

# ***brainlife.io*: A decentralized and open source cloud platform to support neuroscience research**

Hayashi, S.\*, Caron, B.\*, Heinsfeld, A.S., Vinci-Booher, S., McPherson, B.C., Bullock, D., Berto, G., Niso, J.G., Hanekamp, S., Levitas, D., Kitchell, L., Leong, J., Silva, F. N., Koudoro, S., Willis, H., Jolly, J., Pisner, D., Zuidema, T., Kurzwaski, J., Mikellidou, K., Bussalib, A., Rorden, C., Victory, C., Bhatia, D., Aydogan, D.B., Yeh, F.C., Delogu, F., Guaje, J., Veraart, J., Fischer, J., Faskowitz, J., Chaumon, M., Fabrega, R., Hunt, D., McKee, S., Brown, S.T., Heyman, S., Iacovella, V., Mejia, A., Marinazzo, D., Craddock, C., Olivetti, E., Hanson, J., Avesani, P., Garyfallidis, E., Stanzione, D., Carson, J.P., Henschel, R., Hancock, D.Y., Stewart, C.A., Schnyer, D., Eke, D., Poldrack, R.A., George, N., Bridge, H., Sani, I., Freiwald, W., Puce, A., Port, N., and Pestilli, F.

**Competing interests.** The authors declare no competing financial interests.

**Corresponding authors.** Franco Pestilli [pestilli@utexas.edu](mailto:pestilli@utexas.edu)

**Contribution.** S.H. implemented the *brainlife.io* services. B.C. wrote the data analysis code, performed large-scale experiments, and prepared the figures and associated text. A.S.H. improved and implemented some of the services. F.J., C.C., C.C.A., D.H., D.S., D.P., L.K., J.L., C.R., F.N.S., H.W., J.J., Z.T., K.J., S.K., N.A., V.C., B.D., A.D.B., F.D. G.J., S.H., provided assets. All authors edited the manuscript. F.P. invented, designed, and directed *brainlife.io*, wrote the paper, and designed all the experiments, and figures. \*Shared first author's contribution.

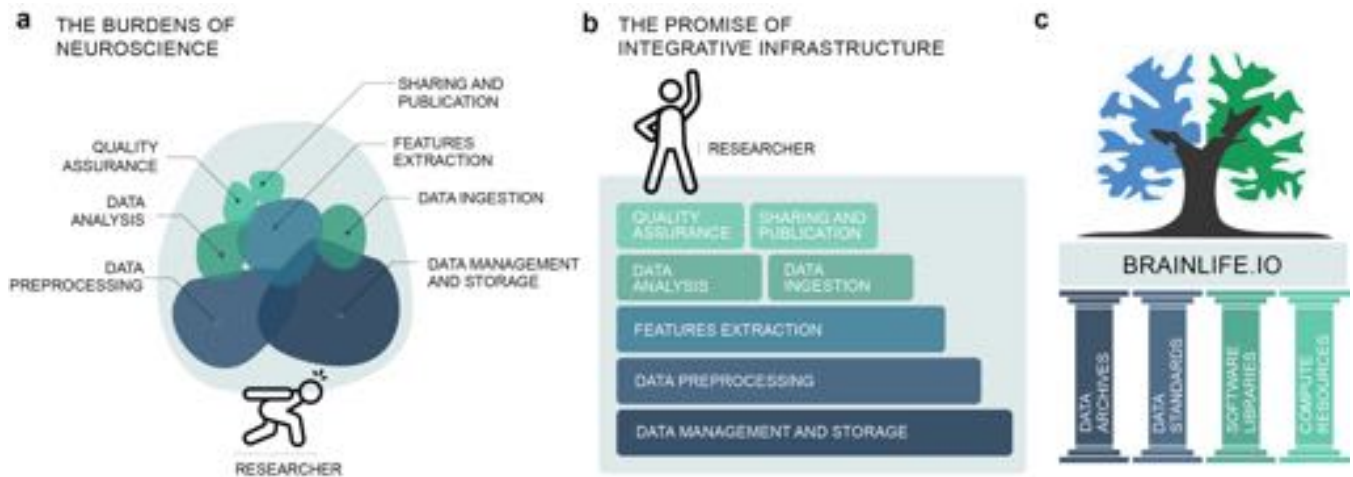
## **ABSTRACT**

Neuroscience research has expanded dramatically over the past 30 years by advancing standardization and tool development to support rigor and transparency. Consequently, the complexity of the data pipeline has also increased, hindering access to FAIR data analysis to portions of the worldwide research community. *brainlife.io* was developed to reduce these burdens and democratize modern neuroscience research across institutions and career levels. Using community software and hardware infrastructure, the platform provides open-source data standardization, management, visualization, and processing and simplifies the data pipeline. *brainlife.io* automatically tracks the provenance history of thousands of data objects, supporting simplicity, efficiency, and transparency in neuroscience research. Here *brainlife.io*'s technology and data services are described and evaluated for validity, reliability, reproducibility, replicability, and scientific utility. Using data from 4 modalities and 3,200 participants, we demonstrate that *brainlife.io*'s services produce outputs that adhere to best practices in modern neuroscience research.

|   |           |
|---|-----------|
| <b>ABSTRACT (149/150 words)</b>                       | <b>1</b>  |
| <b>INTRODUCTION</b>                                   | <b>3</b>  |
| <b>RESULTS</b>  | <b>4</b>  |
| <b>Platform architecture</b>                          | <b>4</b>  |
| <b>Platform evaluation</b>                            | <b>7</b>  |
| Platform utilization                                  | 7         |
| Platform testing                                      | 7         |
| Platform utility for scientific applications          | 9         |
| Replication and generalization of previous results    | 10        |
| Example applications to detecting disease             | 11        |
| A new approach to facilitate quality control at scale | 11        |
| <b>DISCUSSION</b>                                     | <b>12</b> |
| <b>ONLINE METHODS AND MATERIALS</b>                   | <b>15</b> |
| <b>Acknowledgments.</b>                               | <b>19</b> |
| <b>REFERENCES</b>                                     | <b>20</b> |

## INTRODUCTION

Over the last 30 years, neuroimaging research has dramatically expanded our ability to study the structure and function of the living human brain, leading to major advancements in understanding brain-related health and disease<sup>1-4</sup>. Today, neuroimaging modalities and techniques span multiple data types (e.g., magnetic resonance imaging [MRI], positron emission tomography [PET], functional near-infrared spectroscopy [fNIRS], electro-encephalography [EEG], and magnetoencephalography [MEG]), and have increased the feasibility of large-scale, population-level, data collection efforts.<sup>1,5,6</sup> At the same time, the field of neuroimaging has attracted a large and ever-growing community of researchers<sup>7,8</sup>. Furthermore, a process of adopting FAIR principles of data stewardship (Findability, Accessibility, Interoperability, and Reusability<sup>9</sup>), data standardization, open science methods, and increased data size, has been gaining grounds and in turns increasing requirements for rigorous and transparent data analysis and reporting. However, such approaches require significant additional technological support, posing new challenges to many researchers. We refer to these challenges as the burdens of neuroscience (Fig. 1).



**Figure 1. The burdens of neuroscience.** **a.** A figurative representation of the current major burdens of performing neuroimaging investigations. **b.** Our proposal for integrative infrastructure that coordinates services required to perform FAIR, reproducible, rigorous, and transparent neuroimaging research thereby lifting the burden from the researcher. **c.** *brainlife.io* rests upon the foundational pillars of the open science community such as data archives, standards, software libraries and compute resources. Panels **a** and **b** adapted from *Eke et al. (2021)*.

Datasets are growing in size, in large part because they support scientific rigor and reproducibility. Research on the reproducibility of scientific findings indicates that limited sample sizes might have hindered the validity of early, foundational results in hypothesis-driven cognitive neuroscience research,<sup>10-16</sup> but reproducibility issues can be found in biological science,<sup>17,18</sup> psychology,<sup>12</sup> data science, and computational methods,<sup>19,20</sup> cancer biology,<sup>21</sup> and artificial intelligence.<sup>13,22,23</sup> This is largely because small sample sizes increase the probability of reporting spurious effects as statistically significant.<sup>1,24</sup> Recent findings also make the case for increasing sample sizes into the thousands when research focuses on discovery science.<sup>5</sup> Notable examples of large-scale data sharing within neuroscience and neuroimaging include the Human Connectome Project (HCP),<sup>25</sup> the Cambridge Centre for Ageing and Neuroscience study (Cam-CAN),<sup>26,27</sup> the Adolescent Brain Cognitive Development (ABCD) study,<sup>28,29</sup> the UK-Biobank,<sup>30</sup> the Healthy Brain Network (HBN),<sup>31</sup> the Pediatric Imaging Neurocognition and Genetics (PING) study,<sup>32</sup> the Natural Scene Dataset<sup>33</sup> and the thousands of individual brain datasets deposited on OpenNeuro.org.<sup>34</sup> These data-sharing projects not only serve the needs of the neuroscience community with demonstrated impact<sup>35</sup>, but also the incoming generation of AI research.<sup>36-38</sup> However, larger datasets generally entail greater complexity as well. The use of datasets so unprecedented in size requires a substantial scaling up of resources and technical skills, and this in turn results in significant barriers to entry.

Traditionally, neuroimaging researchers have collected a few hours of neuroimaging data on a few dozen subjects and analyzed it using laboratory computers and a single tool-kit or programming environment, often created in-house. Current studies, by contrast, may require the analysis of hundreds (if not thousands) of hours of data,

with an accompanying move of data away from individual laboratory computers toward high-performance computing clusters and cloud systems requiring multiple steps and a variety of scripting and programming languages (e.g., Unix/Linux shell, Python, MatLab, R, C++). The complexity of neuroimaging data pipelines and code development stacks have increased concomitantly.<sup>39,40</sup> To help ensure the reproducibility and rigor of scientific results, the neuroimaging community has also developed data standards<sup>41</sup> and software libraries for data processing and analysis (FSL, Freesurfer, Nibabel, MRTrix, DIPY, DSI-STudio).<sup>42-68</sup> More recently prebuilt data processing pipelines that combine software from multiple libraries into unified partially preconfigured steps have been also developed<sup>69-73</sup>. These pipelines advance data processing standardization but still leave many choices of parameters to users and often require technical input data formats.

As a result of all this progress for data and tools, neuroimaging researchers carry the burden of having to piece together and track multiple processes, such as data ingestion and standardization, storage, and management, preprocessing and feature extraction, all while also attending to tracking quality control, analyses, and publication (**Fig. 1a**). Publication of results requires compliance with the FAIR principles which, though well explained in theory, are often challenging to implement in practice. Submission of manuscripts often necessitates new analyses at a later date, by which point software and data versions may have changed, and data might have been removed from compute clusters or local servers. Existing approaches for managing these steps require manual tracking of data and code versions, along with advanced technical skills.<sup>40,74</sup> Currently, there exists no efficient technology to help piece together and keep track of all of these (ever-changing) technology and data requirements.

As the resources necessary to participate fully in modern neuroscience research have grown, barriers to entry and funding have risen as well. Smaller universities, teaching colleges, undergraduate students, and other settings that lack the resources to support significant investments in infrastructure and training are at a meaningful disadvantage. Lack of resources and infrastructure is a key gap identified in surveys pertaining to both the adoption of FAIR neuroscience<sup>75</sup> and the conduct of neuroscience research in low- and medium-income countries<sup>76,77</sup>. Without added support, FAIR neuroscience might evolve with an ever-increasing bias towards high-resourced teams, institutions, and countries. Such an outcome would not only decrease representation and diversity but would slow scientific progress. In support of simplicity, efficiency, transparency, and equity in big data neuroscience research, our team has developed a community resource, *brainlife.io* (**Fig. 1b**). The *brainlife.io* platform stands on the foundational pillars of the neuroimaging community and the mission of open science (**Fig. 1c**). *brainlife.io* provides free and secure reproducible neuroscience data analysis. *brainlife.io*'s technology works for researchers serving automated tracking of data provenance, preprocessing steps, parameter sets, and analysis versions. Our vision for *brainlife.io* is that of a trusted, interoperable, and integrative platform connecting global communities of software developers, hardware providers, and domain scientists via cloud services.

In the remainder of this article, we describe the technology and utilization of *brainlife.io*. After that, we present the results of our evaluations of the effectiveness of the technology. Experiments focused on the four axes of scientific transparency: external validity, reliability, reproducibility, and replicability. Finally, we demonstrate the platform's potential for scientific utility in identifying human disease biomarkers.

## RESULTS

### Platform architecture

*brainlife.io* is a ready-to-use and ready-to-expand platform. As a ready-to-use system, it allows researchers to upload and analyze data from MRI, MEG, and EEG systems. Data are managed using a secure warehousing system that follows an advanced governance and access-control model. Data can be preprocessed and visualized using version-controlled applications (hereafter referred to as Apps) compliant with major data standards (the Brain Imaging Data Structure, BIDS<sup>41</sup>). As a ready-to-expand system, software developers may contribute or modify existing Apps guided by standard methods and documentation describing how to write Apps ([github.com/brainlife/abcd-spec](https://github.com/brainlife/abcd-spec) and [brainlife.io/docs](https://brainlife.io/docs)). The platform uses a combination of opportunistic computing and publicly funded resources<sup>78-80</sup> that are functionally integrated and can be available for use by a particular project or team of researchers. Computing resource managers can also register computer servers and clusters on *brainlife.io* to make them available either to individual users or projects or to the larger community of *brainlife.io* users (**Fig. 2a** and **Fig. S2a**). The **Supplemental Platform architecture** provides an extended

description of the technology. The platform is available to any type of researcher from students to faculty researchers, either without cost (through opportunistic use of freely contributed resources) or with performance guarantees (through the use of dedicated hardware or payment for use of cloud resources).

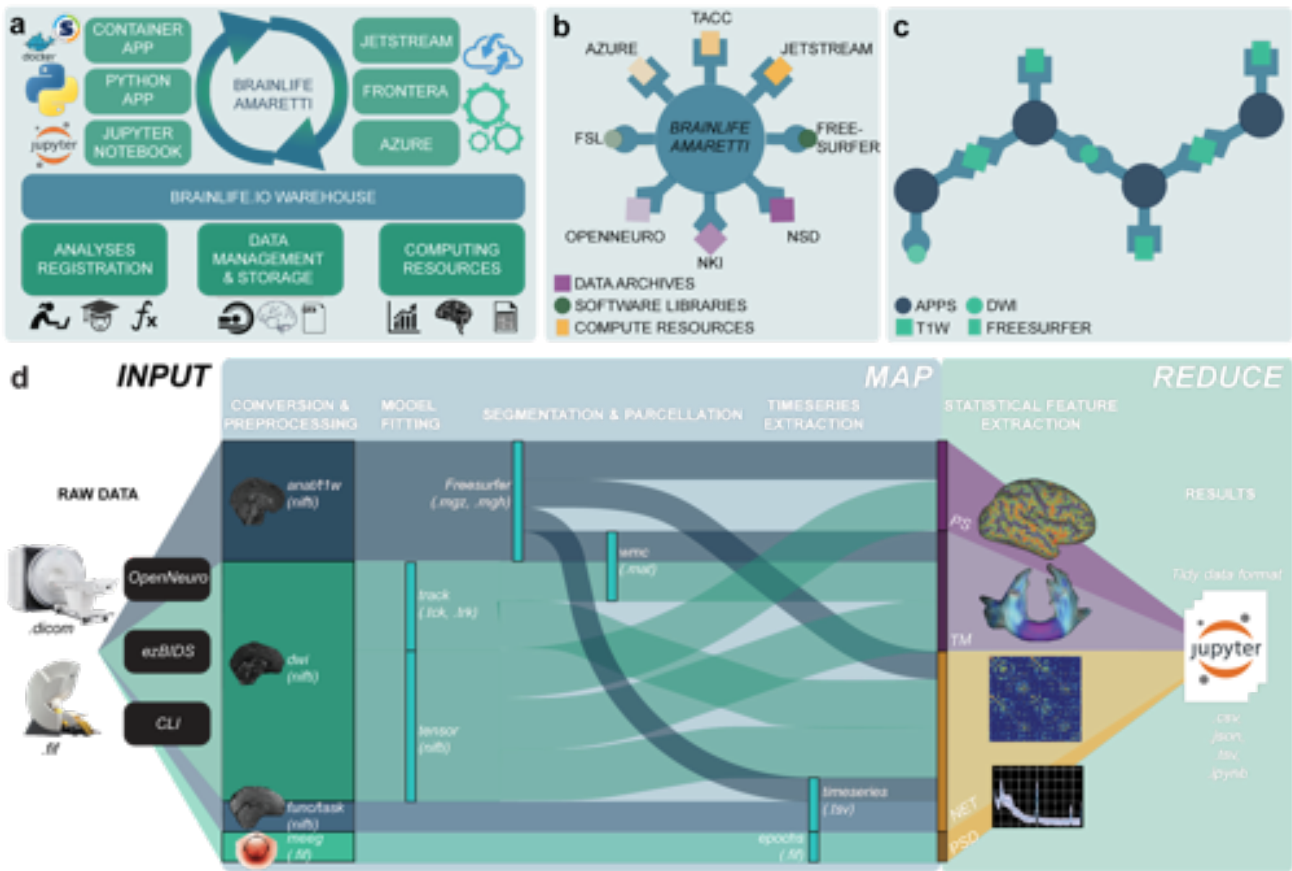
Brainlife.io was founded via an initial investment from the U.S. BRAIN Initiative via a National Science Foundation, followed by support from the National Institutes of Health, the Department of Defense, the Kavli Foundation, and the Wellcome Trust. The platform's geographically distributed computing and storage systems are securely hosted by national supercomputing centers and funded by a combination of institutional, national, and international awards (see [Fig. S2](#)). As of this paper, the Texas Advanced Computing Center, Indiana University Pervasive Technology Institute, Pittsburgh Supercomputing Center, San Diego Supercomputing Center, and the University of Michigan Advanced Research Computing Technology Services have supported the project. The distributed platform is connected with and depends on other major infrastructure and software projects such as OpenNeuro.org, osris.org, DataLad.org, BIDS, Freesurfer, FSL, nibabel, dipy.org, repronim.org, DSI-Studio, jetstream-cloud.org, frontera-portal.tacc.utexas.edu, access-ci.org, and INCF.org.

The architecture of *brainlife.io* is based on an innovative, microservices-based approach, including authentication, preprocessing, warehousing, event handling, and auditing. This architecture allows automated and decentralized data management and processing. Microservices are handled by the meta-orchestration workflow system Amaretti ([Fig. 2a,b](#), and [Table S1](#)). Amaretti can deploy computational jobs on high-performance compute clusters and cloud systems. This allows the utilization of publicly-funded supercomputers and clouds<sup>80</sup>, as well as commercial clouds, such as Google Cloud, AWS, or Microsoft Azure.

Data management on *brainlife.io* is centered around Projects and supported by a databasing and warehousing system ([github.com/brainlife/warehouse](https://github.com/brainlife/warehouse)). Projects are the “one-stop-shop” for data management, processing, analysis, visualization, and publication ([Fig. S3c](#)). Projects are created by users and are private by default, but can also be made publicly visible inside the *brainlife.io* platform. A project can be populated with data using several options ([Fig. 2d](#)). Several major archives and data repositories are currently docked by *brainlife.io*<sup>74</sup> (see [Fig. 2b](#)). Noticeable examples are OpenNeuro.org<sup>34</sup> and the Nathan-Kline data-sharing project.<sup>81-83</sup> Datasets can be imported seamlessly into *brainlife.io* Projects by using either the portal [brainlife.io/datasets](https://brainlife.io/datasets)<sup>74</sup> (see [Video S2](#) and [Video S3](#)), the standardization tool [brainlife.io/ezbids](https://brainlife.io/ezbids) (see [Table S1](#) and [Video S6](#)) or a dedicated Command Line Interface (CLI).

Data processing on *brainlife.io* utilizes an object-oriented and micro workflows service model. Data objects are stored using predefined formats, Datatypes, that allow automated App concatenation and pipelining ([Fig. 2c](#); [brainlife.io/Datatypes](https://brainlife.io/Datatypes)). Apps and Datatypes are the key components of a system that work together to allow automated processing and provenance tracking for millions of data objects. Apps are composable processing units written in a variety of languages using containerization technology.<sup>84,85</sup> Apps are smart, and can automatically identify, accept, or reject datasets before processing ([Fig. 2](#) and [Fig. S2b](#)). Community-developed data visualizers are served by *brainlife.io* to support quality control (see [Table S1](#)). Six new data visualizers have been developed and released as part of the project ([Table S1](#) and [Video S7](#)). Whenever possible, Datatypes are made compatible with BIDS.<sup>41</sup> BIDS Apps can be easily made into *brainlife.io* Apps and multiple examples exist already [brainlife.io/apps](https://brainlife.io/apps).

The data workflow on *brainlife.io* simplifies the complexity of the modern neuroimaging processing pipeline into two steps, akin to Google's MapReduce algorithm.<sup>86</sup> An initial *map step* preprocesses data objects asynchronously and in parallel using Apps, so as to extract features of interest (such as functional activations, white matter maps, brain networks, or time series data; [Fig. 2d](#)). During the *map step*, Datatypes and Apps are synchronized and moved to available compute resources automatically. Apps process data objects in parallel across study participants in a Project. The *map step* is followed by a *reduce step*, wherein features extracted using Apps are made available to pre-configured Jupyter notebooks<sup>87,88</sup> served on the platform to perform statistical analysis, machine-learning applications, and generate figures. Indeed, all statistical analyses and figures in this paper are available in accessible *Jupyter Notebooks* (see [Table S2](#)). *brainlife.io*'s data workflow makes it possible to integrate large volumes of diverse neuroimaging Datatypes into simpler sets of brain features organized into *Tidy data* structures<sup>89</sup> ([Fig. S3c](#)).



**Figure 2. The *brainlife.io* platform concepts, architecture, and approach.** **a.** *brainlife.io*'s Amaretti links data archives, software libraries, and computing resources. Specifically, 'Apps' (containerized services defined on GitHub.com) are automatically matched with data stored in the 'Warehouse' with computing resources. Statistical analyses can be implemented using Jupyter Notebooks. **b.** *brainlife.io* provides efficient docking between data archives, processing apps, and compute resources via a centralized service. **c.** Apps use standardized Datatypes and allow "smart docking" only with compatible data objects. App outputs can be docked by other Apps for further processing. **d.** *brainlife.io*'s Map step takes MRI, MEG and EEG data and processes them to extract statistical features of interest. *brainlife.io*'s reduce step takes the extracted features and serves them to Jupyter Notebooks for statistical analysis. PS: parc-stats Datatype; TM: tractmeasures Datatype; NET: network Datatype; PSD: power-spectrum density Datatype. CLI: Common Line Interface.

A key technological innovation developed for *brainlife.io* is the ability to automatically track all actions performed by platform users on Datatypes and Apps. The platform captures data object IDs, Apps versions, and parameter sets so as to track the full sequence of steps from data import to analysis and publication. A graph describing provenance metadata for each Datatype can be visualized using the provenance visualizer or downloaded (see [Fig. S3d](#) and [Video S10](#)). A shell script is automatically generated to allow the reproduction of full processing sequences ([Video S11](#)). Finally, a single record containing data objects, Apps, and *Jupyter Notebooks* used in a study can be made publicly available outside the platform bundled into a single record addressed by a unique Digital Objects Identifier (DOI)<sup>90</sup>. Whereas all other existing systems provide users with technology to track analysis steps manually or require the use of coding, *brainlife.io* tracks automatically and do not require coding nor user actions to generate a record of everything done by a user for data analysis. This automation technology lowers the barriers of entry and democratizes FAIR, reproducible large-scale neuroimaging data analysis.

## Platform evaluation

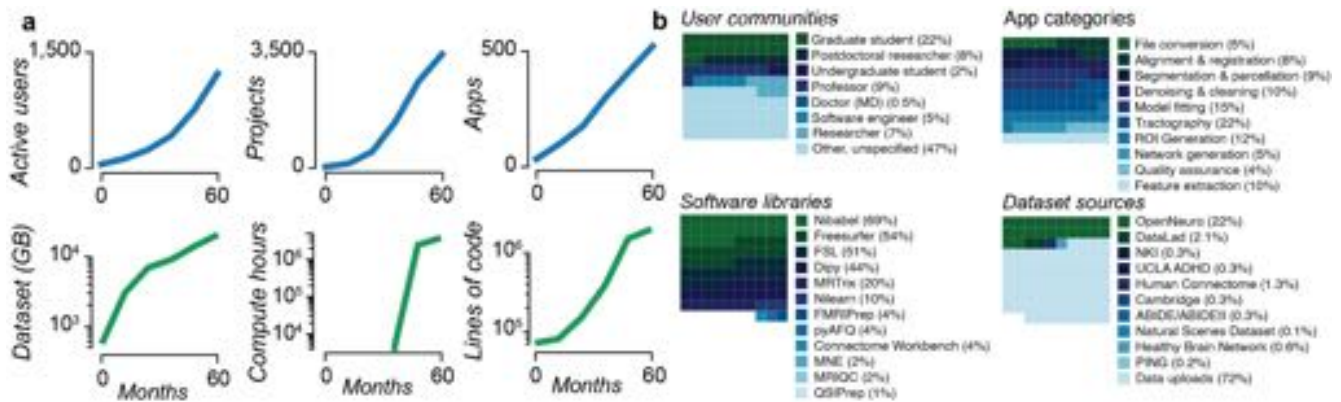
In the following section, we evaluate the utility of *brainlife.io*. To do so, we first present the level of engagement with the platform by the growing community of users. After that, we describe the results of experiments on the

robustness and validity of the platform. A detailed description of each section below describing each App and step used can be found in the corresponding [Supplemental Platform evaluation](#) section.

### Platform utilization

*brainlife.io* is developed following the FAIR principles. It is available worldwide and supports thousands of researchers. First made accessible in Spring 2018, its utilization and assets have grown steadily (**Fig. 3** and [Fig. S2c](#) and [S4](#)). At the time of this writing, over 2,341 users across 43 countries have created a *brainlife.io* account. Over 1,542 of these have been active users (**Fig. 3a**). Over 3,439 data management Projects have been created, and a community of developers has implemented over 530 data processing Apps. Over 270 TBs of data have been stored and processed using *brainlife.io*, for a total of 1,097,603 hours of compute time.

Researchers ranging from undergraduate students to faculty use *brainlife.io* (**Fig. 3b**), and analyses span the full range of the neuroimaging data lifecycle. The most frequently used Apps pertained to diffusion tractography (22%), model fitting (15%), and anatomical ROI generation (12%). Community-developed software libraries provided the foundations for data processing, including Nibabel, Freesurfer, FSL, DIPY, MRtrix, the Connectome Workbench, and MNE-Python. Terabytes of data have been uploaded (72%) or imported from OpenNeuro.org (22%), the Nathan-Kline Institute data sharing projects (3%<sup>31,81,83</sup>), and other sources. This degree of world-wide platform access highlights the global need for technology like *brainlife.io* (see [Fig. S2e](#)). More details can be found in [Supplemental platform utilization](#).



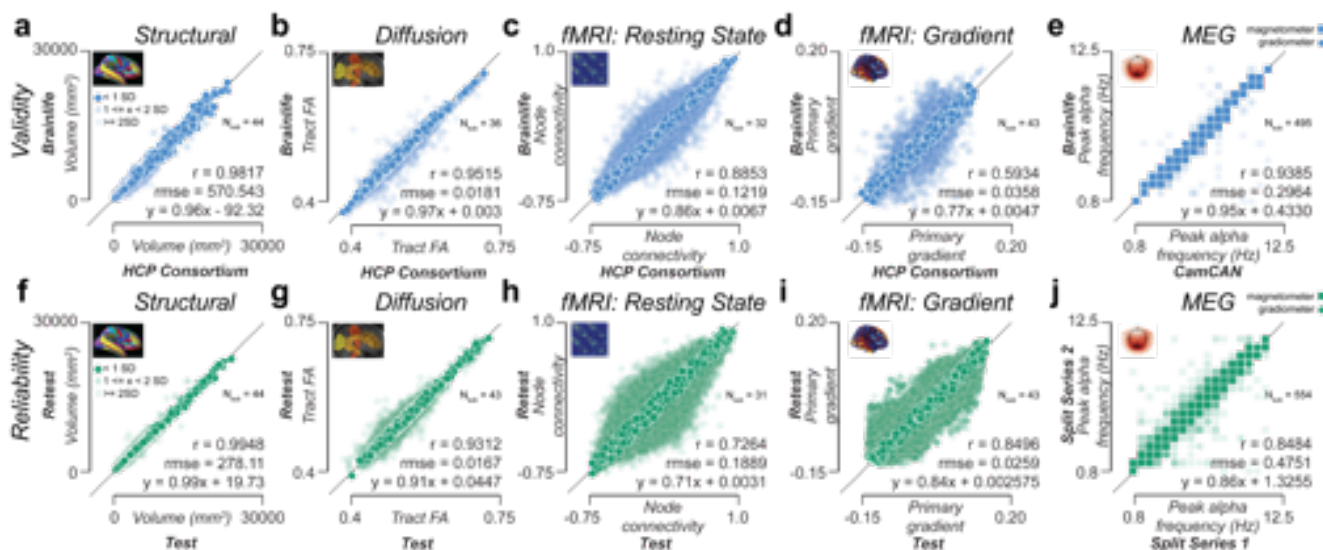
**Figure 3. *brainlife.io* impact (2018-2022).** **a.** *Top left.* Number of users submitting more than 10 jobs per month. *Top middle.* Number of projects over time. *Top right.* Number of Apps over time. *Bottom left.* Data storage across all Projects. *Bottom middle.* Compute hours across all Projects (data only available 6 months post project start). *Bottom right.* Lines of code in the top 50 most-used Apps. **b.** *Top left.* User communities. *Top right.* App categories. *Bottom left.* Percent of total jobs launched with the software library installed (percentage for jobs of top 50 most-used Apps). *Bottom right.* Datasets sources. See also [Fig. S2c](#) for a world-wide distribution of the researchers that have accessed *brainlife.io*.

### Platform testing

Experiments were performed to demonstrate the ability of the platform to provide accurate data processing and analysis at scale. The experiments focused on the four axes of scientific transparency: data processing external validity (DPEV), reliability, reproducibility, and replicability.<sup>91,92</sup> Four data modalities (sMRI, fMRI, dMRI, MEG) were evaluated using, among others, the test-retest HCP<sub>TR</sub>,<sup>93</sup> the Cam-CAN,<sup>27</sup> the HBN,<sup>31</sup> and the ABCD<sup>28</sup> datasets. In total, data from over 3,200 participants across 12 datasets were processed. Extracted brain features included cortical parcel volumes, white matter tract profilometry, functional and structural network properties, functional gradients, and peak alpha frequency (**Fig. 4**). Over 193,000 data objects and 22 Terabytes of data were generated for the experiments. A detailed description of the experiments below can be found in the [Supplemental platform testing](#) section. The *brainlife.io* Apps used for the experiments are reported in [Table S3](#). Post-processing analyses were performed using *brainlife.io*-hosted Jupyter Notebooks (see [Table S2](#)).

Data processing external validity (DPEV) was defined as the ability of data processed on *brainlife.io* to accurately reflect brain properties proficiently processed by other teams. DPEV was estimated for four data modalities (sMRI, dMRI, fMRI, and MEG) and five brain features (brain areas volumes, major white matter tracts fractional

anisotropy, resting state functional connectivity, resting-state function gradients, and MEG peak alpha frequency). Features values obtained using *brainlife.io* Apps were compared against data preprocessed by data originators, specifically the HCP consortium or Cam-CAN project team (Fig. 4, Fig. S4d,e,h). Cortical area volume estimates on 148 parcels were obtained using *brainlife.io* Apps and compared to corresponding estimates provided by the HCP consortium (Fig. 4a;  $r_{\text{validity}}=0.98$ ,  $\text{rmse}_{\text{validity}}=570.54\text{mm}^3$ ). Fractional anisotropy (FA) in 61 white matter tracts was estimated using the raw and minimally preprocessed HCP<sub>TR</sub> dMRI data (Fig. 4b;  $r_{\text{validity}}=0.95$ ,  $\text{rmse}_{\text{validity}}=0.018$ ). Functional connectivity estimates between 117<sup>2</sup> nodes-pairs<sup>94</sup> were compared between raw and minimally preprocessed HCP<sub>TR</sub> dMRI data (Fig. 4c;  $r_{\text{validity}}=0.89$ ,  $\text{rmse}_{\text{validity}}=0.12$ ). In addition, functional gradients<sup>95,96</sup> were computed on 400 nodes estimated on raw and minimally processed HCP<sub>TR</sub> fMRI data (Fig. 4d;  $r_{\text{validity}}=0.59$ ,  $\text{rmse}_{\text{validity}}=0.036$ ). Finally, the peak alpha frequency values were compared between Cam-CAN and *brainlife.io* processed MEG data (Fig. 4e;  $r_{\text{validity}}=0.94$ ,  $\text{rmse}_{\text{validity}}=0.30$  Hz). Overall, the results show strong similarity in feature estimates between data processed on *brainlife.io* versus those processed by external groups (functional gradients demonstrated the lowest validity and data processing-type dependency based on fMRI preprocessing procedures<sup>97</sup>).



**Figure 4. Data processing validity and reliability analysis.** **Top row:** Validity measures derived using the HCP Test-Retest data. Each dot corresponds to the ratio for a given subject between data preprocessed and provided by the HCP Consortium vs data preprocessed on *brainlife.io* in a given measure for a given structure. Pearson's correlation ( $r$ ), root mean squared error ( $\text{rmse}$ ), and a linear fit between the test and retest results were calculated. **a.** Parcel volume ( $\text{mm}^3$ ). **b.** Tract-average fractional anisotropy (FA). **c\***. Node-wise functional connectivity (FC). **d\***. Primary gradient value derived from resting-state fMRI. **e.** Peak frequency (Hz) in the alpha band derived from MEG. Data from magnetometer sensors are represented as squares, and data from gradiometer sensors are represented as circles. **Bottom row:** Test-retest reliability measures derived from derivatives of the HCP<sub>TR</sub> dataset generated using *brainlife.io*. Each dot corresponds to the ratio between a test-retest subject and a given measure for a given structure. Pearson's correlation ( $r$ ), root mean squared error ( $\text{rmse}$ ), and a linear fit between the test and retest results were calculated. **f.** Parcel volume ( $\text{mm}^3$ ). **g.** Tract-average fractional anisotropy (FA). **h\***. Node-wise functional connectivity (FC). **i\***. Primary gradient value derived from resting-state fMRI. **j.** Peak frequency (Hz) in the alpha band derived from MEG using the Cambridge (Cam-CAN) dataset. Data from magnetometer sensors are represented as squares, and data from gradiometer sensors are represented as circles. Dark colors represent data within  $\pm 1$  standard deviation (SD). 50% opacity represents data within 1-2 SD. 25% opacity represents data outside 2 SD. \*A representative 5% of data presented in **c, d, h, i**.

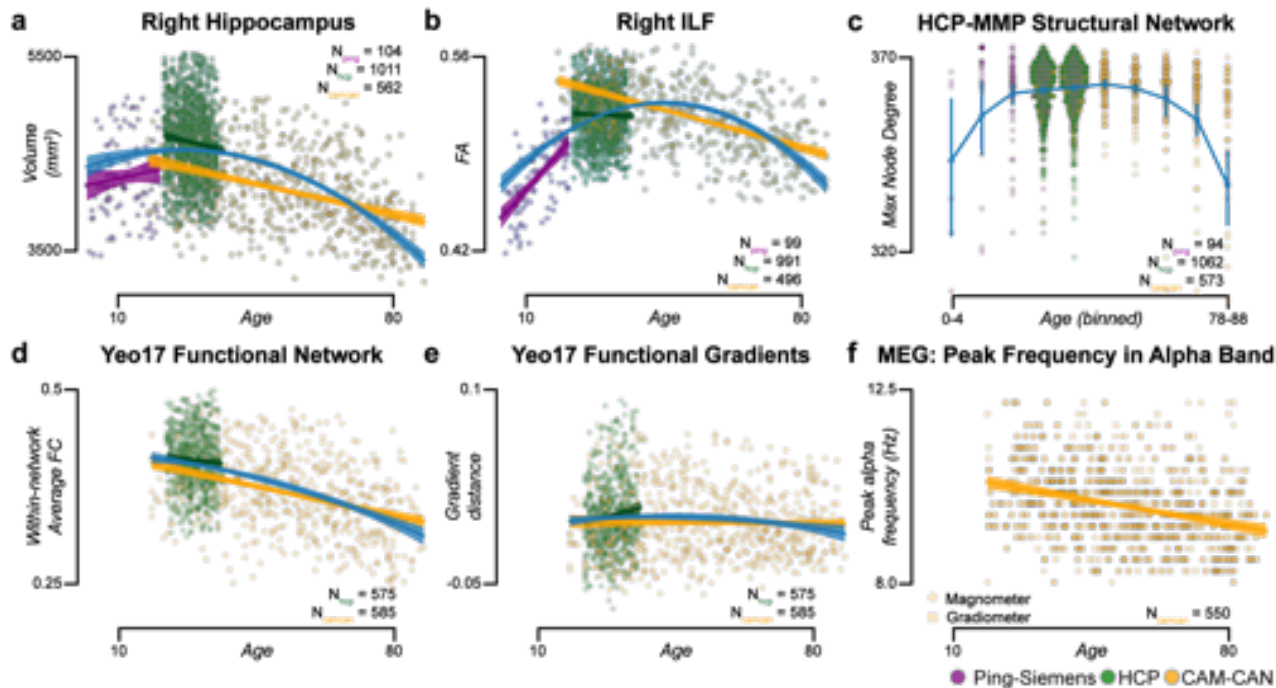
Data processing reliability (DPR) was defined as the ability to produce highly similar results on *test* and *retest* measurements within a study participant. DPR was estimated for the four data modalities and five brain features used above to estimate DPEV. Brain features estimated using *brainlife.io* Apps on *test* and *retest* measurements (HCP<sub>TR</sub> dataset) or median splits data (Cam-CAN MEG) were compared. Reliability estimates of brain area volumes, major tracts FA, networks FC, functional gradients, and Peak Alpha Frequency were obtained (see Fig. 4f-i and associated supplemental text). DPR varied between  $r_{\text{reliability}}=0.99$  and 0.73, with sMRI and dMRI demonstrating the highest reliability ( $r_{\text{reliability}}=0.99$ , 0.93, respectively). See also Fig. S4f-g,i for estimates on



additional brain features and [Table S4](#) for a full report of all correlation values obtained in all brain features. The results show strong reliability of most of all the pipelines with the fMRI reliability being lowest, this is consistent with previous reports <sup>98</sup>. We also performed computational reproducibility (CR) experiments (see [Fig. S4j-n](#) and associated text). These experiments demonstrated the similarity in estimates produced by *brainlife.io* Apps when used twice to process the same dataset. Given the use of containerization technology for the Apps, this test was expected to return high correlation values. Indeed, all correlations were above 0.99, demonstrating high consistency. These experiments demonstrate the ability of the platform to conduct valid, reliable, and reproducible data processing and analysis at scale across multiple data modalities and brain features.

### Platform utility for scientific applications

Next, we evaluated the platform’s potential to support scientific findings. To do so, we evaluated whether data processed using *brainlife.io*’s Apps contained meaningful patterns. We used over 1,800 participants from three datasets: PING (Pediatric Imaging, Neurocognition, Genetics), HCP<sub>s1200</sub>, (HCP Young Adult 1,200), and Cam-CAN. Data were collected across ages, but age ranges differed in each dataset (i.e., 3-20 years for PING, 20-37 years for HCP<sub>s1200</sub>, and 18-88 years for Cam-CAN). The lifelong trajectory was plotted for multiple brain features (e.g., volumes of brain parts, FA of major tracts, network properties. MEG peak frequency, etc; **Fig. 5**). The collated age range spanned 7 decades. Features were combined using *brainlife.io*’s Jupyter Notebooks.



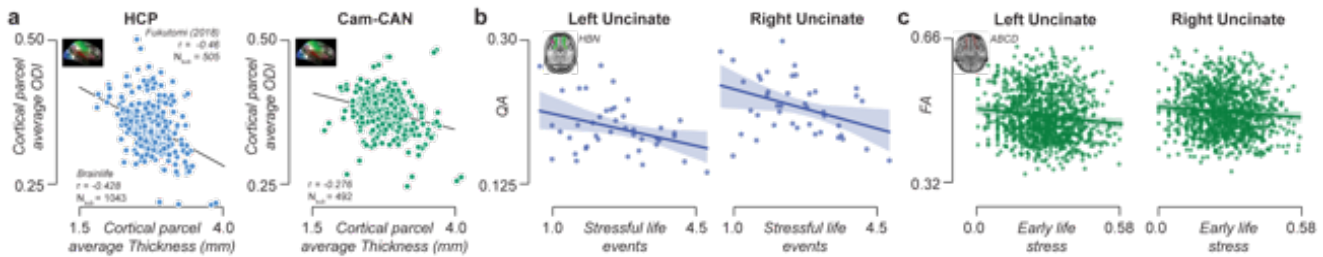
**Figure 5. Lifelong brain maturation estimated across datasets.** Relationship between subject age and **a.** Right hippocampal volume, **b.** Right inferior longitudinal fasciculus (ILF) fractional anisotropy (FA), **c.\*.** maximum node degree of density network derived using the *hcp-mmp* atlas, **d.\*.** Within-network average functional connectivity (FC) derived using the Yeo17 atlas, **e.** Functional gradient distance for visual resting state network derived from the Yeo17 atlas, and **f.** Peak frequency in the alpha band derived from magnetometer (squares) and gradiometers (circles) from MEG data. These analyses include subjects from the PING (*purple*), HCP<sub>1200</sub> (*green*), and Cam-CAN (*yellow*) datasets. Linear regressions were fit to each dataset, and a quadratic regression was fit to the entire dataset (blue). \* All points in **c**, and **d** are presented. See also [Fig. S5](#) and [Supplemental platform utility for scientific applications](#).

Multiple reports have shown inverted U-shaped lifelong trajectories across data modalities.<sup>99–103</sup> We plotted brain features derived for each data modality (sMRI, dMRI, fMRI, and MEG) as a function of age across datasets (**Fig. 5**). Six exemplary lifelong trajectories are shown (additional features are reported in [Fig. S5](#)). For each data

modality, a quadratic model was fit across all three datasets between 3 and 88 years of age:  $y_{feature} = ax_{age}^2 + bx_{age} + c$ , ( $R^2=0.152 \pm 0.0773$  s.d.). Mean quadratic term ( $a$ ) across all data modalities was negative ( $-0.0514 \pm 0.111$  s.d.), demonstrating the expected inverted U-shape trajectory. Results show that, by automatically analyzing data using *brainlife.io* Apps, it is possible to collate across datasets with substantial differences in data acquisition parameters and signal-to-noise profiles. Additional details regarding these experiments can be found in [Supplemental platform utility for scientific applications](#).

### Replication and generalization of previous results

We then evaluated the ability of *brainlife.io* to replicate previous results and generalize findings across datasets. A more detailed description and additional experiments can be found in [Supplemental replication and generalization](#). First, we tested *brainlife.io*'s ability to replicate the results of three previous studies. A negative correlation between cortical thickness and tissue orientation dispersion (ODI;  $r_{original} = -0.46$ ) has been reported in the HCP<sub>s1200</sub> dataset.<sup>104</sup> *brainlife.io* Apps were created to estimate cortical thickness and ODI and analyze HCP<sub>s1200</sub> dataset. A negative relationship between cortical thickness and ODI was estimated, replicating the original study (**Fig. 6a**;  $r_{HCP-brainlife} = -0.43$  vs.  $r_{original}$ ). More examples of replications can be found in [Fig. S6a,b](#).

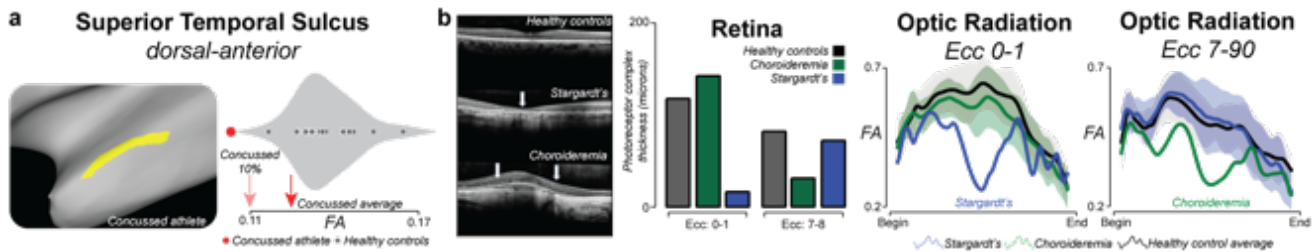


**Figure 6. Replication of previous studies using *brainlife.io*.** **a.** Average cortical hcp-mmp parcel thickness ( $N_{struc} = 322$ ) compared to parcel orientation dispersion index (ODI) from the NODDI model mapped to the cortical surface (*inset*) of the HCP S1200 dataset ( $N_{sub} = 1,043$ ) and Cam-CAN ( $N_{sub} = 492$ ) dataset compared to the parcel-average cortical thickness. **b.** Stressful life events obtained from Negative Life Events Schedule (NLES) survey from Healthy Brain Network participants ( $N_{sub} = 42$ ) compared to Uncinate-average normalized Quantitative Anisotropy (QA). Mean linear regression (*blue line*) fits and standard deviation (*shaded blue*). **c.** Early life stress was obtained from multiple surveys collected from ABCD participants ( $N_{sub} = 1,107$ ) compared to Uncinate-average Fractional Anisotropy (FA). Linear regression (*green line*) fits the data with standard deviation (*shaded green*).

Second, the generalization of the original findings to a different dataset was tested in three ways. The first test was run using the cortical ODI estimated in the Cam-CAN dataset. A negative trend of about half the magnitude of the original was estimated (**Fig. 6a**;  $r_{Cam-CAN-brainlife} = -0.28$  vs.  $r_{original}$ ). The result generalizes the original results and the reduced effect in a new dataset is consistent with reports on the reproducibility of scientific findings.<sup>12</sup> The second generalization test focused on the reported relationship between life stressors and white matter structural organization of the uncinate fasciculus (UF;  $r = -0.057$ ).<sup>105</sup> Two datasets were used to extend the finding to new data, i.e., HBN and ABCD. The number of negative life events (Negative Life Events Schedule; NLES) in the HBN dataset was correlated with subjects' quantitative anisotropy (QA) in the right- and left-hemisphere UF. Results show a negative correlation similar in magnitude as found in the original study (**Fig. 6b**  $r_{HBN\_LEFT} = -0.35$ ,  $p$ -value  $< 0.05$ ;  $r_{HBN\_RIGHT} = -0.39$ ,  $p$ -value  $< 0.05$ ). The third and final attempt at the generalization of the same result was made using the ABCD dataset. Early life stress was estimated as a composite score of traumatic life events, environmental and neighborhood safety, and the family conflict subscale of the Family Environment Scale.<sup>29</sup> A negative relationship between UF FA and the composite score was estimated in the left- and right-UF (**Fig. 6c**  $r_{ABCD\_LEFT} = -0.12$ ,  $p$ -value  $< 0.001$ ;  $r_{ABCD\_RIGHT} = -0.09$ ,  $p < 0.01$ ). Overall, these results demonstrate both the robustness of the original results and the potential of *brainlife.io* services to detect meaningful associations in large, heterogeneous datasets.

## Example applications to detecting disease

The final two tests evaluated the platform's ability to identify human disease biomarkers. Data from individuals with a sports-related concussion, eye disease (Choroideremia and Stargardt's disease), and matched controls were used (**Fig. 7**). A detailed description of the experiments can be found in [Supplemental to detecting disease](#). It has been reported that concussion can alter brain tissue both in cortical and deep white matter tracts.<sup>106</sup> We set out to measure the difference in cortical white matter tissue in concussed and matched controls. FA was estimated from data collected within 24-48 hours post-concussion. The distribution of FA in the superior temporal sulcus (STS) is reported (**Fig. 7a**). One representative athlete showed strong post-concussive symptoms and low STS cortical FA (red). The result demonstrates the potential of *brainlife.io* processed data to report meaningful changes in brain tissue following a concussion.



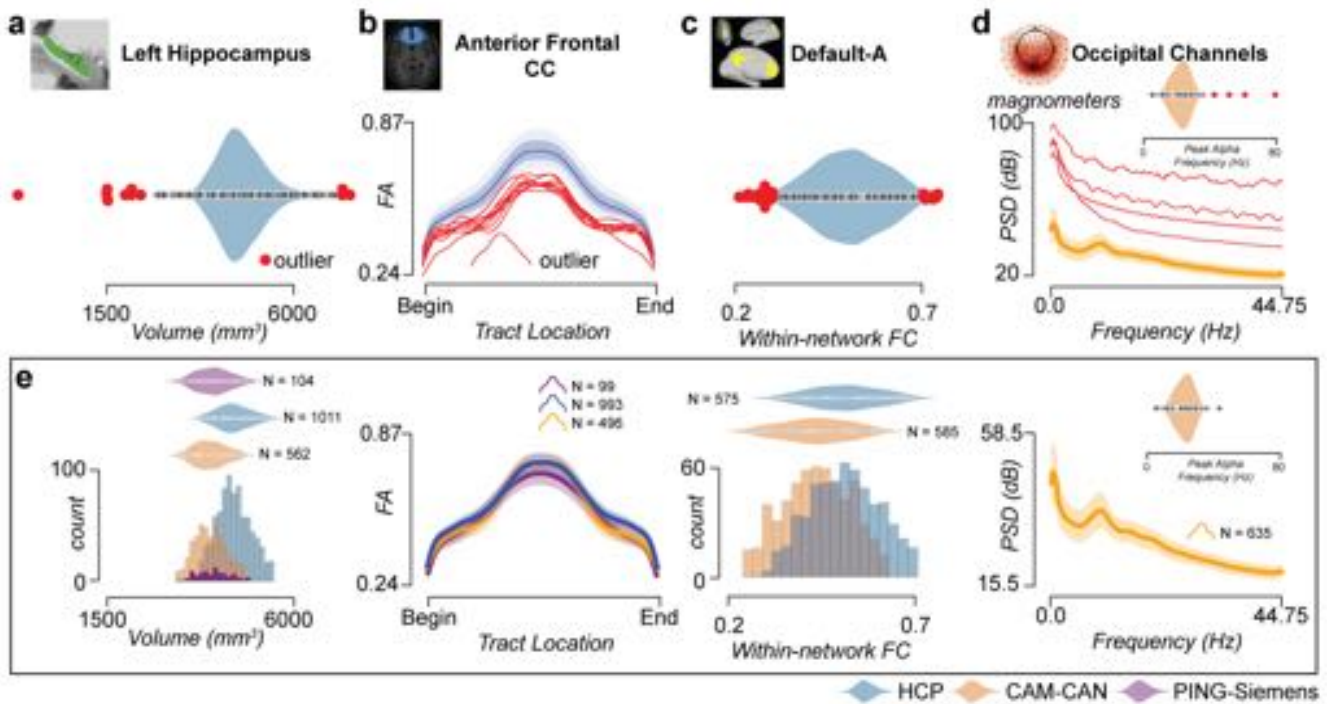
**Figure 7. Using *brainlife.io* to identify and characterize clinical populations from healthy controls.** **a.** Fractional anisotropy (FA) values were estimated within the superior temporal sulcus (da: dorsal anterior) from 20 healthy athlete controls (*gray distribution*) and 10 concussed athletes. Average FA, 10% low FA, and the lowest FA value across all concussed athletes were measured (*red arrows and dot*). **b.** Retinal OCT images from healthy controls (*top row*), Stargardt's disease patients (*middle row*), and Choroideremia patients (*bottom row*). From these images, photoreceptor complex thickness was measured for each group (Controls: gray; Choroideremia: green; Stargardt's: blue) in two distinct areas of the retina: the fovea (eccentricities 0-1 degrees) and the periphery (eccentricities 7-8 degrees). In addition, optic radiations carrying information for each area of the retina were segmented and FA profiles were mapped. Average profiles with standard error (shaded regions) were computed. One Stargardt and one Choroideremia participant were each identified as having FA profiles that deviated from both healthy controls and the opposing retinal disorder.

Changes in the white matter of the optic radiation (OR) as a result of eye disease have been reported.<sup>107-111</sup> We set out to test the ability of *brainlife.io* Apps to detect similar changes in the OR white matter tissue in two eye diseases for which OR white matter changes have not previously been reported. Individuals with Stargardt's disease (a deterioration of the retina initiating in the central fovea), and Choroideremia (retinal deterioration initiating in the visual periphery), were compared to healthy controls. Retina photoreceptor complex thickness was estimated in the fovea and peripheral using optical coherence tomography (0-1 and 7-90 degrees of visual eccentricity, respectively; **Fig. 7b**). Choroideremia patients showed photoreceptor complex thickness comparable to healthy controls in the fovea, but deviated in the periphery (**Fig. 7b**). The trend was opposite for Stargardt's patients. *brainlife.io* Apps were developed to automatically separate OR bundles projecting to different visual eccentricity in cortical area V1. Average FA profiles for each patient group and controls were estimated for OR fibers projecting to the fovea or periphery.<sup>112 113,114</sup> Results show a reduction in FA in the component of the OR projecting to the fovea (but not the periphery) in Stargardt's patients (**Fig. 7b**, blue), and the opposite pattern (OR fibers projecting to the periphery had lower FA than controls) in Choroideremia patients (**Fig. 7b**, blue). These results demonstrate the ability of the platform technology to detect disease biomarkers.

## A new approach to facilitate quality control at scale

*brainlife.io* offers a unique quality assurance (QA) approach to ensure processed data has the quality necessary to serve large user bases. *Reference ranges* are often used in vision science to provide a reference for a measurement,<sup>115</sup> and a similar approach was integrated within the *brainlife.io* data processing interface. To test it, the mean, first, and second SD were estimated (via multiple Apps) for four brain features (tractmeasures, parc-stats, networks, PSD) using the HCP<sub>s1200</sub>, Cam-CAN, and PING datasets. For each of the four brain features, the estimated mean and estimated s.d. (referred to here as *Reference ranges*) are automatically calculated on the

*brainlife.io* platform. That is, when a researcher uses an App to estimate one of the four features, the values of the researcher’s dataset are automatically overlaid on top of the mean, first, and second s.d. marks provided as a reference by *brainlife.io*. In this way, the mean and variability can be used by researchers to efficiently judge whether a recently processed dataset returned appropriate values. For example, reference datasets can be used to detect outlier data (Fig. 8a-d). Example reference datasets for four Datatypes are in Fig. 8e and an example of platform interfaces reporting these reference datasets is shown in Fig. S8. A detailed description of the approach used in this section can be found in [Supplemental to quality control at scale](#). These reference ranges are an additional source for quality assurance, alongside other options for QA such as online data visualization, the automated generation of images and plots from the processed data as well as the detailed technical reports from major BIDS Apps such as fMRIprep, QSIPrep, MRIQC, Freesurfer<sup>69,70,72,116</sup>.



**Figure 8. Reference datasets for quality assurance.** Example workflow for building normative reference ranges for multiple derived statistical products (cortical parcel volume, white matter tract profilometry, within-network functional connectivity, and power-spectrum density (PSD)). **a.** Cortical volumes of the left hippocampus from HCP participants. Red dots indicate outlier data points. **b.** Average fractional anisotropy (FA) profiles (blue line) plotted with two standard deviations (shaded regions). Red lines indicate outlier profiles. **c.** Within-network functional connectivity for the nodes within the Default-A network using the Yeo17 atlas. Red dots indicate outlier data points. **d.** Average PSD from occipital channels using magnetometer sensors from Cam-CAN participants with one standard deviation (shaded regions). Red lines indicate outlier participants. Peak alpha frequency distribution was also computed, and outliers were detected (inset). **e.** Normative reference distributions for each derived statistical product across the PING (purple), HCP (blue), and Cam-CAN (orange) datasets. These distributions have had outliers removed. An example of the *brainlife* visualization for reference datasets can be found in [Fig. S8](#).

## DISCUSSION

The *brainlife.io* platform was developed with public funding to promote the progress of brain science and education and to enable discovery and improve health. The platform connects researchers with publicly available datasets, analysis code, data archives, and compute resources. *brainlife.io* is an end-to-end, turnkey data analysis platform that provides researchers interested in the brain with services for data upload, management, visualization, preprocessing, analysis, and publication—all integrated within a unique cloud environment and web

interface. The platform uses opportunistic computing and publicly-funded resources for storage and computing,<sup>78-80</sup> but it can also use popular commercial clouds. The goal is to advance the democratization of big data neuroscience by lowering the barriers of entry to multimodal data analysis, network neuroscience, and large-scale analysis, all opportunities historically limited to a paucity of highly-skilled, high-profile research teams.<sup>39,99,117-122</sup> The platform supports a rigorous and transparent scientific process spanning the research data lifecycle from after data collection to sharing<sup>123</sup> and automatically tracks complex sequences of interactions between researchers, Apps, analysis notebooks, and data objects to support reproducibility. The FAIR data principles for data stewardship and management<sup>9</sup> are generally used as guidelines for any data-centric project. Recently, it has been proposed that a modern definition of neuroscience data should extend beyond measurements and data to include metadata and software for analysis and management.<sup>123</sup> Each research asset on *brainlife.io* (i.e., data derivatives, analysis software, and software services, as handled by the platform) is aligned with the FAIR data principles (see [Supplement on brainlife.io and the FAIR principles](#)). The following discussion will include descriptions of the resources available for getting started on *brainlife.io*, applications of *brainlife.io* to educational settings, the platform's strict data governance principles, increasing "data gravity" via *brainlife.io*, potential expansion of the platform, and the platform's current limitations.

The *brainlife.io* project provides multiple resources for App developers, computing resource managers, and neuroscience researchers to learn to use the platform or contribute to the project. A comprehensive overview of the platform and tutorials for getting started with developing Apps or using the platform can be found in the integrated documentation ([brainlife.io/docs](https://brainlife.io/docs)), as well as on a YouTube Channel that provides tutorials and demonstrations of concepts ([youtube.com/@brainlifeio](https://youtube.com/@brainlifeio)). A public slack channel is used for managing user communications, requests, feedback, and operations ([brainlife.slack.com](https://brainlife.slack.com)). Users can also ask questions to developers and the community using the topic 'brainlife' on [neurostars.org](https://neurostars.org) and adding [GitHub issues](#). Finally, a quarterly community engagement and outreach newsletter is sent to all users, and a Twitter account (@brainlifeio) informs the wider community on critical events and connects to information relevant to the project.

*brainlife.io* and its user community are highly engaged in providing innovative training and education opportunities for the next generation of students, postdocs, and clinicians interested in the intersection between neuroscience, data science, and information. The platform allows new students and educators to access many complex data files and analysis methods with minimal overhead. Educators have started using *brainlife.io* to teach neuroscience and data science concepts in the classroom, and courses have been organized in Europe, the USA, Canada, and Africa. These courses introduce basic concepts and teach students how to perform neuroimaging investigations without the requirement of programming or computing expertise. The skills that can be learned using the platform include data preprocessing, quality assurance, and statistical analyses. Integrative data management and analysis provide opportunities for educators and students in under-resourced institutions or countries to perform research and teach neuroscience with hands-on experience.

The project leadership and advisory team recognize the importance of ensuring that data processing workflows are ethically responsible, legally compliant, and socially acceptable. Indeed, data governance is considered an integral part of data processing. Data governance is defined as the principles, procedures, technologies, and policies that ensure acceptable and responsible processing of data at each stage of the data life cycle.<sup>123</sup> It comprises the management of the availability, usability, integrity, quality, and security of data.<sup>123</sup> The data governance policies, processes, and technologies within *brainlife.io* cover three key elements: people, processes, and technologies. A comprehensive set of advanced security measures and protocols guarantee that only authorized individuals have access. These measures include end-to-end encrypted communication, strict access control, and support for multi-factor authentication. Datasets uploaded by users using [brainlife.io/ezBIDS](https://brainlife.io/ezBIDS) are pseudonymized,<sup>124</sup> (i.e. direct identifiers are removed) at upload. The platform interface provides fields for project managers to add Data Use Agreements (DUA) in alignment with the nature and context of their data. The platform even provides template DUAs describing data users' responsibilities and liabilities, including becoming the data controller (the person who controls the purposes and means of processing the data). These governance mechanisms comply with available regulations and mandates, such as the European Union's General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) in the United States, which require that personal data be stored and managed in a secure and compliant manner. Cloud systems are designed to provide the level of protection necessary to ensure the privacy and confidentiality of research participants. Finally, the incoming changes to data deposition and sharing mandates (such as that recently released by the National Institutes of Health in the United States<sup>125,126</sup>) are likely to increase the workload

for neuroscience researchers. The *brainlife.io* publication records are compatible with the NIH data sharing mandates (for privacy, sharing, and preservation), and the platform is registered on fairsharing.org, datacite.org, datasetsearch.research.google.com, and nitric.org.

Data gravity is the ability of datasets to attract utilization<sup>127</sup> Neuroimaging research within the larger neuroscience field has led the way in increasing data gravity. A long and growing list of tools orchestrated under a general label of open science are being developed to support and facilitate data utilization and access. These tools can be divided into four primary categories: software library, data archives and database systems, data standards, and computing platforms.<sup>40</sup> The data archives and systems closest to *brainlife.io* are the INDI,<sup>128,129</sup> OpenNeuro.org,<sup>34</sup> DANDI,<sup>130</sup> BossDB,<sup>131</sup> DataLad,<sup>74</sup> NITRC,<sup>132</sup> PING,<sup>32</sup> Can-CAM,<sup>27</sup> the Brain/MINDS project,<sup>133</sup> and LORIS.<sup>134</sup> The web services most related to the current work are NeuroQuery,<sup>135</sup> NeuroScout,<sup>136</sup> CBRAIN,<sup>137</sup> NeuroDesk,<sup>138</sup> XNAT,<sup>139</sup> NEMAR,<sup>140</sup> EBRAINS<sup>141</sup>, LONI,<sup>142,143</sup>, the International Brain Lab data Infrastructure<sup>144</sup>, COINSTAC<sup>145</sup> and CONP<sup>146</sup>. Most projects are open-source and provide various degrees of data access. *brainlife.io* end-to-end integrated environment that brings researchers from raw data to Jupyter Notebooks and Tidy data tables while tracking data provenance automatically is unique. But many other projects exist and given the fast-growing landscape of neuroinformatics projects, we collected a table listing the major ones (see [Table S5](#)). The International Neuroinformatics Coordinating Facility also provides a list of major projects [incf.org/infrastructure-portfolio](http://incf.org/infrastructure-portfolio). *brainlife.io* is one of the approved resources, as it complies with the INCF requirement for FAIR infrastructure. The ability of the platform to utilize data from multiple modalities (MEG, EEG, MRI) is a unique feature, connecting neuroimaging research sectors that have been historically siloed. However, we envision additional opportunities for expanding the types of data managed by the platform, fostering further data integration. For example, other data modalities could be mapped to *brainlife.io* Datatypes, and the mechanism for data Integration with metadata capture toolkits<sup>147</sup> and data models<sup>148</sup> would provide additional facilitation for the analysis domains of data currently not covered by the BIDS standard.

Improving the platform's automation and interoperability is part of the vision and sustainability plan. For example, despite the best efforts of App developers, errors occur (see [Fig. S3d](#)). Currently, researchers only have simple interfaces that report technical output logs and error messages when Apps fail to process data, and parsing these messages requires expertise. Users are required to either contact the *brainlife.io* team or parse the error logs themselves. Planned improvements to *brainlife.io*'s error reporting interfaces will help users understand the sources of errors and find solutions. In addition to error identification, identifying the optimal set of processing steps or parameter sets at the beginning of a project can prove challenging. In addition, currently, researchers identify the optimal data processing steps by looking at existing documentation or videos. In the future, mechanisms that automatically identify processing steps can be implemented to suggest to researchers optimal ways to process their data (e.g. given what other researchers might have already implemented on the platform). Finally, improving connection with major archives and platforms such as OpenNeuro.org, DANDI, NeuroScout, NeuroDesk, and neurosynth.org, would contribute to implementing the vision of a global interoperable ecosystem for a FAIR, accessible, and democratized neuroscience.

In summary, the capabilities of *brainlife.io* are unique, open, accessible, and expandable. The expansion of instrument capabilities in neuroimaging has in the last 30 years revolutionized our ability to collect data about the brain and brain function. As the landscape of neuroscience big-data projects is only expected to grow in the coming years, moving research data management and computing to cloud platforms will become not just a brilliant option, but a serious requirement. Compliance with mandates for data privacy and sharing will ultimately require researchers to move data management and processing to secure and professionally managed to compute and storage systems. Our goal for *brainlife.io* is to facilitate this process and thereby revolutionize the ability to rigorously and reliably make use of the wealth of data now available to understand brain function, leading to new cures for brain disease. In so doing, *brainlife.io* will also make cutting-edge datasets and analysis resources more accessible to students and researchers from traditionally underrepresented groups in high-, medium- and low-income countries.

## ONLINE METHODS AND MATERIALS

**Data collection approval.** Multiple experiments were performed by individuals at various institutions using the platform. Experiments were approved by the local institutional review boards (IRB), and only the personnel approved for a specific study accessed the data in private projects on brainlife.io. Some of the secondary data usages were deemed IRB-exempt.

**Data sources.** Multiple openly available data sources were used for examining the validity, reliability, and reproducibility of brainlife.io Apps and for examining population distributions. All information regarding the specific image acquisitions, participant demographics, and study-wide preprocessing can be found in the following publications <sup>27,28,31,149–153</sup>. Some data sources are currently unpublished. For these, the appropriate information is provided.

### **Validity, reliability, reproducibility, replicability, developmental trends, & reference datasets**

*Human Connectome Project (HCP; Test-Retest, s1200-release)* <sup>149</sup>. Data from these projects were used to assess the validity, reliability, and reproducibility of the platform. They were used to assess the abilities of the platform to identify developmental trends in structural and functional measures, and they were used to generate reference datasets. Structural data (sMRI): The minimally-preprocessed structural T1w and T2w images from the Human Connectome Project (HCP) from 1066 participants from the s1200 and 44 participants from the Test-Retest releases were used. Specifically, the 1.25 mm 'acpc\_dc\_restored' images generated from the Siemens 3T MRI scanner were used for all analyses involving the HCP. For most examinations, the already-processed Freesurfer output from HCP was used. Diffusion data (dMRI): To assess the validity of preprocessing on brainlife.io, the unprocessed dMRI data from 44 participants from the HCP Test dataset was used. For reliability and all remaining analyses, the minimally-preprocessed diffusion (dMRI) images from 1,066 participants from the s1200 and 44 participants from the Test-Retest releases from the 3T Siemens scanner were used. All processes incorporated the multi-shell acquisition data. Functional data (fMRI): For validation, the unprocessed resting-state functional MRI (fMRI) from 44 participants from the HCP Test dataset was compared to the minimally-preprocessed BOLD data provided by HCP. For reliability and all other analyses, the minimally-preprocessed BOLD data from 1,066 participants from the s1200 and 44 participants from the Test-Retest releases from the 3T Siemens scanner were used.

*The Cambridge Centre for Ageing and Neuroscience (Cam-CAN)* <sup>27</sup>. The data from this project were used to assess the validity, reliability, and reproducibility of the platform and to assess the abilities of the platform to identify developmental trends of structural and functional measures, and to generate reference datasets. Structural data (sMRI): The unprocessed 1mm isotropic structural T1w and T2w images from 652 participants from the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study were used. Diffusion data (dMRI): The unprocessed 2mm isotropic diffusion (dMRI) images from 652 participants from the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study were used. Functional data (fMRI): The 3mm x 3mm x 4mm unprocessed resting-state fMRI images from 652 participants from the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study were used. Electromagnetic data (MEG): The 1000 Hz resting-state filtered and unfiltered datasets from 652 participants from the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study were used.

### **Developmental trends & reference datasets**

*Pediatric Imaging, Neurocognition, and Genetics (PING)* <sup>32</sup>. The data from this project were used to assess the abilities of the platform to identify developmental trends of structural measures and to generate reference datasets. Structural data (sMRI): The unprocessed 1.2 x 1.0 x 1.0 mm structural T1w and the 1.0 mm isotropic T2w images from 110 participants from the *Pediatric Imaging, Neurocognition, and Genetics (PING)* study were used. Diffusion data (dMRI): The unprocessed 2mm isotropic diffusion (dMRI) images from 110 participants from the *Pediatric Imaging, Neurocognition, and Genetics (PING)* study were used.

### **Replicability datasets**

*Adolescent Brain Cognitive Development (ABCD)* <sup>28,29</sup>. Structural data (sMRI): The unprocessed 1mm isotropic structural T1w and T2w images from a subset of 1,877 participants from the Adolescent Brain Cognitive Development (ABCD release-2.0.0) study were used. Diffusion data (dMRI): The unprocessed 1.77mm isotropic diffusion (dMRI) images from a subset of 1877 participants from the Adolescent Brain Cognitive Development (ABCD release-2.0.0) study were used. A single diffusion gradient shell was used for these experiments ( $b=3000s/msec^2$ ). Research approved by the University of Arkansas IRB (#2209425822).

*Healthy Brain Network (HBN)* <sup>31</sup>. The data from this project were used to assess the abilities of the platform to replicate previously published findings via the assessment of the relationship between microstructural measures mapped to segmented uncinate fasciculi and self-reported early life stressors. Research approved by the University of Pittsburgh IRB (#PRO17060350). Structural data (sMRI): The 0.8 mm isotropic structural T1w images from 42 participants from the Healthy Brain Network (HBN) study were used. Diffusion data (dMRI): The unprocessed 1.8 mm isotropic diffusion (dMRI) images from 42 participants from the CitiGroup Cornell Brain Imaging Center site of the Healthy Brain Network (HBN) study were used. Research approved by the University of Pittsburgh IRB (#PRO17060350).

UPENN-PMC<sup>154</sup>. The data from this project were used to assess the abilities of the platform to replicate previously published findings via the assessment of the performance of an automated hippocampal segmentation algorithm. All procedures were conducted under the approval of the Institutional Review Board at the University of Texas at Austin. Structural data (sMRI): The T1w and T2w data were provided within the Automated Segmentation of Hippocampal Subfields (ASHS) atlas<sup>154</sup>.

### Clinical-identification datasets

**Indiana University Acute Concussion Dataset.** The data from this project were used to assess the abilities of the platform to identify clinical populations via the mapping of microstructural measures to the cortical surface. Neuroimaging was performed at the Indiana University Imaging Research Facility, housed within the Department of Psychological and Brain Sciences with a 3-Tesla Siemens Prisma whole-body MRI using a 64-channel head coil. Within this study, 9 concussed athletes and 20 healthy athletes were included. Research approved by Indiana University (IRB: 906000405). Structural data (sMRI): High-resolution T1-weighted structural volumes were acquired using an MPRAGE sequence: TI = 900 ms, TE = 2.7 ms, TR = 1800 ms, flip angle = 9°, with 192 sagittal slices of 1.0 mm thickness, a field of view of 256 x 256 mm, and an isometric voxel size of 1.0 mm<sup>3</sup>. The total acquisition time was 4 minutes and 34 seconds. High-resolution T2-weighted structural volumes were also acquired: TE = 564 ms, TR = 3200 ms, flip angle = 120°, with 192 sagittal slices, a field of view of 240 x 256 mm, and an isometric voxel size of 1.0mm<sup>3</sup>. Total acquisition time was 4 minutes 30 seconds. Diffusion data (dMRI): Diffusion data were collected using single-shot spin-echo simultaneous multi-slice (SMS) EPI (transverse orientation, TE = 92.00 ms, TR = 3,820 ms, flip angle = 78 degrees, isotropic 1.5 mm<sup>3</sup> resolution; FOV = LR 228 mm x 228 mm x 144 mm; acquisition matrix MxP = 138 x 138. SMS acceleration factor = 4). This sequence was collected twice, one in the AP fold-over direction and the other in the PA fold-over direction, with the same diffusion gradient strengths and the number of diffusion directions: 30 diffusion directions at b = 1000 s/mm<sup>2</sup>, 60 diffusion directions at b = 1,750 s/mm<sup>2</sup>, 90 diffusion directions at b = 2,500 s/mm<sup>2</sup>, and 19 b = 0 s/mm<sup>2</sup> volumes. The total acquisition time for both sets of dMRI sequences was 25 minutes and 58 seconds.

**Oxford University Choroideremia & Stargardt's Disease Dataset.** The data from this project was used to assess the abilities of the platform to identify clinical populations via mapping retinal-layer thickness via OCT and mapping of microstructural measures along optic radiation bundles segmented using visual field information (eccentricity). Neuroimaging was performed at the Wellcome Centre for Integrative Neuroimaging, Oxford with the Siemens 3T scanner. Research approved by the UK Health Regulatory Authority reference 17/LO/1540. Structural data (sMRI): High-resolution T1-weighted anatomical volumes were acquired using an MPRAGE sequence: TI = 904 ms, TE = 3.97 ms, TR = 1900 ms, flip angle = 8°, with 192 sagittal slices of 1.0 mm thickness, a field of view of 174 mm x 192 mm x 192 mm, and an isometric voxel size of 1.0 mm<sup>3</sup>. The total acquisition time was 5 minutes and 31 seconds. Diffusion data (dMRI): Diffusion data were collected using EPI (transverse orientation, TE = 92.00ms, TR = 3600 ms, flip angle = 78 degrees, 2.019 x 2.019 x 2.0 mm<sup>3</sup> resolution; FOV = 210 mm x 220 mm x 158 mm; acquisition matrix MxP = 210 x 210, SMS acceleration factor = 3). This sequence was collected twice, one in the AP fold-over direction and the other in the PA fold-over direction. The PA fold-over scan contained 6 diffusion directions, 3 at b = 0 s/mm<sup>2</sup> and 3 at b = 2000 s/mm<sup>2</sup>, and was used primarily for susceptibility-weighted corrections. The AP fold-over scan contained 105 diffusion directions, 5 at b = 0 mm/s<sup>2</sup>, 51 at b = 1000 mm/s<sup>2</sup>, and 49 at b = 2000 mm/s<sup>2</sup>. The total acquisition time for both sets of dMRI sequences was 7 minutes and 8 seconds.

### General processing pipelines

**Structural processing.** For the ABCD, Cam-CAN, Oxford University Choroideremia & Stargardt's Disease Dataset, and the Indiana University Acute Concussion datasets, the structural T1w and T2w (sMRI) images (if available) were preprocessed, including bias correction and alignment to the anterior commissure-posterior commissure (ACPC) plane, using [A273](#) and [A350](#) respectively. For PING data, no bias correction was performed but alignment to the ACPC plane was performed using [A99](#) and [A116](#) for T1w and T2w data respectively. For HCP data, this data was already provided. The structural T<sub>1</sub>-weighted images for each participant and dataset were then segmented into different tissue types using functionality provided by *MRTrix3* (Tournier et al, 2019) implemented as [A239](#). For a subset of datasets, this was performed within the diffusion tractography generation step using [A319](#). The gray- and white-matter interface mask was subsequently used as a seed mask for white matter tractography. The processed structural T1w and T2w images were then used for segmentation and surface generation using the *recon-all* function from *Freesurfer*<sup>72</sup> ([A0](#)). Following *Freesurfer*, representations of the cortical 'midthickness' surface were computed by spatially averaging the coordinates of the pial and white matter surfaces generated by *Freesurfer* using the *wb\_command -surface-cortex-layer* function provided by Workbench command for the HCP<sub>TR</sub>, HCP<sub>s1200</sub>, ABCD, Cam-CAN, PING, and Indiana University Acute Concussion datasets. These surfaces were used for cortical tissue mapping analyses. Following *Freesurfer* and midthickness-surface generation, the 180 multimodal cortical nodes (*hcp-mmp*) atlas and the Yeo 17 (*yeo17*) atlas were mapped to the *Freesurfer* segmentation of each participant implemented as *brainlife.io* App [A23](#). These parcellations were used for subsequent cortical, subcortical, and network analyses. In addition, measures for cortical thickness, surface area, volume, and summaries of diffusion models of microstructure were estimated using [A383](#) and [A389](#). To estimate population receptive fields (pRF) and visual field eccentricity properties in the cortical surface in the Oxford University Choroideremia & Stargardt's Disease Dataset, the automated mapping algorithm developed by<sup>155,156</sup> was implemented using [A187](#). To segment thalamic nuclei for optic radiation tracking, the automated thalamic nuclei segmentation algorithm provided by *Freesurfer*<sup>72</sup> was implemented as [A222](#). Finally, visual regions of interest binned by eccentricity were then generated using AFNI<sup>157</sup> functions implemented in [A414](#). To assess the replicability capabilities of the platform, an



automated hippocampal nuclei segmentation app ([A262](#)) was used to segment hippocampal subfields from participants within the UPENN-PMC dataset provided within the ASHS atlas.

**Diffusion (dMRI) processing. Preprocessing & model fitting:** For a majority of the analyses involving the HCP dataset, the minimally-preprocessed dMRI images were used and thus no further preprocessing was performed. However, to assess the validity of the preprocessing pipeline, the unprocessed dMRI data from the HCP Test dataset, dMRI images were preprocessed following the protocol outlined in <sup>158</sup> using [A68](#). The same app was also used for preprocessing the dMRI images for the ABCD, Cam-CAN, PING, Oxford University Choroideremia & Stargardt's Disease Dataset, the Indiana University Acute Concussion, and HBN datasets. Specifically, dMRI images were denoised and cleaned from Gibbs ringing using functionality provided by *MRTrix3* before being corrected for susceptibility, motion, and eddy distortions and artifacts using FSL's *topup* and *eddy* functions <sup>44,159</sup>. Eddy-current and motion correction was applied via the *eddy\_cuda8.0* with the replacement of outlier slices (*i.e. repol*) command provided by FSL <sup>160-163</sup>. Following these corrections, *MRTrix3*'s *dwigradcheck* functionality was used to check and correct for potential misaligned gradient vectors following top-up and eddy <sup>164</sup>. Next, dMRI images were debiased using ANT's *n4* functionality <sup>165</sup> and the background noise was cleaned using *MRTrix3.0*'s *dwidenoise* functionality <sup>166</sup>. Finally, the preprocessed dMRI images were registered to the structural (T1w) image using FSL's *epi\_reg* functionality <sup>167-169</sup>. Following preprocessing, brain masks for dMRI data using *bet* from FSL were implemented as [A163](#).

**DTI, NODDI, and q-sampling model fitting.** Following preprocessing, the diffusion tensor (DTI) model <sup>170</sup> and the neurite orientation dispersion and density imaging (NODDI) <sup>171,172</sup> models were subsequently fit to the preprocessed dMRI images for each participant using either [A319](#) or [A292](#) for DTI model fitting and [A365](#) for NODDI fitting. Note, the NODDI model was only fit on the HCP, Cam-CAN, Oxford University Choroideremia & Stargardt's Disease Dataset, and the Indiana University Acute Concussion datasets. For those datasets, the NODDI model was fit using an intrinsic free diffusivity parameter ( $d_{||}$ ) of  $1.7 \times 10^{-3}$  mm<sup>2</sup>/s for white matter tract and network analyses, and a  $d_{||}$  of  $1.1 \times 10^{-3}$  mm<sup>2</sup>/s for cortical tissue mapping analyses, using AMICO's implementation <sup>172</sup> as [A365](#). The constrained spherical deconvolution (CSD) (Tournier et al, 2007) model was then fit to the preprocessed dMRI data for each run across 4 spherical harmonic orders (*i.e.*  $L_{max}$ ) parameters (2,4,6,8) using functionality provided by *MRTrix3* implemented as brainlife.io App [A238](#). For the PING datasets, the CSD model was fit using the same exact code found in [A238](#), but performed using the tractography App [A319](#). For the HBN dataset, the isotropic spin distribution function was obtained by reconstructing the diffusion MRI data with the Generalized q-sampling imaging method <sup>173</sup> using functionality provided by DSI-Studio<sup>66</sup> ([A423](#)). Quantitative anisotropy (QA) was then estimated from the isotropic spin distribution function.

**Tractography.** Following model fitting, the fiber orientation distribution functions (fODFs) for  $L_{max}=6$  and  $L_{max}=8$  were subsequently used to guide anatomically-constrained probabilistic tractography (ACT; Smith et al, 2012) using functions provided by *MRTrix3* implemented as brainlife.io App [A297](#) or [A319](#). For the HCP<sub>TR</sub>, HCP<sub>s1200</sub>, and Oxford University Choroideremia & Stargardt's Disease datasets,  $L_{max}=8$  was used. For ABCD and Cam-CAN datasets,  $L_{max}=6$  was used. For the HCP, ABCD, Cam-CAN, datasets, a total of 3 million streamlines were generated. For all datasets, a step-size of 0.2 mm was implemented. For the HCP<sub>TR</sub>, HCP<sub>s1200</sub>, ABCD, and Cam-CAN datasets, minimum and maximum lengths of streamlines were set at 25 and 250mm respectively, and a maximum angle of curvature of 35° was used. For the PING dataset, minimum and maximum lengths of streamlines were set at 20 and 220mm respectively, and a maximum angle of curvature of 35° was used.

**Whiter Matter Segmentation and cleaning.** Following tractography, 61 major white matter tracts were segmented for each run using a customized version of the white matter query language (Bullock et al, 2019) implemented as brainlife.io App [A188](#). Outlier streamlines were subsequently removed using functionality provided by Vistasoft and implemented as brainlife.io App [A195](#). Following cleaning, tract profiles with 200 nodes were generated for all DTI and NODDI measures across the 61 tracts for each participant and test-retest condition using functionality provided by Vistasoft and implemented as [A361](#). Macrostructural statistics, including average tract length, tract volume, and streamline count was computed using functionality provided by Vistasoft implemented as [A189](#). Microstructural and macrostructural statistics were then compiled into a single data frame using [A397](#).

**Segmentation of the optic radiation (OR).** To generate optic radiations segmented by estimates of visual field eccentricity in the Oxford University Choroideremia & Stargardt's Disease Dataset, ConTrack <sup>111</sup> tracking was implemented as [A252](#). 500,000 sample streamlines were generated using a step size of 1mm. Samples were then pruned using inclusion and exclusion waypoint ROIs following methodologies outlined in <sup>108,109</sup>.

**Segmentation of uncinate fasciculus (UF).** To assess the relationship between Uncinate tract-average quantitative anisotropy (QA) and fractional anisotropy (FA) and Early Life Stressors within two independent datasets (Healthy Brain Network, ABCD), the tract-average QA for the Left and Right Uncinates were computed from 42 participants from the HBN and the tract-average FA were computed from 1107 participants from the ABCD dataset. For the HBN dataset, a full tractography segmentation pipeline was used to preprocess the dMRI data and segment the uncinate fasciculus using [A423](#). Automatic fiber tracking was then performed to segment the uncinate fasciculus using default parameters and templates from a

population tractography atlas from the Human Connectome Project <sup>174</sup>. A threshold of 16 mm as the maximum allowed threshold for the shortest streamline distance was then applied to remove spurious streamlines. The whole tract average QA was then estimated. To probe stress exposure within the HBN dataset, we used the Negative Life Events Schedule (NLES), a 22-item questionnaire where participants were asked about the occurrence of different stressful life events. For the questions pertaining to early life stressors, the ABCD dataset was used. The tract-average FA for the Left and Right Uncinates were estimated using procedures described previously, then compared to the participant's life stressors behavioral measures by fitting a linear regression to the data.

**Structural networks:** Following tract segmentation, structural networks were generated using the multi-modal 180 cortical node atlas and the tractograms for each participant using MRTrix3's *tck2connectome*<sup>175</sup> functionality implemented as [A395](#). Connectomes were generated by computing the number of streamlines intersecting each ROI pairing in the 180 cortical node parcellation. Multiple adjacency matrices were generated, including count, density (i.e. count divided by the node volume of the ROI pairs), length, length density (i.e. length divided by the volume of the ROI pairs), and average and average density AD, FA, MD, RD, NDI, ODI, and ISOVF. Density matrices were generated using the *-invnodevol* option<sup>176</sup>. For non-count measures (length, AD, FA, MD, RD, NDI, ODI, ISOVF), the average measure across all streamlines connecting and ROI pair was computed using MRTrix3's *tck2scale* functionality using the *-precise* option<sup>177</sup> and the *-scale\_file* option in *tck2connectome*. These matrices can be thought of as the "average measure" adjacency matrices. These files were outputted as the 'raw' Datatype, and were converted to *conmat* Datatype using [A393](#). Connectivity matrices were then converted into the 'network' Datatype using functionality from python functionality implemented as [A335](#).

**Cortical & subcortical diffusion & morphometry mapping.** For the PING, HCP<sub>TR</sub>, HCP<sub>s1200</sub>, Cam-CAN, and Indiana University Acute Concussion datasets, DTI and NODDI (if available) measures were mapped to each participant's cortical white matter parcels following methods found in Fukutomi and colleagues using functions provided by Connectome Workbench<sup>93</sup> implemented as *brainlife.io* App [A379](#). A Gaussian smoothing kernel (FWHM = ~4mm,  $\sigma = 5/3$ mm) was applied along the axis normal to the midthickness surface, and DTI and NODDI measures were mapped using the *wb\_command* -volume-to-surface-mapping function. Freesurfer was used to map the average DTI and NODDI measures within each parcel using functionality from Connectome Workbench using [A389](#) and [A483](#). Measures of volume, surface area, and cortical thickness for each cortical parcel were computed using Freesurfer and [A464](#). Freesurfer was also used to generate parcel average DTI and NODDI measures for the subcortical segmentation (*aseg*) from Freesurfer using [A383](#). Measures of volume for each subcortical parcel were computed using Freesurfer and [A272](#).

**Resting-state Functional (rs-fMRI) preprocessing and functional connectivity matrix generation.** For the HCP<sub>TR</sub> and Cam-CAN datasets, unprocessed rs-fMRI datasets were preprocessed using fMRIPrep implemented as [A160](#). Briefly, fMRIPrep does the following preprocessing steps. First, individual images are aligned to a reference image for motion estimation and correction using *mcfliirt* from FSL. Next, slice timing correction is performed in which all slices are realigned in time to the middle of each TR using *3dTShift* from AFNI. Spatial distortions are then corrected using field map estimations. Finally, the fMRI data is aligned to the structural T1w image for each participant. Default parameters provided by fMRIPrep were used. For a subset of analyses involving the HCP Test and Retest datasets, the preprocessed rs-fMRI datasets provided by the HCP consortium were used. Following preprocessing via fMRIPrep for the volume data, connectivity matrices were generated using the Yeo17 parcellation and [A369](#) and [A532](#). Within-network functional connectivity for the 17 canonical resting state networks was computed by computing the average functional connectivity values within all of the nodes belonging to a single network. These estimates were used for subsequent analyses.

**Resting-state Functional (rs-fMRI) gradient processing.** For the HCP<sub>TR</sub> and Cam-CAN datasets, unprocessed rs-fMRI data from HCP Test and the Cam-CAN datasets were preprocessed using fMRIPrep implemented as [A267](#). Within this app, the same preprocessing steps are undertaken as in [A160](#), except for an additional volume-to-surface mapping using *mri\_vol2surf* from Freesurfer. The surface-based outputs were then used to compute gradients following methodologies outlined in <sup>96</sup> for each participant in the HCP<sub>s1200</sub>, HCP<sub>TR</sub>, and Cam-CAN datasets using [A574](#) using diffusion embedding <sup>178</sup> and functions provided by BrainSpace <sup>179</sup>. More specifically, connectivity matrices were computed from surface vertex values within each node of the Schaffer 1,000 parcellation <sup>180</sup>. Cosine similarity was then computed to create an affinity matrix to capture inter-area similarity. Dimensionality reduction is then used to identify the primary gradients. A normalized-angle kernel was used to create the affinity matrix, from which two primary components were identified. Gradients were then aligned across all participants using a Procrustes alignment and joined embedding procedure <sup>96</sup>. Values from the primary gradient and the cosine distance used to generate the affinity matrices were used for subsequent analyses.

**Magnetoencephalography (MEG) processing.** For some analyses, raw resting-state magnetoencephalography (rs-MEG) time series data from the Cam-CAN dataset was filtered using a Maxwell filter implemented as [A476](#) and median split using [A529](#). For the remainder of the analyses, filtered data provided by the Cam-CAN dataset was used. For all MEG data, power-spectrum density profiles (PSD) were estimated using functionality provided by MNE-Python <sup>181</sup> implemented as [A530](#). Following PSD estimation, peak alpha frequency was estimated using [A531](#). Finally, PSD profiles were averaged across all

nodes within each of the canonical lobes (frontal, parietal, occipital, temporal) using [A599](#). Measures of power-spectrum density and peak alpha frequency were used for all subsequent analyses.

## **DATA AVAILABILITY.**

All data derived and described in this paper are made available via the *brainlife.io* platform as “Publications”. User data agreements are required for some projects, like data from the HCP, Cam-CAN, PING, ABCD, and HBN datasets. The *Indiana University Acute Concussion Dataset* and the *Oxford University Choroideremia & Stargardt’s Disease Dataset* are parts of ongoing research projects and are not being released at this current time. All other datasets are made freely available via the *brainlife.io* platform. See supplementary Table 6 for the *brainlife.io/pubs* [we have added one example data record (<https://doi.org/10.25663/brainlife.pub.40>) for the review process <the DOIs for the remaining data records will be added at publication>].

## **CODE AVAILABILITY.**

As part of the article we are describing a total of 9 platform components. All components are made publicly available open source under MIT License. All the software for the platform components is listed in Supplementary Table 1. In addition, we share the code used for the statistical analyses as Jupyter Notebooks (Supplementary Table 2). Finally, the Apps used and tested in this article are listed in Supplementary Table 3.

## **ACKNOWLEDGMENTS.**

The *brainlife.io* project, development and operations were supported by awards to Franco Pestilli; U.S. National Science Foundation (NSF) awards 1916518, 1912270, 1636893, and 1734853; U.S. National Institutes of Health awards (NIH) R01MH126699, R01EB030896, and R01EB029272; The Wellcome Trust award 226486/Z/22/Z; A Microsoft Investigator Fellowship; A gift from the Kavli Foundation. Additional funding was provided to support data collection used by the team, research that used *brainlife.io* or infrastructure that supported the platform: NSF awards 2004877 (Sophia Vinci-Booher), 1541335 and 2232628 (Shawn McKee), 1445604 and 2005506 (David Hancock), 1341698 (Michael Norman), 1928224 (Michael Norman), 1445606 (Shawn Brown), 1928147 (Sergiu Sanalevici). NIH awards 1U54MH091657 (HCP data, PIs David Van Essen and Kamil Ugurbil), U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147 (ABCD Study, multiple PIs). Multiple philanthropic contributions to the HBN (Michael Milham).

## **ETHICS DECLARATIONS.**

The authors declare no competing financial interests.

## REFERENCES

1. Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
2. Birur, B., Kraguljac, N. V., Shelton, R. C. & Lahti, A. C. Brain structure, function, and neurochemistry in schizophrenia and bipolar disorder—a systematic review of the magnetic resonance neuroimaging literature. *npj Schizophrenia* **3**, 1–15 (2017).
3. Pando-Naude, V. *et al.* Gray and white matter morphology in substance use disorders: a neuroimaging systematic review and meta-analysis. *Transl. Psychiatry* **11**, 29 (2021).
4. Ahmadzadeh, M., Christie, G. J., Cosco, T. D. & Moreno, S. Neuroimaging and analytical methods for studying the pathways from mild cognitive impairment to Alzheimer’s disease: protocol for a rapid systematic review. *Systematic Reviews* vol. 9 Preprint at <https://doi.org/10.1186/s13643-020-01332-7> (2020).
5. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
6. Naselaris, T., Allen, E. & Kay, K. Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences* **40**, 45–51 (2021).
7. Gau, R. *et al.* Brainhack: Developing a culture of open, inclusive, community-driven neuroscience. *Neuron* **109**, 1769–1775 (2021).
8. Thompson, P. M. *et al.* ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl. Psychiatry* **10**, 100 (2020).
9. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
10. Camerer, C. F. *et al.* Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
11. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav* **2**, 637–644 (2018).
12. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

13. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).
14. Klein, R. A. *et al.* Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science* **1**, 443–490 (2018).
15. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1**, 1–9 (2017).
16. Nissen, S. B., Magidson, T., Gross, K. & Bergstrom, C. T. Publication bias and the canonization of false facts. *Elife* **5**, (2016).
17. Stupples, A., Singerman, D. & Celi, L. A. The reproducibility crisis in the age of digital medicine. *NPJ Digit Med* **2**, 2 (2019).
18. National Academies of Sciences, Engineering, and Medicine *et al.* *Enhancing Scientific Reproducibility in Biomedical Research Through Transparent Reporting: Proceedings of a Workshop*. (National Academies Press, 2020). doi:10.17226/25627.
19. McNutt, M. Reproducibility. *Science* **343**, 229 (2014).
20. Stodden, V. *et al.* Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241 (2016).
21. Nosek, B. A. & Errington, T. M. Making sense of replications. *Elife* **6**, (2017).
22. McKinney, S. M. *et al.* Reply to: Transparency and reproducibility in artificial intelligence. *Nature* **586**, E17–E18 (2020).
23. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16 (2020).
24. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
25. Van Essen, D. C. *et al.* The Human Connectome Project: a data acquisition perspective. *Neuroimage* **62**, 2222–2231 (2012).
26. Taylor, J. R. *et al.* The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage* **144**, 262–269 (2017).
27. Shafto, M. A. *et al.* The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol.* **14**, 204 (2014).

28. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
29. Karcher, N. R. & Barch, D. M. The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology* **46**, 131–142 (2021).
30. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
31. Alexander, L. M. *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci Data* **4**, 170181 (2017).
32. Jernigan, T. L. *et al.* The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository. *Neuroimage* **124**, 1149–1154 (2016).
33. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
34. Markiewicz, C. J. *et al.* The OpenNeuro resource for sharing of neuroscience data. *Elife* **10**, (2021).
35. Milham, M. P. *et al.* Assessment of the impact of shared brain imaging data on the scientific literature. *Nat. Commun.* **9**, 2818 (2018).
36. Willeminck, M. J. *et al.* Preparing Medical Imaging Data for Machine Learning. *Radiology* **295**, 4–15 (2020).
37. Zeng, N., Zuo, S., Zheng, G., Ou, Y. & Tong, T. Editorial: Artificial Intelligence for Medical Image Analysis of Neuroimaging Data. *Front. Neurosci.* **14**, 480 (2020).
38. NeuroImage. <https://www.sciencedirect.com/journal/neuroimage/special-issue/101CWG577G3>.
39. Poldrack, R. A., Gorgolewski, K. J. & Varoquaux, G. Computational and Informatic Advances for Reproducible Data Analysis in Neuroimaging. *Annu. Rev. Biomed. Data Sci.* (2019)  
doi:10.1146/annurev-biodatasci-072018-021237.
40. Niso, G. *et al.* Open and reproducible neuroimaging: From study inception to publication. *Neuroimage* **263**, 119623 (2022).
41. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).
42. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *Neuroimage* **62**,

- 782–790 (2012).
43. Woolrich, M. W. *et al.* Bayesian analysis of neuroimaging data in FSL. *Neuroimage* **45**, S173–86 (2009).
  44. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23 Suppl 1**, S208–19 (2004).
  45. Dale, A., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *Neuroimage* **9**, 179–194 (1999).
  46. Fischl, B., Sereno, M. I. & Dale, A. Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. *Neuroimage* **9**, 195–207 (1999).
  47. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).
  48. Fischl, B., Liu, A. & Dale, A. M. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Medical Imaging* **20**, 70–80 (2001).
  49. Fischl, B. *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).
  50. Fischl, B. *et al.* Sequence-independent segmentation of magnetic resonance images. *Neuroimage* **23**, S69–S84 (2004).
  51. Fischl, B., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
  52. Fischl, B. *et al.* Automatically Parcellating the Human Cerebral Cortex. *Cereb. Cortex* **14**, 11–22 (2004).
  53. Han, X. *et al.* Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *Neuroimage* **32**, 180–194 (2006).
  54. Jovicich, J. *et al.* Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *Neuroimage* **30**, 436–443 (2006).
  55. Kuperberg, G. R. *et al.* Regionally localized thinning of the cerebral cortex in Schizophrenia. *Arch. Gen. Psychiatry* **60**, 878–888 (2003).
  56. Reuter, M., Schmansky, N. J., Rosas, H. D. & Fischl, B. Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. *Neuroimage* **61**, 1402–1418 (2012).

57. Reuter, M. & Fischl, B. Avoiding Asymmetry-Induced Bias in Longitudinal Image Processing. *Neuroimage* **57**, 19–21 (2011).
58. Reuter, M., Rosas, H. D. & Fischl, B. Highly Accurate Inverse Consistent Registration: A Robust Approach. *Neuroimage* **53**, 1181–1196 (2010).
59. Salat, D. *et al.* Thinning of the cerebral cortex in aging. *Cereb. Cortex* **14**, 721–730 (2004).
60. Segonne, F. *et al.* A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**, 1060–1075 (2004).
61. Segonne, F., Pacheco, J. & Fischl, B. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans. Med. Imaging* **26**, 518–529 (2007).
62. Brett, M. *et al.* *nipy/nibabel*: (2022). doi:10.5281/zenodo.6658382.
63. Tournier, J.-D., Calamante, F. & Connelly, A. MRtrix: Diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* **22**, 53–66 (2012).
64. Tournier, J.-D. *et al.* MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage* **202**, 116137 (2019).
65. Garyfallidis, E. *et al.* Dipy, a library for the analysis of diffusion MRI data. *Front. Neuroinform.* **8**, 8 (2014).
66. Yeh, F.-C. Shape analysis of the human association pathways. *Neuroimage* **223**, 117329 (2020).
67. Yeh, F.-C. Population-based tract-to-region connectome of the human brain and its hierarchical topology. *Nat. Commun.* **13**, 1–13 (2022).
68. Yeh, F.-C. *et al.* Automatic Removal of False Connections in Diffusion MRI Tractography Using Topology-Informed Pruning (TIP). *Neurotherapeutics* **16**, 52–58 (2019).
69. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
70. Cieslak, M. *et al.* QSIPrep: an integrative platform for preprocessing and reconstructing diffusion MRI data. *Nat. Methods* **18**, 775–778 (2021).
71. Esteban, O. *et al.* MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* **12**, e0184661 (2017).
72. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
73. Adebimpe, A. *et al.* ASLPrep: a platform for processing of arterial spin labeled MRI and quantification of



- regional brain perfusion. *Nat. Methods* **19**, 683–686 (2022).
74. Halchenko, Y. *et al.* DataLad: distributed system for joint management of code, data, and their relationship. *J. Open Source Softw.* **6**, 3262 (2021).
75. Paret, C. *et al.* Survey on Open Science Practices in Functional Neuroimaging. *Neuroimage* **257**, 119306 (2022).
76. Maina, M. B. *et al.* Two decades of neuroscience publication trends in Africa. *Nat. Commun.* **12**, 3429 (2021).
77. Valenzuela-Toro, A. M. & Viglino, M. How Latin American researchers suffer in science. *Nature Publishing Group UK* <http://dx.doi.org/10.1038/d41586-021-02601-8> (2021) doi:10.1038/d41586-021-02601-8.
78. McKee, S. *et al.* OSiRIS: a distributed Ceph deployment using software defined networking for multi-institutional research. *J. Phys. Conf. Ser.* **898**, 062045 (2017).
79. Stanzione, D. *et al.* Frontera: The Evolution of Leadership Computing at the National Science Foundation. in *Practice and Experience in Advanced Research Computing* 106–111 (Association for Computing Machinery, 2020). doi:10.1145/3311790.3396656.
80. Stewart, C. A. *et al.* Jetstream: A self-provisioned, scalable science and engineering cloud environment. (2015) doi:10.1145/2792745.2792774.
81. Nooner, K. B. *et al.* The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Front. Neurosci.* **6**, 152 (2012).
82. Tobe, R. H. *et al.* A longitudinal resource for studying connectome development and its psychiatric associations during childhood. *Sci Data* **9**, 300 (2022).
83. Di Martino, A. *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
84. Developers, S. *SingularityCE* 3.8.3. (2021). doi:10.5281/zenodo.5564915.
85. Merkel, D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* **2014**, 2 (2014).
86. Dean, J. & Ghemawat, S. MapReduce: simplified data processing on large clusters. *Commun. ACM* **51**, 107–113 (2008).
87. Kluyver, T. *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. in

- Positioning and Power in Academic Publishing: Players, Agents and Agendas* (eds. Loizides, F. & Schmidt, B.) 87–90 (IOS Press, 2016). doi: Kluyver, Thomas, Ragan-Kelley, Benjamin, Pérez, Fernando, Granger, Brian, Bussonnier, Matthias, Frederic, Jonathan, Kelley, Kyle, Hamrick, Jessica, Grout, Jason, Corlay, Sylvain, Ivanov, Paul, Avila, Damián, Abdalla, Safia, Willing, Carol and Jupyter development team, (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. Loizides, Fernando and Schmidt, Birgit (eds.) In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press. pp. 87-90 . (doi:10.3233/978-1-61499-649-1-87 <<http://dx.doi.org/10.3233/978-1-61499-649-1-87>>). .
88. Perez, F. & Granger, B. E. IPython: A System for Interactive Scientific Computing. *Comput. Sci. Eng.* **9**, 21–29 (2007).
  89. Wickham, H. Tidy Data. *J. Stat. Softw.* **59**, 1–23 (2014).
  90. Avesani, P. *et al.* The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. *Sci Data* **6**, 69 (2019).
  91. National Academies of Sciences, Engineering *et al.* *Understanding Reproducibility and Replicability*. (National Academies Press (US), 2019).
  92. Kelley, T. L. Interpretation of educational measurements. **353**, (1927).
  93. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
  94. Yeo, B. T. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
  95. Bethlehem, R. A. I. *et al.* Dispersion of functional gradients across the adult lifespan. *Neuroimage* **222**, 117299 (2020).
  96. Margulies, D. S. *et al.* Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12574–12579 (2016).
  97. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
  98. Noble, S., Scheinost, D. & Constable, R. T. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage* **203**, 116157 (2019).

99. McPherson, B. C. & Pestilli, F. A single mode of population covariation associates brain networks structure and behavior and predicts individual subjects' age. *Commun Biol* **4**, 943 (2021).
100. Betzel, R. F. *et al.* Changes in structural and functional connectivity among resting-state networks across the human lifespan. *Neuroimage* **102 Pt 2**, 345–357 (2014).
101. López-Vicente, M. *et al.* White matter microstructure correlates of age, sex, handedness and motor ability in a population-based sample of 3031 school-age children. *Neuroimage* **227**, 117643 (2021).
102. Lebel, C. & Beaulieu, C. Longitudinal development of human brain wiring continues from childhood into adulthood. *J. Neurosci.* **31**, 10937–10947 (2011).
103. Bethlehem, R. A. I. *et al.* Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).
104. Fukutomi, H. *et al.* Neurite imaging reveals microstructural variations in human cerebral cortical gray matter. *Neuroimage* **182**, 488–499 (2018).
105. Hanson, J. L., Knodt, A. R., Brigidi, B. D. & Hariri, A. R. Lower structural integrity of the uncinate fasciculus is associated with a history of child maltreatment and future psychological vulnerability to stress. *Dev. Psychopathol.* **27**, 1611–1619 (2015).
106. McKee, A. C., Daneshvar, D. H., Alvarez, V. E. & Stein, T. D. The neuropathology of sport. *Acta Neuropathol.* **127**, 29–51 (2014).
107. Hanekamp, S. *et al.* White matter alterations in glaucoma and monocular blindness differ outside the visual system. *Sci. Rep.* **11**, 6866 (2021).
108. Yoshimine, S. *et al.* Age-related macular degeneration affects the optic radiation white matter projecting to locations of retinal damage. *Brain Struct. Funct.* **223**, 3889–3900 (2018).
109. Ogawa, S. *et al.* White matter consequences of retinal receptor and ganglion cell damage. *Invest. Ophthalmol. Vis. Sci.* **55**, 6976–6986 (2014).
110. Malania, M., Konrad, J., Jäggle, H., Werner, J. S. & Greenlee, M. W. Compromised Integrity of Central Visual Pathways in Patients With Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **58**, 2939–2947 (2017).
111. Sherbondy, A. J., Dougherty, R. F., Ben-Shachar, M., Napel, S. & Wandell, B. A. ConTrack: finding the most likely pathways between brain regions using diffusion tractography. *J. Vis.* **8**, 15.1–16 (2008).
112. Yeatman, J. D., Dougherty, R. F., Myall, N. J., Wandell, B. A. & Feldman, H. M. Tract profiles of white matter

- properties: automating fiber-tract quantification. *PLoS One* **7**, e49790 (2012).
113. Aydogan, D. B. & Shi, Y. Parallel Transport Tractography. *IEEE Trans. Med. Imaging* **40**, 635–647 (2021).
114. Baran, D. & Shi, Y. A novel fiber-tracking algorithm using parallel transport frames. in *ISMRM* (unknown, 2019).
115. Yanni, S. E. *et al.* Normative reference ranges for the retinal nerve fiber layer, macula, and retinal layer thicknesses in children. *Am. J. Ophthalmol.* **155**, 354–360.e1 (2013).
116. Esteban, O. *et al.* Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines. *Sci Data* **6**, 30 (2019).
117. Calhoun, V. D., Liu, J. & Adali, T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage* **45**, S163–72 (2009).
118. Hériché, J.-K., Alexander, S. & Ellenberg, J. Integrating Imaging and Omics: Computational Methods and Challenges. *Annu. Rev. Biomed. Data Sci.* **2**, 175–197 (2019).
119. Shen, L. & Thompson, P. M. Brain Imaging Genomics: Integrated Analysis and Machine Learning. *Proc. IEEE Inst. Electr. Electron. Eng.* **108**, 125–162 (2020).
120. Deslauriers-Gauthier, S. *et al.* White matter information flow mapping from diffusion MRI and EEG. *Neuroimage* **201**, 116017 (2019).
121. Wirsich, J., Amico, E., Giraud, A.-L., Goñi, J. & Sadaghiani, S. Multi-timescale hybrid components of the functional brain connectome: A bimodal EEG-fMRI decomposition. *Netw Neurosci* **4**, 658–677 (2020).
122. Engemann, D. A. *et al.* Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *Elife* **9**, (2020).
123. Eke, D. O. *et al.* International data governance for neuroscience. *Neuron* (2021)  
doi:10.1016/j.neuron.2021.11.017.
124. Eke, D. *et al.* Pseudonymisation of neuroimages and data protection: Increasing access to data while retaining scientific utility. *Neuroimage: Reports* **1**, 100053 (2021).
125. Kozlov, M. NIH issues a seismic mandate: share data publicly. *Nature Publishing Group UK*  
<http://dx.doi.org/10.1038/d41586-022-00402-1> (2022) doi:10.1038/d41586-022-00402-1.
126. NOT-OD-21-013: Final NIH Policy for Data Management and Sharing.

<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>.

127. Data Gravity – in the Clouds. *Data Gravititas* <https://datagravititas.com/2010/12/07/data-gravity-in-the-clouds/> (2010).
128. Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 4734–4739 (2010).
129. Milham, M. P. Open neuroscience solutions for the connectome-wide association era. *Neuron* **73**, 214–218 (2012).
130. Halchenko, Y. *et al.* *dandi/dandi-cli: 0.46.2.* (2022). doi:10.5281/zenodo.7041535.
131. Hider, R., Jr *et al.* The Brain Observatory Storage Service and Database (BossDB): A Cloud-Native Approach for Petascale Neuroscience Discovery. *Front. Neuroinform.* **16**, 828787 (2022).
132. Kennedy, D. N., Haselgrove, C., Riehl, J., Preuss, N. & Buccigrossi, R. The NITRC image repository. *Neuroimage* **124**, 1069–1073 (2016).
133. Koike, S. *et al.* Brain/MINDS beyond human brain MRI project: A protocol for multi-level harmonization across brain disorders throughout the lifespan. *Neuroimage Clin* **30**, 102600 (2021).
134. Das, S., Zijdenbos, A. P., Harlap, J., Vins, D. & Evans, A. C. LORIS: a web-based data management system for multi-center studies. *Front. Neuroinform.* **5**, 37 (2011).
135. Dockès, J. *et al.* NeuroQuery, comprehensive meta-analysis of human brain mapping. *Elife* **9**, (2020).
136. de la Vega, A. *et al.* Neuroscout, a unified platform for generalizable and reproducible fMRI research. *Elife* **11**, (2022).
137. Sherif, T. *et al.* CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. *Front. Neuroinform.* **8**, 54 (2014).
138. Renton, A. I. *et al.* Neurodesk: An accessible, flexible, and portable data analysis environment for reproducible neuroimaging. *bioRxiv* 2022.12.23.521691 (2022) doi:10.1101/2022.12.23.521691.
139. Marcus, D. S., Olsen, T. R., Ramaratnam, M. & Buckner, R. L. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* **5**, 11–34 (2007).
140. Delorme, A. *et al.* NEMAR: an open access data, tools and compute resource operating on

- neuroelectromagnetic data. *Database* **2022**, (2022).
141. Schirner, M. *et al.* Brain simulation as a cloud service: The Virtual Brain on EBRAINS. *Neuroimage* **251**, 118973 (2022).
142. Rex, D. E., Ma, J. Q. & Toga, A. W. The LONI Pipeline Processing Environment. *Neuroimage* **19**, 1033–1048 (2003).
143. Elam, J. S. *et al.* The Human Connectome Project: A Retrospective. *Neuroimage* 118543 (2021) doi:10.1016/j.neuroimage.2021.118543.
144. International Brain Laboratory *et al.* A modular architecture for organizing, processing and sharing neurophysiology data. *Nat. Methods* (2023) doi:10.1038/s41592-022-01742-6.
145. Plis, S. M. *et al.* COINSTAC: A Privacy Enabled Model and Prototype for Leveraging and Processing Decentralized Brain Imaging Data. *Front. Neurosci.* **10**, 365 (2016).
146. Harding, R. J., Bermudez, P., Beauvais, M., Bellec, P. & Evans, A. C. The Canadian Open Neuroscience Platform – An Open Science Framework for the Neuroscience Community. (2022) doi:10.31219/osf.io/eh349.
147. Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J. Am. Med. Inform. Assoc.* **22**, 1148–1152 (2015).
148. Maumet, C. *et al.* Sharing brain mapping statistical results with the neuroimaging data model. *Sci Data* **3**, 160102 (2016).
149. Van Essen, D. C. *et al.* The Human Connectome Project: a data acquisition perspective. *Neuroimage* **62**, 2222–2231 (2012).
150. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).
151. Caron, B. *et al.* Collegiate athlete brain data for white matter mapping and network neuroscience. *Sci Data* **8**, 56 (2021).
152. Jernigan, T. L. *et al.* The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository. *Neuroimage* **124**, 1149–1154 (2016).
153. Yushkevich, P. A. *et al.* Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: towards a harmonized segmentation protocol. *Neuroimage* **111**,

- 526–541 (2015).
154. Yushkevich, P. A. *et al.* Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Hum. Brain Mapp.* **36**, 258–287 (2015).
155. Benson, N. C. *et al.* The retinotopic organization of striate cortex is well predicted by surface topology. *Curr. Biol.* **22**, 2081–2085 (2012).
156. Benson, N. C., Butt, O. H., Brainard, D. H. & Aguirre, G. K. Correction of distortion in flattened representations of the cortical surface allows prediction of V1-V3 functional organization from anatomy. *PLoS Comput. Biol.* **10**, e1003538 (2014).
157. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
158. Ades-Aron, B. *et al.* Evaluation of the accuracy and precision of the diffusion parameter Estimation with Gibbs and Noise removal pipeline. *Neuroimage* **183**, 532–543 (2018).
159. Andersson, J. L. R., Skare, S. & Ashburner, J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* **20**, 870–888 (2003).
160. Andersson, J. L. R. & Sotiropoulos, S. N. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* **125**, 1063–1078 (2016).
161. Andersson, J. L. R., Graham, M. S., Zsoldos, E. & Sotiropoulos, S. N. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. *Neuroimage* **141**, 556–572 (2016).
162. Andersson, J. L. R., Graham, M. S., Drobnyak, I., Zhang, H. & Campbell, J. Susceptibility-induced distortion that varies due to motion: Correction in diffusion MR without acquiring additional data. *Neuroimage* **171**, 277–295 (2018).
163. Andersson, J. L. R. *et al.* Towards a comprehensive framework for movement and distortion correction of diffusion MR images: Within volume movement. *Neuroimage* **152**, 450–466 (2017).
164. Jeurissen, B., Leemans, A. & Sijbers, J. Automated correction of improperly rotated diffusion gradient orientations in diffusion weighted MRI. *Med. Image Anal.* **18**, 953–962 (2014).
165. Tustison, N. J. *et al.* Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements.

- Neuroimage* **99**, 166–179 (2014).
166. Veraart, J. *et al.* Denoising of diffusion MRI using random matrix theory. *Neuroimage* **142**, 394–406 (2016).
167. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**, 143–156 (2001).
168. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
169. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* **48**, 63–72 (2009).
170. Pierpaoli, C., Jezzard, P., Basser, P. J., Barnett, A. & Di Chiro, G. Diffusion tensor MR imaging of the human brain. *Radiology* **201**, 637–648 (1996).
171. Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* **61**, 1000–1016 (2012).
172. Daducci, A. *et al.* Accelerated Microstructure Imaging via Convex Optimization (AMICO) from diffusion MRI data. *Neuroimage* **105**, 32–44 (2015).
173. Yeh, F.-C., Wedeen, V. J. & Tseng, W.-Y. I. Generalized q-sampling imaging. *IEEE Trans. Med. Imaging* **29**, 1626–1635 (2010).
174. Yeh, F.-C. *et al.* Population-averaged atlas of the macroscale human structural connectome and its network topology. *Neuroimage* **178**, 57–68 (2018).
175. Smith, R. E., Tournier, J.-D., Calamante, F. & Connelly, A. The effects of SIFT on the reproducibility and biological accuracy of the structural connectome. *Neuroimage* **104**, 253–265 (2015).
176. Hagmann, P. *et al.* Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**, e159 (2008).
177. Smith, R. E., Tournier, J.-D., Calamante, F. & Connelly, A. SIFT: Spherical-deconvolution informed filtering of tractograms. *Neuroimage* **67**, 298–312 (2013).
178. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7426–7431 (2005).
179. Vos de Wael, R. *et al.* BrainSpace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Commun Biol* **3**, 103 (2020).



180. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).

181. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7**, 267 (2013).

## Supplementary material.

### ***brainlife.io*: A decentralized and open source cloud platform to support big data neuroscience research**

Hayashi, S.\*, Caron, B.\*, Heinsfeld, A.S., Vinci-Booher, S., McPherson, B.C., Bullock, D., Berto, G., Niso, J.G., Hanekamp, S., Levitas, D., Kitchell, L., Leong, J., Silva, F. N., Koudoro, S., Willis, H., Jolly, J., Pisner, D., Zuidema, T., Kurzwaski, J., Mikellidou, K., Bussalb, A., Rorden, C., Victory, C., Bhatia, D., Aydogan, D.B., Yeh, F.C., Delogu, F., Guaje, J., Veraart, J., Fischer, J., Faskowitz, J., Chaumon, M., Fabrega, R., Hunt, D., McKee, S., Brown, S.T., Heyman, S., Iacovella, V., Mejia, A., Marinazzo, D., Craddock, C., Olivetti, E., Hanson, J., Avesani, P., Garyfallidis, E., Stanzione, D., Carson, J.P., Henschel, R., Hancock, D.Y., Stewart, C.A., Schnyer, D., Eke, D., Poldrack, R.A., George, N., Bridge, H., Sani, I., Freiwald, W., Puce, A., Port, N., and Pestilli, F.

**Competing interests.** The authors declare no competing financial interests.

**Corresponding authors.** Franco Pestilli [pestilli@utexas.edu](mailto:pestilli@utexas.edu)

**Contribution.** S.H. implemented the *brainlife.io* services. B.C. wrote the data analysis code, performed the large-scale experiments and prepared the figures and associated text. A.S.H. improved and implemented some of the services. F.J., C.C., C.C.A., D.H., D.S., D.P., L.K., J.L., C.R., F.N.S., H.W., J.J., Z.T., K.J., S.K., N.A., V.C., B.D., A.D.B., F.D. G.J., S.H., provided assets. All authors edited the manuscript. F.P. invented, designed, and directed *brainlife.io*, and wrote the paper and designed all the experiments, and figures. \*shared first authors contribution.

#### **ABSTRACT**

Neuroscience research has expanded dramatically over the past 30 years by advancing standardization and tool development to support rigor and transparency. Consequently, the complexity of the data pipeline has also increased, hindering access to FAIR data analysis to portions of the worldwide research community. *brainlife.io* was developed to reduce these burdens and democratize modern neuroscience research across institutions and career levels. Using community software and hardware infrastructure, the platform provides open-source data standardization, management, visualization, and processing and simplifies the data pipeline. *brainlife.io* automatically tracks the provenance history of thousands of data objects, supporting simplicity, efficiency, and transparency in neuroscience research. Here *brainlife.io*'s technology and data services are described and evaluated for validity, reliability, reproducibility, replicability, and scientific utility. Using data from 4 modalities and 3,200 participants, we demonstrate that *brainlife.io*'s services produce outputs that adhere to best practices in modern neuroscience research.

|  |           |
|--|-----------|
| <b>ABSTRACT</b>  | <b>1</b>  |
| <b>SUPPLEMENTAL RESULTS</b>  | <b>2</b>  |
| <b>Supplemental platform architecture</b>  | <b>2</b>  |
| Supplemental Table 1: Platform services serving the brainlife.io platform.                   | 5         |
| Supplemental Table 2: Jupyter notebooks for analyses performed.                              | 5         |
| Supplemental Table 3: Preprocessing Apps used for the experiments.                           | 7         |
| Developing processing Apps for the platform  | 9         |
| Using the platform   | 10        |
| End-to-end reproducible scientific workflow  | 23        |
| <b>Supplemental platform evaluation</b>  | <b>25</b> |
| Supplemental platform utilization  | 25        |
| Supplemental platform testing  | 27        |
| Supplementary Table 4. Validity and reliability correlation tables.                          | 33        |
| Supplemental platform utility for scientific applications                                    | 34        |
| Supplemental replication and generalization  | 35        |
| Supplemental to detecting disease  | 36        |
| Supplement to quality control at scale   | 37        |
| <b>Public services for promoting transparency and data gravity in neuroscience research.</b> | <b>39</b> |
| Supplemental Table 5: Resources for data storage, archiving, and computational analysis.     | 39        |

## SUPPLEMENTAL RESULTS

### Supplemental platform architecture

*brainlife.io* is a composition of microservices, including authentication, preprocessing, warehousing, event handling, and auditing. Microservices are handled by a meta-orchestration workflow system, Amaretti (**Fig. S2a,b**, and **Table S1**). Amaretti can deploy computational jobs on high-performance compute clusters and cloud systems. Both jobs needed for platform operations and data analysis are handled by Amaretti. Amaretti is central to *brainlife.io*'s opportunistic computing approach, i.e., the ability to use donated storage or computing resources. Amaretti allows secure access to either clouds or supercomputers managing platform task scheduling, data transfer, and job submission and monitoring. Amaretti's core concepts are data- and resource awareness, i.e., data products or compute resources are specified as objects that the platform has explicit awareness of (e.g., the platform can dock datatypes, or compute resources; **Fig. S2b**). For example, users and resource managers can register a computing resource, making it available via *brainlife.io* either privately (to a specified set of users) or widely (to the entire platform users base). A variety of resource architectures and job submission systems have been tested and docked using Amaretti so far, including SLURM, PBS, OSG Engine, and CONDOR. Currently, Amaretti is hosted by a public cloud<sup>1,2</sup> and connected to major data centers (via [access-ci.org](https://access-ci.org); see **Fig. S2**) and commercial clouds.

Data processing on *brainlife.io* utilizes an object-oriented service model, based on micro workflows. Apps and datatypes work together to allow smart docking and awareness (**Fig. S2a, b**, and **c**; **Fig. S2b**). Apps are modular, composable processing units comprising either full pipelines<sup>3-23</sup> or small steps within a larger data-processing workflow. Apps are written in a variety of languages following a lightweight specification ([github.com/brainlife/abcd-spec](https://github.com/brainlife/abcd-spec)) and using containerization technology<sup>24,25</sup>. Containerization allows deployment on various compute resource architectures ([hub.docker.com/u/brainlife](https://hub.docker.com/u/brainlife)). Apps code is hosted on github.com. Code must be first registered on *brainlife.io* in order to become an App. An App registration process guides

developers to map both input and output data objects to *brainlife.io* datatypes via a graphical interface. For security reasons, platform administrator approval is required to allow Apps on compute resources. A DOI<sup>26-28</sup> is issued for registered Apps to support scientific transparency and credit assignment to developers<sup>29-39</sup>. App specification requires developers to provide an informative readme file on GitHub with proper citations to software and funding used for the App (Fig. S3a). After registration, platform users can access Apps via a graphical (GUI) or command line interface (CLI). Apps can run on multiple resources, and Amaretti has methods for matching Apps to resources based on criteria such as geolocation, performance profiles, and resource queue length.

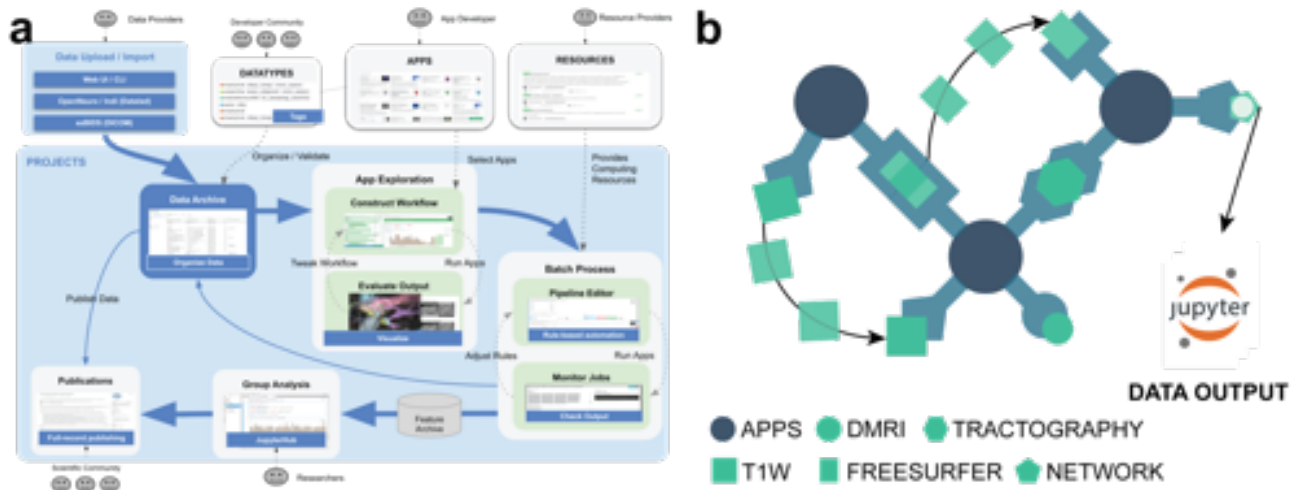
Apps on *brainlife.io* are data-aware and can automatically identify datasets, dock them or send them elsewhere for processing. This is because data objects are stored using predefined formats – datatypes. Datatypes allow App concatenation and automated pipelining (Fig. 2c; [brainlife.io/datatypes](https://brainlife.io/datatypes)). Datatypes comprise collections of files and folders organized into *.tar* archives to limit the number of inodes needed for storage. A platform-side datatype validation service ([github.com/brainlife/?q=validator-](https://github.com/brainlife/?q=validator-)) assures that datatypes comply with their definition. Data are physically stored using S3-like storage buckets organized following the pattern: `<s3bucketName>/<projectID>/<datasetID>.tar`. Buckets can live in multiple geolocations, so as to help with international requirements<sup>40</sup> (Fig. 2b). Datatypes comply with BIDS<sup>41</sup> (if the standard is defined for the data objects).

Data management is centered around Projects and supported by a databasing and warehousing system ([github.com/brainlife/warehouse](https://github.com/brainlife/warehouse)). Projects are the “one-stop-shop” for data management, processing, analysis, visualization, and publication (Fig. S3b). Projects are created independently by users and are private by default, but can be made public within the *brainlife.io* platform. Projects provide stratified access control mechanisms, and data user agreements can be added to the landing page (see Video S1). A project can be populated with data using several options (Fig. 2d). Major archives and data repositories are docked by *brainlife.io*<sup>42</sup> (see Fig. 2b). Noticeable examples are OpenNeuro.org<sup>43</sup>, and the Nathan-Kline data-sharing project<sup>44-46</sup>. Datasets can be seamlessly imported into *brainlife.io* Projects via the portal [brainlife.io/datasets](https://brainlife.io/datasets) (see Video S2 and Video S3). MRI, EEG, and MEG files (e.g., DICOM, *.fif*, *.ctf*) can also be uploaded directly using either a GUI (Video S4) or CLI (Video S5). A DICOM to BIDS conversion service has also been developed for MRI data standardization and importing into Projects ([brainlife.io/ezbids](https://brainlife.io/ezbids); see Table S1 and Video S6).

The data workflow in *brainlife.io* reduces the complexity of the neuroimaging processing pipeline into two steps akin to the MapReduce algorithm<sup>47</sup>. An initial *map step* preprocesses data objects asynchronously, in parallel using Apps, so as to extract features of interest, such as functional or white matter maps, or time series data (Fig. 2d). During the *map step*, datatypes and Apps are synchronized and moved across available compute resources automatically, as optimized by Amaretti. Apps process data objects automatically and in parallel across study participants in a Project. A dedicated web interface exists to explore sequences of Apps and optimize the parameters for each data set (Video S7). In addition, App sequences can be composed using a Pipeline builder interface (Video S8).

The *map step* is followed by a *reduce step*. Features extracted using Apps are synchronized, brought together, and made available to Jupyter notebooks<sup>48,49</sup> for statistical analysis and to generate figures for scientific articles (all figures in the following sections of this paper are available in Jupyter notebooks, see Table S2). App developers can identify datatypes as “statistical features.” Datatypes that are made accessible via Jupyter Lab interfaces hosted inside a Project (Fig. 2d left, Fig. S2a, and Video S9). The statistical features are automatically organized by *brainlife.io* into *Tidy data* formats<sup>50</sup> (*.tsv* and *.json*; Fig. 2d) and can be exported using the *pybrainlife* Python module (<https://pypi.org/project/pybrainlife/>). Jupyter Lab records are tracked for reproducibility and allow data analysis in R, Python, or Octave<sup>48,49</sup>.

The full data workflow (from import to preprocessing to analysis) makes possible the unification of large volumes of diverse neuroimaging datatypes into simpler sets of features organized into *Tidy data* structures<sup>50</sup> (Fig. S3c). The platform provides a variety of methods to visualize data, which aids in performing quality assurance, identifying mistakes, and repeating the processing when needed. Community-developed visualizers are served on the cloud side using docker containers (see Table S1), and six new web visualizers have been developed (Table S1 and Video S7).



**Supplemental Figure 2ab Brainlife architecture and system components.** **a.** Illustrative flowchart of the multi-faceted architecture of brainlife.io. Data providers, developer community (datatypes), app developers, and resources providers. Data providers upload their data to brainlife.io Projects via web UI/CLI, OpenNeuro or DataLad, or ezBIDS. These data are then stored and organized into Projects organized as “datatypes” based on specifications from the developer community. App developers develop self-contained Apps that can then be used to construct workflows from data inputs to final statistical products. Visualizers and QA measures allow for better construction of workflows to ensure the highest statistical quality and fidelity. These Apps run on compute resources provided by many resource providers, including HPC and cloud providers. Once workflows have been optimized, batch processing can be performed via Pipeline rules. These rules automatically track progress, including app completion and failures. Once statistical features of interest have been extracted, they are pushed to the warehouse into JupyterHub, where collaborators can work together on analyzing the data for the entire project. Upon completion, brainlife.io facilitates the publishing of the data and workflows via a Publications mechanism, where data and workflow information are reorganized for easier download and dissemination and given a digital object identifier (DOI) for the scientific community. **b.** An illustrative example of generating a network datatype on brainlife.io starting from structural (T1w) and diffusion (DWI) data on brainlife.io. Datatypes get docked as inputs into apps, from which either modified versions of the input are returned as output or entirely new datatypes are generated. Datatypes can be shared amongst other apps in the workflow (arrows) allowing for the chaining together of Apps into entire workflows. The outputs of this workflow can then be pushed to Jupyter Notebooks for statistical analysis.

**The ABCD specification and brainlife.io Apps.** The Application for Big Computational Data (ABCD; [github.com/brainlife/abcd-spec](https://github.com/brainlife/abcd-spec)) is a lightweight, specification proprietary to *brainlife.io* that enables App developers and resource managers to establish programming interfaces, to facilitate the integration of applications with the job scheduling systems (PBS, CONDOR, SLURM, etc) associated with a resource. The interfaces encompass the "start" entry point, used to initiate a service, the "status" interface, invoked to track the progress of service's job status and the "stop" interface, invoked to conclude the execution of service.

*Amaretti decentralized resource awareness and prioritization.* Amaretti is a meta-orchestration system able to run any App or service published on GitHub and conforming with the ABCD specification. Amaretti is "meta" in the sense that it make use of the underlying batch-scheduler (job-orchestration) mechanism already existing in computing resources. Amaretti has the ability to run services distributedly on multiple computing resources. In the event that a particular service is enabled on multiple resources, Amaretti utilizes a selection mechanism to choose the optimal resource. For example, a data processing workflow can consists of multiple steps, each implemented in a *brainlife.io* App or service. Amaretti allows sending each step in a sequence of processing steps on a different resource. The same step may be sent to different resources everytime it is requested. The outputs resulting from each step are then synchronized after execution is completed. If a user has access to multiple resources on which an App or a service can be executed, Amaretti decides selects a resource using a series of heuristics. At runtime, Amaretti computes the final resource and decides which resource to use for a service by using the following rules:

1. *Resources scoring.* Resource managers enable Apps or services on a resource. The manager can define a default score for the App, the higher score the more likely that the resource will be selected to execute a service. Find the default score configured for the resource. If not configured, the resource is disqualified from being used (resource managers must give explicit permission to run the App)

2. *Inter-resources data transfer minimization.* For each App data dependency, the score is incremented by 5 if the resource is used to run the Apps that generate the prerequisite data. This increases the likelihood of reusing the same resource where App runs produced data that is already available on the resource. This approach mitigates data transfer.
3. *Exclusive resource ownership criteria.* An additional ten points are given to a resource if the user possesses exclusive ownership of the resource. Users can define resources only assigned to them. In such case, rather than utilizing a shared resource, it is advantageous to use the private resource.
4. *Preferred resource ownership criteria.* An increment of fifteen points is added to the score when the resource is designated as the preferred resource to use, as stipulated by the user that submitted the App execution request.
5. *Public resource avoidance.* A project can be configured by users to abstain from using public computing resources. Public resources become ineligible for consideration if the App execution request originates from such a project.
6. *Connection failure.* A resources is disqualified if the resource monitor service detects a connection or server failure.

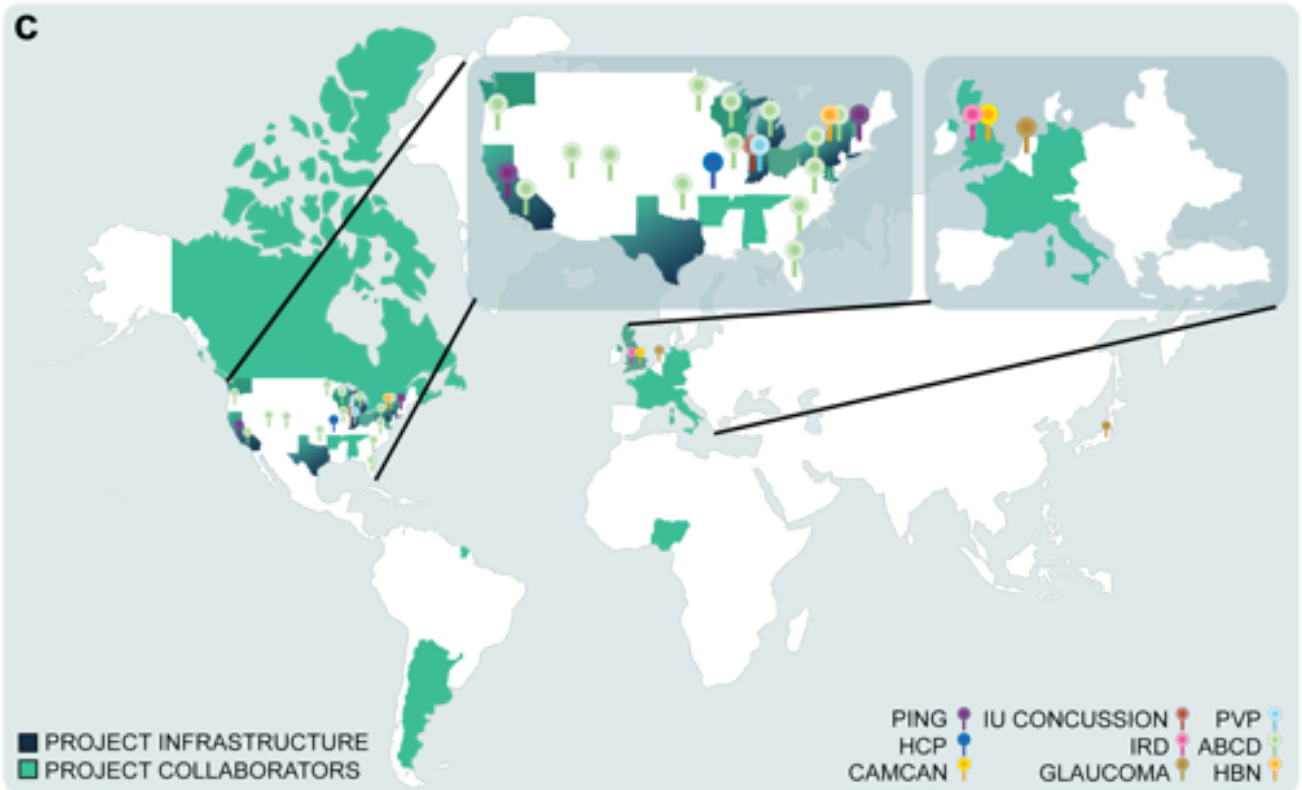
The resource with the highest score is chosen to execute the task, and a report detailing the rationale behind the resource's selection is added to a file within the service working directory

**Tasks.** Tasks are the atomic unit of computational work executed on various compute resources. Examples of Tasks are, a job for batch systems, or a vanilla process running on a vanilla VM. Amaretti keeps track of tasks by assigning each one of them a unique process ID.

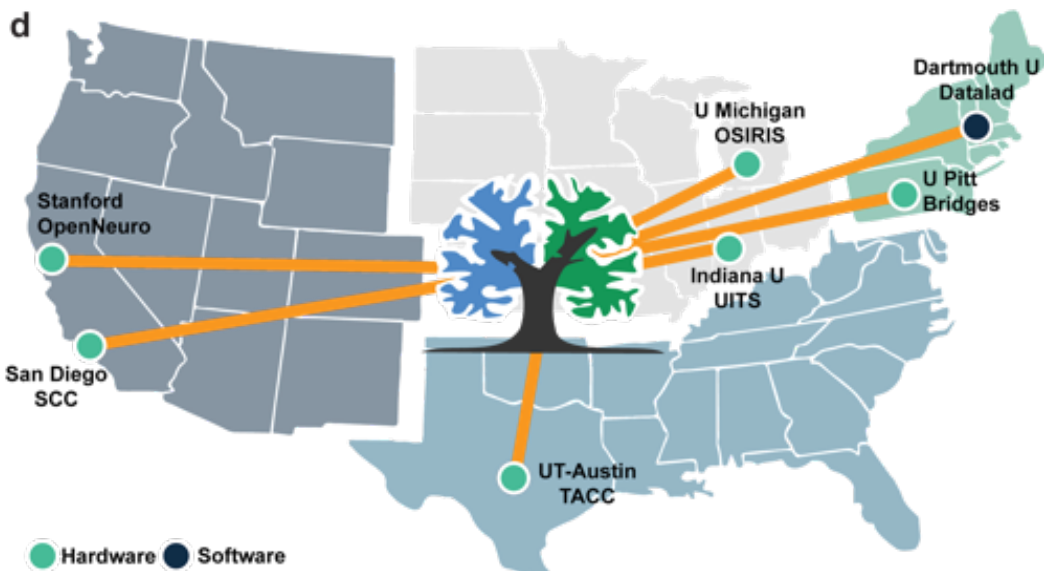
**Service.** Any ABCD-compliant GitHub repository is a service for Amaretti. Apps are Amaretti services. When users or the platform submit a task Amaretti retrieves the code service from GitHub. For example, if the user requests to run the Task specified by `github.com/brainlife/app-life` App, Amaretti will retrieve the code from GitHub, create a copy of the App for that task on a chosen resource and also move.

**(Workflow) Instance.** Amaretti provides DAG workflow capability by establishing dependencies between tasks. Tasks that depend on parent tasks will simply wait for those parent tasks to complete. All Amaretti tasks belong to a workflow instance (or instance for short).

**Resource.** Resource is a remote computing resource where Amaretti can securely connect and set up the App execution through the ABCD interface. The resource can be a single computer, a head node of a large high-performance computing cluster, or a submit node for high-throughput computing clusters. The code for the brainlife.io platform is available at <https://github.com/brainlife/>.



**Supplemental Figure 2c. System components. c.** Map the locations of critical facets of this research, including project infrastructure (i.e. compute resources), collaborators, and data sources. As the United States and Europe are home to many of the infrastructural resources, collaborators, and data sources, more details for these regions are provided (*insets*).



**Supplemental Figure 2d. *brainlife.io* infrastructure geolocation (2023). d.** Map of the locations of critical hubs for *brainlife.io*

## Global User base

~1200 users across +400 institutions / universities.



**Supplemental Figure 2e. brainlife.io user account geolocations. e.** Map of the locations of the users that created an account and accessed *brainlife.io*. This map is a proxy to the level of attention the platform achieved worldwide.

### Supplemental Table 1: Platform services serving the brainlife.io platform.

The brainlife.io platform is an incorporation of many individual services working in concert to increase the efficiency of neuroimaging analyses on the scale of thousands of participants and brain datasets. In addition to our efforts to compile a list of currently available services provided by the greater scientific community, below we provide a list of the many platform services that combine to make the brainlife.io platform (**Table S1**). Because brainlife.io is an open platform for neuroscientific investigations, we provide the individual URLs pointing to the code base of the individual services of the platform.

### Supplemental Table 2: Jupyter notebooks for analyses performed.

The code used to analyze the thousands of datasets processed in this manuscript is openly accessible on GitHub.com. Below we provide a list of the jupyter notebooks for performing the analyses outlined previously (**Table S2**). For this, we provide the jupyter notebook name and the GitHub URL for the respective notebook. Within each notebook, we describe the neuroimaging topic the notebook covers, including structural morphometry (i.e. cortical thickness, surface, area, volume), diffusion profilometry, structural connectivity, functional connectivity, functional gradients, MEEG, and optical coherence tomography (OCT). These notebooks were used to summarize data for different measures and many individual analyses and figures outlined previously. The goal of these notebooks is to document enough information for new users to re-use the notebooks for their own analyses on their own datasets. These notebooks are freely available for use by the greater scientific community.

### Supplemental Table 3: Preprocessing Apps used for the experiments.

In addition to providing documentation to the code servicing brainlife.io, we openly release the App code for each App used to analyze the thousands of datasets processed in this manuscript. Below we provide a list of the Apps used for performing the analyses outlined previously (**Table S3**). For this, we provide the App name listed on brainlife.io, the digital-object identifier (DOI) automatically assigned to each app, and the GitHub Repository where the code for the App resides. The goal of this is to increase the transparency of the processing steps performed in this investigation, and for researchers to validate and incorporate into their currently existing workflows.





**Supplemental Figure 3a. Brainlife App Github template.** 1. App DOI and ABCD specification. 2. App name. 3. Description of the App. 4. Authors and contributors. 5. Funding Acknowledgement. 6. Citations. 7. Instructions for running the app locally, including how to set up the config.json file containing all of the important information for the App including inputs and configuration parameters. 8. Example datasets that can be downloaded to test the app locally. 9. The outputs for the App. 10. The software dependencies subserving the App.

## Developing processing Apps for the platform

Here we describe the requirements for developing Apps on the platform. Despite the over 500 apps currently available on the platform, there still exist possibilities for researchers to develop their own processing Apps for performing specific steps that might not already exist on the platform.

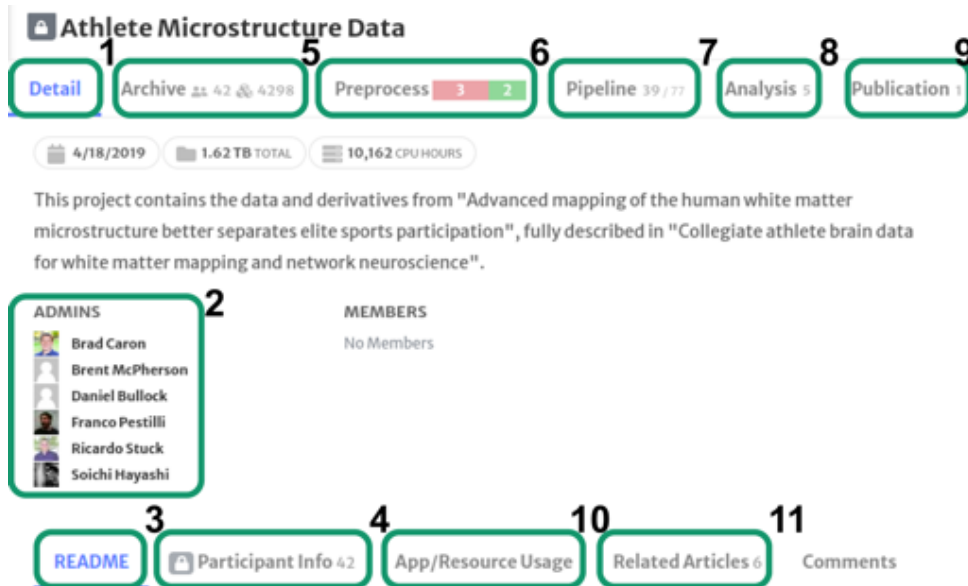
The development process for Apps has been streamlined in order to make it as intuitive as possible. Specifically, each App has a set of requirements necessary for the App to be used on the platform. The most important of these requirements involves the creation of a README file outlining all of the important information needed to describe the contents of an App. On Github, we have developed a set of App README templates for App developers to use (**Fig. S3a**). On the README file, the user must provide information regarding the brainlife.io App DOI and the ABCD specification. In addition, they must also document the app name and a description of what steps the App performs.

Users can also provide information regarding specific authors, coauthors, funding sources, and literature citations in order to provide proper credits for the development of the App. Following these descriptive details, the README should also provide information regarding the usage of the App both on brainlife.io and on local workstations, including descriptions of the inputs, outputs, and software library dependencies of the App. These descriptions found in the README increase the transparency of the App in order to increase the findability and usability of the App.

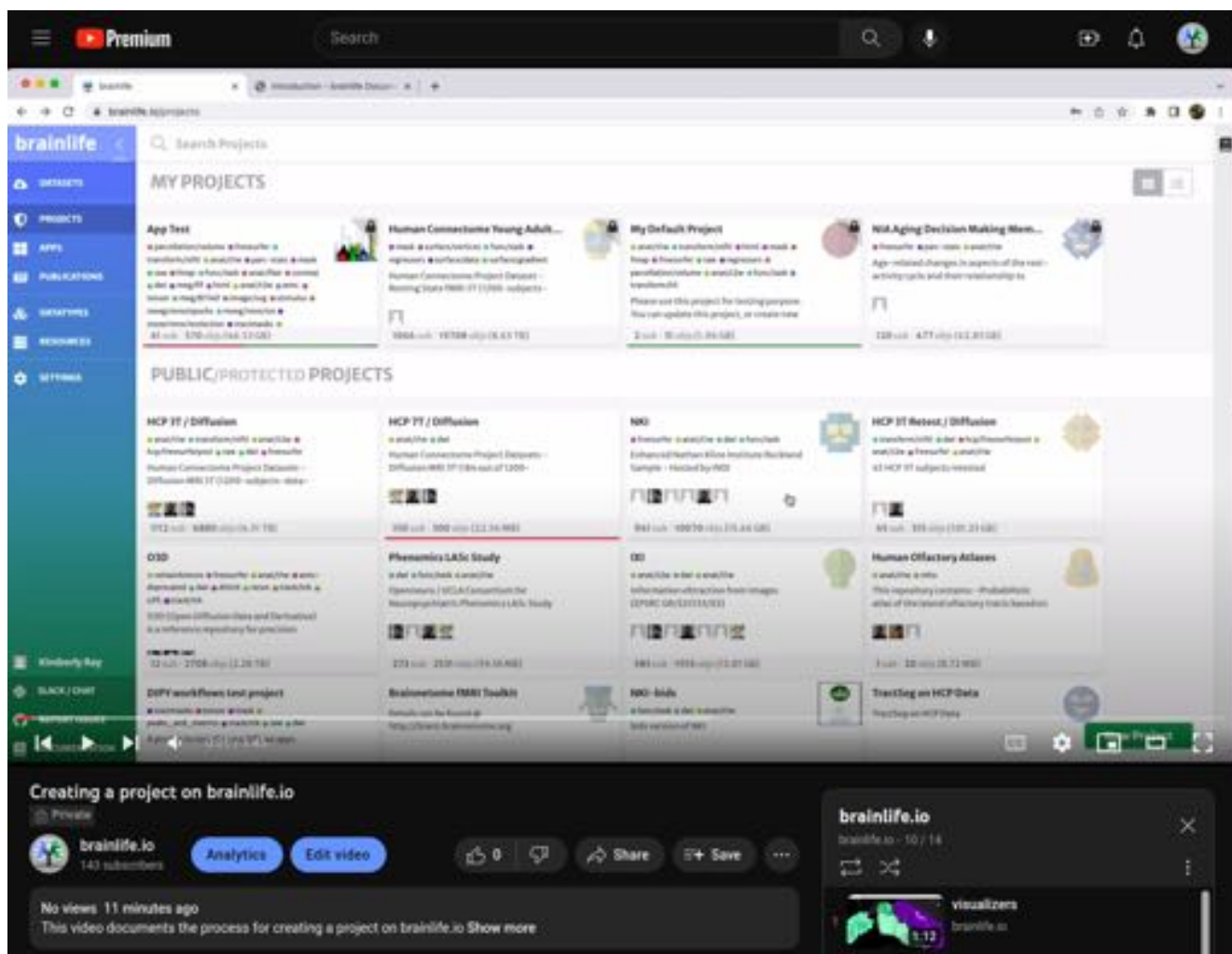
## Using the platform

Here, we describe the user interface of the platform to help introduce the visual interfaces developed as part of the project. These steps will be described in order of how they would be implemented by a typical researcher designing their own set of experiments using the platform. In addition to visual and text descriptions, we also provide a series of videos documenting each step of the process.

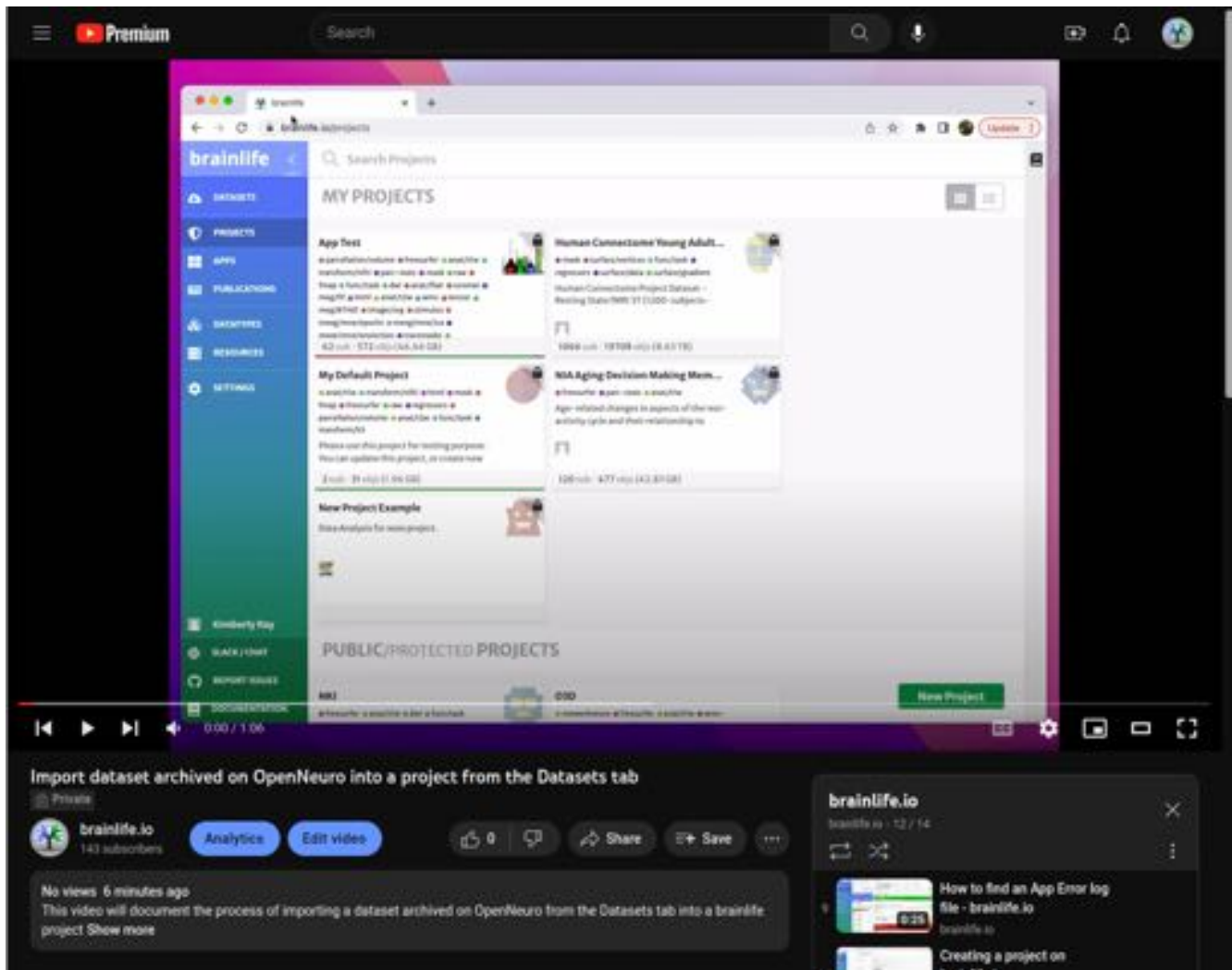
Upon creation of a brainlife.io account, a researcher will first set up a Project within which all of the data processing, storing, and organization will occur (**Fig. S3b; Video S1**). Once their Project is created, users can then update and assign details to the project, including a description of the project, access control to the project, a project README file describing specific information about the project in a machine-readable format, information regarding each participant in the study, and even limit which computing resources the Project will use to process the data.



**Supplemental Figure 3b. Brainlife project landing page.** 1. Detail tab containing all of the important details and information describing the Project. 2. Users can add Admins and members for proper project governance. 3. Projects can have README descriptions, like those on GitHub, to describe important details of the project in a Markdown format. 4. Participant Info contains tables of demographic information that may be helpful for performing an analysis. This is set and defined by the Administrators of the Project. 5. Archive tab is where all of the stored files in the form of brainlife datatypes can be found. 6. The Preprocess tab is where jobs can be launched and monitored. 7. Pipelines allow the investigator to batch process all of the participants in their project for each App they need to run. 8. Once statistical features have been extracted, Administrators can access Jupyter Notebooks within the Analysis tab to perform their statistical investigations across all of the participants in the project. 9. Once the investigators are completed the investigation, they can use the Publication tab to efficiently publish their data and the analysis workflows on brainlife.io. 10. Whenever a job launches, the App/Resource Usage tab is automatically updated in order to provide provenance tracking of what and where the data processing was performed. 11. Brainlife.io will search keywords in your project with previously published studies to identify any related articles to your investigation in the Related Articles tab.



**Supplemental Video 1.** A video documenting the process of creating a project on brainlife.io, including updating access control and participant information. <https://youtu.be/P2kz6E53nIo>

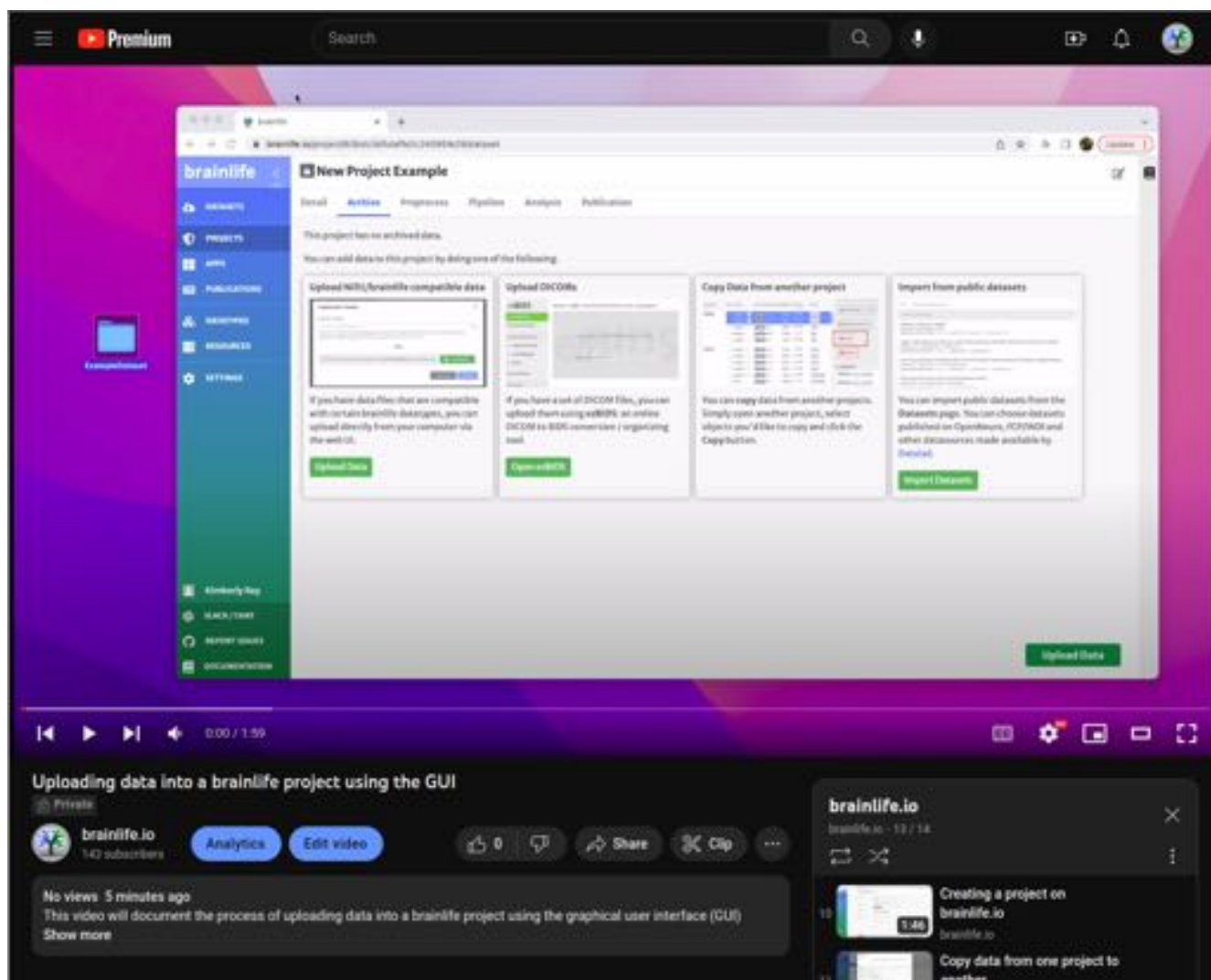


**Supplemental Video 2.** A video documenting the process of pulling datasets from the 'datasets' tab into a Project on brainlife.io. <https://youtu.be/N3UXteQ3tu8>

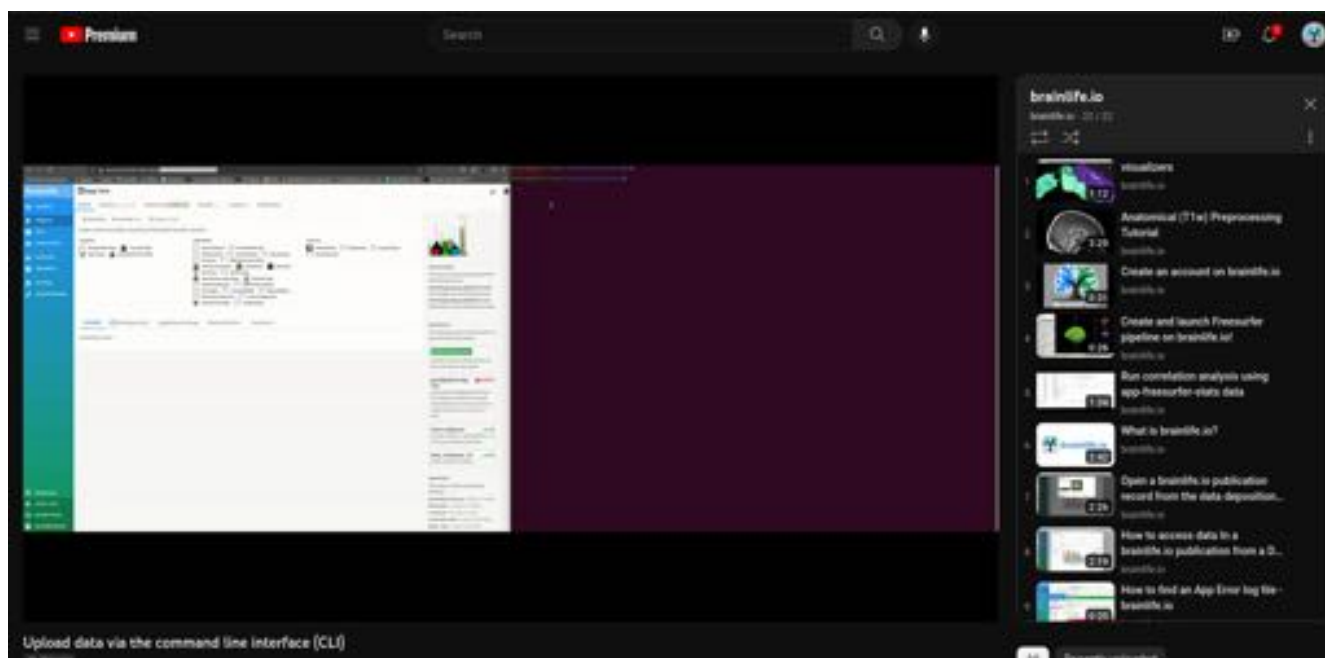
Once this information is defined, users are then ready to either import raw datasets they collected or pull datasets that have been openly released. For openly released datasets, users have a variety of options to pull data from including other projects (**Video S2**), or projects hosted on OpenNeuro (**Video S3**). In a similar fashion, users have a variety of options for uploading any newly collected datasets including a built-in GUI (**Video S4**), a CLI (**Video S5**), or through a newly developed sister technology for automated converting of raw scanner data into BIDS-standardized data files known as ezBIDS (**Video S6**). Each of these methods provide a streamlined, efficient way to import data into a new project for future processing and analysis.



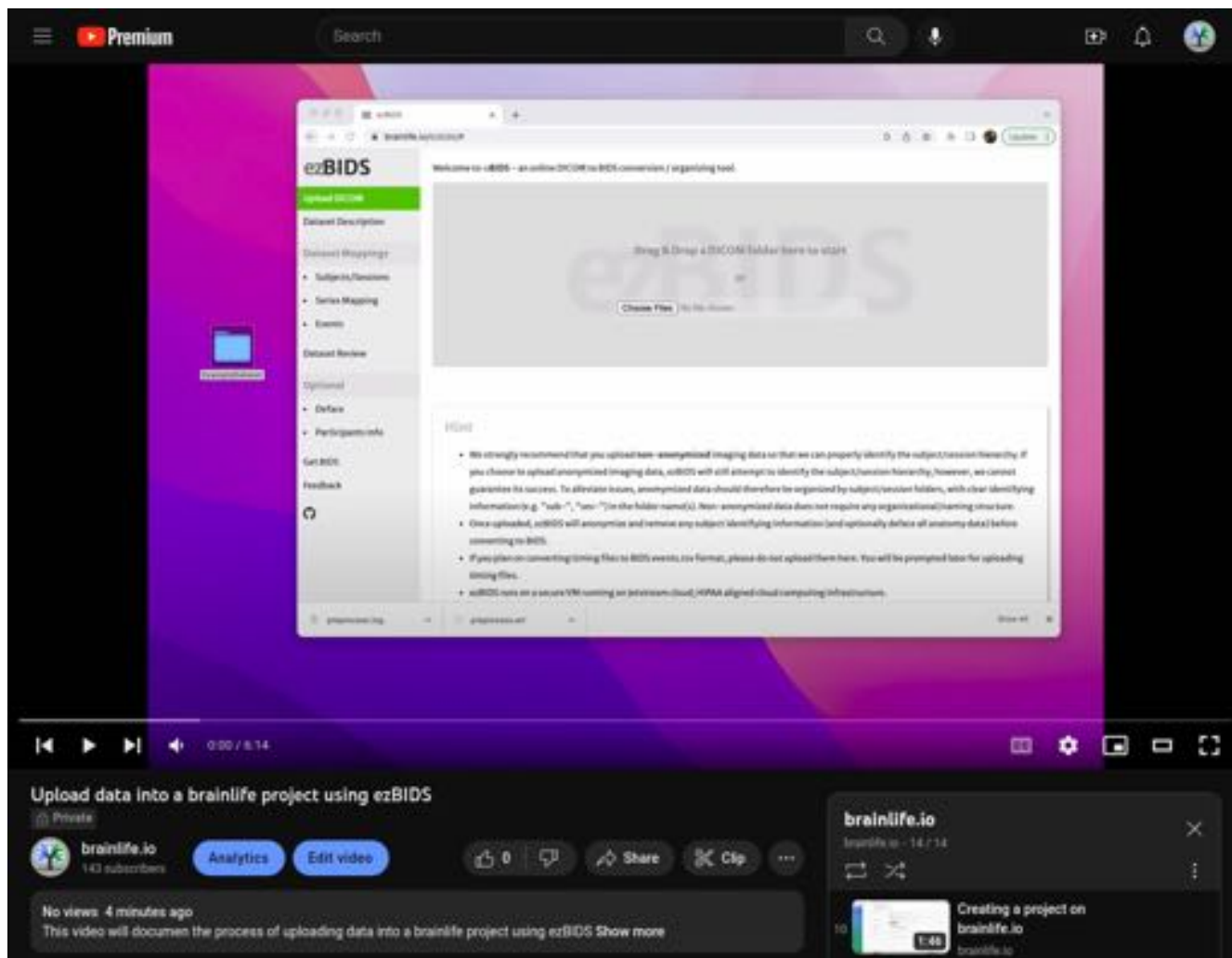
**Supplemental Video 3.** A video documenting the process of pulling data from OpenNeuro into a Project on brainlife.io. <https://youtu.be/OZQyR9jLwYo>



**Supplemental Video 4.** A video documenting the process of uploading data to a brainlife project using the graphical user interface (GUI) directly via the browser. [https://youtu.be/5RGo\\_jY4Oqc](https://youtu.be/5RGo_jY4Oqc)

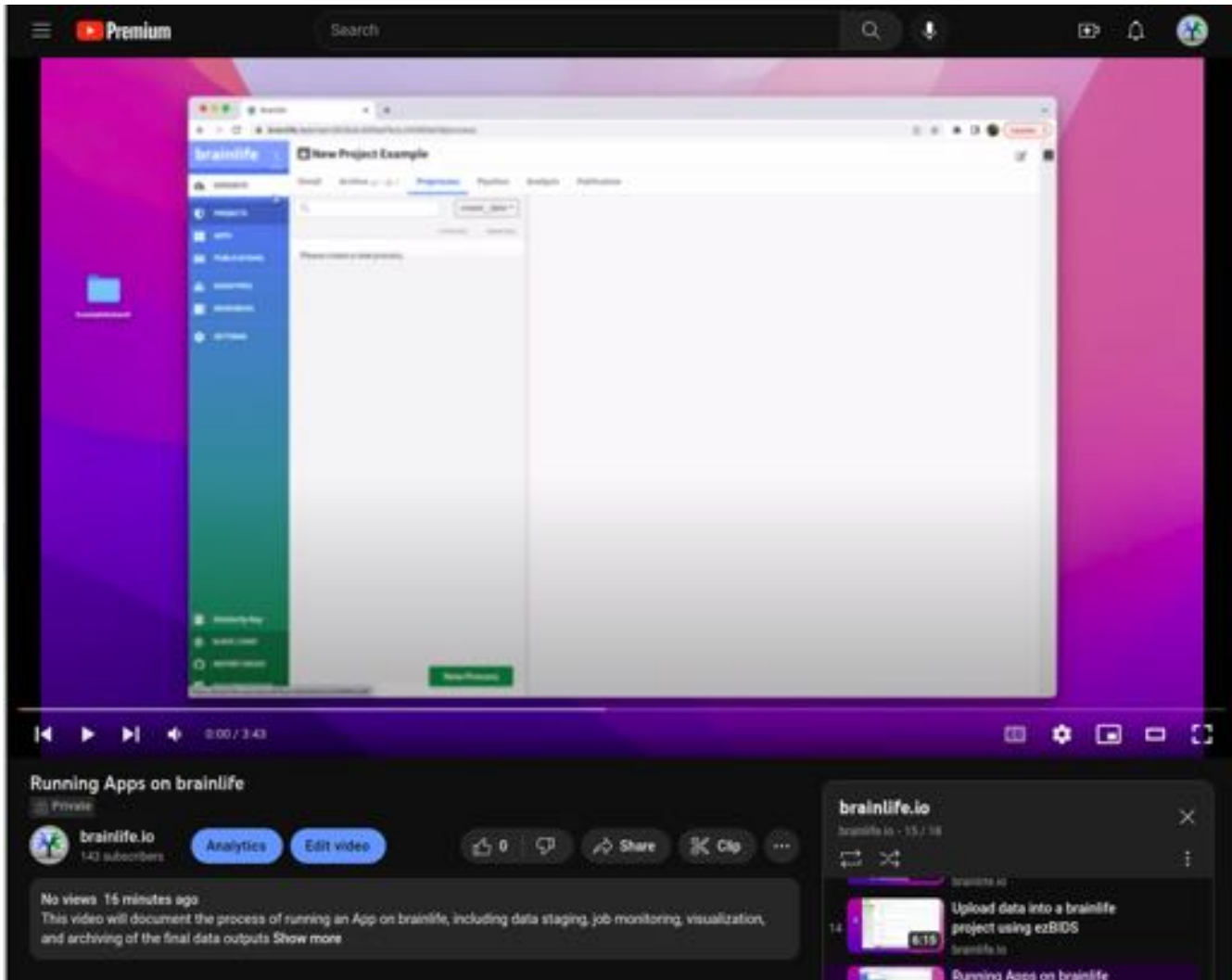


**Supplemental Video 5.** A video documenting the process of uploading data to a brainlife project using *brainlife.io*'s Command Line Interface (CLI). <https://youtu.be/PUTLXJJSBqQ>



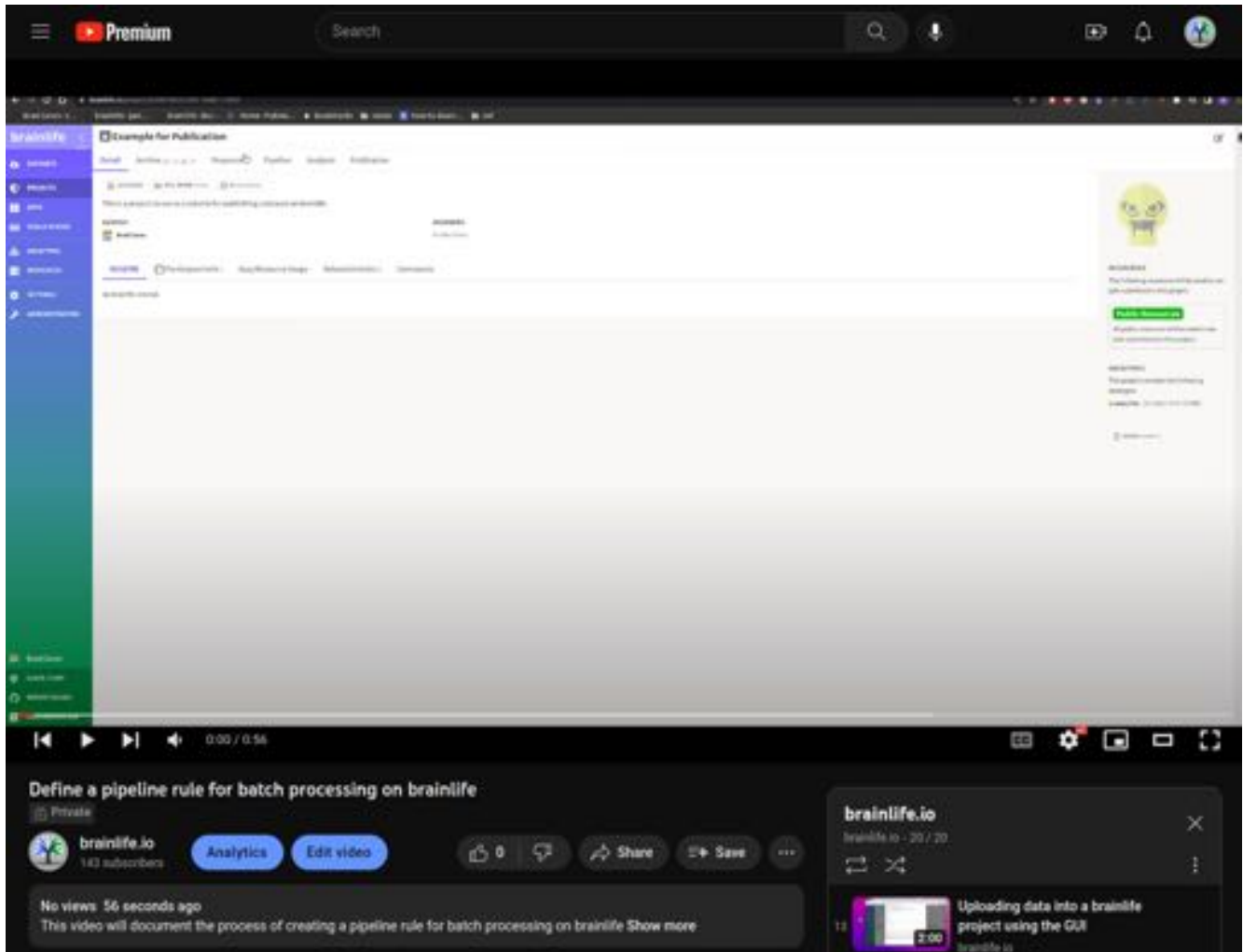
**Supplemental Video 6.** A video documenting the process of uploading data to a *brainlife.io* Project using the DICOM to BIDS converter *brainlife.io/ezBIDS*. <https://youtu.be/KvhlHxzHsI4>





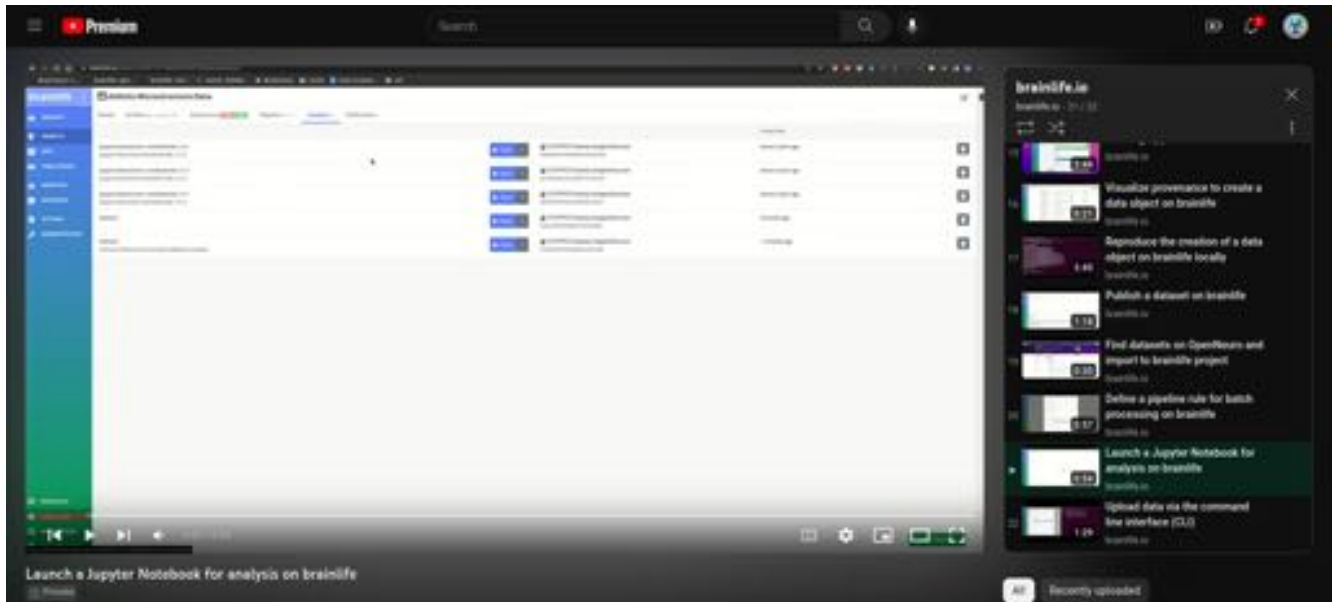
**Supplemental Video 7.** A video documenting the process of running an App on brainlife.io, including data staging, job monitoring, visualization, and archiving of the final data outputs. <https://youtu.be/43yhZ1k6icQ>

Upon importing data into a Project, users can directly interact with the data stored in the Archive tab of the project in multiple ways. First, users can select a data object and visualize the data object using one of the many built-in visualization services for that specific datatype. More importantly, users can then “stage” or move the data from Archive into the Preprocess tab, from which users can select and launch any of the over 400 available Apps (**Video S7**). Because Apps on brainlife are “data aware”, users will only be presented with the Apps that take in the staged datatypes that they are designed to work with as inputs ultimately reducing the potential for user error. From the Preprocess tab, users can monitor the status of the App, interact with the data files generated during the App, and visualize the outputs. Once the user is satisfied with the outputs, data objects can be stored back into the Archive tab directly from the Preprocess tab.



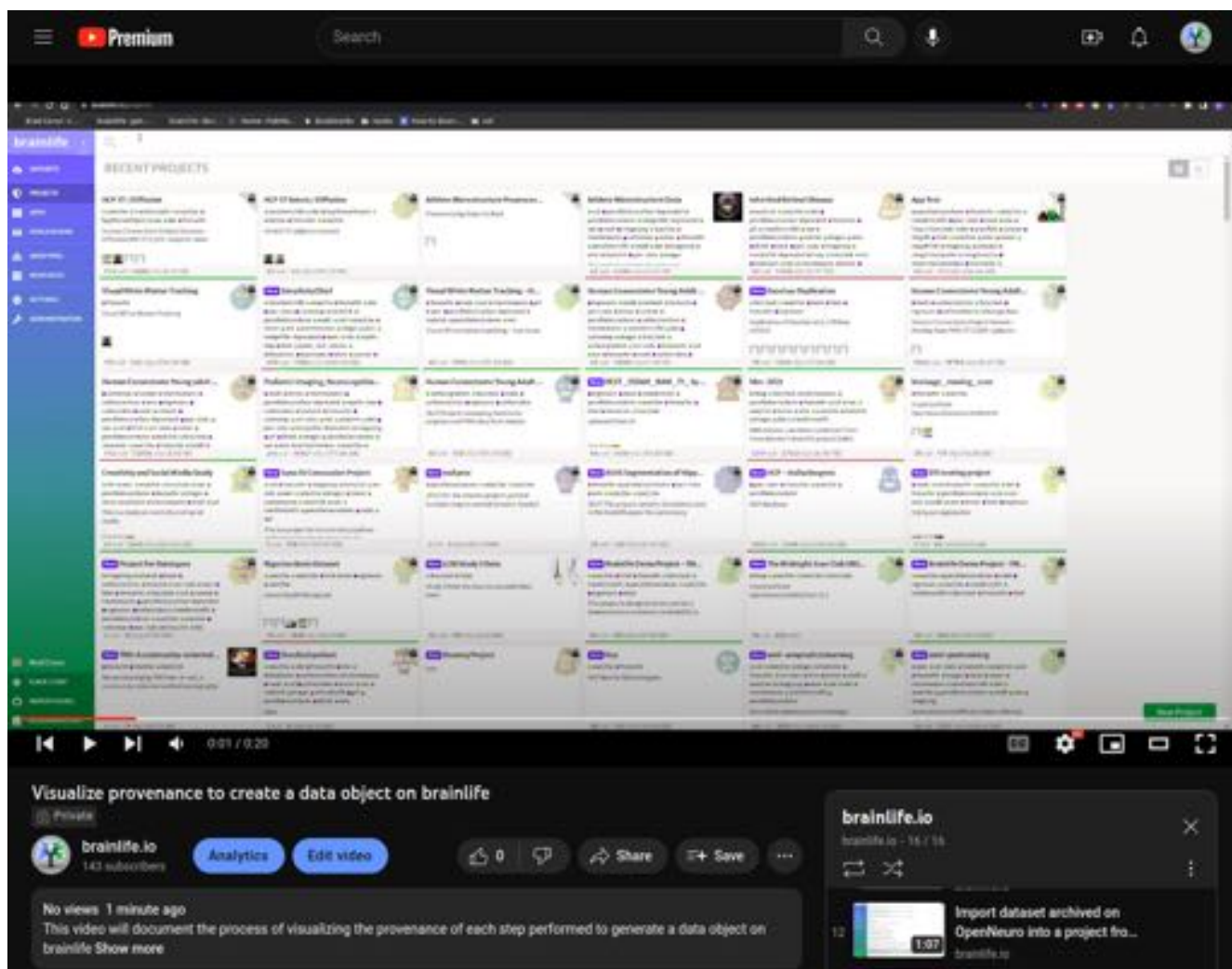
**Supplemental Video 8.** A video documenting the process of defining a Pipeline rule on brainlife.io to perform batch processing. <https://youtu.be/1CSdsf8czL8>

This process for running an App is useful under testing circumstances, but may not be appropriate for batch processing of a large number of participants. To facilitate this, users can define Pipeline rules via the Pipeline tab (**Video S8**). Within these rules, users specify the inputs including which data objects from the Archive to include or exclude, the configuration parameters required by the App, and the archiving of output objects back into the Archive. Upon launching a Pipeline rule, Amaretti will automatically stage all of the data that matches the input criteria, identify the most appropriate compute resource for running the process, and archive the output data objects back into the project Archive for storage. Outputs from one Pipeline can then be set as inputs to another Pipeline, allowing for the chaining of Apps to develop the overall processing workflow required to get from raw data to the final statistical features of interest needed for statistical analysis.



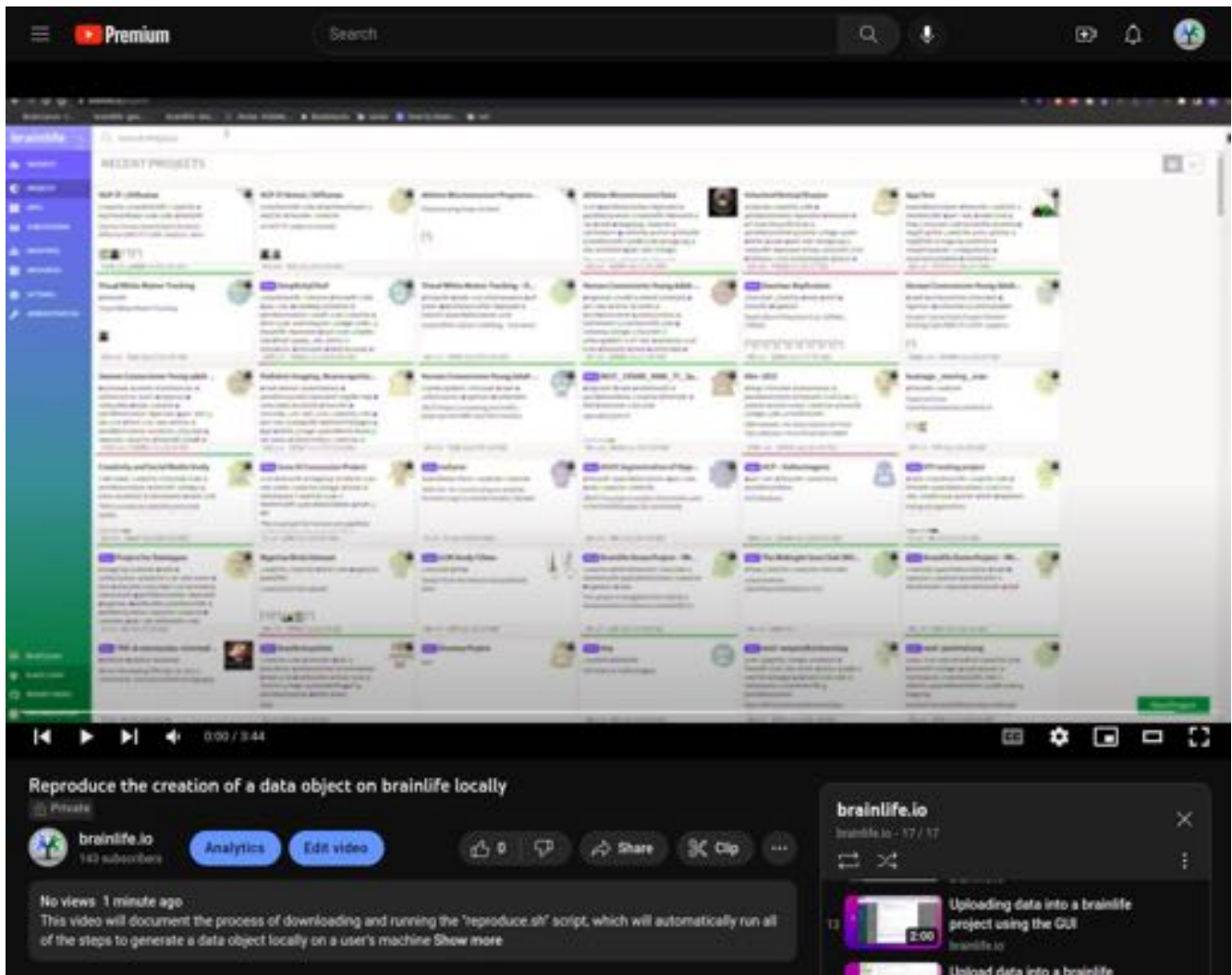
**Supplemental Video 9.** A video documenting the process of launching a Jupyter Notebook for performing statistical analyses within a brainlife project. <https://youtu.be/tJW6374BcpQ>

Once these statistical features of interest are extracted, users can then analyze them directly on the platform via the Jupyter Notebooks provided by brainlife.io (**Video S9**). To facilitate this, a certain subset of all datatypes that correspond to statistical features of interest are stored in a secondary warehouse, which can be directly loaded via the Jupyter Notebooks. This ultimately reduces the number of potential data objects and storage size of the objects required by brainlife.io to move into the Notebooks, ultimately making the process more efficient for users. Common subsets of functions, including those useful for loading data into the Notebooks, have been packaged into a Python package *pybrainlife* that can be imported directly into the Notebooks and used to load and compile an entire study's worth of statistical features. Upon completion of the analyses, these Notebooks can be directly published and/or pushed to Github in order to increase the scientific transparency of the project.

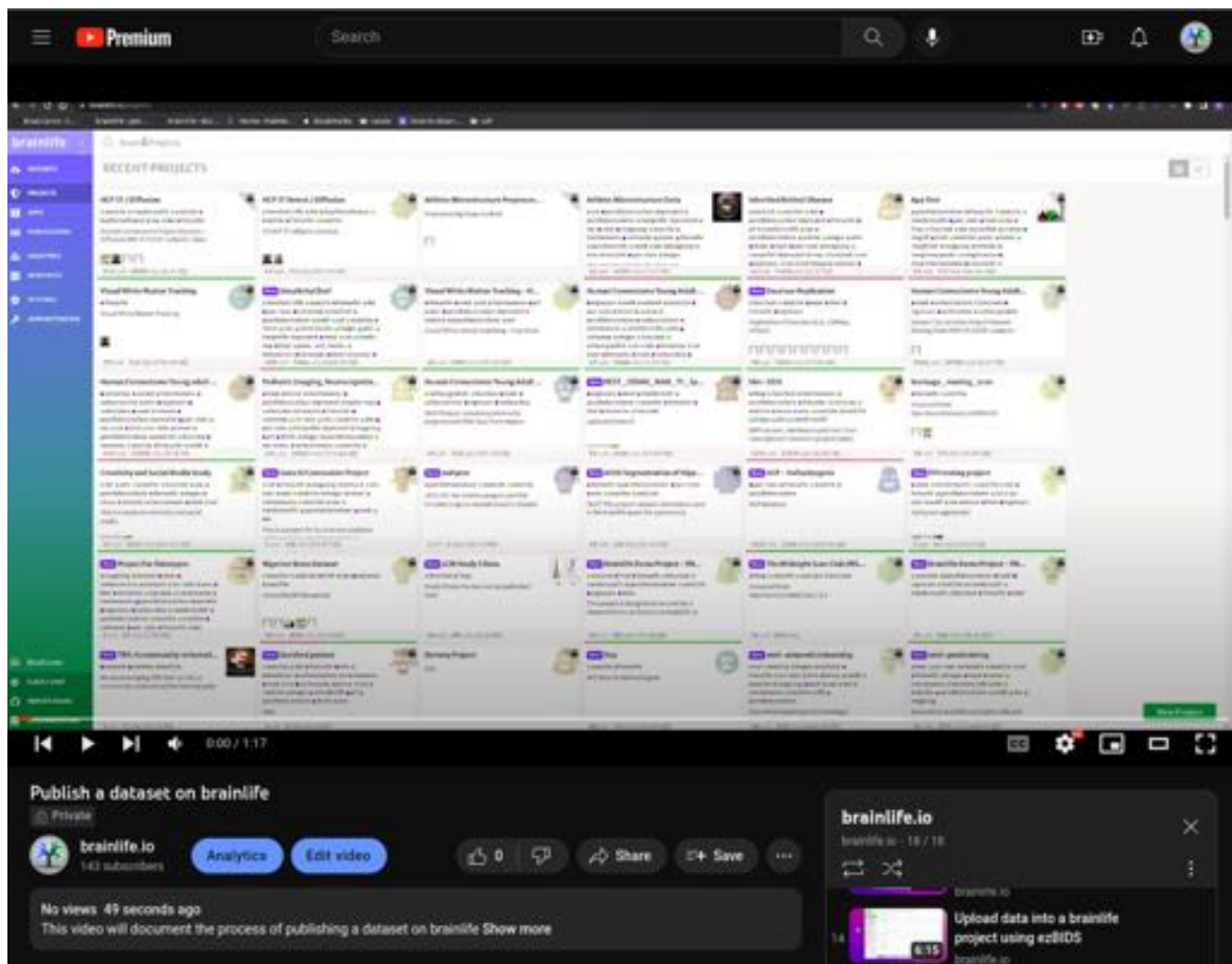


**Supplemental Video 10.** A video documenting the process of monitoring the steps taken to generate a datatype (provenance). [https://youtu.be/NzUObf8\\_x7g](https://youtu.be/NzUObf8_x7g)

In addition to the publication of the Notebooks, brainlife.io automatically keeps track of each individual step performed to obtain a specific datatype (i.e. provenance) (**Video S10**). This visualizer contains all of the information a user might need to validate that the proper steps were taken, and for any outsider users or reviewers to rerun their analysis steps for purposes of replication. With this goal in mind, brainlife.io will also generate a script for any data object to reproduce the individual steps to create that object locally (`reproduce.sh`; **Video S11**). Finally, upon completion of processing and analysis, researchers can Publish their datasets, Pipeline rules, and Analysis notebooks directly on the platform via the Publications tab (**Video S12**). All of these individual features are designed with the goal of increasing the reproducibility of processing and analyses performed via the platform.



**Supplemental Video 11.** Video documenting the process of replicating the generation of a data object via a single bash script that can be run on any machine (reproduce.sh). <https://youtu.be/YMCFU0aQhvl>



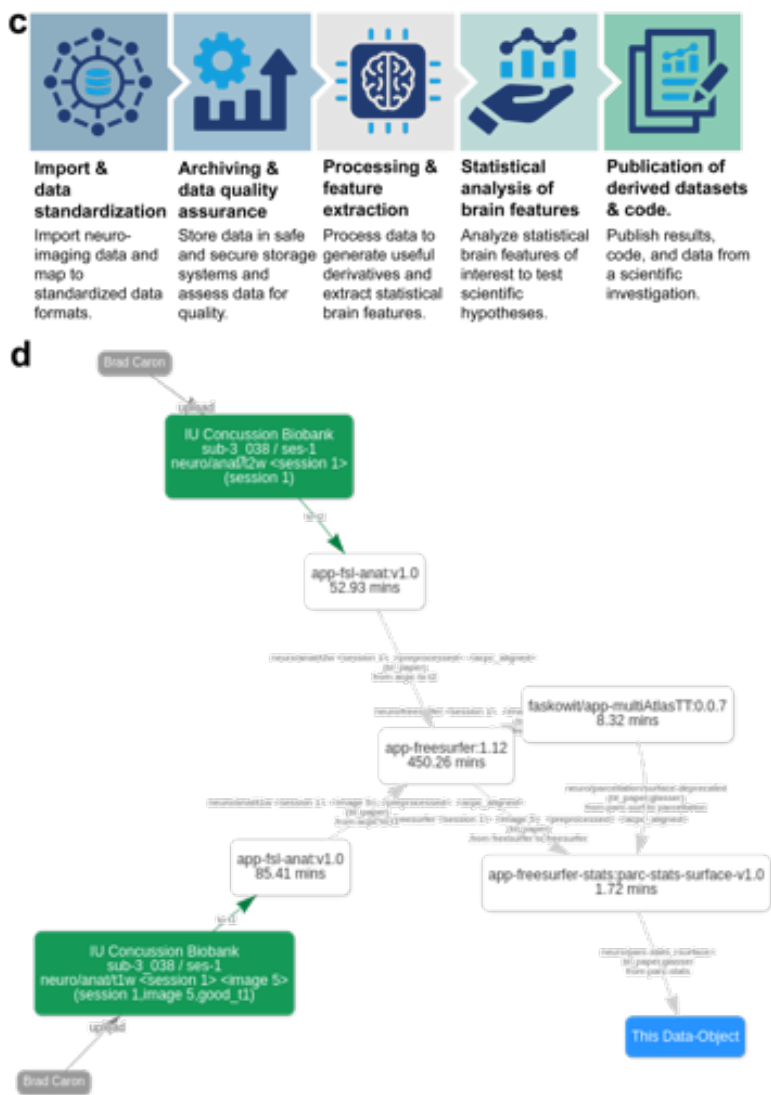
**Supplemental Video 12.** Video documenting the process of creating a Publication on brainlife.io. <https://youtu.be/aUvjuEihWJA>

## End-to-end reproducible scientific workflow

*brainlife.io* automatically tracks all actions performed by researchers during data analysis. Data object IDs, Apps versions, and parameter sets used to launch an App, resources used, error logs, etc are all tracked automatically by *brainlife.io*. The full sequence of steps from data import to preprocessing, analysis and publication is captured by the platform and is used to build a record of all the actions a researchers performed while implementing a data analysis study.

A graph describing provenance metadata for each Datatype can be visualized using the provenance visualizer and downloaded (see [Video S10](#)). Also, a linux shell script is automatically generated to allow the reproduction of full processing sequences ([Video S11](#)).

Finally, a single record containing data objects, Apps, and Jupyter Notebooks used in a study can be made publicly available outside the platform in a single record addressed by Digital Objects Identifiers (DOI)<sup>51</sup>. Whereas all other existing systems provide users with technology to track analysis steps manually or require the use of coding, *brainlife.io* tracks automatically and does not require coding. This automation technology lowers the barriers of entry to reproducible and transparent large-scale neuroimaging data analysis.



**Supplemental Figure 3c,d. End-to-End steps to reproducible computational analysis.** a. Pictorial description of the end-to-end workflow for performing a scientific investigation using neuroimaging data. First, data is collected from

measurement systems including MRI and MEG. Following this, data is converted into workable data formats, including NIFTI, .tsv, and .json files, or into standardized file formats including BIDS. Following conversion, data is preprocessed where common artifacts are removed to increase data quality. Model fitting and brain segmentation can then be performed on this cleaned, preprocessed data. Following this, quality assurance (QA) efforts are usually undertaken to ensure the data is of a high enough standard for publication. If the data does not meet a high standard of quality, adjustments to the preprocessing and model fitting steps can be performed (*circular arrows*). Following this, statistical brain features of interest are extracted in order for statistical analyses to be performed. Once the analyses are finalized, researchers then publish their results, data, and code to the greater scientific community. All of these steps are supported by brainlife.io. **b.** Visualization of data “provenance” automatically generated for each archived data object on brainlife.

Neuroimaging investigations involve a common workflow from data collection to study publication (**Fig. S3c**). Data are first either collected from neuroimaging measurement systems, including MRI and MEG scanners. Following collection, data is then converted to standardized file formats before they can be used by the researcher. From here, common artifacts are removed from the data in a series of preprocessing steps. Once the data is cleaned, models can be fit, brain structures can be segmented, and quality assurance assessments are performed. If any mistakes occurred in the previous steps, adjustments can be made to each individual step in order to increase data quality. Only once the data are of high enough quality are statistical brain features of interest extracted, and statistical analyses are performed on the extracted features. Final results, data, and code are then published to the greater scientific community to increase transparency and data gravity of the investigation. Brainlife.io serves each step following data collection, with each step of the workflow tracked in order to increase reproducibility.

## Supplemental platform evaluation

### Supplemental platform utilization

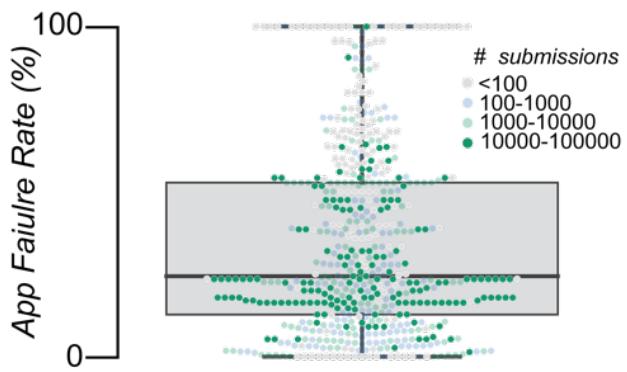
*brainlife.io* was developed with a FAIR model and made available worldwide. Any researcher can create an account on *brainlife.io*, although all new accounts are reviewed by the project team. *brainlife.io* first became publicly available in 2018. We tracked the usage of *brainlife.io* in the past 60 months. The platform community, utilization, and research assets have grown steadily since project inception (**Fig. 3** and **Fig. S2c** and **S4**). At the time of writing, over 2,341 users across 43 countries have created a *brainlife.io* account. Over 1,542 active users submitted more than 10 jobs per month (**Fig. 3a**). There were 3,439 data management Projects. The *brainlife.io* developers' community had implemented 530 data processing Apps comprising 2,438,998 lines of code (top 50 apps), and these had been used to process over 270 TBs of data for a total of 3,951,372,037,289 hours of compute time. Apps success rate on average has been 65.4% across 6,710,091 total job submissions (the estimates contain high-failure rate App test-calls). This level of interest and reach, even prior to a formal publication describing the platform, is a testament to *brainlife.io*'s potential for growth and impact.

Researchers ranging from undergraduate students to faculty have already used *brainlife.io* (**Fig. 3b**). The Apps used spanned various aspects of the neuroimaging data lifecycle. The most frequently used Apps pertained to tractography (22%), model fitting (15%), and ROI generation (12%). Community-developed software libraries provided the foundations for data processing Apps, including Nibabel, Freesurfer, FSL, DIPY, MRTrix, Connectome Workbench, and MNE-Python. Terabytes of data have been uploaded (72%) or imported from OpenNeuro.org (22%), the Nathan-Kline Institute data sharing projects (3%; <sup>44,46,52</sup>), and other sources. Early community attention and adoption preceded this publication describing the project and platform. The worldwide platform access highlights the global need for technology like *brainlife.io* (**Fig. S2e**).

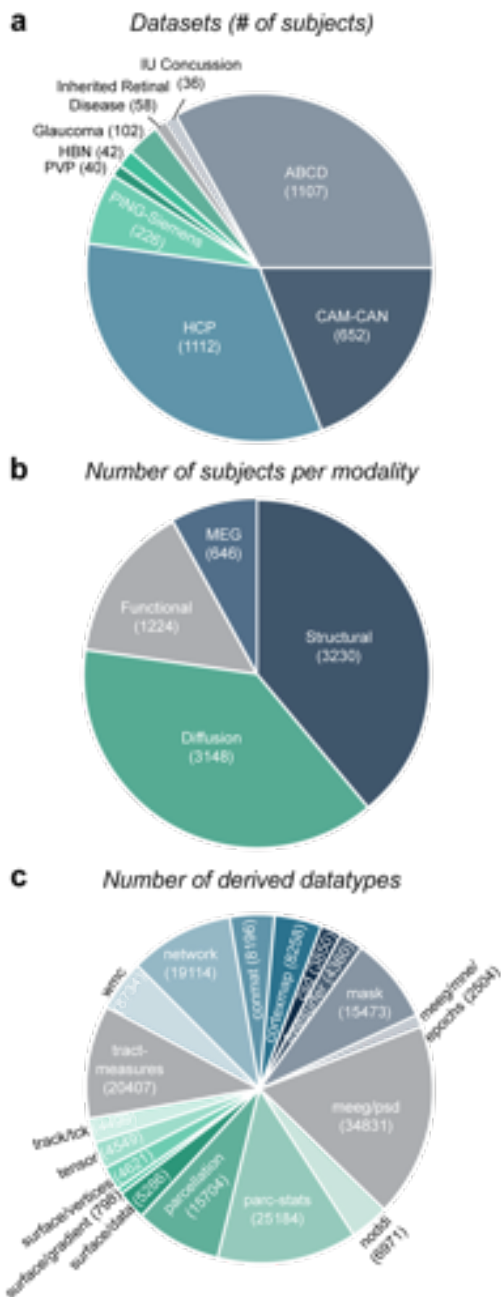
### Apps performance evaluation

Brainlife.io, like any technology, is not *failure-proof*. To examine the rate at which brainlife.io Apps fail, we collected data regarding the failure rates of all Apps across the platform. Since the beginning of the platform, jobs processed on brainlife.io have had a 34.6% failure rate across 6,710,091 submissions, with half estimated to be due to initial App testing and development (**Fig. S3e**).





**Supplemental Figure 3e. Brainlife.io processing is not error-proof.** Distribution of brainlife.io App failure rates (percentage) across all 568 Apps and their respective submissions. Box-and-whisker plot indicates the overall average failure rate across all Apps (*dark black line*), 25th and 75th percentiles (*box*), and overall range (*whiskers*). Each dot is an individual App's failure rate. Colors represent the number of submissions for each App (*grey*: 0-100 submissions, *light blue*: 100-1,000 submissions, *light green*: 1,000-10,000 submissions, *dark green*: 10,000-100,000 submissions).



**Supplemental Figure 4a-c. Overview of data used for the study.** **a.** Number of subjects across all datasets examined. **b.** The number of subjects per imaging modality. **c.** The number of brainlife.io datatypes (i.e. freesurfer, parc-stats, tractmeasures, track/tck, NODDI, tensor, csd, mask, network, conmat, parcellation, cortex map, wmc, meeg/psd, meeg/mne/epochs, surface/data, surface/vertices, surface/gradient) derived across all subjects and datasets examined.

### Supplemental platform testing

The effectiveness of the technology to provide good quality results were evaluated. We performed system load experiments by processing large amounts of data and evaluating the results obtained. These experiments were performed to demonstrate the ability of the platform to serve accurate data processing and analysis at scale.

Our experiments focused on the four axes of scientific transparency,<sup>53</sup> namely: data processing validity<sup>54</sup> (**Fig. 4a-e**; **Fig. S4d-g**), reliability (**Fig. 4f-j**; **Fig. S4h-i**), reproducibility (**Fig. S4j-n**), and replicability (**Fig. 7**; **Fig. S7a,b**). Four data modalities (sMRI, fMRI, dMRI, MEG) were evaluated using the test-retest HCP<sub>TR</sub><sup>55</sup>, the Cam-CAN<sup>56</sup>, the HBN<sup>52</sup>, and the ABCD<sup>57</sup> dataset. For all experiments, the Pearson's correlation ( $r$ ) and root mean-square-error ( $rmse$ ) were computed for each comparison using data products derived from apps on brainlife.io, where high

correlations and low *rmse* would provide strength of evidence in each experiment. Herein we describe the definitions of success for each experiment and the methods used to assess the performance of the platform. For all reported correlations and root mean-square-error values for the data validity and reliability experiments, see [Table S3](#).

A total of 12 different datasets comprising over 4,200 participants were processed ([Fig. S4a](#)), of which 3,200 participants had sMRI, 3,100 had dMRI, 1,200 had fMRI, and 650 had MEG data objects ([Fig. S4b](#)). Derived data products included cortical parcel volumes, white matter profilometry, functional and structural networks properties, functional gradients, and peak alpha frequency ([Fig. 4](#), [Fig. S4c](#)). In sum, over 193,000 objects and 22 Terabytes of data were generated for the experiments, using over 30 Apps ([Table S3](#)).

For each of the four data modalities, data processing validity was defined as the ability of a processing step to accurately reflect ground-truth properties of the brain. Data processing validity was estimated by comparing values obtained using *brainlife.io* Apps (see [Table S3](#)) against data preprocessed by the data originator (HCP or Cam-CAN depending on the data modality). Cortical parcel volumes were estimated from minimally processed HCP<sub>TR</sub> data using *brainlife.io* Apps [A0](#), [A462](#), [A23](#), [A272](#), and [A464](#). Volume estimates were compared to corresponding estimates provided by the HCP consortium ([Fig. 4a](#);  $r_{\text{validity}}=0.98$ ,  $\text{rmse}_{\text{validity}}=570.54\text{mm}^3$ ).

Fractional anisotropy (FA) in 61 white matter tracts was estimated using the raw and minimally preprocessed HCP<sub>TR</sub> dMRI data. The composable processing pipeline comprised of the sequence of Apps: [A68](#), [A238](#), [A297](#), [A305](#), [A188](#), [A195](#), and [A361](#). These Apps were used to process either type of data, with the exception of [A68](#),<sup>30</sup> for which only raw data was used. The average FA for each tract was compared between these two processing methods ([Fig. 4b](#);  $r_{\text{validity}}=0.95$ ,  $\text{rmse}_{\text{validity}}=0.018$ ).

Functional connectivity estimates between 117<sup>2</sup> nodes-pairs<sup>58</sup> were estimated using the raw and minimally preprocessed HCP<sub>TR</sub> fMRI data. [A160](#), [A23](#), [A369](#), and [A532](#) were used to process either dataset, with the exception of [A160](#),<sup>22</sup> which was only used with raw data ([Fig. 4c](#);  $r_{\text{validity}}=0.89$ ,  $\text{rmse}_{\text{validity}}=0.12$ ).

In addition, functional gradients<sup>59,60</sup> were computed on 400 nodes estimated on raw and minimally processed HCP<sub>TR</sub> fMRI data using [A604](#) and [A574](#). The average primary gradient within each node was compared between raw and minimally processed data ([Fig. 4d](#);  $r_{\text{validity}}=0.59$ ,  $\text{rmse}_{\text{validity}}=0.036$ ).

Finally, the peak alpha frequency (Hz) was estimated from the Cam-CAN MEG data filtered by *brainlife.io* apps and data Maxwell-filtered by Cam-CAN using [A476](#) and [A531](#)<sup>61,62</sup>. Peak alpha values were compared between the two differently processed datasets ([Fig. 4e](#);  $r_{\text{validity}}=0.94$ ,  $\text{rmse}_{\text{validity}}=0.30$  Hz).

For each data modality, data preprocessing reliability was defined as the ability to produce the same results given repeated *acquisitions* from within a participant. Data processing reliability was examined by comparing brain features estimated using *brainlife.io* Apps pipelines using either test-retest HCP<sub>TR</sub> data or a median split of Cam-CAN MEG data.

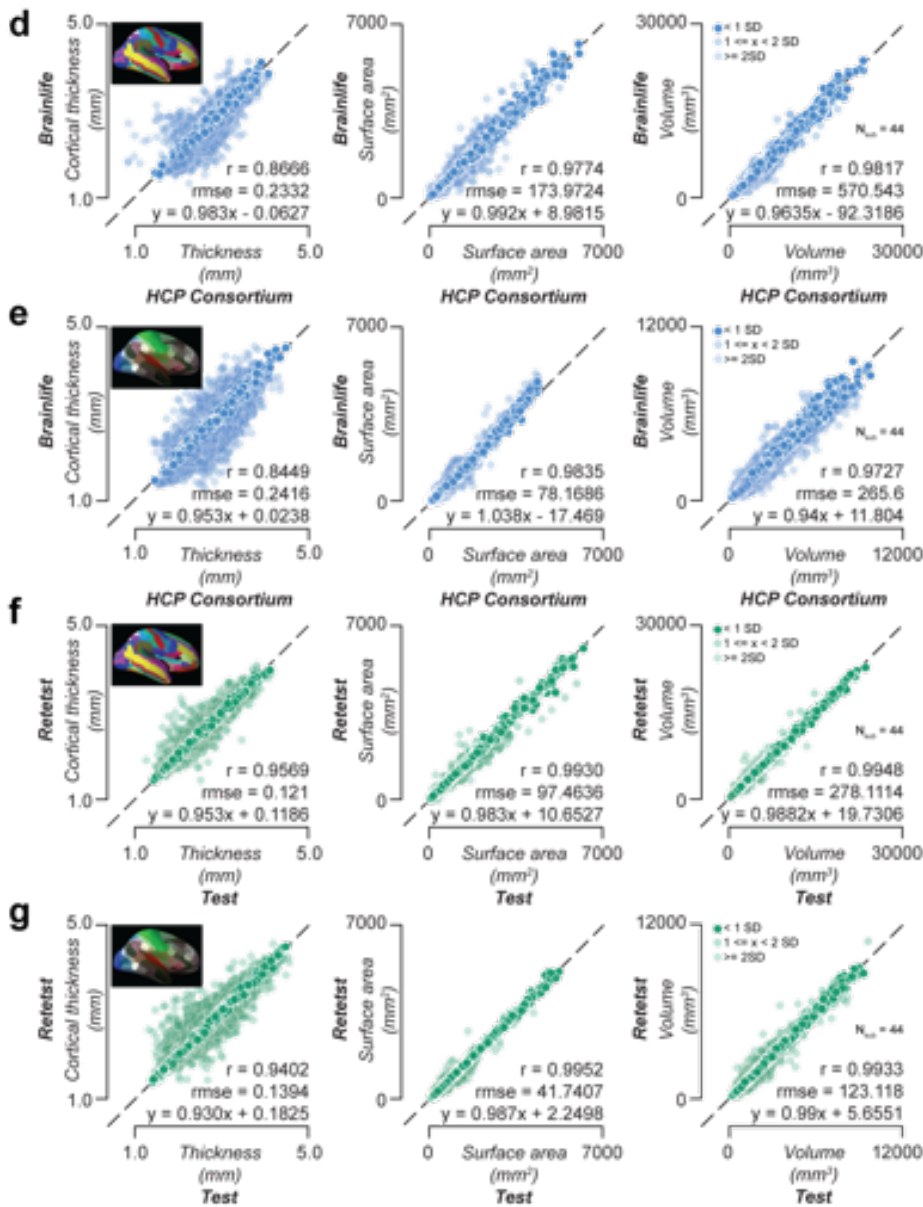
Cortical parcel volumes from the test and retest dataset of HCP<sub>TR</sub> were obtained using [A23](#), [A272](#), and [A464](#) *brainlife.io* Apps (see [Table S3](#)) and compared ([Fig. 4f](#);  $r_{\text{reliability}}=0.99$ ,  $\text{rmse}_{\text{reliability}}=278.11\text{mm}^3$ ).

Mean FA from 61 white matter tracts was estimated independently for *test* and *retest* HCP<sub>TR</sub> dMRI data using [A238](#), [A297](#), [A305](#), [A188](#), [A195](#), and [A361](#). The average FA for each tract was compared between test and retest conditions ( $r_{\text{reliability}}=0.93$ ,  $\text{rmse}_{\text{reliability}}=0.017$ ) ([Fig. 4g](#)).

Functional connectivity estimates between 117<sup>2</sup> nodes-pairs were estimated using the test and retest HCP<sub>TR</sub> fMRI data using [A23](#), [A369](#), and [A532](#) ([Fig. 4h](#);  $r_{\text{reliability}}=0.73$ ,  $\text{rmse}_{\text{reliability}}=0.19$ ).

In addition, functional gradients were computed on 400 nodes estimated on test and retest HCP<sub>TR</sub> fMRI data using [A604](#) and [A574](#). The average primary gradient within each node was compared between datasets ([Fig. 4i](#);  $r_{\text{reliability}}=0.85$ ,  $\text{rmse}_{\text{reliability}}=0.026$ ).

Finally, the frequency of the amplitude peak (between 8 and 13 Hz from the occipital magnetometers and gradiometers) was estimated from two median splits of Maxwell-filtered Cam-CAN MEG data using [A529](#) and [A531](#). Peak alpha frequency values were compared between the two datasets ( $r_{\text{reliability}}=0.85$ ,  $\text{rmse}_{\text{reliability}}=0.48$  Hz; [Fig. 4j](#)). All estimated validity and reliability estimates are reported in [Table S4](#).

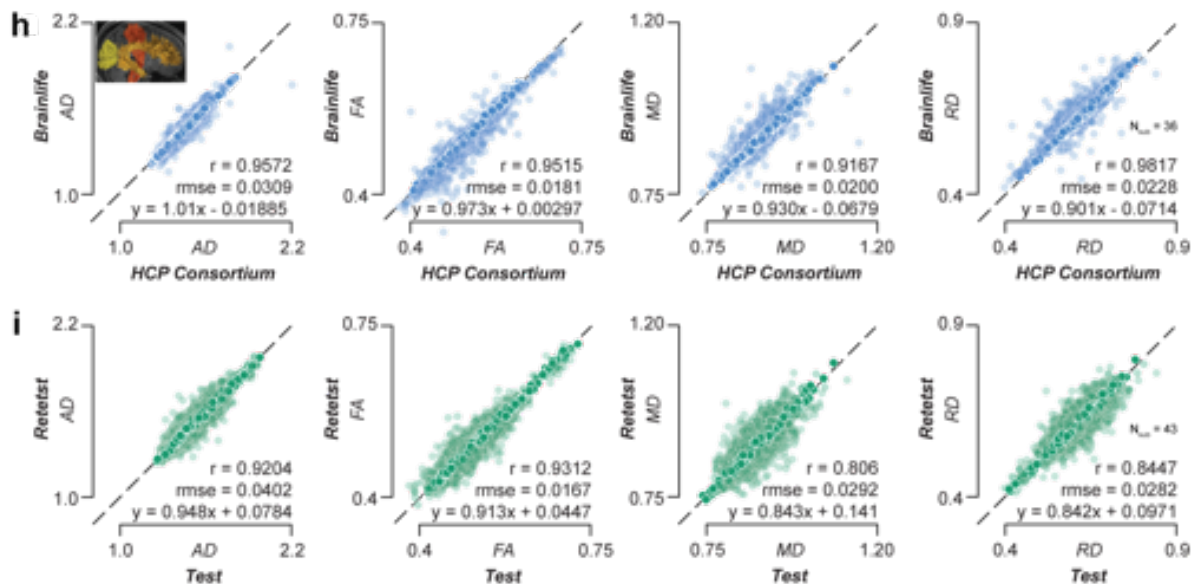


**Supplemental Figure 4d-g. Processing with brainlife.io is valid and test-retest reliability is high. Top rows:** Validity measures derived using the HCP<sub>TR</sub> data preprocessed and provided by the HCP Consortium compared to data preprocessed on brainlife.io. Each dot corresponds to the ratio for a given subject between data preprocessed and provided by the HCP Consortium vs data preprocessed on brainlife.io in a given measure for a given structure. Pearson's correlation ( $r$ ), root mean squared error ( $rmse$ ), and a linear fit between the test and retest results were calculated and provided. **a.** Destrieux Parcel thickness (mm), surface area (mm<sup>2</sup>), and volume (mm<sup>3</sup>). **b.** HPC-mmp Parcel thickness (mm), surface area (mm<sup>2</sup>), and volume (mm<sup>3</sup>). Dark colors represent data within  $\pm 1$  standard deviation. 50% opacity represents data within 1-2 standard deviations. 25% opacity represents data outside 2 standard deviations. **Bottom rows:** Test-retest reliability measures derived from derivatives of the HCP<sub>TR</sub> dataset generated using brainlife.io. Each dot corresponds to the ratio between a test-retest subject and a given measure for a given structure. Pearson's correlation ( $r$ ), root mean squared error ( $rmse$ ), and a linear fit between the test and retest results were calculated and provided. **c.** Destrieux Parcel thickness (mm), surface area (mm<sup>2</sup>), and volume (mm<sup>3</sup>). **d.** HPC-mmp Parcel thickness (mm), surface area (mm<sup>2</sup>), and volume (mm<sup>3</sup>). Dark colors represent data within  $\pm 1$  standard deviation. 50% opacity represents data within 1-2 standard deviations. 25% opacity represents data outside 2 standard deviations.

Computational reproducibility was defined as the ability to produce the same results given repeated *runs* of a processing app. Computational reproducibility was estimated by comparing values obtained from repeated runs of *brainlife.io* Apps.

Cortical parcel volumes were estimated twice from the minimally processed HCP<sub>TR</sub> data ( $N_{\text{sub}} = 44$ ) using [A272](#). Volume estimates between the repeat run were compared (**Fig. S4j**;  $r_{\text{reproducibility}} = 0.99$ ,  $\text{rmse}_{\text{reproducibility}} = 34.22 \text{ mm}^3$ ).

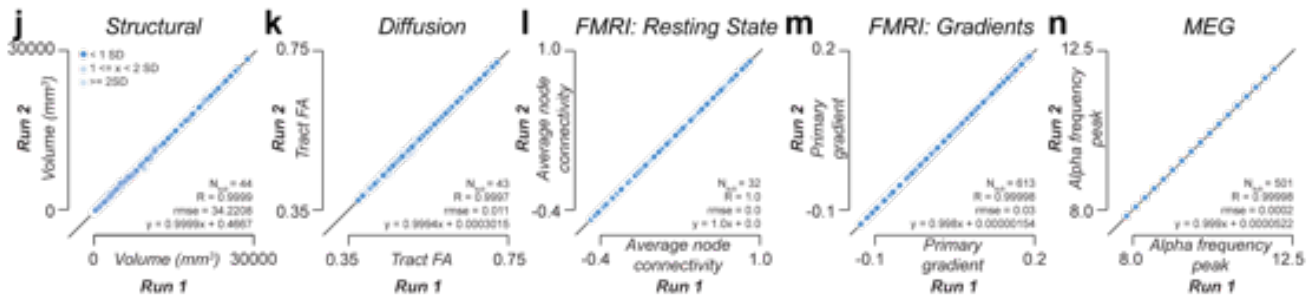
Fractional anisotropy (FA) in 61 white matter tracts was estimated from the minimally processed HCP<sub>TR</sub> data ( $N_{\text{sub}} = 43$ ) using [A361](#). The average FA for each tract was compared between repeated runs (**Fig. S4k**;  $r_{\text{reproducibility}} = 0.99$ ,  $\text{rmse}_{\text{reproducibility}} = 0.011$ ).



**Supplemental Figure 4h-i. Processing with brainlife.io is valid and test-retest reliability. Top row:** Validity measures derived using the HCP<sub>TR</sub> data preprocessed and provided by the HCP Consortium compared to data preprocessed on brainlife.io. Each dot corresponds to the ratio for a given subject between data preprocessed and provided by the HCP Consortium vs data preprocessed on brainlife.io in a given measure for a given structure. Pearson's correlation ( $r$ ), root mean squared error ( $\text{rmse}$ ), and a linear fit between the test and retest results were calculated and provided. **e.** Tract average AD, FA, MD, and RD. Dark colors represent data within  $\pm 1$  standard deviation. 50% opacity represents data within 1-2 standard deviations. 25% opacity represents data outside 2 standard deviations. **Bottom row:** Test-retest reliability measures derived from derivatives of the HCP<sub>TR</sub> dataset generated using brainlife.io. Each dot corresponds to the ratio between a test-retest subject and a given measure for a given structure. Pearson's correlation ( $r$ ), root mean squared error ( $\text{rmse}$ ), and a linear fit between the test and retest results were calculated and provided. **f.** Tract average AD, FA, MD, and RD. Dark colors represent data within  $\pm 1$  standard deviation. 50% opacity represents data within 1-2 standard deviations. 25% opacity represents data outside 2 standard deviations.

Functional connectivity estimates between  $117^2$  node pairs were estimated using the minimally processed test HCP<sub>TR</sub> data ( $N_{\text{sub}} = 32$ ) using [A532](#). Average node connectivity was compared between repeated runs (**Fig. S4l**;  $r_{\text{reproducibility}} = 1.0$ ,  $\text{rmse}_{\text{reproducibility}} = 0.0$ ). In addition, functional gradients were computed on 400 nodes estimated from the Cam-CAN data ( $N_{\text{sub}} = 613$ ) using [A574](#).

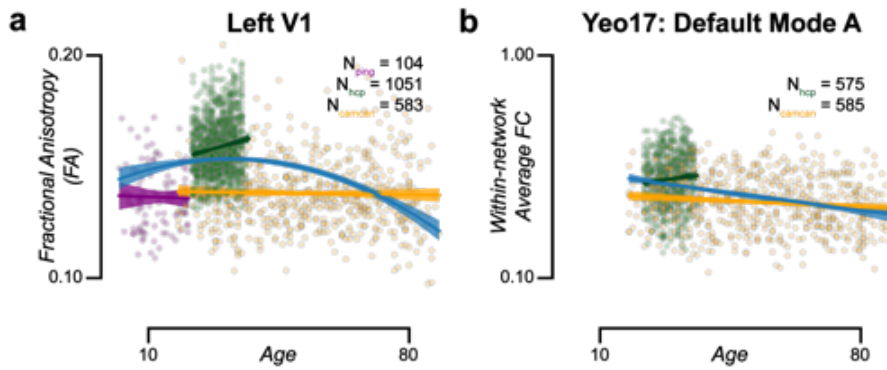
Finally, primary gradient values were compared between repeated runs (**Fig. S4m**;  $r_{\text{reproducibility}} = 0.99$ ,  $\text{rmse}_{\text{reproducibility}} = 0.03$ ). Finally, the peak alpha frequency (Hz) was estimated from the Maxwell-filtered MEEG Cam-CAN data ( $N_{\text{sub}} = 501$ ) using [A531](#). Peak alpha values were compared between repeated runs (**Fig. S4n**;  $r_{\text{reproducibility}} = 0.99$ ,  $\text{rmse}_{\text{reproducibility}} = 0.0002$ ).



**Supplemental Figure 4j-n. App computational reproducibility.** Computational reproducibility values derived by repeating runs of brainlife.io Apps using the HCP<sub>TR</sub> dataset and the CAN dataset. Each dot corresponds to the ratio for a given subject between repeated runs of each App for a given structure. Pearson's correlation ( $r$ ), root mean squared error ( $rmse$ ), and a linear fit between the repeated runs was calculated. **a.** Destrieux Atlas Parcels volume ( $\text{mm}^3$ ). **b.** Tract-average fractional anisotropy (FA). **c.** Node-average functional connectivity (FC). **d.** Primary gradient values derived from resting state fMRI. **e.** Peak alpha frequency (Hz) in the alpha band derived from MEG.

### Supplemental platform utility for scientific applications

Evaluation of the scientific utility of the platform was performed on over 2,000 participants across three large datasets with participant ages spanning over 7 decades—PING (Pediatric Imaging, Neurocognition, Genetics), HCP<sub>s1200</sub>, (Human Connectome Project Young Adult 1,200) and Cam-CAN (Cambridge Center for Ageing Neuroscience). Multiple brain features were derived, including fractional anisotropy of cortical parcels and within-network functional connectivity of individual Yeo17 networks. Specifically, for structural MRI data, the volumes of the cortical and subcortical structures segmented for each participant were compared to their age at the time of scan acquisition on a per-structure basis. Volume measures were estimated using [A464](#), [A462](#), [A272](#), and [A379](#). For diffusion MRI data, the average FA for each of the white matter tracts segmented for each participant was compared to the participant age at scan acquisition on a per-structure basis. Tract average FA values were estimated using [A361](#). In addition to white matter tract FA, average FA within cortical regions was computed using [A383](#). For resting-state functional MRI connectivity, the average within-network connectivity values, defined as the average connectivity values between all of the nodes within each resting state network of the Yeo17 parcellation, was compared to the participant's age at scan acquisition. Network connectivity matrices were estimated using [A532](#). For resting-state functional gradients, the cosine distance of the primary gradient for each of the resting state networks in the Schaffer parcellation was compared to the participant's age at scan acquisition. Gradients were mapped using [A574](#). Finally, for MEG data, the peak frequency in the alpha band across all nodes was compared to the participant's age at the time of acquisition. Peak frequency was estimated using [A531](#). For structural and diffusion MRI data, data from all three data sources (HCP<sub>s1200</sub>, Cam-CAN, PING) was used. For the functional MRI data, data from only the HCP<sub>s1200</sub> and Cam-CAN data sources were used. For the MEG data, only the data from the Cam-CAN data source was used. To assess the relationship between each of the measures and age within each structure investigated, a quadratic model ( $y_{feature} = ax_{age}^2 + bx_{age} + c$ ) was fit across all of the data, and a linear regression was fit within each data source, using functions from scikit-learn<sup>63</sup>. Two additional examples are presented in **Fig. S5**, specifically the average fractional anisotropy (FA) or cortical V1 (**Fig. S5a**) and the within-network average functional connectivity within the default mode (A) network derived from the Yeo17 atlas (**Fig. S5b**). The quadratic model ( $R^2=0.12 \pm 0.015$  s.d.) for these two examples demonstrated the expected inverted U-shape trajectory, with the mean quadratic term ( $a$ ) across each data modality being negative ( $-3.70 \times 10^{-6} \pm 6.60 \times 10^{-6}$  s.d.).

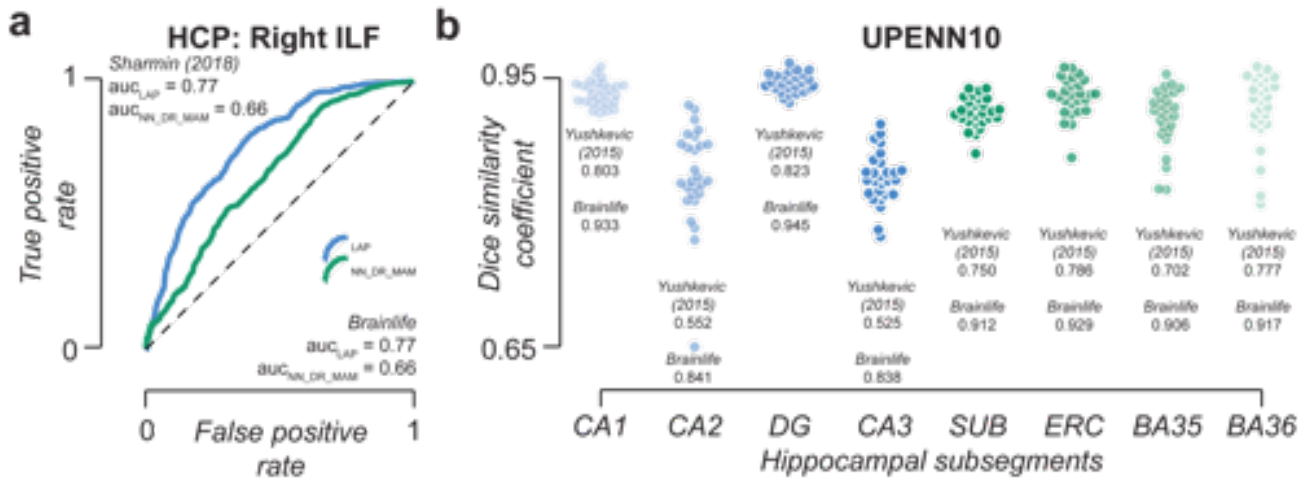


**Supplemental Figure 5. Additional examples of inverted U-shaped trajectories.** Relationship between age of subject and **a.** Cortical fractional anisotropy (FA) of the left V1, **b.** Within-network average functional connectivity (FC) from the Yeo17 Default Mode - A network. These analyses include subjects from the PING (*purple*), HCP<sub>s1200</sub> (*green*), and CAN (*yellow*) datasets. Linear regressions were fit to each dataset, and a quadratic regression was fit to the entire dataset (blue).

### Supplemental replication and generalization

In addition to the replication experiments, five sets of generalization experiments were performed (**Fig. 6; Fig. S6a,b**). First, we tested *brainlife.io*'s ability to replicate scientific results from five previous studies<sup>64-66</sup>. A key finding from each previous study was identified as the target found to be reproduced. We then followed the processing methods as outlined in the original study but performed these processing methods using *brainlife.io* Apps. Post-processing analyses were performed in line with the original study using *brainlife.io*-hosted Jupyter Notebooks (see [Table S2](#)). Replicability success was measured by comparing trends in the data obtained with *brainlife.io* Apps and those reported in the original study.

Replicability was defined as the ability to reproduce individual experiments already published by other members of the scientific community. Within replicability are two pillars: the ability to reproduce results within the *same* dataset, and the ability to generalize results to *new* datasets. Three sets of experiments were performed to assess the ability of the platform to replicate previously published findings. The first experiment attempted to replicate a reported negative correlation between a cortical region's thickness and its tissue orientation organization within the HCP<sub>s1200</sub> dataset. Cortical regions found within the HCP multi-modal parcellation (hcp-mmp) parcellation were first mapped to each participant's Freesurfer surfaces using [A23](#). Brainlife apps [A464](#), [A462](#), [A272](#), and [A379](#) were then used to map and estimate each region's cortical thickness and orientation dispersion index (ODI), respectively. The relationship between ODI and cortical thickness was assessed by computing the correlation between these values across all parcels within the hcp-mmp parcellation (**Fig. 6a**). The second experiment attempted to replicate the improved ability to segment the Inferior Longitudinal Fasciculus from the HCP<sub>s1200</sub> dataset (**Fig. S6a**)<sup>32</sup>. The Right Inferior Longitudinal Fasciculus (ILF) was segmented from the HCP<sub>s1200</sub> dataset using an automated segmentation algorithm ([A174](#)). The same improved ability of tract segmentation was obtained (**Fig. S6a**;  $AUC_{LAP} = 0.77$ ,  $AUC_{NN\_DR\_MAM} = 0.66$ ). The third study used to assess replicability investigated the performance of an automated hippocampal subfield segmentation as compared to hand-drawn regions of interest (ROIs)<sup>67</sup>. The original implementation was performed with a dice coefficient ranging from 0.525-0.823. An App ([A262](#)) was created to implement this segmentation on *brainlife*. The method was implemented on participants from the UPENN-PMC dataset. Improved model performance was obtained for segmenting hippocampal subfields (**Fig. S6b**; dice range = 0.838-0.945).



**Supplemental Figure 6a,b. Replication of previous studies using brainlife.io** **a.** Receiver operator curves (ROC) comparing the performance of segmentation of the Right ILF using two automated segmentation methods (LAP: blue, NN\_DR\_MAM: green) in a subset of the HCP<sub>S1200</sub> dataset ( $N_{\text{sub}} = 15$ ). **b.** Dice coefficients between manual and automated segmentation of the hippocampus using AHSS method in UPENN dataset.

In addition to the replication experiments, three sets of generalization experiments were performed. The first experiment attempted to generalize the same relationship between a cortical region's thickness and orientation dispersion index found within the HCP<sub>S1200</sub> dataset to the Cam-CAN dataset (**Fig. 6a**). *brainlife.io* Apps [A464](#), [A462](#), [A272](#), and [A379](#) were then used to map and estimate each region's cortical thickness and orientation dispersion index (ODI), respectively. The relationship between ODI and cortical thickness was assessed by computing the correlation between these values across all parcels within the hcp-mmp parcellation. A negative trend of about half the magnitude of the original was estimated (**Fig. 6a**;  $r_{\text{Cam-CAN-brainlife}} = -0.28$  vs.  $r_{\text{original}}$ ). The second and third experiments attempted to generalize a relationship between the average quantitative anisotropy (QA) and fractional anisotropy (FA) of the left and right uncinate with the presence of stressful life events as an adolescent (**Fig. 6b,c**). The second experiment assessed tract organization within the UF of 42 participants from within the HBN dataset using [A423](#) to extract the UFs and to map QA to each, respectively. These values were then compared to the number of negative life events as reported on the Negative Life Events Schedule (NLES) collected by the HBN group. A negative relationship between UF QA and number of stressful life events was identified (**Fig. 6b**  $r_{\text{HBN\_LEFT}} = -0.35$ ,  $p\text{-value} < 0.05$ ;  $r_{\text{HBN\_RIGHT}} = -0.39$ ,  $p\text{-value} < 0.05$ ). The third experiment attempted to find the same relationship using FA within 1,107 participants from the ABCD dataset. For this, an end-to-end white matter processing pipeline composed of [A68](#), [A238](#), [A297](#), [A305](#), [A188](#), [A195](#), and [A361](#) was used to extract the UF and to map FA to each tract. These values were then compared to the measure of early life stress was estimated as a composite score by z-scoring separately and then summing across the following questionnaires: traumatic life events reported by the parent, environmental and neighborhood safety reported by both parent and adolescent, and the Family Environment Scale-Family Conflict Subscale Modified from PhenX reported by both parent and adolescent <sup>68</sup>. A negative relationship between UF FA and the composite score was estimated in the left- and right-UF (**Fig. 6c**  $r_{\text{ABCD\_LEFT}} = -0.12$ ,  $p\text{-value} < 0.001$ ;  $r_{\text{ABCD\_RIGHT}} = -0.09$ ,  $p < 0.01$ ).

### Supplemental to detecting disease

The final two tests to demonstrate the platform's potential and scientific utility focused on identifying human disease biomarkers. We examined data from multiple clinical populations including sports-related concussions, glaucoma, Stargardt's, Choroiderema, and healthy populations who have experienced stressful life events to assess the ability to identify unique clinical characteristics using the platform; **Fig. 7**). It has been reported that concussions can alter brain tissue properties both in the cortex and in deep white matter tracts <sup>69</sup>. Here, the differences in cortical white matter tissue in concussed and matched controls were tested. Specifically, for sports-related concussion, 10 concussed athletes and 10 healthy within-sport control athletes from the Indiana University Acute Concussion dataset (in prep) was used. FA was estimated in 358 cortical parcels from the Human Connectome Project multimodal parcellation <sup>70</sup> using a pipeline composed of [A23](#), [A272](#), [A379](#), and [A464](#). The distribution of FA in the superior temporal sulcus (STS) is reported in **Fig. 7a**. One example athlete with strong



post-concussive symptoms and low FA (red arrow) is compared to the distribution of controls (gray) to demonstrate the ability to detect meaningful changes in brain tissue following a concussive event.

Changes in the visual white matter as a result of eye disease have been reported<sup>38,71–74</sup>. Individuals with Stargardt’s disease (a deterioration of the human retina initiating in the central fovea), and Choroideremia (retinal deterioration initiating in the visual periphery), were compared to healthy controls. Optical coherence tomography (OCT) data were processed using [A346](#) (**Fig. 7b**). Photoreceptor complex thickness (microns) was estimated for foveal and peripheral (0-1 and 7-90 degrees of visual eccentricity) regions. Choroideremia patients showed similar levels of photoreceptor complex thickness compared to healthy controls in the foveal bundle but deviated in the peripheral bundle (**Fig. 7b**). This trend was the opposite for Stargardt’s participants. To study the degree to which retinal damage affects the brain’s white matter (the optic radiation, OR), data were processed using a series of Apps ([A273](#), [A462](#), [A187](#), [A414](#), [A233](#), [A361](#), [A68](#), [A238](#), and [A346](#)). Visual eccentricity maps in area V1 were separated between foveal (0-1° of visual angle) and peripheral (7-90° of visual angle) regions<sup>75,76</sup>. Tractography was used to separate OR bundles projecting to the foveal and peripheral maps, and average FA profiles for each group and bundle were computed<sup>77 78,79</sup>. Results show a reduction in FA in the component of the OR projecting to foveal (but not peripheral) V1 in Stargardt’s patients (**Fig. 7b**, blue). Results also show a reduction in FA in the Choroideremia patients’ peripheral (but not foveal) bundle (**Fig. 7b**, blue). Taken together, these results demonstrate the ability of the technology implemented in the platform to measure disease biomarkers.

### Supplement to quality control at scale

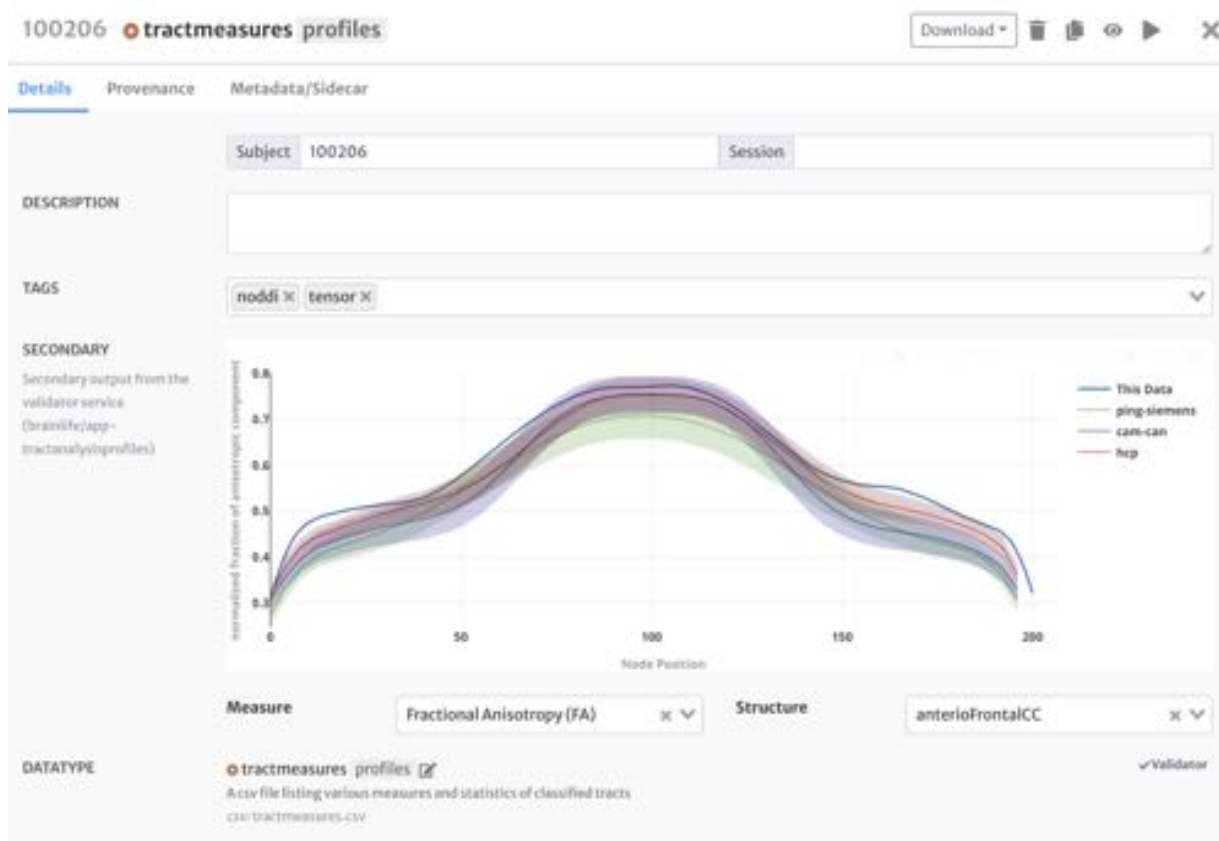
To assure quality in processed data, brainlife.io provides a unique approach to quality assurance (QA). State-of-the-art approaches to QA provide users with the ability to assess quality after data processing is compiled into QA reports<sup>22,23,80–83</sup>. The platform supports QA reports outputted by state-of-the-art processing pipelines ([A160](#), [A246](#), [A160](#), [A462](#), [A423](#), [A399](#)), as well as via QA images, which can be assessed by individuals or groups. Here we propose an additional approach to QA via *normalized reference ranges*, in which brain properties derived from many participants, modalities, and sources of variability are collated together for quick identification of aberrant brain derivatives<sup>84</sup>.

Normalized reference ranges were generated and are served on the brainlife.io platform in addition to the standard QA reports that are generated within Apps for individual datasets. To generate the reference ranges, the brain properties derived from the three datasets (PING, HCP<sub>s1200</sub>, and Cam-CAN) and four data modalities in 1,751 participants generated for the load testing of the platform (as described in the previous sections) were curated (removed of outliers) and collated for *brainlife.io* datatype. For each datatype, a single JSON file was created reporting the mean and  $\pm 1$  and 2 standard deviations of the outlier-removed measure (e.g., the volume of a brain parcel, fractional anisotropy of a white matter tract, functional connectivity of a network, power-spectrum density across MEG sensors, etc). The JSON files were saved on a repository ([github.com/brainlife/reference](https://github.com/brainlife/reference)) and the brainlife.io datatype validator service made use of the JSON to automatically visualize a plot of the data. We call these JSON files reference datasets. Users utilizing Apps ([A272](#), [A463](#), [A483](#), [A361](#), [A530](#), [A531](#), [A532](#)) that generate datatypes for which a reference dataset was created will find the values of the features estimated by the App on any new dataset overlaid on top of the corresponding reference dataset (see **Fig. S8**). We report examples of reference dataset plots for four major datatypes, with outliers data overlaid on top (**Fig. 8a-d**) and the final *reference* datasets for each datatype and data source.

A critical aspect to democratizing big data neuroscience is the ability of investigators to perform quality assurance (QA), because there is no value in increasing dataset size unless quality can be assured for each dataset. State-of-the-art approaches provide users with the ability to assess quality after data processing is compiled into QA reports<sup>22,23,80–83</sup>, or through the use of citizen science<sup>85</sup>. The brainlife.io platform supports visualization of the QA reports outputted by state-of-the-art processing pipelines ([A160](#), [A246](#), [A160](#), [A462](#), [A423](#), [A399](#)), as well as via QA images, which can be assessed by individuals or groups. Here we propose an additional approach to QA via *normalized reference ranges*, in which brain properties derived from many participants, modalities, and sources of variability are collated together for quick identification of abnormal brain derivatives<sup>84</sup>. For more details of the generation of these reference ranges, see **Methods**. Here we provide an example of the brainlife.io visualizations of reference datasets (**Fig. S8**).

The brain properties derived from the three datasets (PING, HCP<sub>s1200</sub>, and CAN) and four data modalities in 1,751 participants generated for the load testing of the platform (as described in the previous sections) were curated (removed of outliers) and collated for brainlife.io datatype. For each datatype, a single JSON file was created reporting the mean and  $\pm 1$  and 2 standard deviations of the outlier-removed measure (e.g., the volume of a brain

parcel, fractional anisotropy of a white matter tract, functional connectivity of a network, power-spectrum density across MEG sensors, etc). The JSON files were saved on a repository (<https://github.com/brainlife/reference>) and the brainlife.io datatype validator service made use of the JSON to automatically visualize a plot of the data. We call these JSON files reference datasets. Users utilizing Apps ([A272](#), [A463](#), [A483](#), [A361](#), [A530](#), [A531](#), [A532](#)) that generate datatypes for which a reference dataset was created will find the values of the features estimated by the App on any new dataset overlaid on top of the corresponding reference dataset (**Fig. S8**).



**Supplemental Figure 8. brainlife.io interface can visualize reference datasets.** Validation services for datatypes containing statistical feature information automatically generate a visualization of newly generated data (*blue line*) overlaid on reference dataset ranges for the three data sources used to generate reference datasets (i.e. HCP<sub>S1200</sub> (*red*), PING (*green*), CAN (*purple*). These reference ranges can be used to quickly assess the quality of the estimated statistical features of interest.

### Public services for promoting transparency and data gravity in neuroscience research.

In the previous section, we described the system architecture for the platform. These components and architectures were implemented in order to reduce barriers of entry to performing neuroimaging investigations and to ultimately increase data gravity and representation in neuroscience. These goals coincide with a push within the neuroimaging community to increase data gravity and representation by providing standardization of data formatting, software libraries, and computing resources. From this push has come an ever-growing list of publicly available services and platforms for increasing data gravity in neuroimaging. However, there currently exists only one compiled list of the services available<sup>86</sup>. To address this, and to help increase transparency in neuroscientific research, we provide a non-comprehensive list of currently available services and platforms for increasing data gravity across the greater neuroimaging community (**Table S5**). This list is not designed to cover all currently available services and platforms, but to provide a sense of the scope of available technologies developed by the neuroscientific community.

## The FAIR principles.

Recently, it has been proposed that platforms should respect the FAIR principle<sup>87</sup>. *brainlife.io* was built with the FAIR principles in mind and below, we pair each FAIR principle with the modern definition of neuroscience data. In the *brainlife.io* project, each principle is applied to multiple research assets, data derivatives, analysis software, and software services.

The three primary research assets pertaining to the *brainlife.io* project are (1) data, derivatives, and metadata, (2) processing applications and data analysis code, and (3) data and analysis management services are each made FAIR via the *brainlife.io* project.

**Findable.** Research data services available on *brainlife.io* such as data sets, processing App, web services and analysis code are either automatic or manual mechanisms to make them findable. *brainlife.io* assigns Digital-Objects-Identifiers (DOI) using DataCite as a partner project. DOIs are automatically assigned to publication records consisting of datasets, as well as versioned preprocessing and analysis software. These *brainlife.io* publication records are compliant with schema.org and as such are also compliant with Google Dataset search (<https://datasetsearch.research.google.com>). DOIs are also assigned to each published App.

**Accessible.** Data and metadata can be retrieved using a number of access methods via Web Interfaces and Command Line Interfaces. Metadata is also accessible programmatically via a web API. Metadata remains available even in the case that data must be removed (e.g., in cases of human subjects concerns). Authentication is necessary to access the data and users' identities are checked by humans to assure compliance with more restrictive data-access policies such as the GDPR. A full record of data management and processing is made accessible. So not just data or analysis streams are accessible but a full record reporting the provenance of each individual data product. The code underlying each processing App is accessible via GitHub, and can be modified or used via common GitHub mechanisms (push requests, pull requests). Previously published datasets can be downloaded to a local machine or copied to a new project.

**Interoperable.** Data can be submitted to *brainlife.io* either using standard file types such as NifTis, but also data from multiple vendors can be used to map the data to the BIDS standard and uploaded on the system using the *brainlife.io/ezBIDS* web tool. The *brainlife.io/ezBIDS* system allows data from multiple vendors and type of sequences to be mapped to the Brain Imaging Data Structure (BIDS) and from there to be pushed to *brainlife.io* Projects, to OpenNeuro.org or downloaded. Furthermore, datasets can be mapped from major archives and projects such as NKI, and OpenNeuro.org using DataLad.org. Finally, *brainlife.io* Apps on their own also use are FAIR, as they are publicly available both as services on *brainlife.io* and code implementing the services on GitHub. The Apps can be stored either on individual user or organization accounts or on the *brainlife.io* team GitHub account depending on the level of commitment of the app developer to maintaining the Apps. The *brainlife.io* team maintains a *bl2bids* (<https://github.com/brainlife/abcd-spec/blob/master/hooks/bl2bids.py>) and the BIDS Walked (<https://github.com/brainlife/cli/blob/master/bids-walker.js>) script that together allow mapping BIDS data types to *brainlife.io* DataTypes. As a result the BIDS standard is the data exchange approach used to increase data interoperability.

**Reusable.** The *brainlife.io* project has multiple aspects of its technology that is developed with a mindset focus of reuse. First, the whole platform is developed as open source and published on GitHub.com. Second, the data processing Applications are developed using a lightweight specification that is compatible with BIDS and can be easily used without *brainlife.io* interfaces on local computers or clusters. Finally, data assets can be shared within the platform across users and projects but also outside of the platform by downloading the data as BIDS-compliant datasets. Data derivatives, processing apps, and analysis notebooks can be accessed in multiple ways via web graphical user interfaces, command line interfaces, or directly via local download. Analysis notebooks in the form of Jupyter notebooks can be pushed to GitHub directly, allowing for instantaneous reuse by the broader community. Data pipelines can be copied and reused within a given project. All configuration parameters for each App are stored, allowing users to reuse previously defined optimal parameters for their given data. The *brainlife.io* publication model is a key component to implementing a vision of an integrated project publication containing data, and preprocessing for future reuse.

**Supplemental Table 1: Platform services serving the brainlife.io platform.**

| Service            | Description   | GitHub Repos  |
|--------------------|---|---|
| UI                 | Platform endpoint, providing an user interface that integrates the diverse services in Brainlife      | <a href="https://github.com/brainlife/warehouse/tree/master/ui">https://github.com/brainlife/warehouse/tree/master/ui</a>   |
| Warehouse          | Data storage and management   | <a href="https://github.com/brainlife/warehouse/">https://github.com/brainlife/warehouse/</a>                               |
| Amaretti           | Automated scheduling service identifying appropriate compute resources and staging and archiving data | <a href="https://github.com/brainlife/amaretti/">https://github.com/brainlife/amaretti/</a>                                 |
| ezBIDS             | DICOM to BIDS conversion  | <a href="https://github.com/brainlife/ezbids/">https://github.com/brainlife/ezbids/</a>                                     |
| Vis                | Services available for running visualizations within the platform                                     | <a href="https://github.com/brainlife/brainlife/tree/master/vis">https://github.com/brainlife/brainlife/tree/master/vis</a> |
| Event              | Event-driven integrator, to provide real-time feedback for users                                      | <a href="https://github.com/brainlife/event/">https://github.com/brainlife/event/</a>                                       |
| Service Monitoring | Monitors individual actions performed by the site   | <a href="https://github.com/brainlife/servicemonitor">https://github.com/brainlife/servicemonitor</a>                       |
| CLI                | Command-line interface for performing data manipulations and data scrubbing                           | <a href="https://github.com/brainlife/cli">https://github.com/brainlife/cli</a>   |
| Auth               | Centralized authentication for the multiple Brainlife services  | <a href="https://github.com/brainlife/auth">https://github.com/brainlife/auth</a>   |

**Supplemental Table 1.** Table with list of all platform services, name, scope, service URL (pointer to brainlife page if available as direct URL) and github URL for code.

**Supplemental Table 2: Jupyter notebooks for analyses performed.**

| Notebook Name                                   | Topic                  | Analysis/Figure   | Datatype(s)         | Measure(s)  | Github URL  |
|---|------------------------|---|---------------------|---|---|
| blp-analysis-structural-mri-volume.ipynb        | Structural morphometry | Validity, reliability, reproducibility, development, references | neuro/parc-stats    | Cortical parcel volume, thickness, surface area, Fractional Anisotropy (FA), Axial Diffusivity (AD), Radial Diffusivity (RD), Mean Diffusivity (MD), Neurite density index (NDI), Orientation dispersion index (ODI), Isotropic volume fraction (IsoVF) | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-structural-mri-volume.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-structural-mri-volume.ipynb</a>               |
| blp-analysis-diffusion-mri-tract-profiles.ipynb | Diffusion profilometry | Validity, reliability, reproducibility, development, references | neuro/tractmeasures | White matter tract Fractional Anisotropy (FA), Axial Diffusivity (AD), Radial Diffusivity (RD), Mean Diffusivity (MD), Neurite density index (NDI), Orientation dispersion index (ODI), Isotropic   | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-diffusion-mri-tract-profiles.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-diffusion-mri-tract-profiles.ipynb</a> |

|   |  |   |   |   |   |
|---|--|---|---|---|---|
|   |  |   |   | volume fraction (IsoVF)   |   |
| blp-analysis-diffusion-mri-structural-connectivity.ipynb  | Structural connectivity                                    | Validity, reliability, reproducibility, development, references | neuro/network                             | Max node degree   | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-diffusion-mri-structural-connectivity.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-diffusion-mri-structural-connectivity.ipynb</a>   |
| blp-analysis-functional-mri-functional-connectivity.ipynb | Functional connectivity                                    | Validity, reliability, reproducibility, development, references | neuro/network                             | Within-network connectivity   | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-functional-mri-functional-connectivity.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-functional-mri-functional-connectivity.ipynb</a> |
| blp-analysis-functional-mri-gradients.ipynb               | Functional gradients                                       | Validity, reliability, reproducibility, development, references | neuro/gradients                           | Distance of primary gradient  | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-functional-mri-gradients.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-functional-mri-gradients.ipynb</a>                             |
| blp-analysis-meeg-power-spectrum-density.ipynb            | MEEG   | Validity, reliability, reproducibility, development, references | neuro/meeg/psd                            | Peak alpha frequency, power spectrum density  | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-meeg-power-spectrum-density.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-meeg-power-spectrum-density.ipynb</a>                       |
| blp-analysis-concussion-structural-mri.ipynb              | Cortical diffusion   | Clinical populations  | neuro/parc-stats                          | Cortical parcel volume, thickness, surface area, Fractional Anisotropy (FA), Axial Diffusivity (AD), Radial Diffusivity (RD), Mean Diffusivity (MD), Neurite density index (NDI), Orientation dispersion index (ODI), Isotropic volume fraction (IsoVF) | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-concussion-structural-mri.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-concussion-structural-mri.ipynb</a>                           |
| blp-analysis-inherited-retinal-disease.ipynb              | Diffusion profilometry, optical coherence tomography (OCT) | Clinical populations  | neuro/tractmeasures, neuro/microperimetry | White matter tract Fractional Anisotropy (FA), Photoreceptor thickness  | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-inherited-retinal-disease.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-inherited-retinal-disease.ipynb</a>                           |
| blp-analysis-usage-statistics.ipynb                       | Platform usage statistics                                  | NA  | NA  | NA  | <a href="https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-usage-statistics.ipynb">https://github.com/bacaron/bp-notebooks/bl_paper/blp-analysis-usage-statistics.ipynb</a>   |

**Supplemental Table 2.** Description and web-links to the open-source code used for each analysis outlined previously in the form of individual Jupyter Notebooks.

**Supplemental Table 3: Preprocessing Apps used for the experiments.**

| Name   | Brainlife DOI   | Github Repository  |
|--|---|--|
| Anatomically Constrained Tractography using precomputed 5tt & CSD                                    | <a href="https://doi.org/10.25663/brainlife.app.297">10.25663/brainlife.app.297</a> | <a href="https://github.com/bacaron/app-mrtrix3-act">bacaron/app-mrtrix3-act</a>   |
| mrtrix3 - WMC Anatomically Constrained Tractography (ACT)  | <a href="https://doi.org/10.25663/brainlife.app.319">10.25663/brainlife.app.319</a> | <a href="https://github.com/brainlife/app-mrtrix3-act">brainlife/app-mrtrix3-act</a>   |
| Compile tract macro-structural and profile data  | <a href="https://doi.org/10.25663/brainlife.app.397">10.25663/brainlife.app.397</a> | <a href="https://github.com/brainlife/app-compile-macro-micro-tract-stats">brainlife/app-compile-macro-micro-tract-stats</a>       |
| Compute summary statistics of diffusion measures from subcortical segmentation                       | <a href="https://doi.org/10.25663/brainlife.app.389">10.25663/brainlife.app.389</a> | <a href="https://github.com/brainlife/app-freesurfer-stats">brainlife/app-freesurfer-stats</a>                                     |
| Compute summary statistics of diffusion measures mapped to the cortical surface - Deprecated Surface | <a href="https://doi.org/10.25663/brainlife.app.383">10.25663/brainlife.app.383</a> | <a href="https://github.com/brainlife/app-cortex-tissue-mapping-stats">brainlife/app-cortex-tissue-mapping-stats</a>               |
| Conmat 2 Network   | <a href="https://doi.org/10.25663/brainlife.app.393">10.25663/brainlife.app.393</a> | <a href="https://github.com/filipinascimento/bl-conmat2network">filipinascimento/bl-conmat2network</a>                             |
| Convert network neuro matrix to conmat   | <a href="https://doi.org/10.25663/brainlife.app.335">10.25663/brainlife.app.335</a> | <a href="https://github.com/brainlife/app-network-matrices-2-mat">brainlife/app-network-matrices-2-mat</a>                         |
| Cortex Tissue Mapping (Native & Template Space)  | <a href="https://doi.org/10.25663/brainlife.app.379">10.25663/brainlife.app.379</a> | <a href="https://github.com/brainlife/app-cortex-tissue-mapping">brainlife/app-cortex-tissue-mapping</a>                           |
| Fit Constrained Deconvolution Model for Tracking   | <a href="https://doi.org/10.25663/brainlife.app.238">10.25663/brainlife.app.238</a> | <a href="https://github.com/bacaron/app-mrtrix3-act">bacaron/app-mrtrix3-act</a>   |
| Freesurfer   | <a href="https://doi.org/10.25663/bl.app.0">10.25663/bl.app.0</a>                   | <a href="https://github.com/brainlife/app-freesurfer">brainlife/app-freesurfer</a>   |
| Freesurfer Statistics  | <a href="https://doi.org/10.25663/brainlife.app.272">10.25663/brainlife.app.272</a> | <a href="https://github.com/brainlife/app-freesurfer-stats">brainlife/app-freesurfer-stats</a>                                     |
| FSL Anat (T1)  | <a href="https://doi.org/10.25663/brainlife.app.273">10.25663/brainlife.app.273</a> | <a href="https://github.com/brainlife/app-fsl-anat">brainlife/app-fsl-anat</a>   |
| Align T1 to ACPC Plane (HCP-based)   | <a href="https://doi.org/10.25663/bl.app.99">10.25663/bl.app.99</a>                 | <a href="https://github.com/brainlife/app-hcp-acpc-alignment">brainlife/app-hcp-acpc-alignment</a>                                 |
| FSL Anat (T2)  | <a href="https://doi.org/10.25663/brainlife.app.350">10.25663/brainlife.app.350</a> | <a href="https://github.com/brainlife/app-fsl-anat">brainlife/app-fsl-anat</a>   |
| FSL Brain Extraction (BET) on DWI  | <a href="https://doi.org/10.25663/brainlife.app.163">10.25663/brainlife.app.163</a> | <a href="https://github.com/brainlife/app-FSLBET">brainlife/app-FSLBET</a>   |
| mrtrix3 preprocess   | <a href="https://doi.org/10.25663/bl.app.68">10.25663/bl.app.68</a>                 | <a href="https://github.com/brainlife/validator-neuro-dwi">brainlife/validator-neuro-dwi</a>                                       |
| Multi-Atlas Transfer Tool (w/surface output)   | <a href="https://doi.org/10.25663/bl.app.23">10.25663/bl.app.23</a>                 | <a href="https://github.com/faskowit/app-multiAtlasTT">faskowit/app-multiAtlasTT</a>   |
| Noddi Amico  | <a href="https://doi.org/10.25663/brainlife.app.365">10.25663/brainlife.app.365</a> | <a href="https://github.com/brainlife/app-noddi-amico">brainlife/app-noddi-amico</a>   |
| Parcellation Statistics - Surface - Deprecated Datatype  | <a href="https://doi.org/10.25663/brainlife.app.464">10.25663/brainlife.app.464</a> | <a href="https://github.com/brainlife/app-freesurfer-stats">brainlife/app-freesurfer-stats</a>                                     |
| Remove Tract Outliers  | <a href="https://doi.org/10.25663/brainlife.app.195">10.25663/brainlife.app.195</a> | <a href="https://github.com/brainlife/validator-neuro-wmc">brainlife/validator-neuro-wmc</a>                                       |
| Tissue-type segmentation   | <a href="https://doi.org/10.25663/brainlife.app.239">10.25663/brainlife.app.239</a> | <a href="https://github.com/brainlife/app-mrtrix3-5tt">brainlife/app-mrtrix3-5tt</a>   |
| Tract Analysis Profiles  | <a href="https://doi.org/10.25663/brainlife.app.361">10.25663/brainlife.app.361</a> | <a href="https://github.com/brainlife/app-tractanalysisprofiles">brainlife/app-tractanalysisprofiles</a>                           |
| Tractography quality check   | <a href="https://doi.org/10.25663/brainlife.app.189">10.25663/brainlife.app.189</a> | <a href="https://github.com/brainlife/app-tractographyQualityCheck">brainlife/app-tractographyQualityCheck</a>                     |
| White Matter Anatomy Segmentation  | <a href="https://doi.org/10.25663/brainlife.app.188">10.25663/brainlife.app.188</a> | <a href="https://github.com/brainlife/validator-neuro-wmc">brainlife/validator-neuro-wmc</a>                                       |
| Align T2 to ACPC Plane (HCP-based)   | <a href="https://doi.org/10.25663/brainlife.app.116">10.25663/brainlife.app.116</a> | <a href="https://github.com/brainlife/app-hcp-acpc-alignment/tree/1.4">brainlife/app-hcp-acpc-alignment/tree/1.4</a>               |
| fMRIPrep - Volume Output   | <a href="https://doi.org/10.25663/brainlife.app.160">10.25663/brainlife.app.160</a> | <a href="https://github.com/brainlife/app-fmriprep/tree/20.2.3-2">brainlife/app-fmriprep/tree/20.2.3-2</a>                         |
| pRFs / Benson14-Retinotopy - Deprecated  | <a href="https://doi.org/10.25663/brainlife.app.187">10.25663/brainlife.app.187</a> | <a href="https://github.com/davhunt/app-benson14-retinotopy/tree/master">davhunt/app-benson14-retinotopy/tree/master</a>           |
| Segment thalamic nuclei  | <a href="https://doi.org/10.25663/brainlife.app.222">10.25663/brainlife.app.222</a> | <a href="https://github.com/brainlife/app-segment-thalamic-nuclei/tree/v1.0">brainlife/app-segment-thalamic-nuclei/tree/v1.0</a>   |
| Track The Human Optic RAdiation (THORA): Contrack - Eccentricity                                     | <a href="https://doi.org/10.25663/brainlife.app.252">10.25663/brainlife.app.252</a> | <a href="https://github.com/brainlife/app-contrack-optic-radiation/tree/v1.1">brainlife/app-contrack-optic-radiation/tree/v1.1</a> |
| Automated Segmentation of Hippocampal Subfields (ASHS)   | <a href="https://doi.org/10.25663/brainlife.app.262">10.25663/brainlife.app.262</a> | <a href="https://github.com/svincibo/app-ashs-segment/tree/master">svincibo/app-ashs-segment/tree/master</a>                       |
| fMRIPrep - Surface Output  | <a href="https://doi.org/10.25663/brainlife.app.267">10.25663/brainlife.app.267</a> | <a href="https://github.com/brainlife/app-fmriprep/tree/20.2.1">brainlife/app-fmriprep/tree/20.2.1</a>                             |

|   |   |   |
|---|---|---|
| FSL DTIFIT  | <a href="https://doi.org/10.25663/brainlife.app.292">10.25663/brainlife.app.292</a> | brainlife/app-fsDTIFIT/tree/v1.1  |
| fMRI Timeseries Extraction  | <a href="https://doi.org/10.25663/brainlife.app.369">10.25663/brainlife.app.369</a> | faskowit/app-fmri-2-mat/tree/0.1.6  |
| Structural Connectome MRTrix3 (SCMRT) - No labels or weights  | <a href="https://doi.org/10.25663/brainlife.app.395">10.25663/brainlife.app.395</a> | brainlife/app-sift2-connectome-generation/tree/no_sift2_v1.2_centers_netneuro |
| Generate Visual Regions of Interest Binned by Eccentricity Estimates (Benson Atlas) - Diffusion Space | <a href="https://doi.org/10.25663/brainlife.app.414">10.25663/brainlife.app.414</a> | brainlife/app-roiGenerator/tree/visual-white-matter-eccentricity-dwi-v1.2     |
| dsi-studio-atk  | <a href="https://doi.org/10.25663/brainlife.app.423">10.25663/brainlife.app.423</a> | frankyeh/dsi-studio-atk/tree/master   |
| Apply Maxwell filter on MEG signals using MNE-python  | <a href="https://doi.org/10.25663/brainlife.app.476">10.25663/brainlife.app.476</a> | brainlife/app-maxwell-filter/tree/master                                      |
| Compute summary statistics of diffusion measures mapped to cortical surface                           | <a href="https://doi.org/10.25663/brainlife.app.483">10.25663/brainlife.app.483</a> | brainlife/app-cortex-tissue-mapping-stats/tree/updated-surface-dtype-v1.1     |
| Split MEG file  | <a href="https://doi.org/10.25663/brainlife.app.529">10.25663/brainlife.app.529</a> | guiomar/app-meg-split-fif/tree/main   |
| PSD: Power Spectral Density (Welch method)  | <a href="https://doi.org/10.25663/brainlife.app.530">10.25663/brainlife.app.530</a> | guiomar/app-psd/tree/main   |
| Find frequency peak of PSD data   | <a href="https://doi.org/10.25663/brainlife.app.531">10.25663/brainlife.app.531</a> | guiomar/app-peak-frequency/tree/master  |
| Time series to network  | <a href="https://doi.org/10.25663/brainlife.app.532">10.25663/brainlife.app.532</a> | filipinascimento/bl-timeseries2network/tree/0.2                               |
| Connectivity Gradients  | <a href="https://doi.org/10.25663/brainlife.app.574">10.25663/brainlife.app.574</a> | anibalsolon/app-connectivity-gradient/tree/main                               |
| Average channels  | <a href="https://doi.org/10.25663/brainlife.app.599">10.25663/brainlife.app.599</a> | guiomar/app-average-channels/tree/main  |

**Supplemental Table 3.** Description and web links to the open-source code and open cloud services used to perform the evaluation experiments described in the main article.

**Supplementary Table 4. Validity and reliability correlation tables.**

| Modality       | Measure               | Analysis    | Parcellation | r      | rmse      |
|----------------|-----------------------|-------------|--------------|--------|-----------|
| Structural MRI | Cortical thickness    | Validity    | Destrieux    | 0.8667 | 0.2332    |
| "              | Cortical surface area | Validity    | Destrieux    | 0.9774 | 173.9724  |
| "              | Cortical volume       | Validity    | Destrieux    | 0.9817 | 570.543   |
| "              | Cortical thickness    | Reliability | Destrieux    | 0.9569 | 0.121     |
| "              | Cortical surface area | Reliability | Destrieux    | 0.9930 | 97.4636   |
| "              | Cortical volume       | Reliability | Destrieux    | 0.9948 | 2378.1114 |
| "              | Cortical thickness    | Validity    | hcp-mmp      | 0.8449 | 0.2416    |
| "              | Cortical surface area | Validity    | hcp-mmp      | 0.9835 | 78.1686   |
| "              | Cortical volume       | Validity    | hcp-mmp      | 0.9727 | 265.6     |
| "              | Cortical thickness    | Reliability | hcp-mmp      | 0.9402 | 0.1394    |
| "              | Cortical surface area | Reliability | hcp-mmp      | 0.9952 | 41.7407   |
| "              | Cortical volume       | Reliability | hcp-mmp      | 0.9933 | 123.118   |
| Diffusion MRI  | Tract AD              | Validity    | wma          | 0.9572 | 0.0309    |
| "              | Tract FA              | Validity    | wma          | 0.9515 | 0.0181    |
| "              | Tract MD              | Validity    | wma          | 0.9167 | 0.0200    |
| "              | Tract RD              | Validity    | wma          | 0.9817 | 0.0228    |
| "              | Tract AD              | Reliability | wma          | 0.9204 | 0.0402    |
| "              | Tract FA              | Reliability | wma          | 0.9312 | 0.0167    |
| "              | Tract MD              | Reliability | wma          | 0.806  | 0.0292    |
| "              | Tract RD              | Reliability | wma          | 0.8447 | 0.0282    |
| Functional MRI | Node connectivity     | Validity    | Yeo17        | 0.8853 | 0.1219    |
| "              | Node connectivity     | Reliability | Yeo17        | 0.7264 | 0.1889    |
| "              | Primary gradient      | Validity    | Shaffer400   | 0.5934 | 0.0358    |
| "              | Primary gradient      | Reliability | Shaffer400   | 0.8496 | 0.0259    |
| MEEG           | Peak alpha frequency  | Validity    | NA           | 0.9385 | 0.2964    |
| "              | Peak alpha frequency  | Reliability | NA           | 0.8484 | 0.4751    |

**Supplemental Table 4.** Pearson correlation (*r*) and root mean square error (*rmse*) for all validity and reliability experiments performed.



**Supplemental Table 5: Resources for data storage, archiving, and computational analysis.**

| Location(s)           | Archive Name   | Web URL  | Type                    | Archive Representative  | Data Modality (-ies)  | Type of access       | Reference (publication) |
|-----------------------|--|--|-------------------------|---|---|----------------------|-------------------------|
| U.S.A                 | <a href="#">BRAIN Initiative Cell Census Network (BICCN)</a> | <a href="http://www.biccn.org/">www.biccn.org/</a>   | service registry        | Multiple; the Allen Institute has an NIH grant to build and host this site, through the Brain Cell Data Center (BCDC) | human, mouse; single cell RNA-Seq, Patch-Seq, cell morphologies, electrophysiological recordings (NWB files), multiple histological image modalities, mFISH |                      |                         |
| US BRAIN              | BICCN Single Cell Portal                                     | <a href="http://singlecell.broadinstitute.org/single-cell">singlecell.broadinstitute.org/single-cell</a> | service registry        | Broad Institute<br>scp-support@broadinstitute.zendesk.com   | Multiple single cell datasets   | N/A                  |                         |
| US BRAIN              | <a href="#">OpenNeuro.org</a>                                | <a href="http://OpenNeuro.org">OpenNeuro.org</a>   | Archive                 | Russ Poldrack   | human MRI, PET, EEG,  |                      |                         |
| US BRAIN              | DABI archive   | <a href="http://dabi.loni.usc.edu/home">dabi.loni.usc.edu/home</a>                                       | Archive                 | TOGA, ARTHUR W  | EEG, MEG, iEEG  |                      |                         |
| US BRAIN              | Allen Brain Map  | <a href="http://portal.brain-map.org">portal.brain-map.org</a>   | service registry        | Allen Institute - multiple teams involved   | human, mouse, rhesus macaque  |                      |                         |
| US BRAIN              | DANDI  | <a href="http://www.dandiarchive.org/">www.dandiarchive.org/</a>   | Archive                 | Satrajit Ghosh  | Neurophysiology (EPhys, ICEphys, Ophys)   |                      |                         |
| US BRAIN              | NeMO   | <a href="http://nemoarchive.org/">nemoarchive.org/</a>   | Archive                 | Owen R. White   | Multi-omics data  |                      |                         |
| US BRAIN              | Brain Image Library (BIL)                                    | <a href="http://www.brainimaginglibrary.org/">www.brainimaginglibrary.org/</a>                           | service registry        | ROPELEWSKI, ALEXANDER J   | Brain imaging data  |                      |                         |
| US BRAIN              | BossDB   | <a href="http://bosssdb.org/">bosssdb.org/</a>   | Archive                 | WESTER, BROCK A.  | EM  |                      |                         |
| US BRAIN              | MiCRONS Explorer   | <a href="http://microns-explorer.org/">microns-explorer.org/</a>   | web-service             | Multiple  | EM  |                      |                         |
| US BRAIN              | [their main site]  | <a href="http://www.braininitiative.org/resources/">www.braininitiative.org/resources/</a>               | service registry        |   | aggregator  |                      |                         |
| US BRAIN              | <a href="#">brainlife.io</a>                                 | <a href="http://brainlife.io">brainlife.io</a>   | computational platforms | Franco Pestilli   | MRI/EEG/MEG   | Governed via license |                         |
| Australian Initiative |  | <a href="http://neurodesk.org">neurodesk.org</a>   | web-service             |   |   |                      |                         |
| Japan Initiative      | SRPBS  | <a href="http://www.cns.atr.jp/decnefpro/">www.cns.atr.jp/decnefpro/</a>                                 | service registry        | Saori Tanaka, Mitsuo Kawato   | Brain imaging data  |                      |                         |
| Japan Initiative      | Brain/MINDS Beyond   | <a href="http://mriportal.umin.jp/">mriportal.umin.jp/</a>   | service registry        | Kiyoto Kasai, Takashi Hanakawa, Saori Tanaka  | Brain imaging data  |                      |                         |

|  |                        |  |                  |                          |   |                       |  |
|--|------------------------|--|------------------|--------------------------|---|-----------------------|--|
| Japan Initiative   | Brain/MINDS            | <a href="http://www.brainminds.riken.jp/">www.brainminds.riken.jp/</a>             | service registry | Alex Woodward            | Marmoset atlas, fMRI, dMRI, tracer, gene expression   | Open to collaborators |  |
| China Initiative   | Linked Brain Data      | <a href="http://www.linked-brain-data.org/">www.linked-brain-data.org/</a>         | service registry |                          |   |                       |  |
| Korea Initiative   | Korea Brain Initiative | <a href="http://kbrain-map.kbri.re.kr:8080/">kbrain-map.kbri.re.kr:8080/</a>       | service registry | Sung-Jin Jeong           | mouse; single cell RNA-Seq, EM data (current); omics data, behavioural data, electrophysiology data (in future) |                       |  |
| European Human Brain Project   | EBRAINS                | <a href="http://ebrains.eu/">ebrains.eu/</a>                                       | service registry | Jan Bjaalie              | Brain imaging data, omics data, behavioural data, electrophysiology data, models etc                            | Closed                |  |
| Canadian Open Neuroscience Platform  | CONP                   | <a href="http://conp.ca/">conp.ca/</a>   | service registry | CONP committee           | Brain imaging data, omics data, behavioural data, electrophysiology data, models etc                            | Governed via license  |  |
| BlueBrainProject   |                        | <a href="http://channelpedia.epfl.ch/">channelpedia.epfl.ch/</a>                   | service registry |                          |   |                       |  |
| DataLad  |                        | <a href="http://datasets.datalad.org/">datasets.datalad.org/</a>                   | service registry |                          |   | Fully open (CC-00)    |  |
| NITRC  |                        |  | service registry |                          |   |                       |  |
| USA  | WebPlotDigitizer       | <a href="http://automeris.io/WebPlotDigitizer/">automeris.io/WebPlotDigitizer/</a> | web-service      | Ankit Rohatgi            |   |                       |  |
| USA  | Brain Map Database     | <a href="http://brainmap.org">brainmap.org</a>                                     | web-service      | Peter Fox                | Brain Imaging data  | Governed via license  |  |
| USA  | NeuroSynth Database    | <a href="http://neurosynth.org">neurosynth.org</a>                                 | web-service      | Alejandro de la Vega     | Brain Imaging data  | Fully open (CC-00)    |  |
| France   | NeuroQuery             | <a href="https://neuroquery.org">https://neuroquery.org</a>                        | web-service      | INRIA/ Jérôme Dockès     | Brain Imaging data  | Fully open (CC-00)    |  |
|  | OSF                    | <a href="http://osf.io">osf.io</a>   | Archive          |                          | Unspecified / Open  | Unspecified           |  |
| U.S.A.   | COINSTAC               | <a href="https://coinstac.org/">https://coinstac.org/</a>                          | Downloadable     | Georgia State University | Brain Imaging Data  | Unspecified           |  |
| <b>Supplemental Table 5.</b> Description and web links to the many available platforms and services for increasing data gravity in the neuroimaging field. |                        |  |                  |                          |   |                       |  |

**Supplemental Table 6: Processed dataset published as part of this article.**

| Project   | DOI   | Brainlife Publication URL   |
|---|---|---|
| Human Connectome Young Adult - Test - Retest  | <a href="https://doi.org/10.25663/brainlife.pub.38">https://doi.org/10.25663/brainlife.pub.38</a> | <a href="https://brainlife.io/pub/640a3da8c538c16a826f912e">https://brainlife.io/pub/640a3da8c538c16a826f912e</a> |
| Human Connectome Young Adult - Full Dataset   | <a href="https://doi.org/10.25663/brainlife.pub.40">https://doi.org/10.25663/brainlife.pub.40</a> | <a href="https://brainlife.io/pub/640a3f9dc538c16a826f9b1a">https://brainlife.io/pub/640a3f9dc538c16a826f9b1a</a> |
| Cambridge Centre for Ageing and Neuroscience - Full Dataset   | <a href="https://doi.org/10.25663/brainlife.pub.39">https://doi.org/10.25663/brainlife.pub.39</a> | <a href="https://brainlife.io/pub/640a3f0cc538c16a826f9648">https://brainlife.io/pub/640a3f0cc538c16a826f9648</a> |
| MEG [fif] Cam-Can   | <a href="https://doi.org/10.25663/brainlife.pub.41">https://doi.org/10.25663/brainlife.pub.41</a> | <a href="https://brainlife.io/pub/640a40fec538c16a826fa468">https://brainlife.io/pub/640a40fec538c16a826fa468</a> |
| MEG [fif] Run1 vs Run2  | <a href="https://doi.org/10.25663/brainlife.pub.42">https://doi.org/10.25663/brainlife.pub.42</a> | <a href="https://brainlife.io/pub/640a4155c538c16a826fa5b9">https://brainlife.io/pub/640a4155c538c16a826fa5b9</a> |
| MEG [fif] CamCan-maxfilt  | <a href="https://doi.org/10.25663/brainlife.pub.43">https://doi.org/10.25663/brainlife.pub.43</a> | <a href="https://brainlife.io/pub/640a41abc538c16a826fa6e6">https://brainlife.io/pub/640a41abc538c16a826fa6e6</a> |
| ASHS Segmentation of Hippocampal Subfields - Replication derivatives  | <a href="https://doi.org/10.25663/brainlife.pub.44">https://doi.org/10.25663/brainlife.pub.44</a> | <a href="https://brainlife.io/pub/640a4267c538c16a826fb09a">https://brainlife.io/pub/640a4267c538c16a826fb09a</a> |
| <b>Supplemental Table 1.</b> Table with list of all platform services, name, scope, service URL (pointer to brainlife page if available as direct URL) and github URL for code. |   |   |

## REFERENCES

1. Stewart, C. A. *et al.* Jetstream: A self-provisioned, scalable science and engineering cloud environment. (2015) doi:10.1145/2792745.2792774.
2. Hancock, D. Y. *et al.* Jetstream2: Accelerating cloud computing via Jetstream. in *Practice and Experience in Advanced Research Computing* 1–8 (Association for Computing Machinery, 2021). doi:10.1145/3437359.3465565.
3. Dale, A., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *Neuroimage* **9**, 179–194 (1999).
4. Fischl, B., Sereno, M. I. & Dale, A. Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. *Neuroimage* **9**, 195–207 (1999).
5. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).
6. Fischl, B., Liu, A. & Dale, A. M. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Medical Imaging* **20**, 70–80 (2001).
7. Fischl, B. *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).
8. Fischl, B. *et al.* Sequence-independent segmentation of magnetic resonance images. *Neuroimage* **23**, S69–S84 (2004).
9. Fischl, B., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
10. Fischl, B. *et al.* Automatically Parcellating the Human Cerebral Cortex. *Cereb. Cortex* **14**, 11–22 (2004).
11. Han, X. *et al.* Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *Neuroimage* **32**, 180–194 (2006).
12. Jovicich, J. *et al.* Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *Neuroimage* **30**, 436–443 (2006).
13. Kuperberg, G. R. *et al.* Regionally localized thinning of the cerebral cortex in Schizophrenia. *Arch. Gen. Psychiatry* **60**, 878–888 (2003).

14. Reuter, M., Schmansky, N. J., Rosas, H. D. & Fischl, B. Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. *Neuroimage* **61**, 1402–1418 (2012).
15. Reuter, M. & Fischl, B. Avoiding Asymmetry-Induced Bias in Longitudinal Image Processing. *Neuroimage* **57**, 19–21 (2011).
16. Reuter, M., Rosas, H. D. & Fischl, B. Highly Accurate Inverse Consistent Registration: A Robust Approach. *Neuroimage* **53**, 1181–1196 (2010).
17. Salat, D. *et al.* Thinning of the cerebral cortex in aging. *Cereb. Cortex* **14**, 721–730 (2004).
18. Segonne, F. *et al.* A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**, 1060–1075 (2004).
19. Segonne, F., Pacheco, J. & Fischl, B. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans. Med. Imaging* **26**, 518–529 (2007).
20. Tournier, J.-D., Calamante, F. & Connelly, A. MRtrix: Diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* **22**, 53–66 (2012).
21. Tournier, J.-D. *et al.* MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage* **202**, 116137 (2019).
22. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
23. Cieslak, M. *et al.* QSIprep: an integrative platform for preprocessing and reconstructing diffusion MRI data. *Nat. Methods* **18**, 775–778 (2021).
24. Developers, S. *SingularityCE* 3.8.3. (2021). doi:10.5281/zenodo.5564915.
25. Merkel, D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* **2014**, 2 (2014).
26. DataCite Schema. *DataCite Schema* <https://schema.datacite.org/meta/kernel-4.1/index.html>.
27. Gonzalez-Beltran, A. N. *et al.* Data discovery with DATS: exemplar adoptions and lessons learned. *J. Am. Med. Inform. Assoc.* **25**, 13–16 (2018).
28. Gonzalez-Beltran, A. & Rocca-Serra, P. *biocaddie/WG3-MetadataSpecifications: DataMed DATS specification v2.2 - NIH BD2K bioCADDIE*. (2017). doi:10.5281/zenodo.438337.

29. Caron, B. *et al.* Collegiate athlete brain data for white matter mapping and network neuroscience. *Sci Data* **8**, 56 (2021).
30. McPherson, B. C. & Pestilli, F. A single mode of population covariation associates brain networks structure and behavior and predicts individual subjects' age. *Commun Biol* **4**, 943 (2021).
31. Bertò, G. *et al.* Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation. *Neuroimage* **224**, 117402 (2021).
32. Sharmin, N., Olivetti, E. & Avesani, P. White Matter Tract Segmentation as Multiple Linear Assignment Problems. *Front. Neurosci.* **11**, 754 (2017).
33. Vinci-Booher, S., Caron, B., Bullock, D., James, K. & Pestilli, F. Development of white matter tracts between and within the dorsal and ventral streams. *Brain Struct. Funct.* **227**, 1457–1477 (2022).
34. Kurzawski, J. W., Mikellidou, K., Morrone, M. C. & Pestilli, F. The visual white matter connecting human area prostriata and the thalamus is retinotopically organized. *Brain Struct. Funct.* **225**, 1839–1853 (2020).
35. Sani, I. *et al.* The human endogenous attentional control network includes a ventro-temporal cortical node. *Nat. Commun.* **12**, 360 (2021).
36. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
37. Puzniak, R. J. *et al.* CHIASM, the human brain albinism and achiasma MRI dataset. *Sci Data* **8**, 308 (2021).
38. Hanekamp, S. *et al.* White matter alterations in glaucoma and monocular blindness differ outside the visual system. *Sci. Rep.* **11**, 6866 (2021).
39. Cheng, H. *et al.* Denoising diffusion weighted imaging data using convolutional neural networks. *PLoS One* **17**, e0274396 (2022).
40. Eke, D. O. *et al.* International data governance for neuroscience. *Neuron* (2021)  
doi:10.1016/j.neuron.2021.11.017.
41. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).
42. Halchenko, Y. *et al.* DataLad: distributed system for joint management of code, data, and their relationship. *J. Open Source Softw.* **6**, 3262 (2021).

43. Markiewicz, C. J. *et al.* The OpenNeuro resource for sharing of neuroscience data. *Elife* **10**, (2021).
44. Nooner, K. B. *et al.* The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Front. Neurosci.* **6**, 152 (2012).
45. Tobe, R. H. *et al.* A longitudinal resource for studying connectome development and its psychiatric associations during childhood. *Sci Data* **9**, 300 (2022).
46. Di Martino, A. *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
47. Dean, J. & Ghemawat, S. MapReduce: simplified data processing on large clusters. *Commun. ACM* **51**, 107–113 (2008).
48. Kluyver, T. *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (eds. Loizides, F. & Schmidt, B.) 87–90 (IOS Press, 2016). doi: Kluyver, Thomas, Ragan-Kelley, Benjamin, Pérez, Fernando, Granger, Brian, Bussonnier, Matthias, Frederic, Jonathan, Kelley, Kyle, Hamrick, Jessica, Grout, Jason, Corlay, Sylvain, Ivanov, Paul, Avila, Damián, Abdalla, Safia, Willing, Carol and Jupyter development team, (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. Loizides, Fernando and Schmidt, Birgit (eds.) In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press. pp. 87-90 . (doi:10.3233/978-1-61499-649-1-87 <<http://dx.doi.org/10.3233/978-1-61499-649-1-87>>). .
49. Perez, F. & Granger, B. E. IPython: A System for Interactive Scientific Computing. *Comput. Sci. Eng.* **9**, 21–29 (2007).
50. Wickham, H. Tidy Data. *J. Stat. Softw.* **59**, 1–23 (2014).
51. Avesani, P. *et al.* The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. *Sci Data* **6**, 69 (2019).
52. Alexander, L. M. *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci Data* **4**, 170181 (2017).
53. National Academies of Sciences, Engineering *et al.* *Understanding Reproducibility and Replicability*. (National Academies Press (US), 2019).
54. Kelley, T. L. Interpretation of educational measurements. **353**, (1927).

55. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
56. Shafto, M. A. *et al.* The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol.* **14**, 204 (2014).
57. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
58. Yeo, B. T. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
59. Bethlehem, R. A. I. *et al.* Dispersion of functional gradients across the adult lifespan. *Neuroimage* **222**, 117299 (2020).
60. Margulies, D. S. *et al.* Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12574–12579 (2016).
61. Taulu, S. & Kajola, M. Presentation of electromagnetic multichannel data: The signal space separation method. *J. Appl. Phys.* **97**, 124905 (2005).
62. Taulu, S. & Simola, J. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Phys. Med. Biol.* **51**, 1759–1768 (2006).
63. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
64. Fukutomi, H. *et al.* Neurite imaging reveals microstructural variations in human cerebral cortical gray matter. *Neuroimage* **182**, 488–499 (2018).
65. Ho, T. C. *et al.* Effects of sensitivity to life stress on uncinate fasciculus segments in early adolescence. *Soc. Cogn. Affect. Neurosci.* **12**, 1460–1469 (2017).
66. Hanson, J. L., Knodt, A. R., Brigidi, B. D. & Hariri, A. R. Lower structural integrity of the uncinate fasciculus is associated with a history of child maltreatment and future psychological vulnerability to stress. *Dev. Psychopathol.* **27**, 1611–1619 (2015).
67. Yushkevich, P. A. *et al.* Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Hum. Brain Mapp.* **36**, 258–287 (2015).



68. Karcher, N. R. & Barch, D. M. The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology* **46**, 131–142 (2021).
69. McKee, A. C., Daneshvar, D. H., Alvarez, V. E. & Stein, T. D. The neuropathology of sport. *Acta Neuropathol.* **127**, 29–51 (2014).
70. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
71. Yoshimine, S. *et al.* Age-related macular degeneration affects the optic radiation white matter projecting to locations of retinal damage. *Brain Struct. Funct.* **223**, 3889–3900 (2018).
72. Ogawa, S. *et al.* White matter consequences of retinal receptor and ganglion cell damage. *Invest. Ophthalmol. Vis. Sci.* **55**, 6976–6986 (2014).
73. Malania, M., Konrad, J., Jäggle, H., Werner, J. S. & Greenlee, M. W. Compromised Integrity of Central Visual Pathways in Patients With Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **58**, 2939–2947 (2017).
74. Sherbondy, A. J., Dougherty, R. F., Ben-Shachar, M., Napel, S. & Wandell, B. A. ConTrack: finding the most likely pathways between brain regions using diffusion tractography. *J. Vis.* **8**, 15.1–16 (2008).
75. Benson, N. C. *et al.* The retinotopic organization of striate cortex is well predicted by surface topology. *Curr. Biol.* **22**, 2081–2085 (2012).
76. Benson, N. C., Butt, O. H., Brainard, D. H. & Aguirre, G. K. Correction of distortion in flattened representations of the cortical surface allows prediction of V1-V3 functional organization from anatomy. *PLoS Comput. Biol.* **10**, e1003538 (2014).
77. Yeatman, J. D., Dougherty, R. F., Myall, N. J., Wandell, B. A. & Feldman, H. M. Tract profiles of white matter properties: automating fiber-tract quantification. *PLoS One* **7**, e49790 (2012).
78. Aydogan, D. B. & Shi, Y. Parallel Transport Tractography. *IEEE Trans. Med. Imaging* **40**, 635–647 (2021).
79. Baran, D. & Shi, Y. A novel fiber-tracking algorithm using parallel transport frames. in *ISMRM* (unknown, 2019).
80. Esteban, O. *et al.* MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* **12**, e0184661 (2017).
81. Yeh, F.-C. Shape analysis of the human association pathways. *Neuroimage* **223**, 117329 (2020).
82. Cameron, C. *et al.* Towards automated analysis of connectomes: The configurable pipeline for the analysis of

- connectomes (C-PAC). *Front. Neuroinform.* **7**, (2013).
83. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
84. Yanni, S. E. *et al.* Normative reference ranges for the retinal nerve fiber layer, macula, and retinal layer thicknesses in children. *Am. J. Ophthalmol.* **155**, 354–360.e1 (2013).
85. Keshavan, A., Yeatman, J. D. & Rokem, A. Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. *Front. Neuroinform.* **13**, 29 (2019).
86. Infrastructure cards. <https://www.incf.org/infrastructure-portfolio>.
87. Sandström, M. *et al.* Recommendations for repositories and scientific gateways from a neuroscience perspective. *Sci Data* **9**, 212 (2022).