projection tracing methods and their interpretation.

The Allen Mouse Brain Connectivity Atlas is used as the subject of a guided experiment in the third section of the

	Main method used, data format	***Data interpretation***	***Online or offline data analysis recommended***	***Computational skills: coding and statistics***
Allen Mouse Brain Atlas	*In situ* hybridization (ISH), images	Relative expression level of genes in whole brain regions	Online for most analyses (built-in ISH image viewer)	No coding or statistics needed for basic analysis
Allen Human Brain Atlas	Mostly microarray as Z scores, some ISH as images	Relative expression level of genes in whole brain regions	Online analysis for one/a few genes, download recommended for analysis of many genes	No coding needed for basic analysis, depends on understanding of Z score calculation and interpretation
Allen Mouse Brain Connectivity Atlas	Anterograde projection tracing with viral tracers (images)	Projection density and locations	Online for most analyses (built-in projection tracing image viewer, online 3D Brain Explorer)	No coding or statistics needed for basic analysis
Allen Cell Types Database - Transcriptomics	RNA sequencing – multiple variants (raw values)	Relative expression levels of genes in single cells	Online for basic analysis, download for more than basic lookups	No coding needed, optional R or Python for advanced users; basic statistics knowledge (mean/quartiles) needed
Allen Brain Observatory	Calcium imaging movies, precomputed statistics available; electrophysiology recordings; behavioral responses	Cellular responses to visual stimuli	Download, or instructions given for using AWS for cloud computing	Python experience and comfort with most basic statistical analyses strongly recommended

Table 1. Properties of selected databases from the Allen Brain Map.

Allen Institute's Exploring Pathways in the Brain lesson plan (Allen Institute, 2020). Experimental questions introduced in that lesson, which students can expand on for larger independent research projects, include identifying and interpreting all the sources for a given target region or all the targets for a given source. For an example using this dataset for a course-based research experience, see Ryan and Casimo (2021).

Allen Brain Observatory

This collection of data resources includes two-photon imaging of calcium responses and electrophysiological recordings of the visual cortex in response to visual stimuli. Mice are presented with a set of standardized visual stimuli, many of which are standardized across the two recording methods. These datasets can be used to investigate visual coding and have applications for computational neuroscience and modeling. Example use cases include examining the relationship between behavioral (running speed and eye movement) and neuronal responses to specific stimuli, generating computational models of neuronal firing or connectivity, or comparing responses to new or familiar images.

While all of the Allen Brain Map resources can be accessed and analyzed programmatically to some degree, meaningful analyses of the Allen Brain Observatory depend especially heavily on coding tools. Institute scientists perform this analysis in Python, and make demonstration code available in this language, accessible at the SDK link on the landing page.

BENEFITS AND LIMITATIONS OF OPEN DATA RESOURCES FOR EDUCATIONAL USERS

Benefits

The benefits of using open data are especially evident in the context of distance learning, but as illustrated by the range of potential and actual example assignments described above, the benefits extend into traditional classroom-based

learning as well as a variety of curricular contexts. These resources support equity in access to neuroscience education and high-quality lab experiences by their free and fully online nature.

First, the data resources and materials for educators are all free to use. The software requirements for offline analysis are relatively minimal, and all are also free. All of the institute's code for analysis is in Python or R, both of which are free, open source, and do not require a computer with any special configurations. Students can use this code to get started with their own more advanced research projects requiring offline analysis. Several of the data resources can be meaningfully analyzed completely online (such as the gene expression atlases) and students who are working with less powerful computers or tablets can concentrate their analysis efforts in these areas.

As described above, the open data resources can be used as the sole data for a student's exploration or experiment or combined with other data collected in class. The data collection process is highly quality controlled, so when used as the primary subject of analysis, students can focus on the data analysis and interpretation rather than data cleanup. The big-data nature of the datasets mean they also lend themselves well to teaching statistics, either by supplementing student-collected data or on their own, as the sample sizes are large enough to support robust effect sizes.

Limitations

While the barriers to access are reduced compared to the equipment, facilities, and time needed for actually collecting the data themselves, students still need access to high-speed internet and a computer in order to view and analyze data. This is particularly true if they are conducting an independent research project that requires more advanced data analysis, which in some cases may require downloading the data for offline analysis and/or installing Python or R. Slower or metered internet may not be sufficient, and depending on the offline analysis in question, a tablet may not support the required software. For students on a campus, school resources are likely adequate, so this is a concern primarily for students who are distance learning.

There are also limitations in using these resources for single labs or longer course-based research experiences. Most importantly, students do not get the hands-on lab experience of collecting data themselves, including learning how to troubleshoot a technique. For some techniques and classes, this may be mitigated by having the students collect a small amount of data themselves using the technique (e.g., *in situ* hybridization) and then using the online resource for analysis to take advantage of the much larger sample sizes and high-quality data (e.g., Allen Mouse Brain Atlas). Further, for many of the open resources, the methods and/or biological samples used are not typically available in undergraduate teaching labs, such as human tissue samples or optogenetic calcium imaging data. For students who do not have lab access, because they are learning remotely or are in classes that do not have a practical component, the ability to do any data analysis without access to a lab can provide some experimental experiences that would not otherwise be possible.

CONCLUSION

The Allen Institute's big, open data resources can be used to enable research experiences for students, introduce them to fundamental concepts and tools widely used in neuroscience, and facilitate the development of critical skills like coding and presenting results. All of the data resources and analysis tools, as well as many lesson plans developed by the Allen Institute or others using its data resources, are free and open access, which promotes equitable access to learning opportunities.

These open resources provide access to types and quantities of high-quality data that are not accessible in an undergraduate teaching lab, such as data from human brains. These resources help reinforce and illustrate core concepts in neuroscience as well as enable students to develop their own research questions to interrogate those concepts. In addition to learning about scientific concepts and developing research questions, using these data resources can allow students to focus on other skills like coding, statistics, and presenting their work. Students with notable findings may even choose to publish them or present them at a conference, further supporting the progression of their education and career.

Online learning with ready-to-use, quality-controlled data cannot fully substitute for learning how to run and troubleshoot lab techniques in person. However, some students do not have the option to be in lab, either temporarily due to distance learning or permanently due to taking online course or a lack of available facilities. Open online data provides an opportunity for these students to learn about methods, data analysis, statistics, coding, and presenting their work. For students who are learning in a lab setting, the open data can be powerfully used in combination with data students collect themselves, which is how they are most commonly used by working neuroscientists, or to provide relevant data that students are not able to collect themselves, at all or in sufficient sample sizes, because of equipment access, time, or cost.

REFERENCES

Akil H, Balice-Gordon R, Cardozo DL, Koroshetz W, Posey Norris SM, Sherer T, Sherman SM, Thiels E (2016) Neuroscience training for the 21st century. Neuron 90:917–926. doi: 10.1016/j.neuron.2016.05.030

Allen Institute (2020) Education Outreach. Seattle, WA: Allen Institute. Available at https://alleninstitute.org/about/education-outreach/.

Allen Institute for Brain Science (2022) Allen Brain Map. Seattle, WA: Allen Institute. Available at https://portal.brain-map.org/.

Bernier R et al. (2014) Disruptive CHD8 mutations define a subtype of autism early in development. Cell 158:263–276. doi: 10.1016/j.cell.2014.06.017

Chu P, Peck J, Brumberg JC (2015) Exercises in anatomy, connectivity, and morphology using neuromorpho.org and the Allen Brain Atlas. J Undergrad Neurosci Educ 13(2):A95-100. Available at https://pubmed.ncbi.nlm.nih.gov/25838808/.

Gilbert T (2018) The Allen Brain Atlas as a resource for teaching undergraduate neuroscience. J Undergrad Neurosci Educ 16. Available at https://pubmed.ncbi.nlm.nih.gov/30254541/.

Gouwens NW, Sorensen S, Baftizadeh F, Budzillo A (2020)

Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. Cell 183. doi: 10.1016/j.cell.2020.09.057

Grisham W, Brumberg JC, Gilbert T, Lanyon L, Williams RW, Olivo R (2017) Teaching with big data: Report from the 2016 Society for Neuroscience Teaching Workshop. J Undergrad Neurosci Educ 16(1):A68-76. Available at https://pubmed.ncbi.nlm.nih.gov/29371844/.

Grisham W, Lom B, Lanyon L, Ramos RL (2016) Proposed training to meet challenges of large-scale data in neuroscience. Front Neuroinformatics 10. doi: 10.3389/fninf.2016.00028

Harris JA et al. (2019) Hierarchical organization of cortical and thalamic connectivity. Nature 575:195–202. doi: 10.1038/s41586-019-1716-z

Ho Y-Y, Roeser A., Law G., and Johnson BR. (2022) Pandemic Teaching: Using the Allen Cell Types Database for Final Semester Projects in an Undergraduate Neurophysiology Lab Course. J of Undergrad Neurosci Educ 20(1): A102-112. Available at https://pubmed.ncbi.nlm.nih.gov/35540944/.

Jenks BG (2009) A Self-Study Tutorial using the Allen Brain Explorer and Brain Atlas to teach concepts of mammalian neuroanatomy and brain function. J Undergrad Neurosci Educ 8(1): A21-25. Available at https://pubmed.ncbi.nlm.nih.gov/23493964/.

Juavinett A (2020) Open neuroscience education. Available at https://sites.google.com/ucsd.edu/neuroedu/.

Juavinett A (2021) Learning how to code while analyzing an open access electrophysiology dataset. J Undergrad Neurosci Educ 19. doi: 10.3389/neuro.11.014.2009

Oh SW et al. (2014) A mesoscale connectome of the mouse brain. Nature 508:207–214. doi: 10.1038/nature13186

Ramos RL, Smith PT, Brumberg JC (2007) Novel *in silico* method for teaching cytoarchitecture, cellular diversity, and gene expression in the mammalian brain. J Undergrad Neurosci Educ 6:A8–A13. Available at https://pubmed.ncbi.nlm.nih.gov/23493835/.

Tasic B et al. (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci 19:335–346. doi: 10.1038/nn.4216

Tasic B et al. (2018) Shared and distinct transcriptomic cell types across neocortical areas. Nature 563:72–78. doi: 10.1038/s41586-018-0654-5

Ryan J and Casimo K (2021) A Course-Based Research Experience Using the Allen Brain Map: From Research Question to Poster Session. J of Undergrad Neurosci Educ 19(2): A260-266. Available at https://pubmed.ncbi.nlm.nih.gov/34552441/.

Weed N, Bakken T, Graddis N, Gouwens N, Millman D, Hawrylycz M, Waters J (2019) Identification of genetic markers for cortical areas using a Random Forest classification routine and the Allen Mouse Brain Atlas. PLOS ONE 14:e0212898. doi: 10.1371/journal.pone.0212898

Wiertelak EP, Ramirez JJ (2008) Undergraduate neuroscience education: Blueprints for the 21st century. J Undergrad Neurosci Educ 6:A34–A39. Available at https://pubmed.ncbi.nlm.nih.gov/23493318/.

Zeisel A et al. (2018) Molecular architecture of the mouse nervous system. Cell 174:999-1014.e22. doi: 10.1016/j.cell.2018.06.021

Received January 8, 2020; revised January 29, 2021; accepted January 19, 2021.

This work was supported by the Allen Institute. The Allen Institute thanks our founder, Paul G. Allen, for his vision, encouragement, and support. Thank you to Rachel Tompa and Nick Hawley for their feedback and support in the development of this article.

Address correspondence to: Dr. Kaitlyn Casimo, Allen Institute, 615 Westlake Avenue North, Seattle, Wa 98109. Email: communications@alleninstitute.org