SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Received: 24 January 2018
Accepted: 11 April 2019
Published online: 23 May 2019

The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services

Paolo Avesani[1,2], Brent McPherson[3], Soichi Hayashi[11], Cesar F. Caiafa[4,5,6], Robert Henschel[12], Eleftherios Garyfallidis[9], Lindsey Kitchell[3], Daniel Bullock[7], Andrew Patterson[7], Emanuele Olivetti[1,2], Olaf Sporns[8], Andrew J. Saykin[10], Lei Wang[13], Ivo Dinov[14], David Hancock[12], Bradley Caron[15], Yiming Qian[4] & Franco Pestilli[16]

We describe the Open Diffusion Data Derivatives (O3D) repository: an integrated collection of preserved brain data derivatives and processing pipelines, published together using a single digital-object-identifier. The data derivatives were generated using modern diffusion-weighted magnetic resonance imaging data (dMRI) with diverse properties of resolution and signal-to-noise ratio. In addition to the data, we publish all processing pipelines (also referred to as open cloud services). The pipelines utilize modern methods for neuroimaging data processing (diffusion-signal modelling, fiber tracking, tractography evaluation, white matter segmentation, and structural connectome construction). The O3D open services can allow cognitive and clinical neuroscientists to run the connectome mapping algorithms on new, user-uploaded, data. Open source code implementing all O3D services is also provided to allow computational and computer scientists to reuse and extend the processing methods. Publishing both data-derivatives and integrated processing pipeline promotes practices for scientific

[1]Neuroinformatics Laboratory, Center for Information Technology, Fondazione Bruno Kessler, via Sommarive 18, 38123, Trento, Italy. [2]Center for Mind/Brain Sciences (CIMeC), University of Trento, via Delle Regole 101, 38123, Trento, Italy. [3]Pestilli Lab. Department of Psychological and Brain Sciences, Program in Cognitive Science, Indiana University Bloomington, 1101 E 10th Street, Bloomington, Indiana, 47405, USA. [4]Pestilli Lab. Department of Psychological and Brain Sciences, Indiana University Bloomington, 1101 E 10th Street, Bloomington, Indiana, 47405, USA. [5]Instituto Argentino de Radioastronomía (CCT-La Plata, CONICET; CICPBA), CC5 V, Elisa, 1894, Argentina. [6]Facultad de Ingeniería, Universidad de Buenos Aires, Buenos Aires, C1063ACV, Argentina. [7]Pestilli Lab. Department of Psychological and Brain Sciences, Program in Neuroscience, Indiana University Bloomington, 1101 E 10th Street, Bloomington, Indiana, 47405, USA. [8]Department of Psychological and Brain Sciences, Programs in Neuroscience and Cognitive Science, and Indiana Network Science Institute, Indiana University Bloomington, 1101 E 10th Street, Bloomington, Indiana, 47405, USA. [9]Department of Intelligent Systems Engineering, Programs in Neuroscience and Cognitive Science, Indiana University Bloomington, 700N Woodlawn Ave, Bloomington, Indiana, 47408, USA. [10]Indiana University School of Medicine, Departments of Radiology and Imaging Sciences and Medical and Molecular Genetics, and the Indiana Alzheimer Disease Center, Indiana University, 355 W 16th St., Indianapolis, Indiana, 46202, USA. [11]Department of Psychological and Brain Sciences and Pervasive Technology Institute, University Information Technology Services, Indiana University, 1101 E 10th Street, Bloomington, IN, 47405, USA. [12]Pervasive Technology Institute, Indiana University Bloomington, 2709 E 10th Street, Bloomington, IN, 47408, USA. [13]Departments of Psychiatry and Behavioral Sciences and Radiology, Northwestern University Feinberg School of Medicine, 710N. Lake Shore Drive, Abbott Hall 1322, Chicago, IL, 60611, USA. [14]Statistics Online Computational Resource (SOCR), Center for Complexity of Self-Management in Chronic Disease (CSCD), Health Behavior and Biological Sciences, Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI, 49109, USA. [15]Pestilli Lab. Indiana University School of Optometry and Program in Neuroscience, Indiana University Bloomington, 1101 E 10th Street, Bloomington, Indiana, USA. [16]Pestilli Lab. Department of Psychological and Brain Sciences, Engineering, Computer Science, Programs in Neuroscience and Cognitive Science, School of Optometry, and Indiana Network Science Institute, Indiana University Bloomington, 1101 E 10th Street, Bloomington, Indiana, 47405, USA. Correspondence and requests for materials should be addressed to F.P. (email: franpest@indiana.edu)

reproducibility and data upcycling by providing open access to the research assets for utilization by multiple scientific communities.

Background & Summary

In the past decade, efforts in large-scale neuroimaging data collection have redirected research attention towards effective practices for data sharing, reuse, standardization and secondary data analyses. Some of the most notable examples include projects such as the Human Connectome Project[1–3], UK Biobank[4,5], ADNI[6], INDI[7], ABCD[8], CamCAN[9] and OpenfMRI (recently rebranded as OpenNeuro)[10]. These projects have served as the bellwether for data-sharing in a growing culture focused on advancing methods for open big-data reproducible science[11–15]. Similar efforts for large-scale shared collections can, in principle, promote the establishment of best practices for measurements standards, and neuroinformatics methods, thereby contributing to a new generation of Big Data neuroscience research[16,17].

The use of the same data can be different across scientific communities. Data sharing can increase data value by promoting reuse for purposes beyond those of the original project; a process we call *data upcycling*[18]. As part of this upcycling process, data derivatives (secondary products generated by the various data analyses processes) can become useful data for scientific communities outside of the community of origin. Succinctly stated, various scientific communities may have different interests in reusing brain data. For example, a white matter segmentation can be used by computer scientists for methods development[19–26] or by neuroscientists to understand the brain[27–32]. Indeed, data can be reused for several applications not foreseen in the original study[33], for example, to develop theoretical frameworks[34], new algorithms[22,35,36], advance data visualization practices[37–39], and even for statistical validation of results[40–42]. The process of upcycling can help to extract additional value from openly available data sets, thereby returning continuing dividends from the initial resource investments.

We propose a unique approach to brain data upcycling by presenting the Open Diffusion Data Derivatives (O3D), a repository composed of both data-derivatives and their associated processing pipelines, bundled together and referenced by a single digital object identifier (DOI[43]). The O3D data were derived from anatomical (T1-weighted), diffusion-weighted magnetic resonance imaging (dMRI) data and tractography methods. The O3D data were obtained from previously published high-quality, high-resolution dMRI data[1,40,44,45] and processing pipelines[22,40,45–47]. The dataset is comprised of (1) the minimally preprocessed dMRI data files (12 brains from three different datasets with different properties of signal-to-noise ratio and resolution) and (2) a large set of diverse data derivatives comprising 360 tractograms, 7,200 segmented major tracts, and 720 connection matrices. The total size of the O3D repository is approximately 1.79 Terabytes of data derivatives.

Diffusion-weighted magnetic resonance imaging (dMRI) and tractography allow measuring structural connectomes, white matter macro-anatomy, and microscopic tissue properties from the living human brain. These techniques have revolutionized our understanding of how brain networks and the brain's white matter impact human behavior, in health and disease[29,30,48–58]. Neurotractography techniques provide fundamental insights about the human brain, and yet there is much work that remains to be done to map the human connectome[40,59–63]. Because of the complexity of these methods, the success of the modern scientific enterprise in mapping the human connectome almost certainly depends on transdisciplinary contributions from multiple communities – from Psychology and Neuroscience to Mathematics and Statistics, as well as to Computer Science and Engineering. For this reason open scientific discovery and collaborative sharing of methods, software, and data are of paramount importance[16,64,65].

In addition to data, the processing pipelines used to generate the O3D data are made available on the brainlife.io platform as a series of "open services," hereafter referred to as brainlife.io Apps or simply Apps. We define open services as self-contained processing applications embedded and reusable in a cloud platform environment. The brainlife.io platform allows running said Apps to process data available within the platform itself[66–68]. The concept of open service is akin to that of the Brain Imaging Data Structure Applications[69] as also introduced previously by others[70]. The brainlife.io Apps used below follow a generalized and light-weight specification as to allow usage with diverse combinations of software from multiple libraries, such as FSL[71], FreeSurfer[72], DIPY[73], Nipype[74], LiFE[22,40], AFQ[75], MRtrix[76], and AFNI[77]. These Apps can be containerized and made reproducible using technologies such as Docker[78] and Singularity[79]. Alternatively, brainlife.io Apps can also run without containerization on software environments compatible with NeuroDebian[80]. The brainlife.io platform currently utilizes a mixture of public (jetstream-cloud.org[81,82]; opensciencegrid.org[68]), commercial (azure.microsoft.com and cloud.google.com), as well as institutional (carbonate.uits.iu.edu) computing resources. The platform is a registered DataCite center (search.datacite.org/data-centers/brainl.iu), member of the fairsharing.org catalogue (see[83]), as well as registered project on both the NeuroImaging Tools and Resources Collaboratory (http://www.nitrc.org/projects/brainlife_io) and scicrunch.org (RRID: SCR_016513).

Publication records on brainlife.io/pubs, such as O3D, are preserved for at least ten years since latest use, and comply with the schema.org metadata specification to promote maximum discoverability and respect of the FAIR principles[84]. The complete list of brainlife.io Apps used to generate the O3D data are preserved as part of the repository[43]. These Apps are both provided as preserved files to allow accessing of the code version used to generate the specific O3D data, and can be reused for future research. The brainlife.io publication and preservation strategy is resilient to version changes likely to occur over time for each App or dataset. A full description of the brainlife.io platform and Apps is beyond the scope of this data descriptor; more information can be found here: brainlife.io/docs.

Using the O3D project, investigators can either process new data using the same pipelines used to generate the core O3D repository or they can download data derivatives processed at different stages along the series of steps

taken to generate tractography, white matter tracts, and connectivity matrices. Additionally, new data can in principle be uploaded using the brainlife.io web-portal and used to generate new results. Data can also be downloaded using a simple web or command line interfaces format as BIDS (Brain Imaging Data Structure)[85,86]. Finally, open source code and containers implementing the processing pipelines can be found at github.com/brainlife and hub.docker.com/u/brainlife.

The O3D repository is unique in that it focuses on publishing repeated-measures data-derivatives for tractography, white matter tracts, and structural connectome matrices–all associated with open services publishing reproducible data processing pipelines and workflows. The O3D dataset provides a means for computational test-retest quantification[41,87–89] and reproducibility. To generate the data derivatives, three tractography algorithms were used ten times on the same data source (individual brain). Due to stochasticity of such algorithms, the results for each of these are slightly different. The number of repeats has been previously shown to allow measuring variability and reliability of connectome mapping methods[21,22,40]. The tractography results were evaluated using state-of-the-art methods[22,40] and compared against classical neuroanatomy atlases used to segment the major human white matter tracts[75,90]. Finally, a series of connection matrices (i.e. brain networks) were generated using standard cortical parcellation methods[91]. Three example scenarios can be used to demonstrate transdisciplinary applications and show how investigators from different communities can utilize the O3D core set. First, investigators developing network science algorithms[35,63,92–94] might have an interest in demonstrating the applicability or efficacy of their methods on brain network data, but lack skills to process the raw diffusion data into connectivity matrices. The data derivatives provide an easily accessible point of entry by making available unthresholded brain connection matrices built using data from multiple individuals and different tracking methods. Second, investigators studying white matter neuroanatomy, or developing software for automated segmentation of white matter tracts, can use the data derivatives as complex test objects to compare the results of new algorithms with the state of the art reference set represented by O3D[25,95–97]. Finally, the data derivatives can be an essential education and training resource. It may be used by students and trainees in the neural and clinical sciences to learn about neuroanatomy or to develop practical analytic skills. All O3D data is compatible with most major neuroimaging software packages and can be conveniently loaded, processed and visualized[40,71–73,75,76].

The present descriptor introduces the O3D repository and some of the brainlife.io publication mechanisms, as necessary to describe the repository. The O3D reference repository will allow investigators from multiple scientific communities to explore brain data, perform visualization experiments, and replicate the data derivatives without having to first learn a full processing pipeline. This lowers the barrier of entry to computational neuroimaging, with the potential to advance algorithmic development, increase the involvement of underrepresented scholars, and to facilitate training and validation[16,98]. The repeated measure data derivatives we plan to distribute as part of O3D will appeal to a diverse range of research interests because of the extensive know-how necessary to generate them. Consequently, they can be used by communities of basic, clinical, translational and computational scientists including neuroscientists, students and trainees early in their careers[16,98,99].

Methods

Data sources. Three diffusion-weighted Magnetic Resonance Imaging datasets (dMRI) were used to generate all the derivatives in the initial repository layout, from publicly available sources (https://purl.stanford.edu/rt034xr8593[45], https://purl.stanford.edu/ng782rw8378[40,45], https://purl.stanford.edu/bb060nk0241[27] and https://www.humanconnectome.org/data[2,100]).

Stanford dataset (STN). We used data collected in four subjects at the Stanford Center for Cognitive and Neurobiological Imaging with a 3T General Electric Discovery 750 MRI (General Electric Healthcare), using a 32-channel head coil (Nova Medical). dMRI data had whole-brain coverage and were acquired with a dual-spin echo diffusion-weighted sequence, using 96 diffusion-weighting directions and gradient strength of 2,000 s/mm^2 (TE = 96.8 ms). Data spatial resolution was set at 1.5 mm isotropic. Each dMRI is the average of two measurements (NEX = 2). Ten non-diffusion-weighted images (b = 0) were acquired at the beginning of each scan[40,45,47].

Human connectome project datasets (HCP3T and HCP7T). We used data collected in 8 subjects from the Human Connectome Project, using Siemens 3T and 7T MRI scanners. Only measurements from the 2,000 s/mm^2 shell were extracted from these data and used to generate the data derivatives in our repositories. Data from the 3T and 7T scanners have different properties of resolution (e.g., HCP3T, 90 gradient directions, 1.25 *mm* isotropic resolution and HCP7T, 60 gradient directions, 1.05 *mm* isotropic resolution) and have been described before along with the processing methods used for data preprocessing[44,100–102].

Data preprocessing. We developed a series of steps to process the anatomical and dMRI data files in a standardized manner for publication as part of the O3D repository. All original data were oriented to the plane defined by the Anterior and Posterior Commissure and the 2,000 s/mm^2 shell was selected and utilized for the subsequent analyses. All MRI data were oriented in Neurological coordinates (Left-Anterior-Superior) and the bvecs files were oriented accordingly. The brainlife.io Apps implementing these operations can be found at[103–105] (see also Tables 1 and 2). No additional denoising, eddy current or head movement correction was applied beyond that performed by the data originators.

Voxel signal reconstruction and tractography. White matter fascicles tracking was performed using MRtrix 0.2.12[76]. White- and gray-matter tissues were segmented with Freesurfer[72] using the T1-weighted MRI images associated to each individual brain, and then resampled at the resolution of the dMRI data. Only voxels identified primarily as white-matter tissue were used to constrain tracking. We used three different tracking methods: (A) tensor-based deterministic tracking[106,107], (B) Constrained Spherical Deconvolution (CSD) -based

App goal	DOIs of each O3D App as service on brainlife.io
1. ACPC alignment of T1	https://doi.org/10.25663/bl.app.16[103]
2. Split dMRI shells	https://doi.org/10.25663/bl.app.17[104]
3. dMRI data preprocessing	https://doi.org/10.25663/bl.app.3[105]
4. Brain parcellation	https://doi.org/10.25663/bl.app.0[112]
5. Tractography	https://doi.org/10.25663/bl.app.59[113]
6. Tractography evaluation	https://doi.org/10.25663/bl.app.1[115]
7. Network neuroscience	https://doi.org/10.25663/bl.app.47[121]
8. White matter classification (WMC)	https://doi.org/10.25663/bl.app.13[116]
9. Refine white matter classification	https://doi.org/10.25663/bl.app.11[117]
10. WMC file format conversion	https://doi.org/10.25663/brainlife.app.127[118]
11. Tractogram file format conversion	https://doi.org/10.25663/brainlife.app.132[139]

Table 1. List of the current Apps implementing the processing steps used to generate O3D to be re-used on brainlife.io as open services.

App goal	URLs of each O3D App code repository on GitHub.com
1. ACPC alignment of T1	https://github.com/brainlife/app-acpcART
2. Split dMRI shells	https://github.com/brainlife/app-splitshells
3. dMRI data preprocessing	https://github.com/brainlife/app-dtiinit
4. Brain parcellation	https://github.com/brainlife/app-freesurfer
5. Tractography	https://github.com/brainlife/app-tracking
6. Tractography evaluation	https://github.com/brainlife/app-life
7. Network neuroscience	https://github.com/brainlife/app-networkneuro
8. White matter classification (WMC)	https://github.com/brainlife/app-tractclassification
9. Refine white matter classification	https://github.com/brainlife/app-AFQclean
10. WMC file format conversion	https://github.com/brainlife/app-wmctotrk
11. Tractogram file format conversion	https://github.com/brainlife/app-convert-tck-to-trk

Table 2. List of software repositories with the code version of the scripts implementing the O3D Apps.

deterministic tracking[76,108], and (C) CSD-based probabilistic tracking[108,109]. Maximum harmonic orders $L_{max} = 10$ (STN, HCP3T) and $L_{max} = 8$ (HCP7T) were used[110,111]. Other parameters settings used to perform tracking were: step size: 0.2 mm; maximum length, 200 mm; minimum length, 10 mm. The fiber orientation distribution function (f_{ODF}) amplitude cutoff, was set to 0.1, and for the minimum radius of curvature we adopted the default values, fixed by MRtrix for each kind of tracking: 2 mm (DTI deterministic), 0 mm (CSD deterministic), 1 mm (CSD probabilistic). We generated repeated measures of tractography derivatives by computing 10 candidate whole-brain fascicles groups for each individual brain using 500,000 fascicles each. Apps implementing the methods can be found at[112–114].

Tractography evaluation. We used the Linear Fascicle Evaluation method (LiFE)[40] to optimize whole-brain tractograms implemented using the recently proposed ENCODE model[22]. The LiFE method identifies fascicles that successfully contribute to prediction of the measured dMRI signal. It has been shown that only a percentage of the total number of fascicles generated through a single tractography method is supported by the properties of given dataset[40,47]. Because of this we removed all fascicles making no significant contribution to explaining the diffusion measurements. The percentage of streamlines retained in these optimized fascicles groups ranged between 10–20% (STN), 15–35% (HCP3T) and 20–40% (HCP7T). Apps implementing the method can be found at[115].

White matter tracts segmentation. Twenty major human white matter tracts were segmented using the Automating Fiber-tract Quantification (AFQ) method[75]. An additional step refined the segmented tracts by removing the fiber outliers. The following tracts were segmented: left and right Anterior Thalamic Radiation (ATRl and ATRr), left and right corticospinal tract (CSTl and CSTr), left and right Cingulum - Cingulate gyrus (CCgl and CCgr), left and right Cingulum - Hippocampus portion (CHil and CHir), left and right Inferior Fronto-Occipital Fasciculus (IFOFl and IFOFr), left and right Inferior Longitudinal Fasciculus (ILFl and ILFr), left and right Superior Longitudinal Fasciculus (SLFl and SLFr), left and right Uncinate Fasciculus (UFl and UFr), left and right Superior Longitudinal Fasciculus - Temporal part (often referred to as the "arcuate fasciculus", SLFTl and SLFTr), Forceps Major (FMJ), and Forceps Minor (FMI). Each tract was stored in trackvis file format. Apps implementing the method can be found at[116–118].

Connection matrix construction. We used tractograms evaluated by the LiFE method to build connectivity matrices. Connectivity matrices were built for each fascicle groups using the 68 cortical regions from the Desikan Killiany atlas, segmented in each individual using T1w MRI images and FreeSurfer[72,91]. Fascicles

terminations were mapped onto each of the 68 regions. All fibers connecting pairs of brain regions were identified and collected. Adjacency matrices were built using two measures: (A) *count*[119], by computing the number of fascicles connecting each unique pair of regions, (B) *density*, by computing the density of fibers connecting each unique pair – computed as twice the number of fascicles between regions divided by sum of the number of voxels in the two atlas regions[88,94,119,120]. Apps implementing the method can be found at[121].

Open service for reproducible neuroscience: brainlife.io/apps. We provide the full set of scripts used to process the O3D repository, both as open services, also referred to as Apps, that can be run on the brainlife.io platform (Table 1), as well as, code, scripts used to implement each App available on github.com/brainlife (Table 2). Whereas the code can be downloaded for running locally the scripts, the Apps are embedded in the brainlife.io platform and can be reused to directly process data avoiding the needs of installing software.

Brainlife.io Apps can be improved over time by users or developers and for this reason their implementation can change. As such, brainlife.io uses github.com to keep track of App versions. We note that whereas the DOIs for the Apps reported in Table 1 direct users to the most recent version of each App available on the platform, the URLs in Table 2 direct users to the specific version of the code used for the preprocessing used to generate the published O3D dataset. To fully support the reproducibility of the O3D publication we preserve for each release both the data and a snapshot of the code for each App. The O3D Apps preserved with the original code version used to generate the repository is reported in[43].

Data Records

Preserved O3D data and Apps can be downloaded at the web URL reported in[43]. Upon download, data will automatically be organized as brainlife.io DataTypes (brainlife.io/docs/user/datatypes and brainlife.io/datatypes) as well as according to the specification defined by the Brain Imaging Data Structure (BIDS)[85]. We note that, currently, BIDS does not officially provide a complete specification for diffusion-weighted magnetic resonance imaging and tractography derivatives.

According to the provisional BIDS specification for data derivatives (https://goo.gl/aFJ6vS), we have organized the files within folders, where each folder name refers to the name of the brainlifle.io App used to generate the files. The file naming convention adopted for the folders is based on three tokens: (A) The name of the github.com organization (e.g., `brainlife`); (B) the name of the repository of the App (e.g., `app-life`). All files generated by an App are aggregated in subfolders, one for each subject. Following the BIDS convention: (1) each file name includes a descriptor (`_desc-`) referring a unique brainlife.io identifier, (2) additional information on the brainlife.io DataType reported in filename by tags (`_tag-`), (3) the repeated measures are denoted by the keyword run (`_run-`), (4) the last token of the file names indicates the BIDS datatype (e.g., `_dwi-`), (5) the suffix denotes the file format (e.g., `.nii.gz`), and (6) metadata are recorded as a JSON file[122].

Source data. The source files of anatomy uploaded to brainlife.io are stored as follow:

```
upload/sub-{}/anat/
       sub-{}_tag-acpcaligned_desc-{}_T1w.json
       sub-{}_tag-acpcaligned_desc-{}_T1w.nii.gz
```

The source files of diffusion MRI uploaded from Stanford to brainlife.io are stored as follow:

```
upload/sub-{}/dwi/
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.json
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.nii.gz
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.bvals
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.bvecs
```

The source files of diffusion MRI uploaded from HCP to brainlife.io are stored as follow

```
brain-life.app-splitshells/sub-{}/dwi/
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.json
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.nii.gz
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.bvals
       sub-{}_tag-normalized_tag-singleshell_desc-{}_dwi.bvecs
```

Data preprocessing. The diffusion data after normalization (alignment and orientation) are stored as follows:

```
brain-life.app-dtiinit/sub-{}/dwi/
       sub-{}_tag-normalized_tag-singleshell_tag-dtiinit_desc-{}_dwi.json
       sub-{}_tag-normalized_tag-singleshell_tag-dtiinit_desc-{}_dwi.nii.gz
       sub-{}_tag-normalized_tag-singleshell_tag-dtiinit_desc-{}_dwi.bvals
       sub-{}_tag-normalized_tag-singleshell_tag-dtiinit_desc-{}_dwi.bvecs
```

Tractography. The diffusivity signal reconstruction models generated the following volumetric images as NifTI files: fractional anisotropy (`_FA.nii.gz`), the diffusion tensor model (`model-DTI`) and the constrained spherical deconvolution model (`model-CSD`). A brain mask and a white matter mask are also distributed at the dMRI data resolution (`type-Brain, type-Whitematter`). To increase impact and compatibility of the O3D data files, two copies of each tractogram are distributed, one in MRtrix format (tck) and the other TrackVis format (trk). One file is outputted per repeated-measure tractogram, and tractography method (`tag-dtstream, tag-sdstream, tag-sddprob`).

```
brain-life.app-tracking/sub-{}/dwi/
       sub-{}_run-{}_desc-{}_FA.nii.gz
```

```
        sub-{}_run-{}_desc-{}_model-DTI_diffmodel.nii.gz
        sub-{}_run-{}_desc-{}_model-CSD_diffmodel.nii.gz
        sub-{}_run-{}_desc-{}_type-Brain_mask.nii.gz
        sub-{}_run-{}_desc-{}_type-Whitematter_mask.nii.gz
        sub-{}_run-{}_tag-dtstream_desc-{}_tractography.json
        sub-{}_run-{}_tag-dtstream_desc-{}_tractography.tck
        sub-{}_run-{}_tag-sdstream_desc-{}_tractography.json
        sub-{}_run-{}_tag-sdstream_desc-{}_tractography.tck
        sub-{}_run-{}_tag-sdprob_desc-{}_tractography.json
        sub-{}_run-{}_tag-sdprob_desc-{}_tractography.tck
brainlife.app-convert-tck-to-trk/sub-{}/dwi/
        sub-{}_run-{}_tag-dtstream_tag-dwi_desc-{}_tractography.json
        sub-{}_run-{}_tag-dtstream_tag-dwi_desc-{}_tractography.trk
        sub-{}_run-{}_tag-sdstream_tag-dwi_desc-{}_tractography.json
        sub-{}_run-{}_tag-sdstream_tag-dwi_desc-{}_tractography.trk
        sub-{}_run-{}_tag-sdprob_tag-dwi_desc-{}_tractography.json
        sub-{}_run-{}_tag-sdprob_tag-dwi_desc-{}_tractography.trk
```

Tractography evaluation. We used the Linear Fascicle Model to evaluate the quality of fit of the dMRI signal. The output of LiFE tractography evaluation process is stored as an *encode* model structure[22]. The encode brainlife.io DataType is stored as matlab structure (`_life.mat`). A detailed documentation of encode model structure is available at github.com/brain-life/encode. One encode model structure per repeated-measure tractogram is distributed, for a total of ten runs, one for each tractography algorithm.

```
brain-life.app-life/sub-{}/dwi/
        sub-{}_run-{}_tag-dtstream_desc-{}_life.json
        sub-{}_run-{}_tag-dtstream_desc-{}_life.mat
        sub-{}_run-{}_tag-sdstream_desc-{}_life.json
        sub-{}_run-{}_tag-sdstream_desc-{}_life.mat
        sub-{}_run-{}_tag-sdprob_desc-{}_life.json
        sub-{}_run-{}_tag-sdprob_desc-{}_life.mat
```

White matter classification. Twenty human major white matter tracts were classified for each Tractogram and are distributed using the TRK file format. A json file for each tractogram records for each tract the enumeration ID, the label of the tract and the number of fibers.

```
brainlife.app-wmctotrk/sub-{}/dwi/
        sub-{}_run-{}_tag-dtstream_tag-afq_tag-cleaned_tag_wmc_\
              desc-run-{}_tractography.trk
        sub-{}_run-{}_tag-dtstream_tag-afq_tag-cleaned_tag_wmc_\
              desc-run-{}_tractography.json
        sub-{}_run-{}_tag-sdstream_tag-afq_tag-cleaned_tag_wmc_\
              desc-run-{}_tractography.trk
        sub-{}_run-{}_tag-sdstream_tag-afq_tag-cleaned_tag_wmc_\
              desc-run-{}_tractography.json
        sub-{}_run-{}_tag-sdprob_tag-afq_tag-cleaned_tag_wmc_\
              desc-run-{}_tractography.trk
        sub-{}_run-{}_tag-sdprob_tag-afq_tag-cleaned_tag_wmc_\
              desc-run-{}_tractography.json
```

Connectome matrices. Connection matrices were built using the aforementioned tractograms and the Desikan-Killiany Atlas from FreeSurfer[91]. A connection matrix was computed for each repeated-measure tractogram, processed using the LiFE method, and for each tractography method {dtstream, sdstream, sddprob}. Two measures of connectivity were computed: fiber count and fiber density (fiber count divided by the volume of the two termination areas[119,123]). Connection matrices are stored as pairs of.csv and.json files. A NifTI file records the cortical parcellation used to define the ROIs of the networks.

```
brain-life.app-networkneuro/sub-{}/dwi/
        sub-{}_run-{}_tag-{}_desc-{}_connectivity.json
        sub-{}_run-{}_tag-{}_desc-{}_tag-count_connectivity.csv
        sub-{}_run-{}_tag-{}_desc-{}_tag-density_connectivity.csv
        sub-{}_run-{}_tag-{}_desc-{}_label-GM_dseg.nii.gz
```

Technical Validation

In this section we provide both a qualitative and quantitative evaluation of the data derivatives made available at[43]. We show data SNR in each dataset used, demonstrate quality of alignment between dMRI and anatomy files, and show the diffusion signal in the voxel reconstruction, several properties of the tractography models and of the major white matter tracts segmented.

Data preprocessing. Data preprocessing was performed using a combination of previously published pipelines[22,40,45–47] (see Methods for additional details). Diffusion weighted MRI data were aligned to the T1-weighted anatomical images (Fig. 1a left-hand columns, see Methods for additional details). The T1w images were used

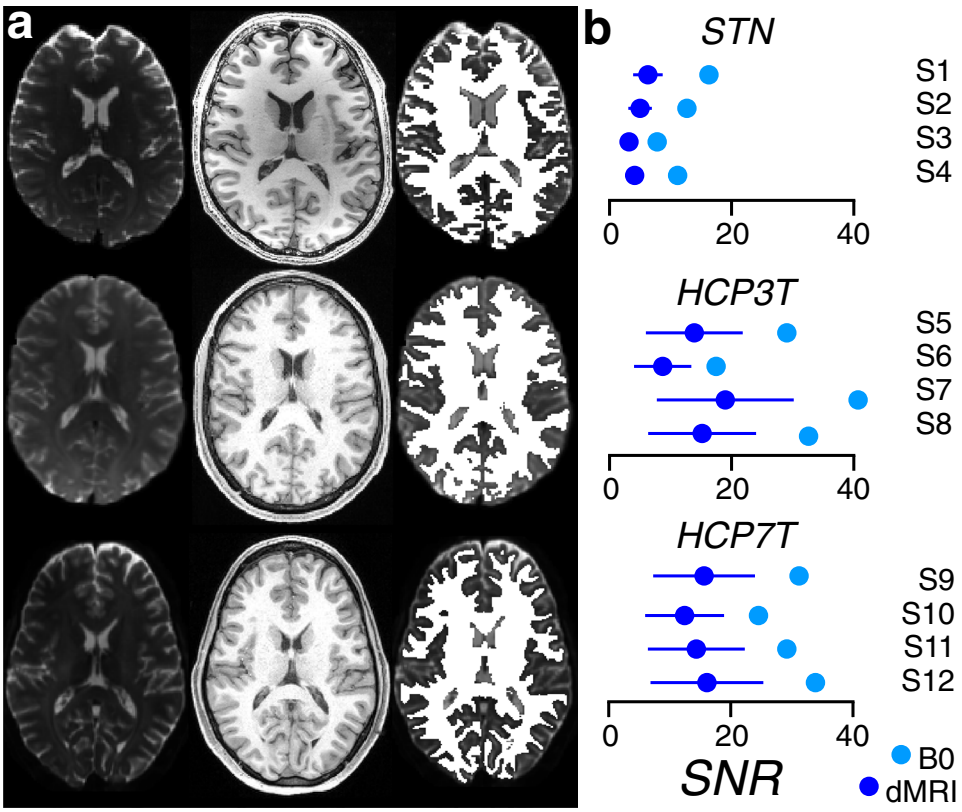

Fig. 1 Data quality and preprocessing. (**a**) Axial view of dMRI (left, non-diffusion weighted volume, B0), aligned anatomical image (center) and white matter mask obtained from the anatomy (white), overlaid on the B0 to show the quality of the white matter volume delineation. One example subject is reproduced from the Stanford (top), Human Connectome 3T (middle) and Human Connectome 7T (bottom) data. (**b**) Mean and ±1 sd across diffusion-weighted measurements of the signal-to-noise (SNR) for each subject and dataset in the O3D distribution as implemented at[126].

to segment the brain into different tissue types and brain regions[72]. The total white matter volume was identified using the previously generated white matter tissue segmentation and all subsequent analyses were performed within the white matter volume. Figure 1a shows how the white matter volume (mask) defined on the anatomical image (middle) aligns with the non diffusion-weighted signal (B_0) image of the diffusion MRI data (left-hand panel) in three example subjects one per dataset.

To compare dMRI data quality across datasets we computed the signal-to-noise ratio (SNR) comparing the mean attenuated dMRi signal to the background noise for both diffusion-weighted and B_0 measurements (Fig. 1b), as described by[124,125]. The brainlife.io App implementing this SNR method can be found at[126].

White matter microstructure reconstruction within the voxel. The dMRI signal within each voxel was reconstructed using the two dominant models, namely the diffusion tensor (DTI[127]) and constrained-spherical deconvolution (CSD[110,111]). Specifically, when applying CSD, we utilized an L_{max} parameter of 10 for STN and 8 for HCP. These models provide different opportunities as well as limitations to characterize the dMRI signal and brain fibers. Figure 2 shows the quality of the estimated deconvolution kernel (a) and the fit of the CSD model in three representative axial brain slices, one per dataset (b). The kernel estimation is important for effective fiber distribution estimation and long-range tracking[128]. Both dMRI reconstructions (DTI and CSD) have been manually curated by visual inspection to assure quality in the O3D dataset.

Tractography. Tractography was reconstructed using two established methods: deterministic and probabilistic[76,106–108,129–131] tractography. We used Deterministic tractography either in combination with DTI or CSD models. Probabilistic tractography was only used in combination with the CSD model. It has been established that application of these different methods result in the generation of white matter fascicles with different anatomical properties[29,40,47,54,132–134]. The O3D dataset provides three tractography reconstructions for each individual brain. Tractography outputs were stored using common file formats (.tck and.trk) to allow investigators to compare, reuse and improve upon current tracking methods.

Figure 3a provides a qualitative depiction of the whole-brain tractography reconstruction in a subject from each dataset. Figure 3b reports a quantitative comparison of the fascicles length distribution for whole brain tractograms in the three example subjects in Fig. 3a.

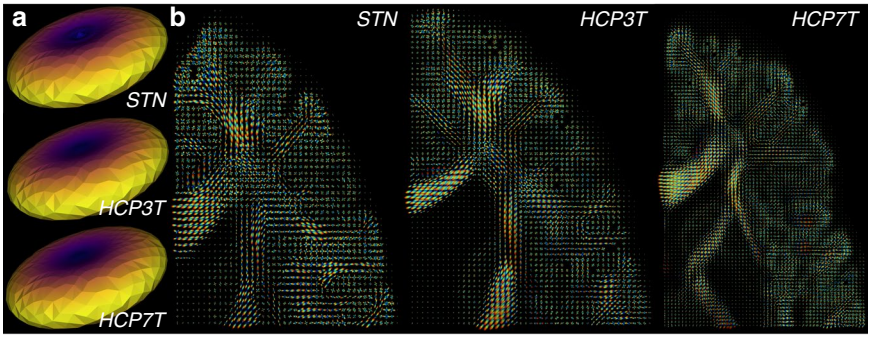

Fig. 2 Estimated fiber orientation distribution functions (fODF). (**a**) Examples of estimated single-fiber response function used to compute the fODF individually in each subject. The similarity and flat shape of the response functions ensures model-fit quality[110,111]. (**b**) Axial brain views from three example subjects in each dataset depicting the estimated fODF (fiber orientation distribution functions) in a series of voxels covering the corpus callosum and the central-semiovale. Coverage of the response functions and orientation are consistent with major anatomical understanding.

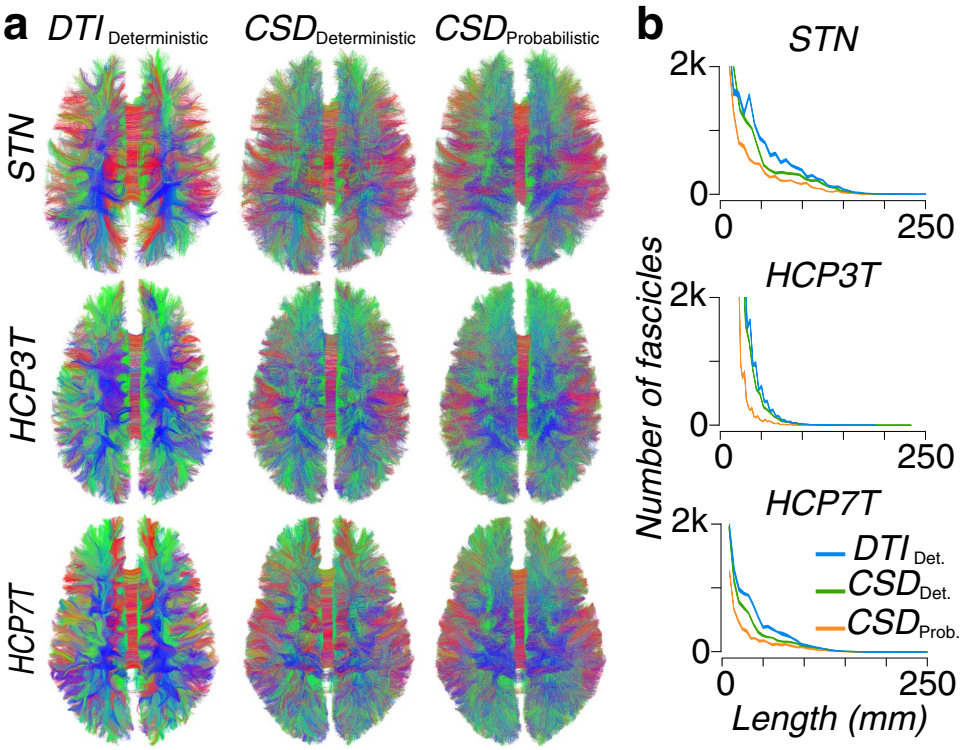

Fig. 3 Visualization of whole-brain tractograms and fascicle length distribution. (**a**) The full brain tractography for each of the three datasets, as generated using $DTI_{deterministic}$, $CSD_{deterministic}$ and $CSD_{probabilistic}$ Models. (**b**) The whole-brain connectome streamline count for each of the three tractography models applied to the STN, HCP3T and HCP7T datasets.

Human major white matter tracts. We report a qualitative visualization of the eleven major white matter tracts which were segmented from each connectome. These correspond to nine major tracts in the left and right hemispheres and two cross-hemispheric tracts. These tracts were segmented using a standardized methodology and atlases[75,90,135]. Files are saved as.tck and.trk file formats. Previous work has shown that the application of different tractography models results in anatomical tracts with different morphologies, volumes and streamline counts[22,29,40,47,54]. Figure 4a depicts these tracts as segmented for each subject, using each diffusion model, with colors corresponding to specific tracts. Figure 4b plots the number of streamlines, from the source whole brain tractogram, identified as constituting each of these major tracts.

Network neuroscience. The aforementioned whole brain tractograms represent a model of how the white matter of the brain connects cortical regions to one another. Together with a cortical parcellation, this rich body

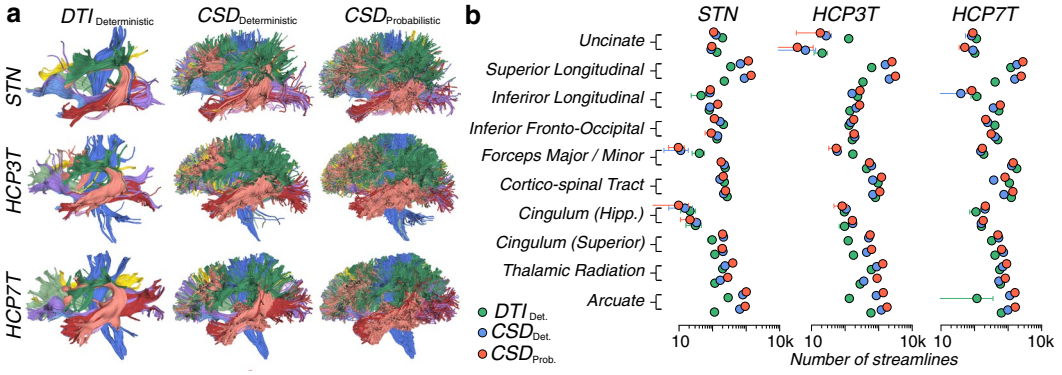

Fig. 4 Anatomy of tracts and number of fascicles per tract. (**a**) The morphologies of several major tracts, overlaid with one another, as segmented from whole brain connectomes. Tractography generated for each dataset using $DTI_{deterministic}$, $CSD_{deterministic}$ and $CSD_{probabilistic}$ models. Colors correspond to individual tracts. (**b**) The streamline counts associated with several major tracts. Marker color corresponds to tractography model. Error bars generated from standard deviation across ten replications.

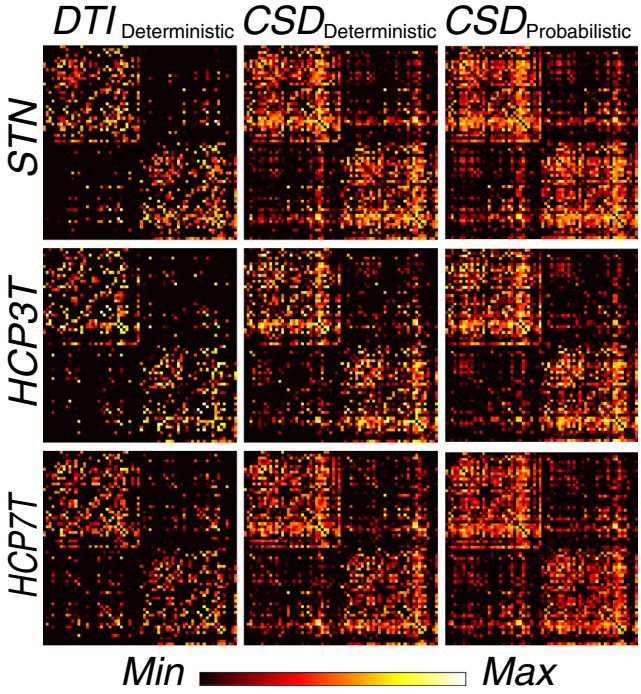

Fig. 5 Brain network matrices. Nine representative matrices of connectivity between anatomical regions defined in the Desikan-Killiany atlas[91]. Matrices report fiber density computed as twice the number of streamlines touching a pair of regions divided by the combined size of the two regions (in number of brain voxels). Density is normalized across matrices, brighter colors indicate higher density. Networks depicted were generated for three representative subjects, one per dataset, using $DTI_{deterministic}$, $CSD_{deterministic}$ and $CSD_{probabilistic}$ tractography.

of connectivity information can be summarized into a network matrix, with brain regions or regions of interest representing network nodes, and measures related to connection weight or density corresponding to network edges. Graphical summaries like those presented in Fig. 5 provide a common way to visualize these connectivity patterns. This graph or network representation of connectomes enables a large array of analytic and modeling tools to probe connectivity motifs, modularity, centrality, vulnerability and other network or graph-theoretic measures[63,136–138]. The O3D dataset features structural connectivity data, arranged as matrices, along with the numeric key indicating the cortical parcels names for each network node. Connectivity matrices were computed using two edge metrics: streamline count and streamline density[88,119,123].

Usage Notes

The O3D dataset is publicly available at the link provided in[43]. Data files can be downloaded organized according to the BIDS[85] standard. Different data derivatives are distributed with formats, such as NifTI, TCK, TRK or plain text. Access to the published data is currently supported via (i) web interface and (ii) Command Line Interface (CLI).

The brainlife.io CLI can be installed on most Unix/Linux systems using the following command: `npm install brainlife -g`. The CLI can be used to query and download partial of full datasets. The following example shows the CLI command to download all T1w datasets from a subject in the publication data Release 2:

```
bl pub query # this will return the publication IDs
bl bids download --pub 5c0ff604391ed50032b634d1 --subject 0001 --datatype
neuro/anat/t1w
```

The following command downloads the data in the entire project (from Release 2) into BIDS format:

```
bl bids download --pub 5c0ff604391ed50032b634d1
```

Additional information about the brainlife.io CLI commands can be found at https://github.com/brainlife/cli In addition, https://brainlife.io/project/5a022fc99c0d250055709e9c/detail is the project page with read-only data supporting browsing, visualization, download or additional processing. O3D uses the data originated from projects with different license and user terms. The four datasets (subject 1–4) originated from the Stanford University project are distributed with CC-BY license (creativecommons.org/licenses/by/4.0/). Access to the eight datasets originated from the Human Connectome Project (subject 5–12) require that users agree to the HCP Data Use Terms humanconnectome.org/study/hcp-young-adult/data-use-terms.

References

1. Glasser, M. F. *et al*. The Human Connectome Project's neuroimaging approach. *Nat. Neurosci.* **19**, 1175–1187 (2016).
2. Van Essen, D. C. *et al*. The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
3. Marcus, D. S. *et al*. Informatics and data mining tools and strategies for the human connectome project. *Front. Neuroinform.* **5**, 4 (2011).
4. Miller, K. L. *et al*. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
5. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. on Behalf of. UK Biobank Data: Come and Get It. *Sci. Transl. Med.* **6**, 224ed4–224ed4 (2014).
6. Weiner, M. W. *et al*. The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers. Dement.* **6**, 202–11.e7 (2010).
7. Biswal, B. B. *et al*. Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. USA* **107**, 4734–4739 (2010).
8. Jernigan, T. L., Brown, S. A. & Dowling, G. J. The Adolescent Brain Cognitive Development Study. *J. Res. Adolesc.* **28**, 154–156 (2018).
9. 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).
10. Poldrack, R. A. *et al*. Toward open sharing of task-based fMRI data: the Open fMRI project. *Front. Neuroinform.* **7**, 12 (2013).
11. Nichols, T. E. *et al*. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* **20**, 299–303 (2017).
12. Eglen, S. J. *et al*. Toward standard practices for sharing computer code and programs in neuroscience. *Nat. Neurosci.* **20**, 770–773 (2017).
13. Nosek, B. A. *et al*. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
14. Pernet, C. & Poline, J.-B. Improving functional magnetic resonance imaging reproducibility. *Gigascience* **4**, 15 (2015).
15. Halchenko, Y. O. & Hanke, M. Open is Not Enough. Let's Take the Next Step: An Integrated, Community-Driven Computing Platform for Neuroscience. *Front. Neuroinform.* **6**, 22 (2012).
16. Poldrack, R. A. & Gorgolewski, K. J. Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* **17**, 1510–1517 (2014).
17. Focus on big data. *Nat. Neurosci.* **17**, 1429 (2014).
18. Vearncombe, J., Riganti, A., Isles, D. & Bright, S. Data upcycling. *Ore Geol. Rev.* **89**, 887–893 (2017).
19. Sharmin, N., Olivetti, E. & Avesani, P. White Matter Tract Segmentation as Multiple Linear Assignment Problems. *Front. Neurosci.* **11**, 754 (2017).
20. Kitchell, L., Bullock, D., Hayashi, S. & Pestilli, F. Shape Analysis of White Matter Tracts via the Laplace-Beltrami Spectrum. *Miccai Shapemi* **11167**, 195–206 (2018).
21. Caiafa, C. F., Sporns, O. & Saykin, A. Unified representation of tractography and diffusion-weighted MRI data using sparse multidimensional arrays. *Adv. Neural Inf. Process. Syst*, 4340–4351 (2017).
22. Caiafa, C. F. & Pestilli, F. Multidimensional encoding of brain connectomes. *Sci. Rep.* **7**, 11491 (2017).
23. Glozman, T. *et al*. Framework for shape analysis of white matter fiber bundles. *Neuroimage* **167**, 466–477 (2018).
24. Glozman, T., Solomon, J. & Pestilli, F. Shape-Attributes of Brain Structures as Biomarkers for Alzheimer's Disease. *Journal of Alzheimer's* **56**(1), 287–295 (2017).
25. Garyfallidis, E., Brett, M., Correia, M. M., Williams, G. B. & Nimmo-Smith, I. QuickBundles, a Method for Tractography Simplification. *Front. Neurosci.* **6**, 175 (2012).
26. Garyfallidis, E. *et al*. Recognition of white matter bundles using local and global streamline-based registration and clustering. *Neuroimage* **170**, 283–295 (2018).
27. Takemura, H. *et al*. A Major Human White Matter Pathway Between Dorsal and Ventral Visual Cortex. *Cereb. Cortex* **26**, 2205–2214 (2016).
28. 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).
29. Rokem, A. *et al*. The visual white matter: The application of diffusion MRI and fiber tractography to vision science. *J. Vis.* **17**, 4 (2017).
30. Leong, J. K., Pestilli, F., Wu, C. C., Samanez-Larkin, G. R. & Knutson, B. White-Matter Tract Connecting Anterior Insula to Nucleus Accumbens Correlates with Reduced Preference for Positively Skewed Gambles. *Neuron* **89**, 63–69 (2016).
31. Leong, J. K., MacNiven, K. H., Samanez-Larkin, G. R. & Knutson, B. Distinct neural circuits support incentivized inhibition. *Neuroimage* **178**, 435–444 (2018).
32. de Schotten, M. T. *et al*. A lateralized brain network for visuospatial attention. *Nat. Neurosci.* **14**, 1245 (2011).
33. Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E. & Martone, M. E. Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat. Neurosci.* **17**, 1442–1447 (2014).

34. Sejnowski, T. J., Churchland, P. S. & Movshon, J. A. Putting big data to good use in neuroscience. *Nat. Neurosci.* **17**, 1440–1441 (2014).
35. Betzel, R. F. *et al*. Generative models of the human connectome. *Neuroimage* **124**, 1054–1064 (2016).
36. Mejia, A. F. *et al*. Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators. *Neuroimage* **112**, 14–29 (2015).
37. Goldstone, R. L., Pestilli, F. & Börner, K. Self-portraits of the brain: cognitive science, data visualization, and communicating brain structure and function. *Trends Cogn. Sci.* **19**, 462–474 (2015).
38. Margulies, D. S., Böttger, J., Watanabe, A. & Gorgolewski, K. J. Visualizing the human connectome. *Neuroimage* **80**, 445–461 (2013).
39. Yeatman, J. D., Richie-Halford, A., Smith, J. K., Keshavan, A. & Rokem, A. A browser-based tool for visualization and analysis of diffusion MRI data. *Nat. Commun.* **9**, 940 (2018).
40. Pestilli, F., Yeatman, J. D., Rokem, A., Kay, K. N. & Wandell, B. A. Evaluation and statistical inference for human connectomes. *Nat. Methods* **11**, 1058–1063 (2014).
41. Pestilli, F. Test-retest measurements and digital validation for *in vivo* neuroscience. *Sci. Data* **2**, 140057 (2015).
42. Fukushima, M. *et al*. Fluctuations between high- and low-modularity topology in time-resolved functional connectivity. *Neuroimage* **180**, 406–416 (2018).
43. Hayashi, S., Avesani, P. & Pestilli, F. Open Diffusion Data Derivatives. *Brainlife.io*, https://doi.org/10.25663/BL.P.3 (2017).
44. Vu, A. T. *et al*. High resolution whole brain diffusion imaging at 7T for the Human Connectome Project. *Neuroimage* **122**, 318–331 (2015).
45. Rokem, A. *et al*. Evaluating the accuracy of diffusion MRI models in white matter. *PLoS One* **10**, e0123272 (2015).
46. Glasser, M. F. *et al*. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).
47. Takemura, H., Caiafa, C. F., Wandell, B. A. & Pestilli, F. Ensemble Tractography. *PLoS Comput. Biol.* **12**, e1004692 (2016).
48. Bassett, D. S. & Bullmore, E. T. Human brain networks in health and disease. *Curr. Opin. Neurol.* **22**, 340–347 (2009).
49. Bassett, D. S. & Gazzaniga, M. S. Understanding complexity in the human brain. *Trends Cogn. Sci.* **15**, 200–209 (2011).
50. Ajina, S., Pestilli, F., Rokem, A., Kennard, C. & Bridge, H. Human blindsight is mediated by an intact geniculo-extrastriate pathway. *Elife* **4**, e08935 (2015).
51. Allen, B., Spiegel, D. P., Thompson, B., Pestilli, F. & Rokers, B. Altered white matter in early visual pathways of humans with amblyopia. *Vision Res.* **114**, 48–55 (2015).
52. Thomason, M. E. & Thompson, P. M. Diffusion Imaging, White Matter, and Psychopathology. *Annu. Rev. Clin. Psychol.* **7**, 63–85 (2011).
53. Gomez, J. *et al*. Functionally defined white matter reveals segregated pathways in human ventral temporal cortex associated with category-specific processing. *Neuron* **85**, 216–227 (2015).
54. Bastiani, M., Shah, N. J., Goebel, R. & Roebroeck, A. Human cortical connectome reconstruction from diffusion weighted MRI: the effect of tractography algorithm. *Neuroimage* **62**, 1732–1749 (2012).
55. Wandell, B. A. Clarifying Human White Matter. *Annu. Rev. Neurosci.* **39**, 103–128 (2016).
56. Libero, L. E., Burge, W. K., Deshpande, H. D., Pestilli, F. & Kana, R. K. White Matter Diffusion of Major Fiber Tracts Implicated in Autism Spectrum Disorder. *Brain Connect.* **6**, 691–699 (2016).
57. Main, K. L. *et al*. DTI measures identify mild and moderate TBI cases among patients with complex health problems: A receiver operating characteristic analysis of U.S. veterans. *NeuroImage: Clinical*, https://doi.org/10.1016/j.nicl.2017.06.031 (2017).
58. Fornito, A., Zalesky, A. & Breakspear, M. The connectomics of brain disorders. *Nat. Rev. Neurosci.* **16**, 159–172 (2015).
59. Maier-Hein, K. H. *et al*. The challenge of mapping the human connectome based on diffusion tractography. *Nat. Commun.* **8**, 1349 (2017).
60. Zalesky, A. *et al*. Connectome sensitivity or specificity: which is more important? *Neuroimage* **142**, 407–420 (2016).
61. Thomas, C. *et al*. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. *Proc. Natl. Acad. Sci. USA* **111**, 16574–16579 (2014).
62. Reveley, C. *et al*. Superficial white matter fiber systems impede detection of long-range cortical connections in diffusion MR tractography. *Proc. Natl. Acad. Sci. USA* **112**, E2820–8 (2015).
63. Bassett, D. S. & Sporns, O. Network neuroscience. *Nat. Neurosci.* **20**, 353–364 (2017).
64. Freeman, J. Open source tools for large-scale neuroscience. *Curr. Opin. Neurobiol.* **32**, 156–163 (2015).
65. Toga, A. W. & Dinov, I. D. Sharing big biomedical data. *Journal of Big Data* **2** (2015).
66. The Open Services Gateway Initiative: an introductory overview. *IEEE Communications Magazine*. **39**, 110–114 (2001).
67. Brebner, P. & Emmerich, W. Deployment of Infrastructure and Services in the Open Grid Services Architecture (OGSA). In *Lecture Notes in Computer Science* 181–195, https://doi.org/10.1007/11590712_15 (2005).
68. Pordes, R. *et al*. New science on the Open Science Grid. *J. Phys. Conf. Ser.* **125**, 012070 (2008).
69. Gorgolewski, K. J. *et al*. BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Comput. Biol.* **13**, e1005209 (2017).
70. Kiar, G. *et al*. Science in the cloud (SIC): A use case in MRI connectomics. *Gigascience* **6**, 1–10 (2017).
71. Smith, S., Bannister, P. R., Beckmann, C. & Brady, M. FSL: New tools for functional and structural brain image analysis. *Neuroimage* **13**, 249 (2001).
72. Fischl, B. & Bruce, F. Free Surfer. *Neuroimage* **62**, 774–781 (2012).
73. Garyfallidis, E. *et al*. Dipy, a library for the analysis of diffusion MRI data. *Front. Neuroinform*. **8** (2014).
74. Gorgolewski, K. *et al*. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* **5**, 13 (2011).
75. 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).
76. Tournier, J.-D., Calamante, F. & Connelly, A. MRtrix: Diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* **22**, 53–66 (2012).
77. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
78. Merkel, D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* **2014** (2014).
79. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* **12**, e0177459 (2017).
80. Halchenko, Y. O., Hanke, M. & Alexeenko, V. NeuroDebian: an integrated, community-driven, free software platform for physiology. *Proc. Aust. Physiol. Pharmacol. Soc* **31**, PCA100 (2014).
81. Stewart, C. A. *et al*. Jetstream: a self-provisioned, scalable science and engineering cloud environment. In *XSEDE* 29–21 (2015).
82. Towns, J. *et al*. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* **16**, 62–74 (2014).
83. FAIR sharing Team. Brainlife.io *FAIR sharing*, https://doi.org/10.25504/FAIRsharing.by3p8p (2017).
84. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
85. 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).
86. Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* **264**, 47–56 (2016).

87. Huang, L., Huang, T., Zhen, Z. & Liu, J. A test-retest dataset for assessing long-term reliability of brain morphology and resting-state brain activity. *Sci. Data* **3**, 160016 (2016).
88. Buchanan, C. R., Pernet, C. R., Gorgolewski, K. J., Storkey, A. J. & Bastin, M. E. Test–retest reliability of structural brain networks from diffusion MRI. *Neuroimage* **86**, 231–243 (2014).
89. Gorgolewski, K. J. *et al*. A test-retest fMRI dataset for motor, language and spatial attention functions. *Gigascience* **2** (2013).
90. Mori, S., Wakana, S., van Zijl, P. C. M. & Nagae-Poetscher, L. M. *MRI Atlas of Human White Matter*. (Elsevier Science, 2005).
91. 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).
92. van den Heuvel, M. P. & Sporns, O. Rich-Club Organization of the Human Connectome. *Journal of Neuroscience* **31**, 15775–15786 (2011).
93. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198 (2009).
94. Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**, 1059–1069 (2010).
95. Garyfallidis, E., Ocegueda, O., Wassermann, D. & Descoteaux, M. Robust and efficient linear registration of white-matter fascicles in the space of streamlines. *Neuroimage* **117**, 124–140 (2015).
96. Olivetti, E., Sharmin, N. & Avesani, P. Alignment of Tractograms As Graph Matching. *Front. Neurosci.* **10**, 554 (2016).
97. Wassermann, D. *et al*. The white matter query language: a novel approach for describing human white matter anatomy. *Brain Struct. Funct*., https://doi.org/10.1007/s00429-015-1179-4 (2016).
98. BRAINS (Brain Imaging in Normal Subjects) Expert Working Group. *et al*. Improving data availability for brain image biobanking in healthy subjects: Practice-based suggestions from an international multidisciplinary working group. *Neuroimage* **153**, 399–409 (2017).
99. Brakewood, B. & Poldrack, R. A. The ethics of secondary data analysis: considering the application of Belmont principles to the sharing of neuroimaging data. *Neuroimage* **82**, 671–676 (2013).
100. Thanh Vu, A. *et al*. Tradeoffs in pushing the spatial resolution of fMRI for the 7T Human Connectome Project. *Neuroimage*, https://doi.org/10.1016/j.neuroimage.2016.11.049 (2016).
101. Sotiropoulos, S. N. *et al*. Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *Neuroimage* **80**, 125–143 (2013).
102. Uğurbil, K. *et al*. Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage* **80**, 80–104 (2013).
103. Hayashi, S. & Kitchell, L. ACPC alignment via ART. *Brainlife.io*, https://doi.org/10.25663/BL.APP.16 (2017).
104. Hayashi, S. & Kitchell, L. Split Shells. *Brainlife.io*, https://doi.org/10.25663/BL.APP.17 (2017).
105. Hayashi, S., Avesani, P., Kitchell, L. & Pestilli, F. dtiInit. *Brainlife.io*, https://doi.org/10.25663/BL.APP.3 (2017).
106. Basser, P. J., Pajevic, S., Pierpaoli, C., Duda, J. & Aldroubi, A. *In vivo* fiber tractography using DT-MRI data. *Magn. Reson. Med.* **44**, 625–632 (2000).
107. Lazar, M. *et al*. White matter tractography using diffusion tensor deflection. *Hum. Brain Mapp.* **18**, 306–321 (2003).
108. Descoteaux, M., Deriche, R., Knösche, T. R. & Anwander, A. Deterministic and probabilistic tractography based on complex fibre orientation distributions. *IEEE Trans. Med. Imaging* **28**, 269–286 (2009).
109. Tournier, J. D., Calamante, F. & Connelly, A. Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions. In *Proc. 18th Annual Meeting of the Intl. Soc. Mag. Reson. Med. (ISMRM)* 1670 (2010).
110. Tournier, J.-D., Calamante, F., Gadian, D. G. & Connelly, A. Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *Neuroimage* **23**, 1176–1185 (2004).
111. Descoteaux, M., Angelino, E., Fitzgibbons, S. & Deriche, R. Apparent diffusion coefficients from high angular resolution diffusion imaging: estimation and applications. *Magn. Reson. Med.* **56**, 395–410 (2006).
112. Hayashi, S., Kitchell, L. & Pestilli, F. Freesurfer 6.0. *Brainlife.io*, https://doi.org/10.25663/BL.APP.0 (2017).
113. Hayashi, S., Kitchell, L. & Pestilli, F. MRtrix2 Tracking with dtiInit. *Brainlife.io*, https://doi.org/10.25663/BL.APP.59 (2017).
114. Hayashi, S. & Kitchell, L. Convert tck + dwi to trk (MRtrix 2). *Brainlife.io*, https://doi.org/10.25663/BL.APP.22 (2017).
115. Hayashi, S., Avesani, P., Kitchell, L. & Pestilli, F. LiFE with dtiInit. *Brainlife.io*, https://doi.org/10.25663/BL.APP.1 (2017).
116. Hayashi, S. & Kitchell, L. AFQ Tract Classification. *Brainlife.io*, https://doi.org/10.25663/BL.APP.13 (2017).
117. Hayashi, S., Kitchell, L. & Bullock, D. Clean WMC output. *Brainlife.io*, https://doi.org/10.25663/BL.APP.11 (2017).
118. Hayashi, S. & Avesani, P. Convert wmc to multiple trk. *Brainlife.io*, https://doi.org/10.25663/BRAINLIFE.APP.127 (2017).
119. Cheng, H. *et al*. Characteristics and variability of structural networks derived from diffusion tensor imaging. *Neuroimage* **61**, 1153–1164 (2012).
120. Qi, S., Meesters, S., Nicolay, K., Ter Haar Romeny, B. M. & Ossenblok, P. Structural Brain Network: What is the Effect of LiFE Optimization of Whole Brain Tractography? *Front. Comput. Neurosci.* **10**, 12 (2016).
121. Hayashi, S., Avesani, P., Pestilli, F. & McPherson, B. Network Neuro. *Brainlife.io*, https://doi.org/10.25663/BL.APP.47 (2017).
122. Bray, T. *The javascript object notation (json) data interchange format* (2017).
123. Hagmann, P. *et al*. Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**, e159 (2008).
124. Descoteaux, M., Deriche, R., Le Bihan, D., Mangin, J.-F. & Poupon, C. Multiple q-shell diffusion propagator imaging. *Med. Image Anal.* **15**, 603–621 (2011).
125. Jones, D. K., Knösche, T. R. & Turner, R. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. *Neuroimage* **73**, 239–254 (2013).
126. Hunt, D. Compute SNR on Corpus Callosum. *Brainlife.io*, https://doi.org/10.25663/BRAINLIFE.APP.120 (2018).
127. Basser, P. J. & Pierpaoli, C. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J. Magn. Reson. B* **111**, 209–219 (1996).
128. Tournier, J.-D. *et al.* Resolving crossing fibres using constrained spherical deconvolution: validation using diffusion-weighted imaging phantom data. *Neuroimage* **42**, 617–625 (2008).
129. Tuch, D. S., Belliveau, J. W. & Wedeen, V. J. Probabilistic tractography using high angular resolution diffusion imaging. *Neuroimage* **11**, S913 (2000).
130. Sherbondy, A., Dougherty, R. & Wandell, B. Identification of optic radiation *in-vivo* using diffusion tensor imaging and fiber tractography. *J. Vis.* **8**, 958–958 (2010).
131. Behrens, T. E. J., Berg, H. J., Jbabdi, S., Rushworth, M. F. S. & Woolrich, M. W. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *Neuroimage* **34**, 144–155 (2007).
132. Fillard, P. *et al*. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *Neuroimage* **56**, 220–234 (2011).
133. Côté, M.-A. *et al*. Tractometer: towards validation of tractography pipelines. *Med. Image Anal.* **17**, 844–857 (2013).
134. Neher, P. F., Descoteaux, M., Houde, J.-C., Stieltjes, B. & Maier-Hein, K. H. Strengths and weaknesses of state of the art fiber tractography pipelines – A comprehensive *in-vivo* and phantom evaluation study using Tractometer. *Med. Image Anal.* **26**, 287–305 (2015).
135. Zhang, W., Olivi, A., Hertig, S. J., van Zijl, P. & Mori, S. Automated fiber tracking of human brain white matter using diffusion tensor imaging. *Neuroimage* **42**, 771–777 (2008).

136. Fornito, A., Zalesky, A. & Bullmore, E. *Fundamentals of Brain Network Analysis*. (Academic Press, 2016).
137. Sporns, O. *Networks of the Brain*. (MIT Press, 2010).
138. Bassett, D. S. Brain network analysis: a practical tutorial. *Brain* **139**, 3048–3049 (2016).
139. Avesani, P. Convert tck to trk in DWI space. *Brainlife.io*, https://doi.org/10.25663/BRAINLIFE.APP.132 (2017).

Acknowledgements

This research was supported by NSF IIS-1636893, NSF BCS-1734853, NIH NIMH ULTTR001108, NIH NIMH U01MH097435, NIH NIMH 5 T32 MH103213, a Microsoft Research Award, a Google Cloud Award, the Indiana University Areas of Emergent Research initiative "Learning: Brains, Machines, Children", and Pervasive Technology Institute to F.P. We thank Aman Arya, Steven O'Riley and David Hunt for contributing to the development of https://brainlife.io, Craig Stewart, Winona Snapp-Childs, Charles A. McClary, and Jeremy Fischer for support with jetstream-cloud.org (NSF ACI-1445604), Melissa Cragin at the NSF Midwest Big Data Hub for coordination and support (NSF IIS-1550320). Data provided in part by the Human Connectome Project (NIH 1U54MH091657) and Brian Wandell (NSF BCS-1228397). Additional support to I.D. provided under NSF 1734853, NSF 1636840, P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760. A.J.S. received support from the following NIH grants: P30 AG010133, R01 AG019771, R01 LM011360, R01 CA129769 and U01 AG024904.

Author Contributions

P.A., F.P., S.H. processed the data. S.H., P.A., A.P., L.K., B.C., Y.Q., B.M. and F.P. contributed analysis software. P.A., S.H., E.O., curated the data. F.P. and S.H. designed and implemented https://brainlife.io. F.P. and P.A. wrote the manuscript. O.S., A.S., R.H., E.G., C.F.C., B.M., I.D., D.B., L.W. edited the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

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