The Brain Image Library:
A Community-Contributed Microscopy Resource for Neuroscientists

Mariah Kenney [1,*], Iaroslavna Vasylieva[2,3,*], Greg Hood [1], Ivan Cao-Berg[1], Luke Tuite[1], Rozita Laghaei [1], Alan M. Watson [2,3], Alexander J. Ropelewski [1]

[1] Pittsburgh Supercomputing Center, Carnegie Mellon University, Pittsburgh PA.
[2] Center for Biologic Imaging, University of Pittsburgh, Pittsburgh PA
[3] Department of Cell Biology, University of Pittsburgh, Pittsburgh PA
*These authors contributed equally to this work.

Keywords: data archive, neuroscience, microscopy, big data, metadata, FAIR

Correspondence should be addressed to:
AJR: ropelews@psc.edu
AMW: alan.watson@pitt.edu

Abstract

Advancements in microscopy techniques and computing technologies have enabled researchers to digitally reconstruct brains at micron scale. As a result, community efforts like the BRAIN Initiative Cell Census Network (BICCN) have generated thousands of whole-brain imaging datasets to trace neuronal circuitry and comprehensively map cell types. This data holds valuable information that extends beyond initial analyses, opening avenues for variation studies and robust classification of cell types in specific brain regions. However, the size and heterogeneity of these imaging data have historically made storage, sharing, and analysis difficult for individual investigators and impractical on a broad community scale. Here, we introduce the Brain Image Library (BIL), a public resource serving the neuroscience community that provides a persistent centralized repository for brain microscopy data. BIL currently holds thousands of brain datasets and provides an integrated analysis ecosystem, allowing for exploration, visualization, and data access without the need to download, thus encouraging scientific discovery and data reuse.

Introduction

In the era of big data and open science, the efficient management and sharing of research data have become crucial for scientific progress. The Brain Initiative Cell Census Network (BICCN)

collaborative research initiative aims to comprehensively catalog and understand the diversity of cell types in the brain. Multiple groups of investigators funded by BICCN focused their collaboration on (i) understanding brain structure, (ii) cell type classification, (iii) mapping brain connectivity, (iv) creating a comprehensive reference cell type atlas, (v) advancing neuroscience research and (vi) data sharing and collaboration. This recent effort created numerous publications, with several more expected over the next few years. Based on this initial success, the follow-on BRAIN Initiative Cell Atlas Network (BICAN) is expected to image physically larger volumes such as whole brains from primates (including humans).

The vast amounts of data produced by these collaborative efforts must be shared easily and made available as rapidly as possible - promoting transparency and scientific discovery. To support these data-sharing goals and to facilitate the reuse and accessibility of data to scientists, the National Institutes of Health (NIH) BRAIN Initiative[1] established several archives to retrieve, store, and make available modality-specific data being produced by or of interest to the BRAIN Initiative (Table 1). Each archive houses modality-specific data, with the Brain Image Library (BIL) focused on optical microscopy. The mission of BIL is to be a national public resource enabling researchers to deposit, analyze, mine, share, and interact with microscopy datasets of the brain by providing (i) a permanent repository for high-quality brain microscopy datasets, (ii) an analysis ecosystem with desktop visualization and high-performance computing (HPC) capability and (iii) user access, training, and support. BIL is housed at the Pittsburgh Supercomputing Center (PSC) and sits adjacent to the center's flagship HPC system, Bridges-2[2], which is a uniquely capable petascale resource for empowering diverse communities by bringing together HPC, Artificial Intelligence, and Big Data. BIL provides data submission and data inventory search portals, RESTFul APIs, and visualization tools that are run using computational analysis resources available at the PSC. The scope of data at BIL includes whole and partial brain microscopy image datasets, their accompanying derived data, and other historical collections of value to the community. BIL accepts optical image data that can include images directly from the imaging apparatus in a format that is open and accessible, and processed data that has been computed upon or transformed. There is no limit on the amount of data deposited per dataset or investigator.

Dozens of research teams across the globe, including those from BICCN and BICAN have deposited their data at BIL. BIL currently contains brain datasets across multiple species (mouse, marmoset, macaque, human, fruit fly, and ant), including experiments focused on high-resolution volumetric microscopy, cell morphology, connectivity, receptor mapping, cell counting/population, and spatial transcriptomics (Fig. 1). The microscope technologies used to create the datasets include serial two-photon tomography (STPT)[3], fluorescence micro-optical sectioning tomography (fMOST)[4], light-sheet fluorescence microscopy (LSFM), and confocal among others.

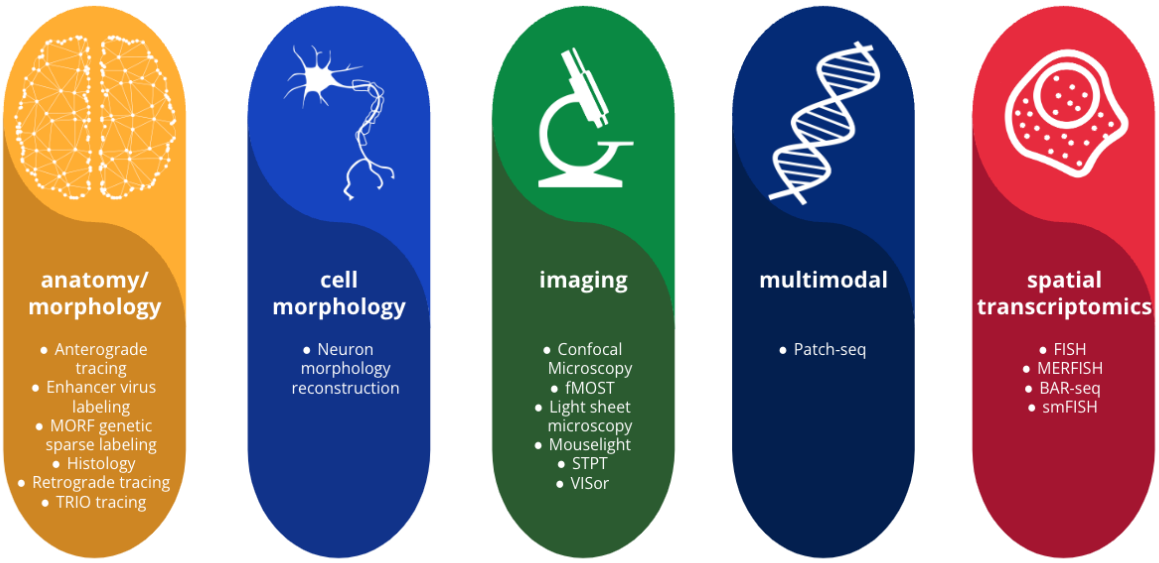

Fig. 1. An overview of the modalities and techniques of the data accepted and available at BIL adapted from the BICCN Data Catalog Glossary[5].

We believe BIL to be the largest archive of its kind and the first petascale brain microscopy data resource. Other related efforts include the Image Data Resource[6] EBRAINS[7], The Cell Vision[8], The Cell Image Library[9], and The BioImage Archive[10]. However, these resources are or were smaller in scale, limited in the size of the imagery they could accept, are based internationally, or may include images for many different fields and are not specifically tailored for neuroscience.

Data Type	Archive	Website
Optical microscopy	Brain Image Library (BIL)	http://www.brainimagelibrary.org
Multi-omics	Neuroscience Multi-Omic Data Archive (NeMO)[11]	https://nemoarchive.org
Invasive device	Data Archive for the Brain Initiative (DABI)[12]	https://dabi.loni.usc.edu
Magnetic Resonance Imaging	OpenNeuro[13]	https://openneuro.org
Positron Emission Tomography	OpenNeuroPet	https://openneuropet.github.io
Electron microscopy and X-ray Microtomography	Block and Object Storage Service & Database(BossDB)[14]	https://bossdb.org
Cellular neurophysiology	Distributed Archives for Neurophysiology DataIntegration (DANDI)	https://www.dandiarchive.org
Human EEG and MEG	Neuroelectromagnetic Data Archive and ToolsResource (NEMAR)[15]	https://nemar.org

Table 1. The BRAIN Initiative Data Archives and their data focus

Methods

BIL is designed to serve the scientific community at all stages of the scientific process. Its purpose is to provide researchers with a platform that enables them to access and analyze data, collaborate with other researchers, and publish their findings. Data published at BIL undergoes validation and curation to ensure file validity, data reusability, and interpretability. Data contributed to BIL is distributed in a way that allows for the broadest use and reuse.

Image File Formats

BIL contains image file formats that are accessible and best suited for reuse. Most deposited images are volumetric stacks in native TIFF and JPEG 2000 image file formats. Currently, we

are encouraging the use of high-performance next-generation file formats (NGFF)[16] such as multiscale OME-Zarr[17], which represent multi-resolution image pyramids optimized for rapid visualization and scalable analysis. While BIL continues to accommodate historically accepted file formats, datasets in these formats may not fully leverage the complete range of visualization capabilities now provided through BIL.

Metadata

To maximize the value and impact of data, it is essential to ensure that datasets are Findable, Accessible, Interoperable, and Reusable (FAIR)[18]. The implementation of FAIR standards requires the development and adoption of standardized metadata. Having detailed standardized metadata promotes data reuse by enabling researchers to understand the context and limitations of the data, ensuring proper interpretation, and facilitating the reproducibility of scientific experiments. Structured information describing datasets plays a fundamental role in enabling data discoverability, understanding, and integration. BIL initially started with a simple metadata model that included ~14 flexible fields for broad descriptions of the datasets. In 2021, BIL transitioned from this initial model to the 3D Microscopy Metadata Standards (3D-MMS) developed through a BRAIN Initiative standardization project[19]. The standard is in the process of being adopted by the International Neuroinformatics Coordinating Facility (INCF)[20].

The 3D-MMS model is aligned with DataCite metadata[21] and includes information about the data contributors, project funders, and associated publications. In addition, the model contains descriptive metadata about the dataset, the specimen, and the instrumentation used to generate the data. Only 31 metadata fields of the standard are required fields with 19 needed to assign Digital Object Identifiers (DOI). There are also 41 fields within the standard that align with the OME metadata schema[22]. Expansions to the 3D-MMS include new fields specific to BIL to better describe neuron tracing datasets deposited at BIL.

Finding and downloading data

At BIL, data findability is ensured by assigning a unique DOI to each dataset upon publication. This provides a unique stable identifier, easy tracking of data use, and simple citations for data.

Each dataset has a landing page (Fig. 2) that provides detailed metadata, data descriptions, related datasets, and publications. BIL uses several persistent identifiers to identify data:

- Submission Identifiers. A unique ID is assigned to all data that is part of a submission to the archive. The submission identifier is a 16-digit hexadecimal string. An example of a BIL submission identifier is abcdef0123456789.
- Dataset Identifier. All datasets are assigned unique identifiers. These identifiers consist of a string of (English) pronounceable triplets. An example of a BIL dataset identifier is "ace-cot-new".
- Traced Neuron Identifier. BIL assigns unique identifiers to traced neurons when additional metadata is available for the neuron. These identifiers are similar to Dataset Identifiers with the exception that they all begin with the triplet "swc".
- DOI (Dataset). When a DOI is issued for a BIL dataset, it will be assigned a URL with the BIL doi prefix (10.35077) plus the data identifier. An example of a BIL dataset DOI is: "https://doi.org/10.35077/ace-and"
- DOI (Collection/Group). Upon request, a DOI can be issued which groups datasets together into a collection. For example, all datasets associated with a publication can be grouped into a single DOI. These DOIs begin with a "g." followed by a number. An example of a BIL group DOI is: "https://doi.brainimagelibrary.org/doi/10.35077/g.948"

Inventory Search

A metadata API and web portal that uses the API are available to search for datasets and return metadata along with links to the data. The metadata API is extremely flexible and is capable of full-text searching and searching all individual metadata fields. The API has two primary user-facing endpoints:

1. *query* - the query endpoint will query the system and return a JavaScript Object Notation (JSON) document with all matching dataset entries.
2. *retrieve* - the retrieve endpoint will return the metadata as a JSON document for the given Dataset Identifiers.

In addition to the API, BIL provides a web search interface that can query the metadata API in a variety of different ways, and display selectable results based on the query.

Bulk Download

While individual links to data are available through the web search interface described above, all BIL data is available for authenticated access using standard Unix tools (rsync, sftp, scp, etc.). Access without authentication is available through web URLs (https://download.brainimagelibrary.org) and a Globus/GridFTP endpoint "Brain Image Library Download". To facilitate the download of individual datasets, each dataset has an associated manifest file in JSON format which contains checksums and related metadata. Downloading data from BIL is optional as BIL provides a comprehensive Analysis Ecosystem (described in the next section), which includes a wide range of computational resources and tools to explore data in place.

Data Submission Process

Data generators who wish to submit their own data to BIL can do so by using the data submission process outlined in Fig. 3. The submission process includes (i) data upload through the submission portal, (ii) file validation, (iii) curation of data and metadata by a dedicated data curator, and ultimately, (iv) the publication of the data (Fig. 3). Data submitters have the option to select a limited embargo period for their data. During the embargo period, only the associated metadata will be accessible to the public.

The submission portal provides a bridge to link metadata with uploaded data. Three different ways to transfer data are supported: over the network, using Linear Tape-Open tapes, or portable shippable drives. The majority of data contributors transfer data over the network. A unique service that BIL provides is network analysis and diagnostic support to help resolve transfer bottlenecks between the BIL data center and the contributing site. In addition to linking metadata, the submission portal provides an administrative view for project data management.

Whole-brain LSFM imaging of a E15.5 mouse for DevCCF project (Subject: E15.5_BB0428Technique: Background)

[BIL ID: ace-cab-ear]

Dataset Citation:

Zhang, Jiangyang, Lee, Choong Heon, Jiangyang Zhang Lab, Betty, Rebecca, Liwang, Josephine, Yongsoo Kim, Kronman, Fae, Manjila, Steffy Babu, Yongsoo Kim, Yongsoo Kim Lab, Minteer, Jennifer. (2022). Whole-brain LSFM imaging of a E15.5 mouse for DevCCF project (Subject: E15.5_BB0428Technique: Background) [Dataset/Microscopy]. Brain Image Library.

Abstract:

This dataset is part of a BICCN project to create developmental CCFswith associated ontology and true 3D anatomical labels while also demonstrating the application of our CCFsby generating quantitative mappings of GABa Ergic neurons in the developing mouse brain. This sample consist of LSFM imaging of a E15.5 mouse (Subject: E15.5_BB0428Technique: Background Autofluorescence

Methods:

Images were processed using a modified LifeCanvas Protocol

Contributors and Roles:

Zhang, Jiangyang (ProjectLeader), New York University; Kronman, Fae (DataCurator), Pennsylvania State University ; Lee, Choong Heon (DataCollector), New York University; Jiangyang Zhang Lab (ResearchGroup) ; Betty, Rebecca (DataCollector), Pennsylvania State University; Liwang, Josephine (DataCollector), Pennsylvania State University; Yongsoo Kim (ProjectLeader), Pennsylvania State University; Kronman, Fae (ContactPerson), Pennsylvania State University; Manjila, Steffy Babu (DataCollector), Pennsylvania State University; Yongsoo Kim (ContactPerson), Pennsylvania State University; Yongsoo Kim Lab (ResearchGroup) ; Minteer, Jennifer (DataCollector), Pennsylvania State University; Beck, Kira (ProjectMember), Pennsylvania State University

Funding:

National Institutes of Health
(1-RF1-MH124605-01) Establishing Common Coordinate Framework for Quantitative Cell Census in Developing Mouse Brains

Technical Information:

other (Modality) / light sheet microscopy (Technique)

Images were stitched and compressed using in house code with MATLAB and Python, respectively

Instrument Metadata

microscopetype	Light sheet microscope (LSFM)
microscopemanufacturerandmodel	Life Canvas Tech
objectivename	
objectiveimmersion	
objectivena	

License:

Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

Data:

Data location on the Brain Image Library Analysis Ecosystem:

/bil/data/91/aa/91aad194ce577ebe/E15.5_BB0428/LSFM/stitched_00

Dataset download link:

https://download.brainimagelibrary.org/91/aa/91aad194ce577ebe/E15.5_BB0428/LSFM/stitched_00

Metadata download link:

https://api.brainimagelibrary.org/retrieve?bildid=ace-cab-ear

Manifest download link:

https://download.brainimagelibrary.org/inventory/122a14b8-df8a-59e4-8abf-7c1aea6cc47b.json

Fig. 2. The landing page for each dataset at BIL provides a brief description of the dataset, detailed information about the images that would be needed to reuse them, contributor and licensing information, as well as links to the data themselves and visualizations if available. The landing page for this example dataset is available at BIL with the dataset ID ace-cab-ear.

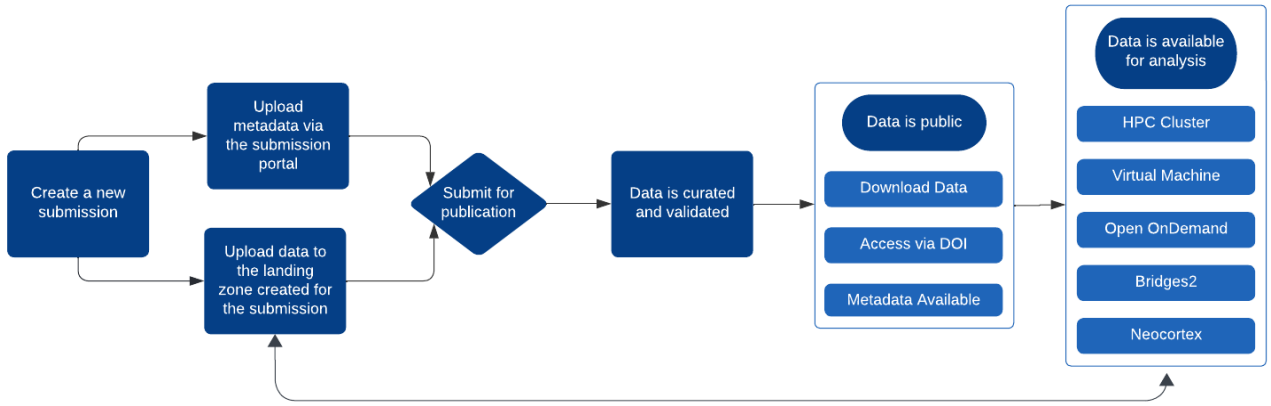

Fig. 3. A flow diagram of the data submission and publication process at BIL.

This view offers a comprehensive overview of the data submitted by the research project (and/or subproject) along with the submission status.

Data submission and exploration workshops are held regularly to guide the community through the data submission process and provide training on using the BIL Analysis Ecosystem. The recordings of prior workshops are available on the BIL website, and the supporting materials can be found in the BIL GitHub repository. A guide for the data submission process is also available online.

BIL Analysis Ecosystem

Architecture

The BIL Analysis Ecosystem offers computational resources designed for visualizing and processing BIL data and high-speed networking systems, eliminating the need for downloading. This ecosystem is equipped for desktop visualization and high-performance computing to handle pre-submission data processing and post-submission exploration. The architecture seamlessly integrates mass storage, networking, and HPC components within a unified environment (Fig. 4). While each component is briefly described here, readers interested in thorough technical architecture details of the ecosystem are referred to Benninger et al[23]. Data transfer components (top 3 boxes in Fig. 4) are a fundamental component of BIL and allow the capability to efficiently support large file and data transfers. The data transfer infrastructure includes redundant data transfer nodes integrated with wide area and local area network connections. The

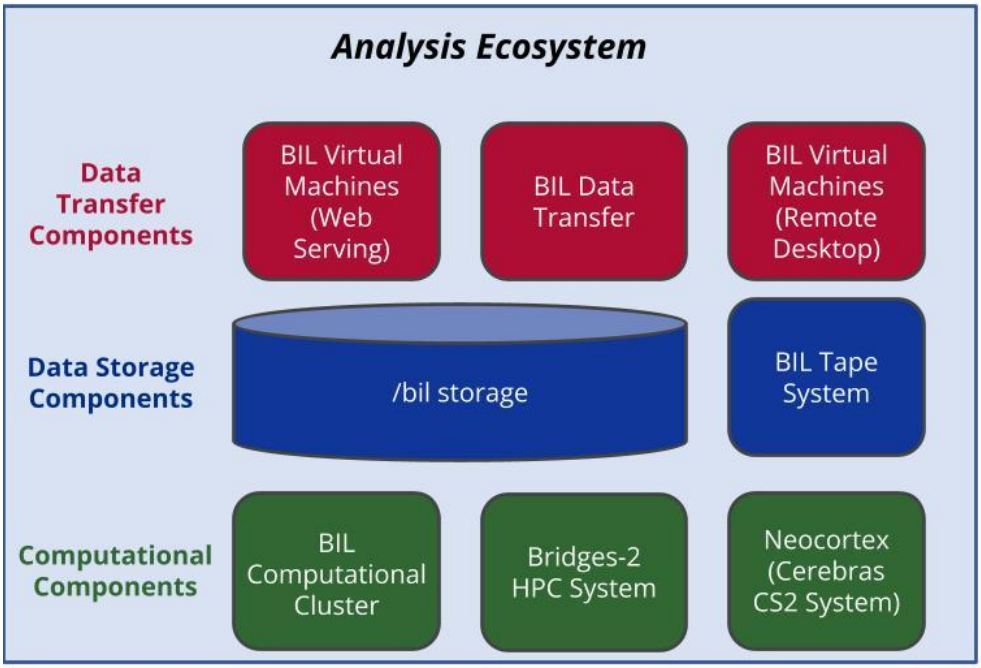

Fig. 4. Key components of the BIL Analysis Ecosystem include data transfer and visualization (top 3 boxes), data storage (middle 2 boxes), and computation (bottom 3 boxes).

virtual machine system provides a remote desktop environment to run interactive applications for visualization and analysis such as Fiji[24], Napari[25], and Vaa3d[26], and to serve web portals dedicated to tasks such as data ingestion, and APIs. The data storage components consist of a fault-tolerant Lustre multi-petabyte scalable filesystem which is mounted on all of our computing platforms. Tape is used as a medium both for making archival backups for all public data for the data filesystem and for bulk import of data when necessary. Computational components include a variety of large memory nodes, GPU nodes, and access to high-performance computing resources, including PSC's Bridges2[2] and Neocortex[27] systems for extensive data exploration.

Software

Various software options are available for users, including traditional software development languages, higher-level scripting languages, and applications. Most application software at BIL uses modules[28] that provide a uniform and stable method to access multiple software versions for reproducible workflows. Software available through the module system at BIL include image analysis tools, Artificial Intelligence toolkits, file conversion tools, and other open-source packages. BIL staff can work with users to enable new open-source software upon request.

Batch and Interactive Processing

BIL provides command-line access to explore data in place. The ecosystem uses the SLURM[29] scheduler, which can be used interactively and for batch processing. Access to resources is provided through login nodes which are capable of providing SLURM access to a variety of node types including HPC resources which include large memory nodes with up to 4 TB of RAM, GPU nodes with up to 8 GPUs, and thousands of smaller memory nodes.

Containers and workflows

Singularity[30] containers are supported on the analysis ecosystem. Singularity is a standard tool on HPC clusters that provides a secure way to enable users to create reproducible applications. While users can create their own containers, BIL provides several public singularity definition files, which users will be able to use as models for their own containers or pull for use on local infrastructure or public cloud computing environments. Most existing Docker containers can be

converted to Singularity containers. Workflows developed using the Common Workflow Language (CWL)[31] standard or SnakeMake[32] workflows can be run on BIL systems.

Web Computational Gateway through Open OnDemand

The ecosystem also supports an OpenOnDemand (OOD)[33] instance to explore the data in place without downloading. OOD is an intuitive, innovative, and interactive web-based interface to remote computing resources. OOD enables BIL users to create and run Jupyter notebooks[34] and create and run R scripts through RStudio. Example Jupyter notebooks for brain image analysis are available at BIL GitHub.

Results

Community Use

BIL currently contains about 7,000 datasets (approximately 80% of these datasets are from BICCN), contributed by over 268 data contributors from more than 45 unique affiliations from at least 6 different countries. This includes data from over 68 uniquely funded grants. Modalities of the data at BIL include anatomy, cell counting, cell morphology, connectivity, histology imaging, morphology, population imaging, receptor mapping, and spatial transcriptomics. Techniques used to acquire the data as well as species involved are outlined in Fig. 5.

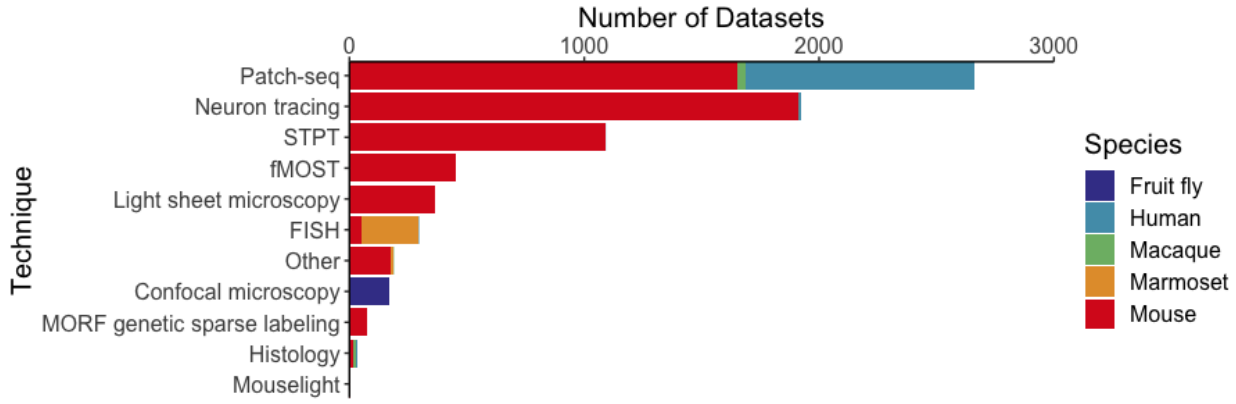

Fig. 5 Distribution of datasets contributed to BIL by technique and species.

User Support

BIL provides office hours regularly and for general inquiries, user support is also offered through an email helpdesk at bil-support@psc.edu. Typical questions received are about data usage, tools, or software - and for new data contributors, walkthroughs of the data submission process including metadata assistance. Training on the data submission process and using the data ecosystem is provided regularly. Data Submission workshops are half-day hands-on online sessions that outline all aspects of the data publication process including data submission, data structure, and metadata. The Data Ecosystem workshops involve tutorials for running analysis jobs at BIL and using tools such as SLURM, Open OnDemand, and Jupyter Notebooks to interact with datasets and images housed in the archive.

Visualization

The data stored within BIL is inherently visual. The advantages of visualizing and exploring BIL data are crucial to the scientific process, as they enable rapid validation of experiments, feedback on the appropriateness of feature extraction algorithms, and ultimately promote post hoc reuse by the scientific community. Scientists who seek to reuse BIL data aim to visually explore datasets to determine whether they meet their needs. For instance, they may investigate whether the data contain features relevant to their studies, such as particular cell types or brain structures. Alternatively, they may develop applications that enable the display, annotation, and exploration of information derived from the original imagery. Additionally, visualization is an excellent educational tool for teachers and students who wish to explore brain structure, compare and contrast imaging modalities, and understand the scientific process. The inherent beauty of microscopic datasets should not be overlooked as a motivating factor to propel students into science, technology, engineering, arts, and mathematics fields and compel citizen scientists to get involved in the process of discovery.

We are aware that the size and diversity of data at BIL can pose a stumbling block for data reuse. While small datasets may be downloaded and visualized locally, this approach is impractical for large multi-terabyte datasets. We address these concerns through the use of Napari and Neuroglancer viewers. Each viewer offers methods to visualize data over the internet

without the need to download whole datasets and, in the case of Neuroglancer, is easily accessible even on a smartphone.

Napari

Napari[25] is an open-source multi-dimensional image viewer that can lazily load large image stacks or multi-scale data. We developed a plugin for napari called napari-bil-data-viewer[35] that enables users to visualize publicly available datasets at BIL over the web without registration, making the data more easily accessible and reusable in line with FAIR[18] principles. The napari-bil-data-viewer can be installed on a low-end laptop, and the software requirements and technical expertise required are minimal. The current version of the plugin is pre-loaded with a comprehensive set of downsampled (summary) whole brain fMOST datasets (Fig 6.) and neuron morphology SWC[36] files associated with these brains. Additionally, arbitrary image stacks,

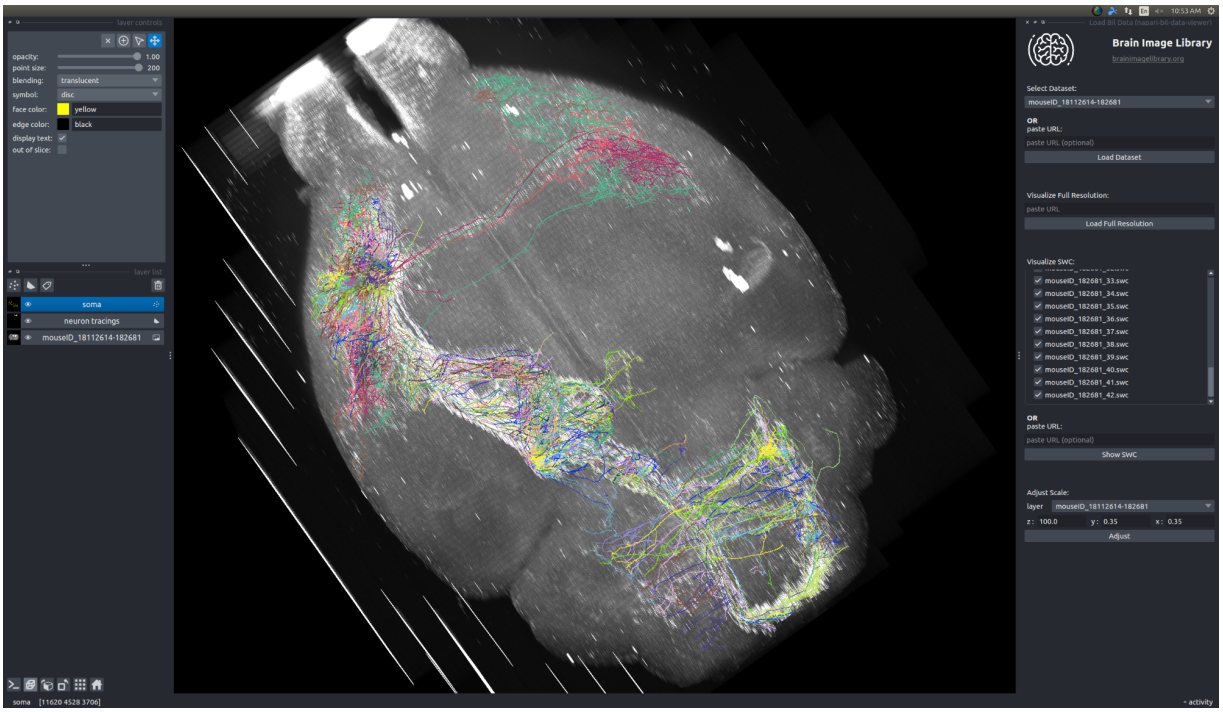

Fig. 6. Visualization of BIL data using the napari-bil-data-viewer plugin. The 3D view of an fMOST dataset with overlaid neuron morphologies extracted from the SWC files are shown along with the plugin interface. The neuron morphology dataset shown in this example is available at the BIL with the dataset ID web and the fMOST images are available with the dataset ID ace-ban-ear.

multi-scale data in OME-Zarr format, and neuron morphology files can be visualized by providing a URL to the data. The data loaded through the napari-bil-data-viewer is represented as standard napari layers, which allows use as an input to other plugins from the rich napari ecosystem, which includes tools for imaging, image analysis, and which cater to the needs of neuroscientists.

Neuroglancer

Neuroglancer[37] is a popular open-source 3-dimensional image viewer created and maintained by Google and the broader community. Neuroglancer is widely used by the neuroscience community, including the customizations[38,39] made by different labs meeting their specific needs. These needs include the ability to quickly and easily visualize cloud-based datasets, overlay annotations, as well as share specific views of the data (camera parameters, lookup table settings, zoom, etc). Neuroglancer scales well for visualizing petascale datasets. Several multiscale file formats are supported including the latest OME-Zarr standards. As discussed below, the data in other multiscale formats (for example, Imaris[40]) can be represented as OME-Zarr or the natively supported neuroglancer precomputed format on the fly, without duplication of data, which therefore allows BIL to make neuroglancer compatible with multiple image formats that would otherwise not be supported. Users can visualize arbitrarily large datasets even on a smartphone, at the speed of mobile internet, without registration or downloading data. Fig. 7 shows examples of visualizing BIL data with Neuroglancer.

Discussion

Solutions for FAIR image data

Public data resources focused on molecular biosciences have existed for over fifty years[41]. By contrast, the culture of collecting and sharing image data is relatively speaking in its infancy. Data has often been stored at individual laboratories and not made public. In the process of moving or transforming the data, the accompanying metadata can be lost, making data impossible to reuse or data analysis hard to reproduce. However, with the publication of FAIR

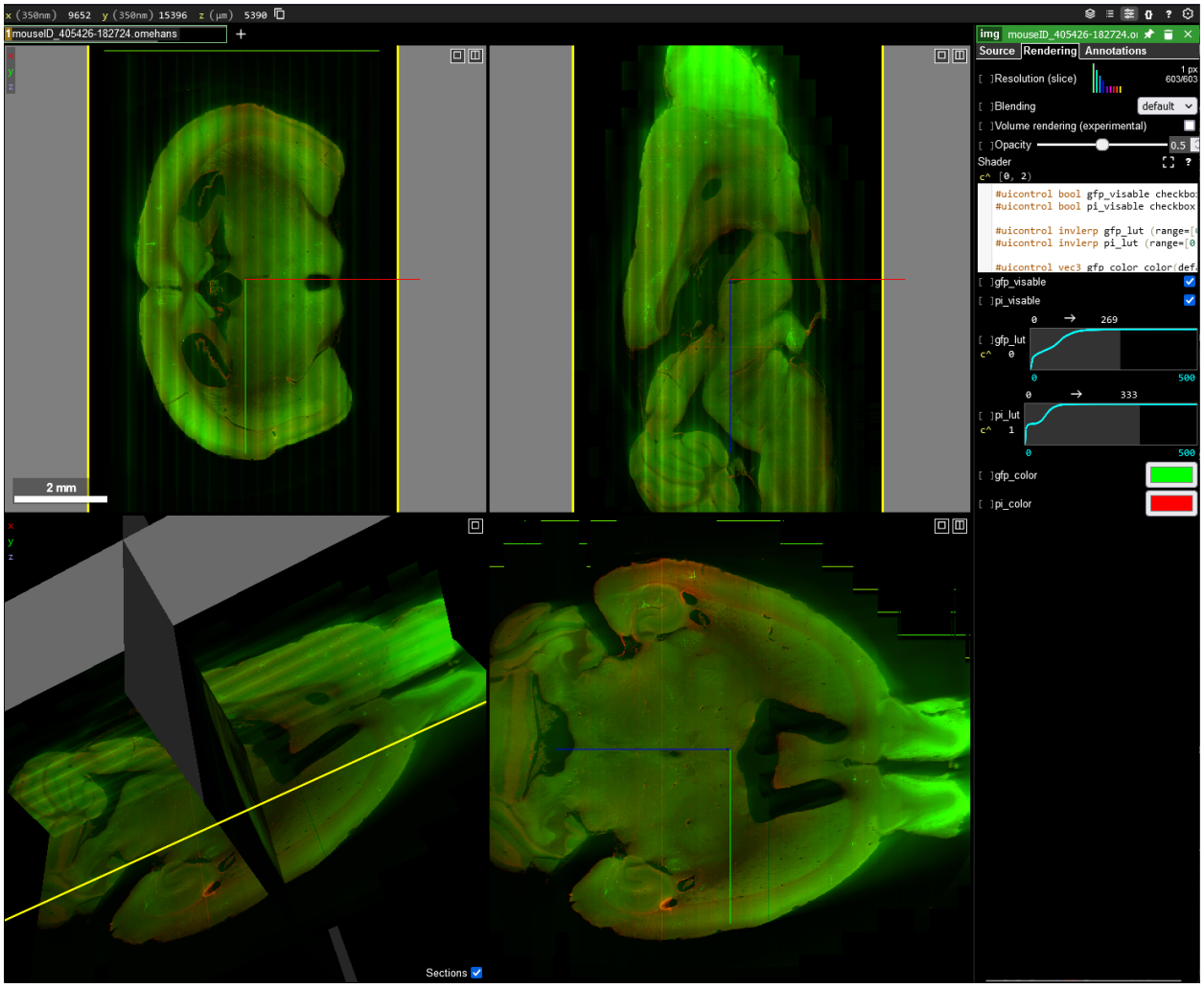

Fig. 7. Visualization of whole-brain fMOST using Neuroglancer. The dataset shown in this example is available at BIL with the dataset ID ace-bag-kit.

principles in 2016[18] came the increased emphasis by funders on preserving and making all types of research data publicly available. For example, the BRAIN Initiative's data-sharing policy became active in 2019 and includes all data being produced with BRAIN Initiative funding to be deposited in one of several designated archives[42]. In January 2023, NIH adopted a similar institution-wide data sharing policy[43].

We continuously work towards increasing the FAIRness of BIL data so it can be more useful to the research community and reused beyond its initial purpose. As new imaging modalities are developed and widely adopted, BIL will continue to work with its community of collaborators to expand and improve the metadata provided with new technologies. For example,

BIL will continue to work with BICAN members and subject matter experts to expand metadata to better describe spatial transcriptomics datasets as more of these datasets will be submitted in the next several years.

Beyond its primary objectives, BIL data can be used for developing novel analysis tools, benchmarking algorithms, improving visualization tools, teaching, and citizen science projects. The potential value of these data, especially if cross-referenced and integrated across projects, is enormous, and the first examples of this type of multi-modal, cross-scale imaging data integration now exist[44]. While this effort focused on mapping cell types in a mouse brain, the more recent BICAN consortium is focusing on mapping brain cell types in primates including humans, which will potentially allow for cross-species comparisons.

Challenges of data integration and data formats

Even if imaging data is organized according to FAIR principles, the experiments can produce a multitude of formats that present challenges for data interoperability and reuse. There are likely hundreds of imaging data formats currently in use including several standardized formats along with vendor-specific formats that do not conform to community-accepted standards. Researchers who wish to integrate data often must possess domain-specific knowledge of the imaging formats in which the data is stored, which can be a substantial barrier to reuse. Complicating this is that tools typically support a limited number of file formats[45], which will require that the data be available in the formats selected by the tool developer.

Community-driven efforts like the open microscopy environment (OME)[22] have developed excellent tools like Bio Formats[46] which has a spectrum of support for some 150 file formats. However, direct integration of Bio Formats into image processing applications is relatively rare and it is most often used to convert from non-standard to standard formats for data reuse. Data conversion is the traditional approach used to deal with incompatibility issues, and BIL does provide access to Bio-Formats. However, as hundreds of terabytes to petabyte-sized primate brain images start to be deposited in BIL, the traditional data conversion approach becomes extremely costly and impractical in terms of computational and storage resources required.

Transforming data formats on demand

To optimize the accessibility and promote the reuse of the diverse datasets at BIL, we have been developing an on-demand transformer. This transformer facilitates the efficient delivery of data residing on BIL file servers to end-users, either through local computing or over the internet in various formats. The motivation behind this effort stems from the vast amount of data available at BIL, the expected future data sizes, and the impracticality of converting and storing data in multiple formats. Thus, we aim to provide flexibility to the neuroscience community while maintaining storage efficiency within BIL by serving the data in useful ways.

The objective of this approach is twofold: visualization and data interaction. Through an API interface, arbitrary requests at the voxel level can be made, ranging from individual voxels to entire datasets, at any scale. Since data can be delivered in standardized formats, it is compatible with visualization tools like Napari or Neuroglancer as well as programmatic access to the computational infrastructures. For instance, it allows the extraction of specific pixel coordinates within multiple resolution versions of a multi-terabyte whole brain dataset, allowing for the extraction of regions of interest at different scales. Although still in development, this capability is currently being used to serve our neuroglancer data viewer described in the results section offering single-click access to hundreds of brain datasets. Data can currently be served as NumPy[47] arrays for Python-based image analysis applications and for visualization in OME-Zarr and neuroglancer pre-computed formats. We plan to offer a software development kit (SDK) for use remotely or locally on BIL infrastructure.

To enable this flexible delivery of the data, the underlying file formats must store large datasets to disk as many small chunks and at multiple scales. This enables efficient reading of targeted regions of the dataset at multiple scales and allows us to provide flexible dynamic delivery of image data. Several file formats offer these characteristics including over 80,000 Imaris files already publicly available at BIL. For future datasets, we are utilizing a multiscale chunked OME-Zarr-like file format that offers greater flexibility, improved parallel data access, and is becoming widely adopted by the bioimaging community[16,17]. Currently, we are in the process of converting legacy datasets into multi-scale OME-Zarr. Legacy datasets have been received by BIL in a variety of formats. Predominantly, volumetric imaging data have been

received in image stacks, usually TIFF, where each image file represents a z-plane of a single channel. An example includes fMOST datasets, many of which include tens of thousands of individual images comprising complete datasets as large as 25 terabytes if stored uncompressed. This method of representing volumetric data is relatively universal, as most computing systems and image analysis tools can open individual image files. However, these datasets only represent a single scale, and pixel-level access is highly inefficient due to the way data is written to disk. Thus conversion to a multiscale OME-Zarr data store makes sense to facilitate visualization and computation.

Enabling flexible efficient data access is critical to the future of BIL. For example, in the next year, we anticipate receiving several human whole-brain microscopy datasets at cellular resolution, with each dataset potentially reaching three petabytes. The ability to efficiently compute on and serve these datasets in multiple formats becomes crucial to avoid data duplication. For datasets that have not yet been submitted to BIL, we are encouraging data providers to submit datasets that can be easily converted to OME-Zarr or provide data in a format that is compatible with our on-demand transformer for direct access in OME-Zarr. More information on the on-demand transformer and our plan to serve OME-Zarr dynamically at BIL can be found in the paper describing community adoption of OME-Zarr[17]. We plan to detail the capabilities of the on-demand transformer in a separate publication.

Integration of resources to promote FAIRness

As we build the interactive visualization and analysis tools and make more datasets available, it becomes more important to improve the information content of metadata and link datasets with outside resources such as ontological definitions, relevant anatomical atlases, and related data housed in other repositories. This would help to make routine connections between related data, for example, 3D MRI of the brain, 2D histology sections, and single-cell atlases, more straightforward than what is possible today.

The straightforward way to accomplish this is through verbose metadata using common ontologies, which can be searched for desired properties and provide connections between datasets and high-level annotations. Complicating harmonization is that (i) each resource has evolved metadata schemas independently that serve their specific charters, thus there is no

agreed-upon metadata standard that describes one-to-one relationships between resources, and (ii) annotation is an evolving target as cell types and atlases are refined and more fundamental knowledge is acquired.

As BIL is a neuroscience resource, we will be focusing our efforts on linking our data with other BRAIN Initiative resources and archives along with anatomical atlases produced by the community[45]. We plan to systematically enhance our metadata with both automated and user-contributed annotations which can be indexed and queried using an API in a reasonable time. Ultimately, ongoing work in the directions discussed above will further facilitate collaboration, cross-archive integration, data access, and reuse, making BIL even more useful for data contributors and users.

Overall, BIL is a unique microscopy data resource that provides the neuroscience community with an extensive collection of brain images representing a variety of microscopy-enabled experiments along with rich metadata to facilitate data reuse. BIL also provides a platform for analyzing and sharing imaging data that would otherwise be difficult to access. For data generators, BIL provides a vehicle for preserving valuable brain data being generated by the research community. To facilitate data reuse, BIL provides data users with a visualization and analysis ecosystem that enables data to be visualized over the web and analyzed in place on HPC systems. These features are already essential but will be even more important as the collection of microscope-enabled experiments of human brain data grows. BIL is well-positioned to accept even larger and more complex data as the neuroscience community continues to innovate.

Data availability

All public BIL data is available and searchable at https://brainimagelibrary.org, RRID: SCR_017272.

Code availability

All software products for BIL are developed in the open and stored at our GitHub repository https://github.com/brain-image-library and https://github.com/CBI-PITT. Jupyter notebooks for analyzing BIL data presented at workshops are available at

https://github.com/brain-image-library/workshops. Source code for napari-bil-data-viewer is at https://github.com/brain-image-library/napari-bil-data-viewer.

Funding

The Brain Image Library is supported by the National Institutes of Mental Health of the National Institutes of Health under award number R24-MH-114793. The napari visualization plugin is funded by the Chan Zuckerberg Initiative Donor Advised Fund grant #DAF2022-309651. This work utilizes Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it uses the Bridges system, which was supported by NSF award number ACI-1445606, and the Bridges-2 system, which is supported by NSF award number ACI-1928147. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies mentioned.

References

1. Insel, T. R., Landis, S. C. & Collins, F. S. The NIH BRAIN Initiative. *Science* **340**, 687–688 (2013).
2. Brown, S. T. *et al.* Bridges-2: A Platform for Rapidly-Evolving and Data Intensive Research. in *Practice and Experience in Advanced Research Computing* 1–4 (ACM, 2021). doi:10.1145/3437359.3465593.
3. Ragan, T. *et al.* Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nat. Methods* **9**, 255–258 (2012).
4. Gong, H. *et al.* Continuously tracing brain-wide long-distance axonal projections in mice at a one-micron voxel resolution. *NeuroImage* **74**, 87–98 (2013).
5. BICCN Data Catalog Glossary https://dashboard.brain-bican.org/glossary_view/biccn/.
6. Williams, E. *et al.* Image Data Resource: a bioimage data integration and publication platform. *Nat. Methods* **14**, 775–781 (2017).
7. EBRAINS https://ebrains.eu.
8. Masinas, M. P. D., Usaj, M. M., Usaj, M., Boone, C. & Andrews, B. J. TheCellVision.org: A Database for Visualizing and Mining High-Content Cell Imaging Projects. *G3 GenesGenomesGenetics* **10**, 3969–3976 (2020).
9. The Cell Image Library https://cellimagelibrary.org.
10. Hartley, M. *et al.* The BioImage Archive – Building a Home for Life-Sciences Microscopy Data. *J. Mol. Biol.* **434**, 167505 (2022).
11. Ament, S. A. *et al.* The Neuroscience Multi-Omic Archive: a BRAIN Initiative resource for single-cell transcriptomic and epigenomic data from the mammalian brain. *Nucleic Acids Res.* **51**, D1075–D1085 (2023).
12. Duncan, D. *et al.* Data Archive for the BRAIN Initiative (DABI). *Sci. Data* **10**, 83 (2023).
13. Markiewicz, C. J. *et al.* The OpenNeuro resource for sharing of neuroscience data. *eLife* **10**, e71774 (2021).

14. Hider, R. *et al.* The Brain Observatory Storage Service and Database (BossDB): A Cloud-Native Approach for Petascale Neuroscience Discovery. *Front. Neuroinformatics* **16**, 828787 (2022).
15. Delorme, A. *et al.* NEMAR: an open access data, tools and compute resource operating on neuroelectromagnetic data. *Database* **2022**, baac096 (2022).
16. Moore, J. *et al.* OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies. *Nat. Methods* **18**, 1496–1498 (2021).
17. Moore, J. *et al.* OME-Zarr: a cloud-optimized bioimaging file format with international community support. *Histochem. Cell Biol.* **160**, 223–251 (2023).
18. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
19. Ropelewski, A. J. *et al.* Standard metadata for 3D microscopy. *Sci. Data* **9**, 449 (2022).
20. International Neuroinformatics Coordinating Facility https://www.incf.org/sbp/brain.
21. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4. 82 pages (2021) doi:10.14454/3W3Z-SA82.
22. Goldberg, I. G. *et al.* The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biol.* **6**, R47 (2005).
23. Benninger, K. *et al.* Cyberinfrastructure of a Multi-Petabyte Microscopy Resource for Neuroscience Research. in *Practice and Experience in Advanced Research Computing* 1–7 (ACM, 2020). doi:10.1145/3311790.3396653.
24. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
25. Ahlers, J. *et al.* napari: a multi-dimensional image viewer for Python. (2023) doi:10.5281/ZENODO.8115575.
26. Peng, H., Ruan, Z., Long, F., Simpson, J. H. & Myers, E. W. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nat. Biotechnol.* **28**, 348–353 (2010).
27. Buitrago, P. A., Uran, J. A. & Nystrom, N. A. System Integration of Neocortex, a Unique, Scalable AI Platform. in *Practice and Experience in Advanced Research Computing* 1–4 (ACM, 2021). doi:10.1145/3437359.3465604.
28. McLay, R., Schulz, K. W., Barth, W. L. & Minyard, T. Best practices for the deployment and management of production HPC clusters. in *State of the Practice Reports* 1–11 (ACM, 2011). doi:10.1145/2063348.2063360.
29. Yoo, A. B., Jette, M. A. & Grondona, M. SLURM: Simple Linux Utility for Resource Management. in *Job Scheduling Strategies for Parallel Processing* (eds. Feitelson, D., Rudolph, L. & Schwiegelshohn, U.) vol. 2862 44–60 (Springer Berlin Heidelberg, 2003).
30. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLOS ONE* **12**, e0177459 (2017).
31. Crusoe, M. R. *et al.* Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language. (2021) doi:10.48550/ARXIV.2105.07028.
32. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33 (2021).
33. Open OnDemand https://openondemand.org/.
34. Kluyver, T. *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* 87–90 (IOS Press, 2016). doi:10.3233/978-1-61499-649-1-87.
35. Napari-bil-data-viewer https://github.com/brain-image-library/napari-bil-data-viewer/.
36. SWC specification https://github.com/BICCN/swc_specification.
37. Maitin-Shepard, J. *et al.* google/neuroglancer: (2021) doi:10.5281/ZENODO.5573294.
38. Neurodata Cloud https://neurodata.io/help/visualization/.
39. MICrONS Explorer https://www.microns-explorer.org/.
40. Imaris file format https://imaris.oxinst.com/support/imaris-file-format.

41. Strasser, B. J. Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965. *J. Hist. Biol.* **43**, 623–660 (2010).
42. Notice of Data Sharing Policy for the BRAIN Initiative https://grants.nih.gov/grants/guide/notice-files/NOT-MH-19-010.html.
43. Final NIH Policy for Data Management and Sharing https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.
44. BRAIN Initiative Cell Census Network (BICCN) *et al.* A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).
45. Tyson, A. L. & Margrie, T. W. Mesoscale microscopy and image analysis tools for understanding the brain. *Prog. Biophys. Mol. Biol.* **168**, 81–93 (2022).
46. Linkert, M. *et al.* Metadata matters: access to image data in the real world. *J. Cell Biol.* **189**, 777–782 (2010).
47. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

Acknowledgments

The authors would like to thank the BIL development team (Kathy Benninger, Derek Simmel, Art Wetzel, Elizabeth Pantalone), Pittsburgh Supercomputing Centers Advanced Systems and Operations Group, the BIL Advisory Board (Jason Swedlow, Lydia Ng, Yongsoo Kim, Matt McCormick), Jesse F. Prentiss, Adam Tyson, BICCN and BICAN collaborators and our many data contributors.

Author Contributions

Mariah Kenney: Writing, Editing, Reviewing, Data Curation
Iaroslavna Vasylieva: Writing, Editing, Reviewing, Visualization
Greg Hood: Writing, Editing, Reviewing
Ivan Cao-Berg: Software, Writing, Editing, Reviewing
Luke Tuite: Software, Writing, Editing, Reviewing
Rozita Laghaei: Writing, Editing, Reviewing
Alan M. Watson: Conceptualization, Writing, Editing, Reviewing, Funding Acquisition, Visualization, Methodology
Alexander J. Ropelewski: Conceptualization, Supervision, Writing, Editing, Reviewing, Funding Acquisition

Competing interests

The authors declare no competing interests.