A Principled Approach to Human Connectome Estimation and Meganalysis

Gregory Kiar*, Eric W. Bridgeford*, Vikram Chandrashekhar, Disa Mhembere, Consortium for Reliability and Reproducibility (CoRR), Sephira Ryman, Rex Jung, Randal Burns, Zhi Yang, Xi-Nian Zuo, Michael P. Milham, William R. Gray Roncal†, Joshua T. Vogelstein†
Johns Hopkins University, University of New Mexico, Child Mind Institute, & University of Chinese Academy of Sciences

Abstract

Visualizing and quantifying connectivity within the human brain is fundamental to understanding connectome coding: the organizational principles governing neural connectivity. Critical to this effort are pipelines capable of processing data from multimodal magnetic resonance imaging (including structural, functional, and diffusion) to generate estimates of brain connectivity. Pipelines developed to date vary according to the unique needs of the studies at hand, creating challenges for reproducibility and hampering the pace of scientific discovery. We propose a set of algorithmic and implementation principles to guide the development of reference pipelines that can be used by different investigators using different hardware, acquisition protocols, and sample demographics, to address disparate scientific and clinical goals. Neurodata's MRI to Graphs (NDMG) pipeline was built in accordance with these principles. Running NDMG on 18 functional datasets and 10 diffusion datasets (a total of 2,295 subjects and 4,507 scans) and evaluating the results demonstrates that NDMG meets or exceeds the standards of excellence in each principle across multiple imaging modalities. Using NDMG on the largest multi-site dataset collected to date, we discover that while certain connectome properties are preserved across all sites, cross-site variability ameliorates the value of pooling data across sites, suggesting the need for data acquisition harmonization to achieve reproducible statistical connectomics in both basic and translational research.

1 Introduction

Neuroimaging datasets from magnetic resonance imaging (MRI) are becoming increasingly available in clinical and research populations [17; 50; 78]. This includes both Diffusion Weighted MRI (DWI), a technique that provides high contrast for the connective tissue of the brain (i.e. white matter) [64], and resting-state functional MRI (fMRI), a technique that provides estimates of the correlations between gray matter regions of the brain [6; 78]. These multimodal MRI data provide the scientific community with a unique opportunity to identify clinically useful biomarkers and discover the principles of connectome coding.

Yet, discovery from these datasets requires algorithmically and computationally sophisticated pipelines to process the data. Previously, a number of pipelines have been developed for either DWI [13; 15] or fMRI [5; 12; 19; 26] datasets. However, while the algorithms that these pipelines leverage have become standard, the pipelines have not. Indeed, there is a distinct lack of reference pipelines in the field. This has resulted in each publication using (sometimes subtly) different parameters, often without specifying precise details and dependencies, making reproducibility difficult. Moreover, the lack of reference pipelines also creates inefficiencies in the collective scientific process because pipelines have to be designed and tuned for each study, essentially. Perhaps even more problematic is that such pipelines cannot readily be used for meganalysis (an approach that pools data across datasets, centers, or sites) because most are idiosyncratically designed for a particular study [11]. This stifles our ability to understand and quantify batch effects in neuroimaging, which are extremely problematic in -omics data [27; 43; 53; 60], though relatively understudied [20; 21; 55].

To address these problems with reproducibility and efficiently, we sought to develop a "reference" pipeline. In so doing, we established several prin-

ciples and metrics by which novel pipelines can be evaluated. The principles are organized into two categories: algorithmic and implementational. Methods that empirically perform well along the algorithmic principles—statistical accuracy, reliability, robustness, and expediency—can be trusted by other investigators to yield results with scientific and clinical utility. Methods that perform well along the implementation principles—scalable, portable, turnkey, and open—are easy to use by investigators with different sets of domain expertise (e.g., neuroimagers or machine learners).

Our approach, "Neuro Data MRI to Graphs" (NDMG), meets or exceeds standards of excellence along each of the above mentioned principles. We validated our pipeline by running NDMG on DWI data from 10 sites, comprising 2,295 subjects with 2,861 scans, and fMRI data from 18 sites, comprising 714 subjects with 1,646 scans. For each scan NDMG estimated a connectomes at multiple spatial resolutions, yielding a total of over 75,000 estimated connectomes, all of which are now publicly available. This is the largest database of connectomes to date [9], and, to our knowledge, the largest meganalysis of connectomics data [71].

These data motivated the development of several statistical connectomics methods to quantify a number of coarse-scale connectome properties, such as the relative probability of ipsilateral vs. contralateral connections. We were able to apply these methods to each site, therefore providing perhaps the largest body of evidence in favor of many of these principles of connectome coding. Moreover, we determined that, even after optimizing and harmonizing the connectome estimation pipelines, there are significant quantitative differences across sites that are apparent at finer scales. This suggests that further work is required for statistical connectomics to produce accurate and reliable p-values or clinical biomarkers across datasets.

2 Results

Table 1 depicts NDMG's performance with regard to each of the below described principles. For each principle, we outline a procedure for evaluating the degree to which an approach adheres to that principle, enabling the principle-based evaluation of disparate approaches. Moreover, for each comparable pipeline, it receives a ✓ if it adheres to the principle, an ✓ if it partially adheres, and a ✗ if it does not adhere. Pipelines tied for best in a particular column are denoted with a *.

Algorithmic Principles The algorithmic principles are designed to evaluate the empirical quality of the method on real data. Better quality means the method can be trusted for subsequent investigations.

Accuracy quantifies the distance between "the truth" and an estimate, and is closely related to the statistical concepts of "unbiased" and "consistent". Because the truth is unknown for these data, instead we developed an extensive set of quality assurance (QA) metrics and figures, to subjectively evaluate the accuracy of each processing stage. For example, users can visualize fibers to confirm that they all stay within the skull. NDMG incorporates the QA from Craddock et al. [12], and other pipelines, and also adds several novel QA figures, yielding a total of 11 QA figures per diffusion scan and 18 per functional scan.

Reliability colloquially refers to methods that produce a similar result given a similar input, also called "stability" in the statistics literature [76]. To evaluate a method's reliability, Wang et al. [73] developed a metric called "discriminability" that quantifies the fraction of measurements from the same subject that are closer to one another than they are to the measurement of any other subject (details below). NDMG's discriminability over all scans was nearly 0.98 for DWI data and over 0.88 for fMRI. Other pipelines have been evaluated using intrac-class correlation, for example, but on fewer datasets.

Robustness quantifies accuracy and reliability across a wide range of datasets with different properties, including different experimental design, measurement devices, etc. We therefore ran NDMG on 10 DWI datasets and 18 fMRI datasets using different hardware and acquisition parameters (see Table 2 for details). In some of these datasets samples were not filtered to discard outliers or samples with poor signal-to-noise properties. Nonetheless, for each site, NDMG's QA confirmed accuracy, and each dataset achieved a score of discriminability above 0.8 for each site. Other pipelines have not demonstrated accuracy and reliability on as many datasets; some can only be applied to datasets using particular scan-

Table 1: **Comparing M3R Processing Pipelines.** NDMG is designed with both algorithmic and implementation principles in mind. This table compares existing pipelines along these principles, demonstrating that for each, NDMG performs at least as well as the current state of the art. The parentheses denotes the modality of the pipeline being considered, DWI (d) or fMRI (f). A checkmark ✓ is given for pipelines that satisfy the respective desiderata, a ✓ for pipelines that partially satisfy the respective desiderata, and a ✗ is given for pipelines that do not satisfy the respective desiderata. A technology evaluation showing how the columns were chosen along with an explanation of each cell not receiving a ✓ can be found in Appendix A.

pipeline	accurate	reliable	robust	expedient	multimomdal	end-to-end	scalable	portable	turn-key	open
NDMG (d & f)	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*
CPAC[12] (f)	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓
HCP[26] (f)	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓
fmriprep[19] (f)	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓
NIAK[5] (f)	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓
PANDA[13] (d)	✓	✗	✗	✓	✗	✓	✓	✗	✗	✗
CMTK[15] (d)	✓	✗	✗	✓	✗	✓	✓	✗	✗	✓

ners and/or scanner parameters (for example, the HCP pipeline [26]).

Expediency refers to the time it takes to run an approach on a typical sample. The NDMG runtime is about 20 minutes for a functional scan, and < 1 hour for diffusion scan. Pipelines that require about a day are not expedient.

Implementation Principles Adhering to the implementation principles lowers the barrier for use. In practice, this means that both domain experts and researchers from other disciplines (such as machine learning and statistics), can more easily use the tools.

Multimodal pipelines operate on both functional and diffusion weighted MRI data.

End-to-End refers to the pipeline performing all steps of analysis required to acquire connectomes given raw, unprocessed M3R scans with no user input. The NDMG-d pipeline was built to take raw DWI and T1w images and produce a diffusion connectome, and the NDMG-f pipeline takes raw fMRI and T1w images and produces a functional connectome. Other functional pipelines do not provide algorithms for actually estimating connectomes.

Scalability refers to the ability of the method to parallelize the code across multiple computers. Indeed, NDMG enables parallelization across scans—using for example AWS Batch or a high-performance cluster—requiring only an hour to run many thousands of scans.

Portability—meaning the ability to be run on different platforms, from laptop and cloud, with minimal installation and configuration energy—enables different analysts using different hardware resources to seamlessly use the code. We have tested NDMG on Windows, OSX, and Linux laptops, multi-core workstations, singularity clusters, and the Amazon cloud. Moreover, we have deployed NDMG on both openneuro [**?**] and CBRAIN [61], making it possible for anybody to run NDMG for free on their own or other's computational resources.

Turn-key methods do not require the user to specify parameters and settings for each stage of processing or for each new dataset. This feature reduces the time for researchers to get a pipeline running, and enables pooling data across multiple pipelines because the analysis is harmonized. NDMG parameters have been tuned to yield accurate and reliable connectome estimates across nearly 30 different datasets. Moreover, NDMG is fully compliant with Brain Imaging Data Structure (BIDS)—a recently proposed specification for organizing multi-scan, multi-subject, multi-modality datasets [28; 29].

Openness, referring to both open source code and open access data derivatives processed using the code, enables anybody with Web-access to con-

tribute to the scientific process. NDMG leverages open source packages with permissive licenses, and is released under the Apache 2.0 open source license. Our website, `http://m2g.neurodata.io` contains links to download all of the data derivatives and quality assurance figures from each scan, the largest database of connectomes and other data derivatives to our knowledge.

By developing NDMG according to these algorithmic and computational design principles, and running it on many diverse datasets, we can evaluate subjects, sites, and the collection of sites at an unprecedented scale. Below, we describe the nuts and bolts of the pipeline, followed by a set of NDMG -enabled scientific findings. Our hope is that NDMG and the data products derived from it will be useful for a wide variety of discoveries.

2.1 Subject-Level Analysis

In the subject-level analysis, a subject is a particular human. The input for a given session includes a structural scan (T1w/MPRAGE), and either or both of (1) a diffusion scan (DWI), including the diffusion parameters files (b-values, b-vectors), and (2) a functional scan, including the slice acquisition sequence. The subject-level of NDMG analysis leverages existing open source tools, including FSL [37; 63; 75], Dipy [25], the MNI152 atlas [46], and a variety of parcellations defined in the MNI152 space [18; 39; 41; 45; 54; 65; 69]. All algorithms requiring hyperparameter selection were initially set to the suggested parameters for each tool, and tuned to improve the accuracy, reliability, expediency, and robustness of the results. The output of each processing stage includes data derivatives and QA figures to enable subjective accuracy assessments. The QA figures at many stages include cross-sectional images at different depths in the three canonical planes (sagittal, coronal, and axial) of images or overlays. Example QA figures are provided in Appendix C.1.

Subject-Level Diffusion Analysis The NDMG-d pipeline consists of four key components: (1) registration, (2) tensor estimation, (3) tractography, and (4) graph generation (see **??** for further details):

Registration uses FSL's "standard" linear registration pipeline to register the structural and diffusion images to the MNI152 atlas [37; 46; 63; 75].

Tensor Estimation uses DiPy [25] to obtain an estimated tensor in each voxel.

Tractography uses DiPy's *EuDX* [24], a deterministic tractography algorithm closely related to *FACT* [49] to obtain a streamline from each voxel.

Graph Generation uses our own custom Python code that places an edge between a pair of regions of interest (ROIs) whenever a streamline passes through both, yielding a weighted voxel-wise graph. Downsampling and QA are performed as in the functional pipeline, and are described below.

Subject-level analysis in NDMG-d takes approximately one hour to complete using 1 CPU core and 12 GB of RAM at 1mm resolution. The subject-level analysis was run on 10 datasets, including 2,295 subjects and 2,861 scans. Each dataset generated connectomes across each of the 24 parcellations in NDMG-d, resulting in 68,664 total brain-graphs.

2.1.1 Subject-Level Functional Analysis

The NDMG-f pipeline can be broken up into four key components: (1) preprocessing, (2) registration, (3) nuisance correction, and (4) graph generation (see Appendix C for further details). The NDMG-f pipeline was constructed starting with the optimal processing pipeline identified in Wang et al. [73] using Craddock et al. [12]. Hyperparameters and further algorithm selection was optimized for reliability based on multiple measurement datasets (including test-retest). Below, we provide a brief description of each step.

Preprocessing uses AFNI [1] for brain extraction, and FSL [35; 74] for slicetiming correction and motion correction. QAX overlays pre- and post-correction images.

Registration uses FSL [3; 33; 63] to non-linearly register the fMRI volume to the MNI152 atlas [46]. The registration pipeline implemented is "standard" when working with functional data and FSL's tools.

Nuisance Correction uses custom Python code to implement a general linear model incorporating regressors for quadratic detrending [62;

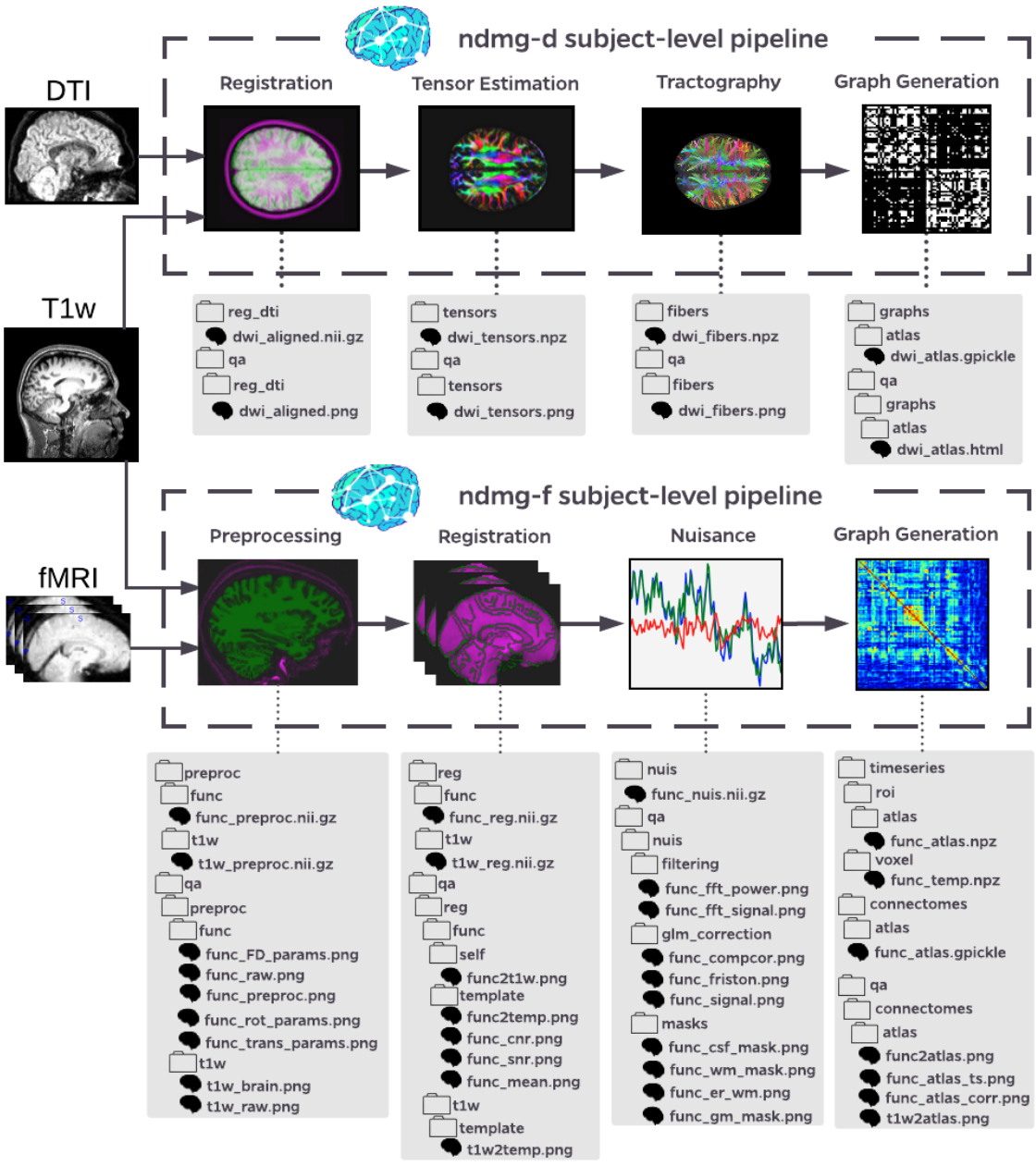


Figure 1: The subject-level NDMG-d pipeline transforms raw DWI data into sparse structural connectomes, whereas the NDMG-f pipeline transforms raw fMRI data into dense functional connectomes. Each pipeline consists of four key steps, and each step generates both data derivatives and quality assurance figures to enable both qualitative assessments and quantitative comparisons (see Appendix B and Appendix C for details).

66], top 5 white-matter and cerebrospinal fluid Principal Components (CompCor) [4; 10], and the Friston 24-parameter regressors [23]. Low-frequency drift is then removed below 0.01 Hz [44], and the first 15 seconds of the fMRI sequence are discarded [8].

Graph Generation uses custom Python code to compute the average timeseries for all voxels within an ROI, then compute correlations be-

tween all pairs of ROIs. The functional connectome is then rank-transformed by replacing the magnitude of the correlation with its relative rank, from smallest to largest [73].

Subject-level analysis in NDMG-f takes approximately 20 minutes to complete using 1 CPU core and 3 GB of RAM at 2mm resolution. The subject-level analysis was run on 714 subjects and 1,646 from 18 datasets; each generating connectomes across each of the 24 parcellations in NDMG-d, resulting in 39,504 total brain-graphs.

Multi-Scale Multi-Connectome Analysis For both diffusion and functional MRI, NDMG downsamples the voxel-wise graphs to obtain weighted graphs for many different parcellation schemes. This includes: (1) neuroanatomically delineated parcellations, such as the HarvardOxford cortical and sub-cortical atlases [45], JHU [54], Talairach [41], Desikan [18], and AAL [69] atlases; (2) algorithmically delineated parcellations, such as slab907 [65], Slab1068 [39], CC200 [12]; and (3) 16 downsampled (DS) parcellations [47] ranging from 70 to 72,783 nodes that we developed. The QA for graph generation includes a heatmap of the adjacency matrix, the number of non-zero edges, and several multivariate graph statistics (one statistic per vertex in the graph): betweenness centrality, clustering coefficient, hemisphere-separated degree sequence, edge weight, eigenvalues of the graph laplacian, and locality statistic-1 [47]. We developed the hemisphere-separated degree sequence to indicate the ipsilateral degree and contralateral degree for each vertex, which we found quite useful for QA. Appendix C.4 includes definitions and implementation details for each of the statistics. Supplementary Figure 14 shows, for a single subject, the graph summary statistics for the multi-connectome (including both functional and diffusion) across the set of atlases.

2.2 Group-Level Analysis

We ran NDMG on the 10 diffusion and 18 functional datasets, listed in Table 2. For each, NDMG group-level analysis computes and plots group-level graph summary and reliability statistics.

Graph Summary Statistics Each scan's estimated connectome can be summarized by a set of graph statistics, as described above. For group-level analysis, we visualize each session's summary statistics overlaid on one another. For example, Figure 2 demonstrates that each diffusion graph from the BNU3 dataset using the Desikan atlas has relatively similar values for the statistics (we use the Desikan atlas for the remainder of the analyses unless otherwise specified). It is clear from both the degree plot and the mean connectome plot that the DWI connectomes from this site tend to have more connections within a hemisphere than across a hemisphere, as expected. Supplementary Figure **??** shows example group-level analysis for the same site's functional data.

Reliability Group-level results from NDMG that include repeated measurements are quantitatively assessed using a statistic called discriminability [73]. The group's sample discriminability estimates the probability that two observations within the same class are more similar to one another than to objects belonging to a different class:

$$D = Pr(||a_{ij} - a_{ij'}|| \leq ||a_{ij} - a_{i'j'}||). \quad (1)$$

In the context of reliability in NDMG, each connectome, a_{ij}, is compared to other connectomes belonging to the same subject, $a_{ij'}$, and to all connectomes belonging to other subjects, $a_{i'j'}$. A perfect discriminability score indicates that for all observations within the group, each connectome is more alike to connectomes from the same subject than to others. Table 2 lists the discriminability score of each dataset with repeated measurements. NDMG-d achieves a discriminability score of nearly 0.99 or greater on most datasets (the lowest scoring was nearly 0.9). NDMG-f achieves a discriminability score of around 0.9 on all datasets. Note that we invested significant effort in both pipelines to achieve these high discriminability scores, and to our knowledge, no other pipelines have been assessed with regards to discriminability.

2.3 Meganalysis

Many sources of variability contribute to the observed summary statistics, including subject- or

Table 2: NDMG **pipeline robustness and reliability.** We ran NDMG on over 20 different datasets, including both fMRI and DWI data, spanning multiple different scanners, acquisition parameters, and population demographics. Nonetheless, for both fMRI and DWI data, across all datasets with multiple measurements, NDMG always achieved > 0.8 discriminability, and NDMG-d's discriminability was typically > 0.98 on the DWI data.

						DWI		fMRI	
Dataset	Scanner	Age	Portion Male	Subjects	Sessions	Total	TRT	Total	TRT
BNU1 [78]	Siemens	23.0 ± 2.3	0.53	57	2	114	0.984	108	0.906
BNU2 [78]	Siemens	$20.9 \pm .9$	0.54	61	2	–	–	121	0.863
BNU3 [78]	Siemens	22.5 ± 2.1	0.50	48	1	47	NA	47	NA
HNU1 [78]	GE	24.4 ± 2.3	0.50	30	10	300	0.993	300	0.956
KKI2009 [42]	Philips	31.8 ± 9.4	0.52	21	2	42	1.0	–	–
MRN1313	–	–	–	1313	1	1299	NA	–	–
NKI1 [78]	Siemens	34.4 ± 12.8	0.0	24	2	40	0.984	–	–
NKI-ENH [52]	Siemens	42.5 ± 19.6	0.40	198	1	198	NA	–	–
SWU1 [78]	–	21.55 ± 1.7	0.3	20	3	–	–	60	0.935
SWU2 [78]	–	21.0 ± 1.6	0.33	27	2	–	–	54	0.874
SWU3 [78]	–	20.4 ± 1.6	.35	23	2	–	–	46	0.986
SWU4 [78]	–	20.0 ± 1.3	0.51	235	2	454	0.884	466	0.891
Templeton114	Siemens	21.8 ± 3.0	0.58	114	1	114	NA	–	–
Templeton255	Siemens	22.1 ± 3.9	0.51	255	1	253	NA	–	–
IPCAS1 [78]	Siemens	20.9 ± 1.7	0.3	30	2	–	–	60	0.893
IPCAS2 [78]	Siemens	13.4 ± 0.96	0.35	34	2	–	–	68	0.868
IPCAS5 [78]	Siemens	18.3 ± 0.5	1.0	22	2	–	–	44	0.819
IPCAS6 [78]	Siemens	23.0 ± 2.0	0.5	2	9	–	–	18	0.994
IPCAS8 [78]	Siemens	57.6 ± 3.6	0.46	13	2	–	–	26	0.958
IBATRT [78]	Siemens	28.02 ± 7.5	0.54	36	2	–	–	50	0.974
NYU1 [78]	Siemens	29.44 ± 8.4	0.4	25	3	–	–	75	0.931
UWM [78]	–	24.96 ± 3.2	0.56	25	2	–	–	50	0.849
XHCUMS [78]	Siemens	–	–	23	5	–	–	115	0.823
Pooled dti				2295		2861	0.979	–	–
Pooled fMRI				714		–	–	1646	0.881

population-level variation, or different types of measurement or analysis techniques. By virtue of harmonizing the analysis across subjects and groups, we are able assess the remaining degrees of variability due to measurement and population-specific effects. Although population-level effects are expected when comparing two different populations with different demographics, variability across measurements must be relatively small for inferences based on neuroimaging to be valid. Therefore, we conducted meganalysis to address these remaining sources of variation.

Coarse Grained Connectome Codes Preserved Across Disparate Groups Figure 11d (top) shows the mean estimated connectome computed from each dataset for which we have both DWI and fMRI data. We also calculate the mega-mean connectomes, derived by pooling the datasets. Several group-specific properties are readily apparent simply by visualizing the connectomes:

a. The DWI connectomes have significantly stronger ipsilateral connections than contralateral connections.

b. The fMRI connectomes have significantly stronger bilateral connections than non-bilateral connections.

To formally test these initial assessments we developed statistical connectomics models and methods. Specifically, we developed a structured, independent edge random graph model that generalizes the stochastic block model (SBM) [38] and the independence edge random graph model [7]. In this new model, each edge is sampled independently, but there are only K possible probabilities, and we have *a priori* knowledge of which edges are in which groups. Unlike the SBM model, in which each vertex is in a group, here each edge is in a group. We then developed test statistics that are consistent for this model, meaning that with sufficient data, power (the probability of correctly rejecting a false null hypothesis) approaches unity (see Appendix **??** for details).

Using the approach described above, we first test whether ipsilateral connections tend to be stronger

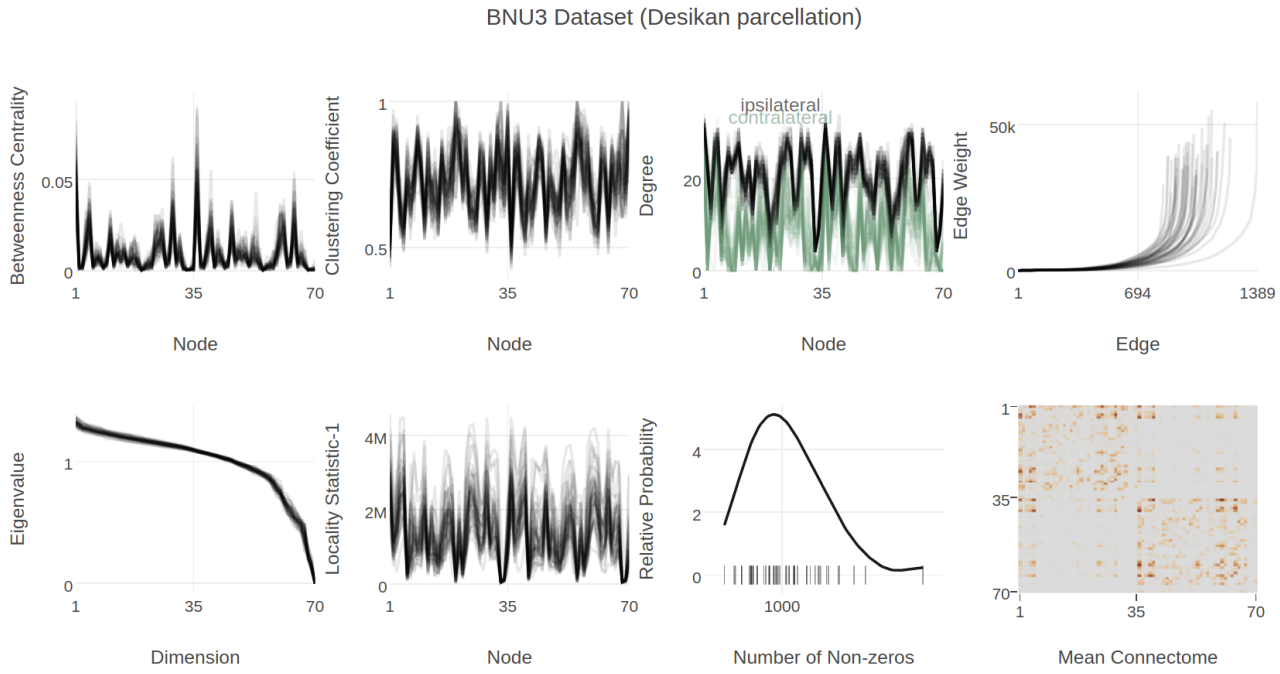


Figure 2: **Graph summary statistics**. NDMG computes and plots of connectome-specific summary statistics after estimating graphs for each sscan, providing immediate quality assurance for the entire site. In theory, any connectome could be an outlier for any of these statistics, so plotting all of them together is particularly useful. Details about algorithm choice for the summary statistics are provided in **??**.

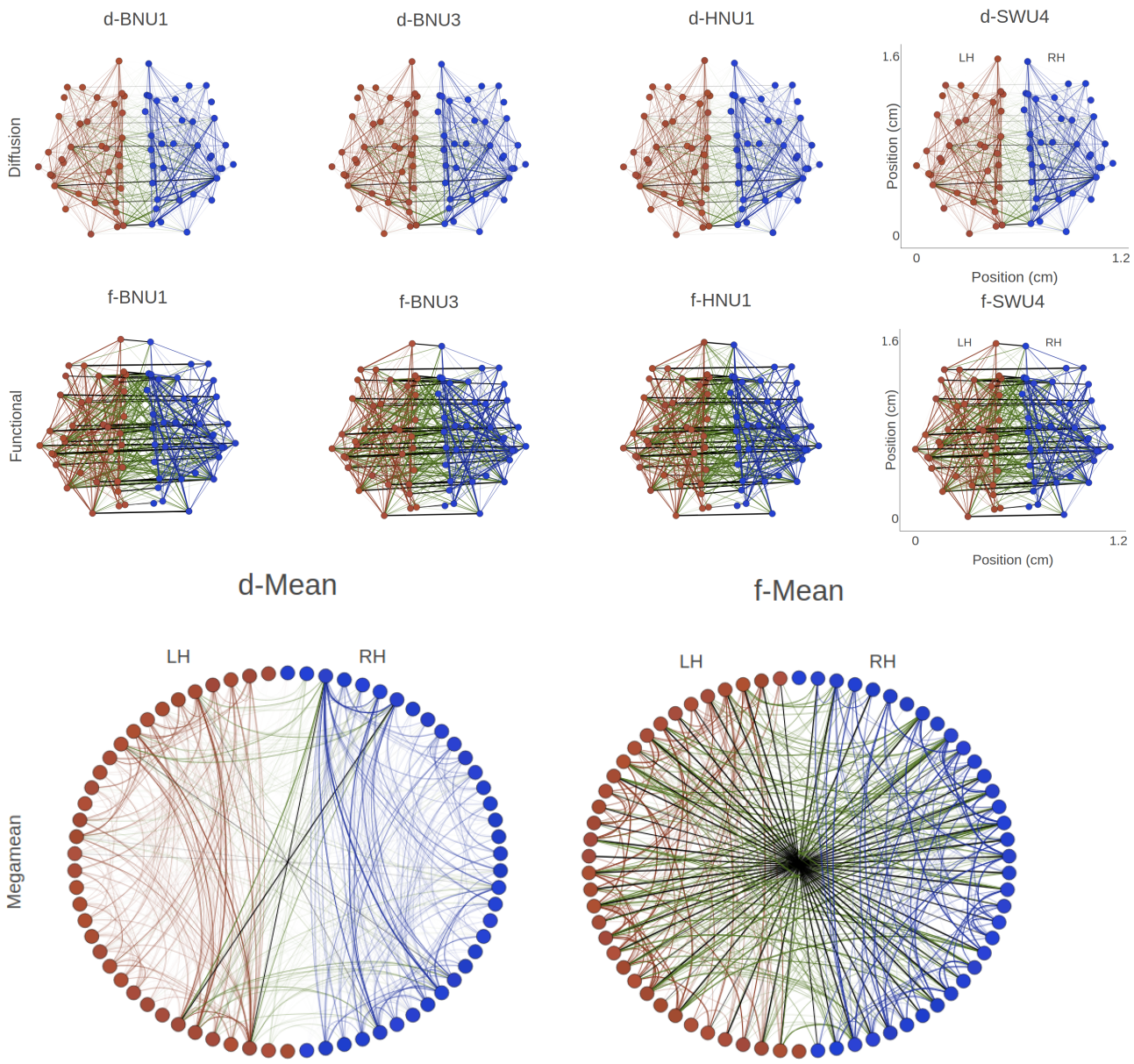


Figure 3: **Multi-Study Mean Connectomes**. Site-specific mean connectomes (top two rows) and megamean (bottom row), using the Desikan parcellation, using all the data for which we have both functional and diffusion datasets (> 900 scans of each), with edges color coded as follows: **blue**: edges within the left hemisphere; **red**: edges with right hemisphere; **black**: bilateral connections; **green**: non-bilateral contralateral connections. In the top rows, graphs are shown with vertex position determined by the coronally-projected center of mass for each region in the desikan parcellation. The bottom shows radial plots organized by hemisphere. Same-modality connectomes appear qualitatively similar to one another across sites, but some differences across modalities are apparent. For example, bilateral and contralateral connections both seem stronger in functional than diffusion mean connectomes, both within site and after pooling all sites.

than contralateral connections. As can be seen in Figure 4, for 99.9% of DWI connectomes and 6.4% of fMRI connectomes, ipsilateral connections are stronger at the $\alpha = 0.05$ level. Note that this re-

sult uses data pooled from 17 different datasets for fMRI, and 9 different datasets for DWI. We were surprised to see the extent to which the functional connectomes did not exhibit this property, and therefore subsequently investigated whether for a given subject, the difference between ipsilateral and contralateral was stronger in the DWI data than the fMRI data.

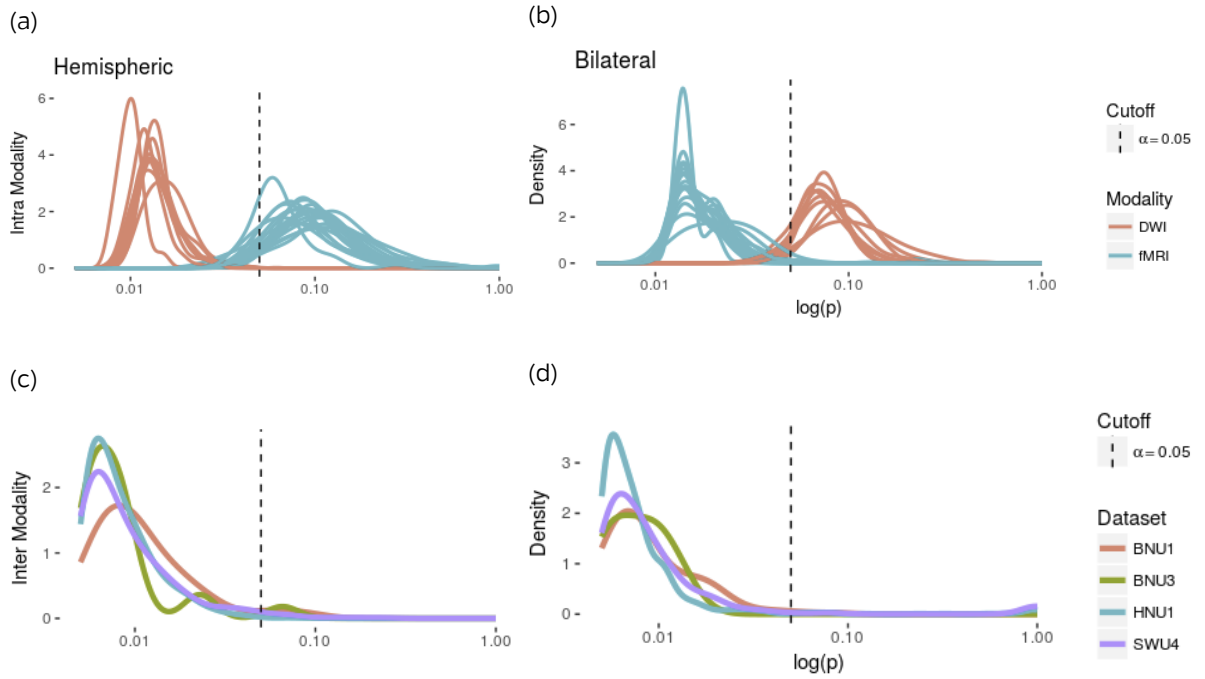
As shown in Figure 4, we determined that out of 907 subjects with both DWI and fMRI, 99.5% exhibit significantly stronger ipsilateral versus contralateral connections in the DWI dataset as compared to their corresponding fMRI dataset. The mega-mean connectomes also indicated that bilateral connections are stronger than non-bilateral connections in the fMRI datasets. To investigate this, we modified our model such that one group corresponds to only bilateral connections, and the other group corresponds to the remaining edges. For the 907 subjects with both DWI and fMRI, 99.0% exhibit significantly stronger bilateral versus non-bilateral connections in the fMRI data than in their corresponding DWI dataset.

These few examples demonstrate that coarse-scale properties of connectomes, across both DWI and fMRI, are preserved across batches with widely varying data collection strategies and processes.

Fine-Grained Difference Across Sites with Implications At coarse resolution, each site's connectomes exhibited similar properties. However, the above analysis is insufficient to determine the extent of "batch effects"— sources of variability such as scanner, acquisition sequence, and operator, none of which are of neurobiological interest. If the batch effects are larger than the signal of interest (for example, whether a particular subject is suffering from a particular psychiatric disorder), then inferences based on individual studies are prone to be irreplaceable, thereby creating inefficiencies in the collective scientific process. We therefore use discriminability to quantify the degree to which different groups differ from one another. More specifically, using the discriminability framework, and keeping only a single session per subject, we compute the discriminability across groups, rather than subjects. On the subject level, high discriminability indicates that subject variability is larger than other sources of variability within a group. However, on the group level, low discriminability is desirable as it indicates that group variability (i.e., the batch effect) is smaller than biological variability. Chance discriminability, defined as the level at which connectomes are equally similar regardless of dataset, can be calculated using $C = \frac{1}{N^2}\sum_{i\in k}^{k} M_i^2$, where k is number of classes, M_i is the number of elements in per class, and N is the total number of observations. The discriminability across a random subsample of single session per subject is 0.632, as compared to chance levels which are 0.259, a significant difference at the $p < 0.0001$ level (see [73] for details). Figure 5 shows the average discriminability both within and across datasets.

The KKI2009 dataset was acquired on a Philips scanner, whereas the other groups predominantly used Siemens scanners (HNU1 alone used GE). Removing either KKI2009 or HNU1, or both, from the group of diffusion connectomes that used non-Siemens scanner did not meaningfully change discriminability (0.626 with p<0.0001 for removing KKI2009, and 0.627 with p<0.0001, for removing both KKI2009 and HNU1). Templeton114 and Templeton255 were acquired at the same site, using different scanner sequences, and exhibited a significant difference.

The fact that there are significant batch effects does not on its own indicate that downstream inference tasks, such as calculating p-values or developing biomarkers on the basis of the estimated connectomes, will be impossible. Rather, it is the relative size of the batch effect, as compared to the effect of the signal of interest, that determines the impact of batch effects. To assess the extent to which batch effects are influencing these data, we built two-class classifiers to differentiate subject sex on the basis of their connectomes. Previous work has demonstrated performance significantly above chance for this task, with accuracy typically in the 80% to 90% range [72]. Figure 5 depicts the leave-one-out (LOO) subject out classification error for each of the seven DWI datasets and eighteen fMRI datasets that contain sex information; the accuracy ranges from around 40% to around 80% using a k-nearest neighbor classifier, and post-hoc selecting the optimal k. We then pooled the subjects across datasets, and computed both the LOO subject and LOO dataset accuracies. If the batch effect was smaller than the sex effect, then pooling the data would increase the effective sample size, and therefore improve accuracy. However, the LOO-pooled data exhibited poor accu-



(e)

Test	DWI	fMRI	inter-modality
Ipsi-Lateral vs. Contra-Lateral Connectivity	99.9%	6.4%	99.5% of DWI exceed fMRI
Bilateral vs. Non-Bilateral Connectivity	99.4%	7.7%	99.0% of fMRI exceed DWI

Figure 4: **Structured Independent Edge Model**. In 4a, we look at the density estimates of the per-graph p-values that ipsi-lateral connectivity exceeds contra-lateral connectivity. One line is reported per dataset, with the modality as indicated by the legend. Low p-values indicate that we are less-likely to falsely report a difference in connectivity ipsi-laterally vs. contra-laterally unless one is present. As we can see in 4e, we determine that 99.9% of the DWI connectomes exhibit this property, whereas only 6.4% of the fMRI connectomes exhibit this property. In 4c, we look at the density estimates of the 2-graph p-values that this difference in connectivity is stronger in the DWI connectomes for a particular subject than in the fMRI connectomes for that subject. Low p-values indicate taht we are less-likely to falsely report a difference in connectivity in DWI connectomes vs. fMRI connectomes. 99.5% of the graph-pairings exhibit this property. In 4b, we look at the density estimates of the per-graph p-values that bilateral connectivity exceeds non-bilateral connectivity. Low p-values indicate that we are less-likely to falsely report a difference in connectivity bilaterally vs. non-bilaterally. 99.4% of the fMRI connectomes exhibit this property, whereas only 7.7% of the DWI connectomes exhibit this property. In 4d, we look at the density estimates of the 2-graph p-values that this difference in connectivity is stronger in the fMRI connectomes for a particular subject than in the DWI connectomes for that subject. Low p-values indicate that we are less-likely to falsely report a difference in connectivity in fMRI connectomes vs. DWI connectomes. 99.0% of the graph-pairings exhibit this property. For more details, see Appendix E.

racy, with LOO-dataset performing even worse than many of the individual datasets, indicating that the batch effect was indeed eclipsing the sex effect in these data.

It is possible that a large batch effect could be mitigated by some type of normalization scheme.

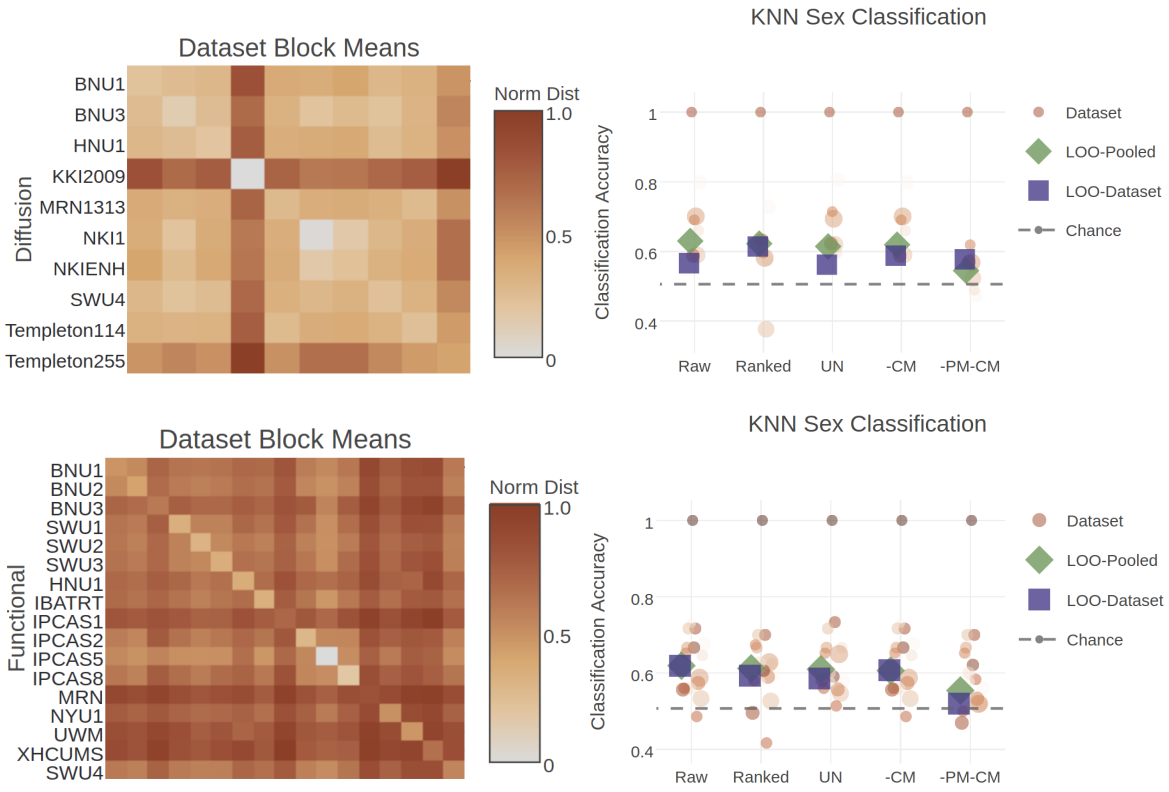


Figure 5: **Prevalence of batch effects**. **(A)** Discriminability was computed across all sites (Table 2) using the site as the class label for diffusion (top) and functional (bottom) connectomes. If no significant difference between sites exists, the discriminability score for the connectomes would not be significantly different from chance (0.362 for the diffusion connectomes and 0.138 for the functional connectomes). The discriminability score was 0.632 for the diffusion connectomes and 0.619 for the functional connectomes. A permutation test demonstrates that these scores are significantly different from chance (both p-values < 0.0001), suggesting there is significant site-specific signal present for both modalities. **(B)** K-Nearest Neighbors (KNN) classifiers were trained to classify sex on the basis of either diffusion (top) or functional (bottom) data, both using individual sites and then pooling sites. If different sites batch effects were not important for subsequent inference, pooling should improve accuracy by virtue of obtaining larger sample sizes. We compared several different normalization strategies. For both functional nor diffusion, no strategy that we adopted was able to mitigate the site-specific batch effects, so pooling data did not improve performance, using either leave-one-out (LOO) subject or site.

We considered several, including: converting the edge weights to relative ranks, unit-normalizing each connectome, subtracting the cohort-mean, and subtracting the population-mean and then the residual cohort-mean. Converting to relative ranks improved performance on the DWI data, but none of the normalization schemes we tested improved performance on the fMRI data. Collectively, these results suggest that the batch effect present in these data is large, and that signals found in one dataset were idiosyncratic to that dataset, rather than representing true neurobiological signals of interest.

3 Discussion

The NDMG pipeline is a reliable tool for structural connectome estimation that has a low barrier to entry for neuroscientists, and is capable of producing accurate brain-graphs across scales and datasets. NDMG abstracts hyper-parameter selection from users by providing a default setting that is robust across a variety of datasets, achieving equal or improved discriminability when performing either single- or multi-dataset analysis compared to alternatives [57; 58]. Though this generalizability means that NDMG may not use the optimal parameters for a given dataset, it provides a consistent estimate of connectivity across a wide range of datasets. This makes it trivial to compare graphs across studies and avoids overfitting of the pipeline to a specific dataset. NDMG has been optimized with respect to discriminability, yet one can always further improve the pipeline via incorporating additional algorithms, datasets, or metrics. For example, one could further optimize to reduce the batch effect. Alternately, one could incorporate probabilistic tractography to compare with deterministic approaches in a principled meganalysis using the open source data derivatives generated here.

Previous efforts have developed pipelines for DWI data. For example, PANDAS [14] and CMTK [16] are flexible pipelines enabling users to select hyper-parameters for their dataset. This is a useful feature, but they do not provide a reference pipeline that is optimized for any particular criteria across datasets. MRCAP [31] and MIGRAINE [32] provide reference pipelines, but are difficult to deploy, and also lacked vetting across datasets.

Other efforts have focused on multi-site data. Specifically, Abraham et al. [2] used fMRI-derived connectomes from the ABIDE dataset, and demonstrate an impressive ability to minimize batch effects. Unfortunately, most ABIDE datasets lack DWI data, so a similar strategy for NDMG-d is not currently possible. Similarly, concomitant with our manuscript is Noble et al. [51], who conducted a multi-scanner analysis with harmonized fMRI acquisition on 12 subjects. The compelling results suggest that pooling across datasets with harmonized acquisition can potentially improve reliability. Additionally, a variety of studies propose methods for post-acquisition data harmonization using either minimally pre-processed or raw MRI data [22; 48; 56] that could be explored within the context of the NDMG pipeline.

Author Information CoRR members contributed most of the data used. GK[1] led experimental design and writing of the manuscript, performed analyses, and co-designed NDMG-d; EB[1] wrote the experiments, analysis, portion of the paper corresponding to fMRI, and designed NDMG-f; VC[1] assisted with conducting experiments; DM[1] wrote the large graph analysis, advised by RB[1]; Rex Jung[7] and Sephira Ryman[7] provided DWI data; WGR[1] developed the initial prototype for creating and assessing DWI connectomes, ran preliminary experiments, co-designed NDMG-d, and co-advised GK; JTV[1,2] oversaw everything and is the corresponding author: ⟨jovo@jhu.edu⟩.

CoRR Members Jeffrey S. Anderson, Pierre Bellec, Rasmus M. Birn, Bharat B. Biswal, Janusch Blautzik, John C.S. Breitner, Randy L. Buckner, Vince D. Calhoun, F. Xavier Castellanos, Antao Chen, Bing Chen, Jiangtao Chen, Xu Chen, Stanley J. Colcombe, William Courtney, R. Cameron Craddock, Adriana Di Martino, Hao-Ming Dong, Xiaolan Fu, Qiyong Gong, Krzysztof J. Gorgolewski, Ying Han, Ye He, Yong He, Erica Ho, Avram Holmes, Xiao-Hui Hou, Jeremy Huckins, Tianzi Jiang, Yi Jiang, William Kelley, Clare Kelly, Margaret King, Stephen M. LaConte, Janet E. Lainhart, Xu Lei, Hui-Jie Li, Kaiming Li, Kuncheng Li, Qixiang Lin, Dongqiang Liu, Jia Liu, Xun Liu, Guangming Lu, Jie Lu, Beatriz Luna, Jing Luo, Daniel Lurie, Ying Mao, Daniel S. Margulies, Andrew R. Mayer, Thomas Meindl, Mary E. Meyerand, Weizhi Nan, Jared A. Nielsen, David O'Connor, David Paulsen, Vivek Prabhakaran, Zhigang Qi, Jiang Qiu, Chunhong Shao, Zarrar Shehzad, Weijun Tang, Arno Villringer, Huiling Wang, Kai Wang, Dongtao Wei, Gao-Xia Wei, Xu-Chu Weng, Xuehai Wu,

Ting Xu, Ning Yang, Zhi Yang, Yu-Feng Zang, Lei Zhang, Qinglin Zhang, Zhe Zhang, Zhiqiang Zhang, Ke Zhao, Zonglei Zhen, Yuan Zhou, Xing-Ting Zhu.

Acknowledgements The authors would like to graciously thank: NIH, NSF, DARPA, IARPA, Johns Hopkins University, Johns Hopkins University Applied Physics Lab, and the Kavli Foundation for their support. Specific information regarding awards can be found at `https://neurodata.io/about`.

References

[1] AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, 29(3):162–173, jun 1996. ISSN 0010-4809. doi: 10.1006/CBMR.1996.0014. URL `http://www.sciencedirect.com/science/article/pii/S0010480996900142`.

[2] Alexandre Abraham, Michael P Milham, Adriana Di Martino, R Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017.

[3] Jesper L R Andersson, Mark Jenkinson, Stephen Smith, and Jesper Andersson. Non-linear registration aka Spatial normalisation. 2007. URL `https://www.fmrib.ox.ac.uk/datasets/techrep/tr07ja2/tr07ja2.pdf`.

[4] Y Behzadi, K Restom, J Liau, and T T Liu. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, aug .

[5] P Bellec, F M Carbonell, V Perlbarg, C Lepage, O Lyttelton, V Fonov, A Janke, J Tohka, and A C Evans. A neuroimaging analysis kit for Matlab and Octave. In *Proceedings of the 17th International Conference on Functional Mapping of the Human Brain*, pages In Press+, 2011.

[6] Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995. ISSN 1522-2594. doi: 10.1002/mrm.1910340409. URL `http://dx.doi.org/10.1002/mrm.1910340409`.

[7] Béla Bollobás, Svante Janson, and Oliver Riordan. Sparse random graphs with clustering. 2009. URL `https://arxiv.org/pdf/0807.2040.pdf`.

[8] Molly G Bright, Christopher R Tench, and Kevin Murphy. Potential pitfalls when denoising resting state fMRI data using nuisance regression. 2016. doi: 10.1016/j.neuroimage.2016.12.027. URL `www.elsevier.com/locate/neuroimage`.

[9] Jesse A Brown and John D Van Horn. Connected brains and minds—the umcd repository for brain connectivity matrices. *Neuroimage*, 124:1238–1241, 2016.

[10] Rastko Ciric, Daniel H. Wolf, Jonathan D. Power, David R. Roalf, Graham L. Baum, Kosha Ruparel, Russell T. Shinohara, Mark A. Elliott, Simon B. Eickhoff, Christos Davatzikos, Ruben C. Gur, Raquel E. Gur, Danielle S. Bassett, and Theodore D. Satterthwaite. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154:174 – 187, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.03.020. URL `http://www.sciencedirect.com/science/article/pii/S1053811917302288`. Cleaning up the fMRI time series: Mitigating noise with advanced acquisition and correction strategies.

[11] Sergi G Costafreda. Pooling FMRI data: meta-analysis, mega-analysis and multi-center studies. *Frontiers in neuroinformatics*, 3:33, 2009. ISSN 1662-5196. doi: 10.3389/neuro.11.033.2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19826498http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2759345`.

[12] R Cameron Craddock, G Andrew James, Paul E Holtzheimer, Xiaoping P Hu, and Helen S Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8): 1914–1928, 2012.

[13] Zaixu Cui, Suyu Zhong, Pengfei Xu, Yong He, and Gaolang Gong. PANDA: a pipeline toolbox for analyzing brain diffusion images. *Frontiers in human neuroscience*, 7:42, 2013. ISSN 1662-5161. doi: 10.3389/fnhum.2013.00042. URL `http://www.ncbi.nlm.nih.gov/pubmed/23439846http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3578208`.

[14] Zaixu Cui et al. PANDA: a pipeline toolbox for analyzing brain diffusion images. *Frontiers in human neuroscience*, 7:42, jan 2013. ISSN 1662-5161. URL `http://journal.frontiersin.org/article/10.3389/fnhum.2013.00042/abstract`.

[15] Alessandro Daducci, Stephan Gerhard, Alessandra Griffa, Alia Lemkaddem, Leila Cammoun, Xavier Gigandet, Reto Meuli, Patric Hagmann, and Jean-Philippe Thiran. The Connectome Mapper: An Open-Source Processing Pipeline to Map Connectomes with MRI. *PLoS ONE*, 7(12):e48121, dec 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0048121. URL `http://dx.plos.org/10.1371/journal.pone.0048121`.

[16] Alessandro Daducci et al. The connectome mapper: an open-source processing pipeline to map connectomes with MRI. *PloS one*, 7(12):e48121, jan 2012. ISSN 1932-6203. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0048121`.

[17] Samir Das, Alex P Zijdenbos, Dario Vins, Jonathan Harlap, and Alan C Evans. Loris: a web-based data management system for multi-center studies. *Frontiers in neuroinformatics*, 5:37, 2012.

[18] Rahul S Desikan et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.01.021.

[19] Oscar Esteban, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Russell A. Poldrack, and Krzysztof J. Gorgolewski. poldracklab/fmriprep: 1.0.0-rc10, November 2017. URL `https://doi.org/10.5281/zenodo.1044752`.

[20] Jean-Philippe Fortin, Elizabeth M Sweeney, John Muschelli, Ciprian M Crainiceanu, Russell T Shinohara, and Alzheimer's Disease Neuroimaging Initiative. Removing inter-subject technical variability in magnetic resonance imaging studies. *Neuroimage*, 132:198–212, May 2016.

[21] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, Robert T Schultz, Ragini Verma, and Russell T Shinohara. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, August 2017.

[22] Jean-Philippe Fortin, Drew Parker, Birkan Tunc, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, et al. Harmonization of multi-site diffusion tensor imaging data. *bioRxiv*, page 116541, 2017.

[23] Karl J. Friston, Steven Williams, Robert Howard, Richard S. J. Frackowiak, and Robert Turner. Movement-related effects in fmri time-series. *Magnetic Resonance in Medicine*, 35(3):346–355, 1996. ISSN 1522-2594. doi: 10.1002/mrm.1910350312. URL `http://dx.doi.org/10.1002/mrm.1910350312`.

[24] Eleftherios Garyfallidis, Matthew Brett, Marta Morgado Correia, Guy B Williams, and Ian Nimmo-Smith. Quickbundles, a method for tractography simplification. *Frontiers in neuroscience*, 6:175, 2012.

[25] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian Nimmo-Smith. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8, 2014.

[26] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, David C Van Essen, and Mark Jenkinson. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80 (Supplement C):105–124, 2013. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.04.127. URL `http://www.sciencedirect.com/science/article/pii/S1053811913005053`.

[27] Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.*, 35(6):498–507, June 2017.

[28] Krzysztof Gorgolewski, Tibor Auer, Vince Calhoun, Cameron Craddock, Samir Das, Eugene Duff, Guillaume Flandin, Satrajit Ghosh, Tristan Glatard, Yaroslav Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments.

[29] Krzysztof J. Gorgolewski, Fidel Alfaro-Almagro, Tibor Auer, Pierre Bellec, Mihai Capotă, M. Mallar Chakravarty, Nathan W. Churchill, Alexander Li Cohen, R. Cameron Craddock, Gabriel A. Devenyi, Anders Eklund, Oscar Esteban, Guillaume Flandin, Satrajit S. Ghosh, J. Swaroop Guntupalli, Mark Jenkinson, Anisha Keshavan, Gregory Kiar, Franziskus Liem, Pradeep Reddy Raamana, David Raffelt, Christopher J. Steele, Pierre-Olivier Quirion, Robert E. Smith, Stephen C. Strother, Gaël Varoquaux, Yida Wang, Tal Yarkoni, and Russell A. Poldrack. Bids apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology*, 13(3):1–16, 03 2017. doi: 10.1371/journal.pcbi.1005209. URL `http://dx.doi.org/10.1371%2Fjournal.pcbi.1005209`.

[30] Günther Grabner, Andrew L. Janke, Marc M. Budge, David Smith, Jens Pruessner, and D. Louis Collins. Symmetric Atlasing and Model Based Segmentation: An Application to the Hippocampus in Older Adults. pages 58–66. Springer, Berlin, Heidelberg, 2006. doi: 10.1007/11866763_8. URL `http://link.springer.com/10.1007/11866763{_}8`.

[31] William R Gray, John A Bogovic, Joshua T Vogelstein, Bennett A Landman, Jerry L Prince, and R Jacob Vogelstein. Magnetic resonance connectome automated pipeline. *IEEE Pulse*, 3(2):42–48, 2011.

[32] William Gray Roncal et al. MIGRAINE: MRI Graph Reliability Analysis and Inference for Connectomics. *Global Conference on Signal and Information Processing*, 2013.

[33] Douglas N. Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, oct 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.06.060. URL `http://www.ncbi.nlm.nih.gov/pubmed/19573611http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2733527http://linkinghub.elsevier.com/retrieve/pii/S1053811909006752`.

[34] M Jenkinson and S Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–56, jun 2001. ISSN 1361-8415. URL `http://www.ncbi.nlm.nih.gov/pubmed/11516708`.

[35] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–41, oct 2002. ISSN 1053-8119. URL `http://www.ncbi.nlm.nih.gov/pubmed/12377157`.

[36] Mark Jenkinson, Mickael Pechaud, and Stephen Smith. BET2 -MR-Based Estimation of Brain, Skull and Scalp Surfaces. 2005. URL `http://mickaelpechaud.free.fr/these/HBM05.pdf`.

[37] Mark Jenkinson et al. FSL. *NeuroImage*, 62(2):782–90, aug 2012. ISSN 1095-9572. URL `http://www.ncbi.nlm.nih.gov/pubmed/21979382`.

[38] Brian Karrer and M E J Newman. Stochastic blockmodels and community structure in networks. URL `https://pdfs.semanticscholar.org/6780/823d309a4a96fd4fcab53544bb8724bf461b.pdf`.

[39] Daniel Kessler et al. Modality-spanning deficits in attention-deficit/hyperactivity disorder in functional networks, gray matter, and white matter. *The Journal of Neuroscience*, 34(50):16555–16566, 2014.

[40] Gary G. Koch. *Intraclass Correlation Coefficient*. John Wiley and Sons, Inc., 2004. ISBN 9780471667193. doi: 10.1002/0471667196.ess1275. URL `http://dx.doi.org/10.1002/0471667196.ess1275`.

[41] JL Lancaster. The Talairach Daemon, a database server for Talairach atlas labels. *NeuroImage*, 1997. ISSN 1053-8119.

[42] Bennett A Landman, Alan J Huang, Aliya Gifford, Deepti S Vikram, Issel Anne L Lim, Jonathan AD Farrell, John A Bogovic, Jun Hua, Min Chen, Samson Jarso, et al. Multi-parametric neuroimaging reproducibility: a 3-t resource study. *Neuroimage*, 54(4):2854–2866, 2011.

[43] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, 11(10):733–739, October 2010.

[44] Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *NeuroImage*, 45(1 Suppl):S187–98, mar 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2008.10.065. URL `http://www.ncbi.nlm.nih.gov/pubmed/19084070http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3318970`.

[45] Nikos Makris, Jill M Goldstein, David Kennedy, Steven M Hodge, Verne S Caviness, Stephen V Faraone, Ming T Tsuang, and Larry J Seidman. Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia research*, 83(2):155–171, 2006.

[46] John Mazziotta et al. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association*, 8(5):401–430, 2001.

[47] Disa Mhembere, William Gray Roncal, Daniel Sussman, Carey E Priebe, Rex Jung, Sephira Ryman, R Jacob Vogelstein, Joshua T Vogelstein, and Randal Burns. Computing scalable multivariate glocal invariants of large (brain-) graphs. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 297–300. IEEE, 2013.

[48] Hengameh Mirzaalian, Lipeng Ning, Peter Savadjiev, Ofer Pasternak, Sylvain Bouix, Oleg Michailovich, Sarina Karmacharya, Gerald Grant, Christine E Marx, Rajendra A Morey, et al. Multi-site harmonization of diffusion mri data in a registration framework. *Brain Imaging and Behavior*, pages 1–12, 2017.

[49] Susumu Mori et al. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of neurology*, 45(2):265–269, 1999.

[50] Jared A Nielsen, Brandon A Zielinski, P Thomas Fletcher, Andrew L Alexander, Nicholas Lange, Erin D Bigler, Janet E Lainhart, and Jeffrey S Anderson. Multisite functional connectivity mri classification of autism: Abide results. 2013.

[51] Stephanie Noble, Marisa N Spann, Fuyuze Tokoglu, Xilin Shen, R Todd Constable, and Dustin Scheinost. Influences on the Test–Retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cereb. Cortex*, pages 1–15, 12 September 2017.

[52] Kate Brody Nooner, Stanley Colcombe, Russell Tobe, Maarten Mennes, Melissa Benedict, Alexis Moreno, Laura Panek, Shaquanna Brown, Stephen Zavitz, Qingyang Li, et al. The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, 6:152, 2012.

[53] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, January 2016.

[54] Kenichi Oishi et al. *MRI atlas of human white matter*. Academic Press, 2010.

[55] Emanuele Olivetti, Susanne Greiner, and Paolo Avesani. ADHD diagnosis from multiple data sources with batch effects. *Front. Syst. Neurosci.*, 6:70, January 2012.

[56] Emanuele Olivetti, Susanne Greiner, and Paolo Avesani. Adhd diagnosis from multiple data sources with batch effects. *Frontiers in systems neuroscience*, 6:70, 2012.

[57] Julia P Owen, Etay Ziv, Polina Bukshpun, Nicholas Pojman, Mari Wakahiro, Jeffrey I Berman, Timothy PL Roberts, Eric J Friedman, Elliott H Sherr, and Pratik Mukherjee. Test–retest reliability of computational network measurements derived from the structural connectome of the human brain. *Brain connectivity*, 3(2):160–176, 2013.

[58] Dmitry Petrov, Alexander Ivanov, Joshua Faskowitz, Boris Gutman, Daniel Moyer, Julio Villalon, Neda Jahanshad, and Paul Thompson. Evaluating 35 methods to generate structural connectomes using pairwise classification. *arXiv preprint arXiv:1706.06031*, 2017.

[59] Russell A. Poldrack. Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, 2 (1):67–70, mar 2007. ISSN 1749-5016. doi: 10.1093/scan/nsm006. URL `https://academic.oup.com/scan/article-lookup/doi/10.1093/scan/nsm006`.

[60] Simon Rosenfeld. Do DNA microarrays tell the story of gene expression? *Gene Regul. Syst. Bio.*, 4:61–73, June 2010.

[61] Tarek Sherif, Pierre Rioux, Marc-Etienne Rousseau, Nicolas Kassis, Natacha Beck, Reza Adalat, Samir Das, Tristan Glatard, and Alan C Evans. Cbrain: a web-based, distributed computing platform for collaborative neuroimaging research. *Recent Advances and the Future Generation of Neuroinformatics Infrastructure*, page 102, 2015.

[62] Anne M Smith, Bobbi K Lewis, Urs E Ruttimann, Frank Q Ye, Teresa M Sinnwell, Yihong Yang, Jeff H Duyn, and Joseph A Frank. Investigation of Low Frequency Drift in fMRI Signal. 1999. doi: 10.1006/nimg.1999.0435.

[63] Stephen M Smith et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23 Suppl 1:S208–19, jan 2004. ISSN 1053-8119. URL `http://www.ncbi.nlm.nih.gov/pubmed/15501092`.

[64] Stamatios N. Sotiropoulos, Saad Jbabdi, Junqian Xu, Jesper L. Andersson, Steen Moeller, Edward J. Auerbach, Matthew F. Glasser, Moises Hernandez, Guillermo Sapiro, Mark Jenkinson, David A. Feinberg, Essa Yacoub, Christophe Lenglet, David C. Ven Essen, Kamil Ugurbil, Timothy EJ Behrens, and for the WU-Minn HCP Consortium. Advances in diffusion mri acquisition and processing in the human connectome project. *Neuroimage*, 80:125–143, Oct 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2013.05.057. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3720790/`. 23702418[pmid].

[65] Chandra S Sripada et al. Lag in maturation of the brain's intrinsic functional architecture in attention-deficit/hyperactivity disorder. *Proceedings of the National Academy of Sciences*, 111(39):14259–14264, 2014.

[66] Jody Tanabe, David Miller, Jason Tregellas, Robert Freedman, and Francois G Meyer. Comparison of Detrending Methods for Optimal fMRI Preprocessing. doi: 10.1006/nimg.2002.1053. URL `http://ecee.colorado.edu/{~}fmeyer/pub/tanabe-meyer2002.pdf`.

[67] J-Donald Tournier, Fernando Calamante, David G Gadian, and Alan Connelly. Direct estimation of the fiber orientation density function from diffusion-weighted mri data using spherical deconvolution. *NeuroImage*, 23(3):1176–1185, 2004.

[68] David S Tuch, Timothy G Reese, Mette R Wiegell, and Van J Wedeen. Diffusion mri of complex neural architecture. *Neuron*, 40(5):885–895, 2003.

[69] Nathalie Tzourio-Mazoyer et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[70] Koene R A Van Dijk, Mert R Sabuncu, and Randy L Buckner. The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage*, 59(1):431–8, jan 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.07.044. URL `http://www.ncbi.nlm.nih.gov/pubmed/21810475http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3683830`.

[71] Gaël Varoquaux and R Cameron Craddock. Learning and comparing functional connectomes across subjects. *NeuroImage*, 80:405–415, 2013.

[72] Joshua T Vogelstein, William Gray Roncal, R Jacob Vogelstein, and Carey E Priebe. Graph classification using signal-subgraphs: Applications in statistical connectomics. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1539–1551, 2013.

[73] Shangsi Wang, Zhi Yang, Michael Milham, Cameron Craddock, Xi-Nian Zuo, Carey E. Priebe, and Joshua T. Vogelstein. Optimal experimental design for generating reference connectome datasets. In *Organization for Human Brain Mapping, 21st Annual Meeting*, 2015.

[74] Mark W. Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M. Smith. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45(1):S173–S186, 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2008.10.055. URL `http://www.sciencedirect.com/science/article/pii/S1053811908012044`.

[75] Mark W Woolrich et al. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45(1 Suppl):S173–86, mar 2009. ISSN 1095-9572. URL `http://www.sciencedirect.com/science/article/pii/S1053811908012044`.

[76] Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, September 2013.

[77] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, jan 2001. ISSN 02780062. doi: 10.1109/42.906424. URL `http://www.ncbi.nlm.nih.gov/pubmed/11293691http://ieeexplore.ieee.org/document/906424/`.

[78] Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John CS Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1:140049, 2014.

Appendix A Pipeline Comparison Technology Evaluation

The principles (the columns of our chart) of a connectome acquisition pipeline were evaluated as follows:

- Accuracy: whether the pipeline produces QAX figures for each step of processing.
- Reliability: whether the pipeline has been quantitatively vetted using the discriminability [73] or the intraclass-correlation [40].
- Robust: whether the pipeline has been evaluated on M3R from feature-rich datasets collected from many sites.
- Expedience: whether the pipeline runs in < 1 hour on a typical computer (4 logical cores @ 3.3 GHz, 32 GB RAM).
- End-to-End: whether the pipeline can produce human connectomes from raw M3R data.
- Scalability: whether the pipeline can be deployed to parallelize locally (through multi-threading) or in the cloud (AWS EC2, AWS Batch).
- Portability: whether the pipeline is made available through the use of scientific containers (docker, singularity) or processing services (openneuro, cbrain).
- Turn-key: whether the pipeline is pre-optimized to run an optimal processing pipeline without user algorithm selection.
- Openness was evaluated as whether users could acquire code or pre-processed derivatives associated with the pipeline.

And the cells receiving a non-✓ were evaluated as follows:

- The CPAC pipeline for fMRI [12] is not end-to-end: the closest derivative to a functional connectome are the parcellated timeseries, resulting in a ✓as the users are only required to perform pairwise correlations of the timeseries. Also, the CPAC pipeline requires users to select the algorithms to run rather than having a pre-optimized choice; it is not turn-key and receives a ✗.
- The HCP Pipelines for minimally preprocessing fMRI [26] are not end-to-end (its farthest processing step is a preprocessed voxelwise timeseries), resulting in a ✗. Additionally, the pipelines do not provide any of the criterion for scalability nor portability, resulting in ✗ for both criterion.
- FMRIPREP [19] takes several hours per subject on a standard computer, resulting in a ✗ for expedience. The pipeline outputs only preprocessed voxelwise timeseries, resulting in a ✗ for end-to-end.
- The NIAK Pipeline [5] is not end-to-end (its farthest processing step is a preprocessed voxelwise time-series), resulting in a ✗ . The NIAK pipeline cannot leverage cloud infrastructure for scaling, resulting in a ✓ for scalability. The NIAK pipeline is not available on neither openneuro nor cbrain, resulting in a ✓ for portability.
- The PANDA Pipeline [13] has not been used nor quantitatively evaluated on a wide-variety of datasets, indicating it has not been vetted for reliability nor robustness, resulting in ✗. The PANDA pipeline has no extensions for cloud infrastructure, resulting in a ✓ for scalability. The PANDA pipeline is not available on traditional container services nor processing services, resulting in a ✗ for portability. The pipeline provides many possible configurations, resulting in a ✗ for turn-key. Finally, the code is not open source, and pre-computed derivatives are not organized anywhere, resulting in a ✗ for openness.

- The CMTK Pipeline [15] has not been checked for reliability nor robustness across a wide variety of datasets, resulting in ✗. The pipeline allows for probabilistic tractography, which can take several hours on a standard computer, resulting in a ✓ for expedience. The CMTK pipeline cannot be used on cloud infrastructure, resulting in a ✓ for scalability. The CMTK pipeline is not available in scientific containers nor pipeline processing services.The pipeline provides many possible configurations, resulting in a ✗ for turn-key. The CMTK pipeline does not provide a collection of pre-computed derivatives, resulting in a ✓ for openness.
- All the pipelines investigated were for a single modality. NDMG is the only multimodal pipeline analyzed.

Appendix B Diffusion Pipeline

Here we take a deep-dive into each of the modules of the NDMG-d pipeline. We will explain algorithm and parameter choices that were implemented at each step and the justification for why they were used over alternatives.

Appendix B.1 Registration

Registration in NDMG leverages FSL and the Nilearn Python package. The primary concern in development of NDMG was the discriminability and robustness of each step. Additionally, we wanted the pipeline to run on non-

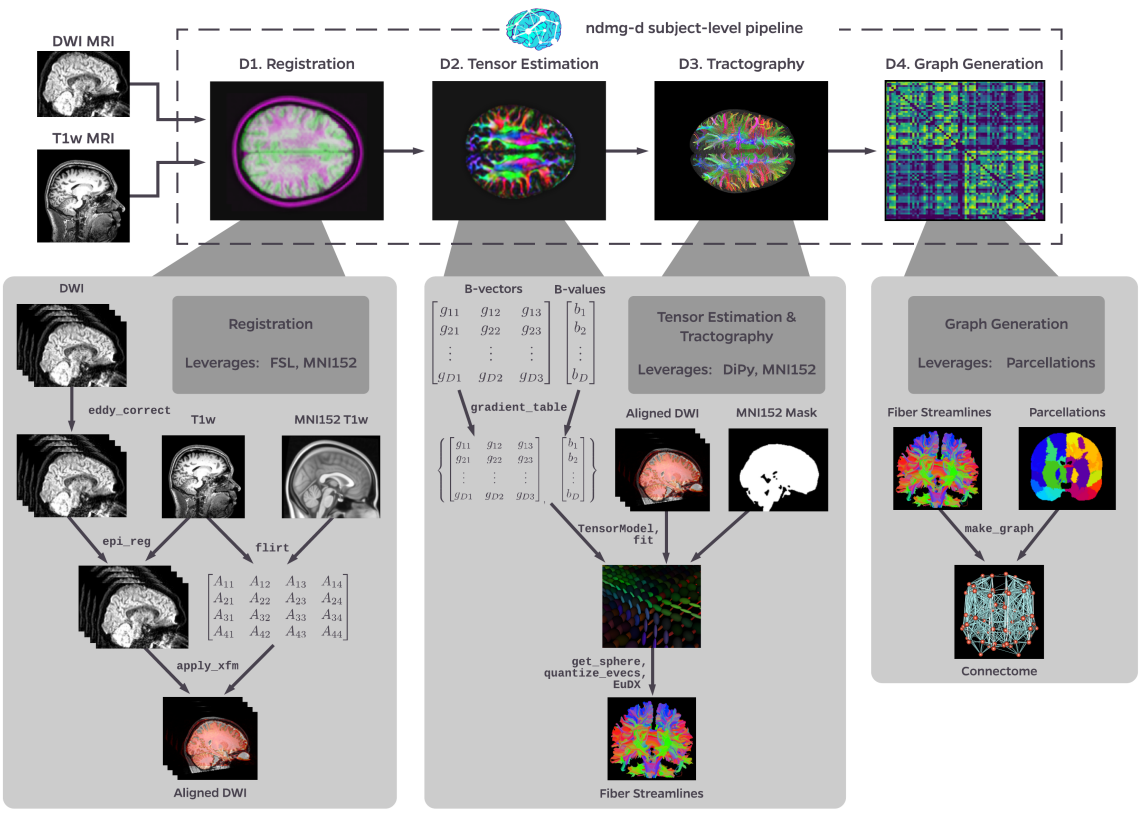


Figure 6: NDMG-d **detailed pipeline**. The NDMG-d pipeline consists of 4 main steps: Registration (**D1**), Tensor Estimation (**D2**), Tractography (**D3**), and Graph Generation (**D4**). Each of these sections leverages pubicly available tools and data to robustly produce the desired derivative of each step. Alongside derivative production, NDMG produces QA figures at each stage, as can be seen in **D1-4**, that enable qualitative evaluation of the pipeline's performance.

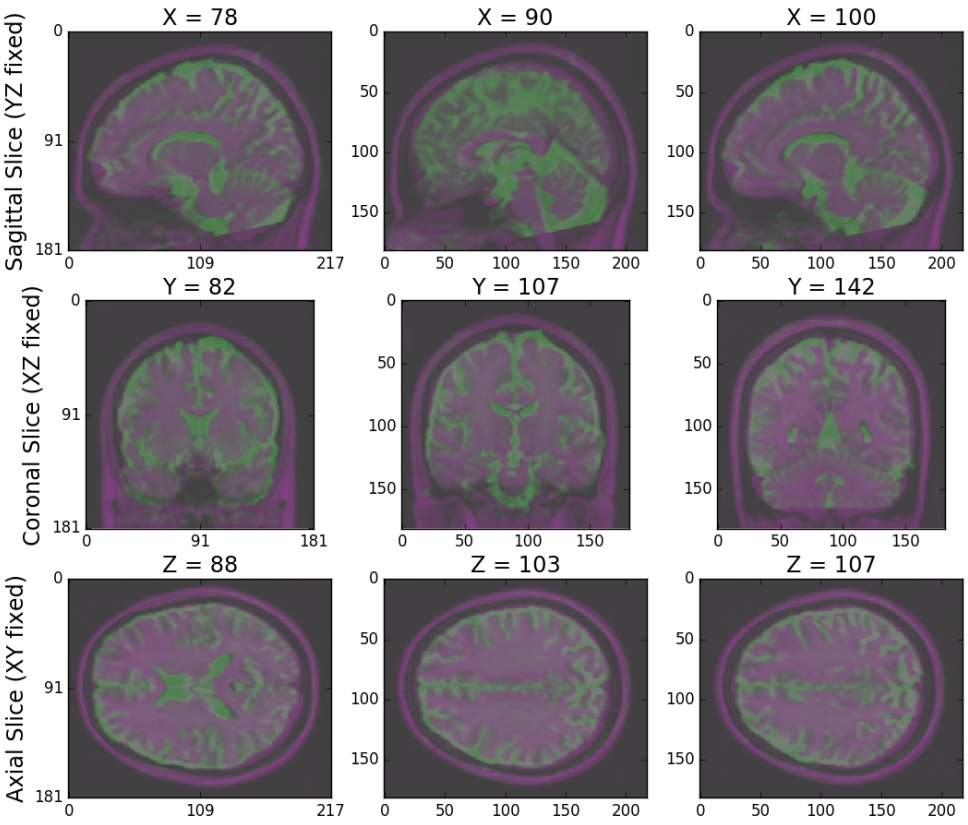


Figure 7: NDMG-d **Registration QAX**. NDMG-d produces registration QAX showing the zeroth slice of the DWI sequence in green overlaid on the template brain in purple.

specialized hardware in a timeframe that didn't significantly hinder the rate of progress of scientists who wish to use it. As such, NDMG uses linear registrations; non-linear methods had higher variability across datasets and increased the resource and time requirements of the pipeline (not shown).

As is seen in Figure 9B1, the first step in the registration module is eddy-current correction and DWI self-alignment to the volume-stack's B0 volume. FSL's `eddy_correct` was used to accomplish this. The `eddy_correct` function was chosen over the newer `eddy` function. The `eddy` function, while providing more sophisticated de-noising, takes significantly longer to run or relies on GPU acceleration, which would reduce the accessibility of NDMG.

Once the DWI data is self-aligned, it is aligned to the same-subject T1w image through FSL's `epi_reg` mini-pipeline. This tool performs a linear alignment between each image in the DWI volume-stack and the T1w volume.

The T1w volume is then aligned to the MNI152 template using linear registration computed by FSL's `flirt`. This alignment is computed using the 1 millimeter (mm) MNI152 atlas, to enable higher freedom in terms of the parcellations that may be used, such as near-voxelwise parcellations that have been generated at 1 mm. FSL's non-linear registration, `fnirt`, is not used in NDMG as the performance was found to vary significantly based on the collection protocol of the T1w images, often resulting in either slightly improved or significantly deteriorated performance.

The transform mapping the T1w volume to the template is then applied to the DWI image stack, resulting in the DWI image being aligned to the MNI152 template in *stereotaxic*-coordinate space. However, while `flirt` aligns the images in stereotaxic space, it does not guarantee an overlap of the data in voxelspace. Using Nilearn's `resample`, NDMG ensures that images are aligned in both voxel- and stereotaxic-coordinates so that all analyses can be performed equivalently either with or without considering the image affine-transforms mapping the data matrix to the real-world coordinates.

Finally, NDMG produces a QA plot showing three slices of the first B0 volume of the aligned DWI image overlaid on the MNI152 template in the three principle coordinate planes. This provides nine plots in total which enable qualitative assessment of the quality of alignment.

Appendix B.2 Tensor Estimation

Once the DWI volumes have been aligned to the template, NDMG begins diffusion-specific processing on the data. All diffusion processing in NDMG is performed using the Dipy Python package [25]. The diffusion processing in NDMG is performed after alignment to facilitate cross-connectome comparisons.

While high-dimensional diffusion models, such as orientation distribution functions (ODFs) or q-ball, enable reconstruction of crossing fibers and complex fiber trajectories, these methods are designed for images with a large number of diffusion volumes/directions for a given image [67; 68]. Because NDMG is designed to be robust across a wide range of DWI datasets, including diffusion tensor imaging, NDMG uses a lower-dimensional tensor model. The model, described in detail on Dipy's website[1], computes a 6-component tensor for each voxel in the image. This reduces the DWI image stack to a single 6-dimensional image that can be used for tractography. Once tensor estimation has been completed, NDMG generates a QA plot showing slices of the FA map derived from the tensors in nine panels, as above.

Appendix B.3 Tractography

In keeping with the theme of computationally efficient and robust methods, NDMG uses DiPy's deterministic tractography algorithm, `EuDX` [24]. Integration of tensor estimation and tractography methods is minimally complex with this tractography method, as it has been designed to operate on the tensors produced by Dipy in the previous step. Probabilistic tractography would be significantly more computationally expensive, and it remains unclear how well it would perform on data with a small number of diffusion directions. A subset of the resolved streamlines are visualized in an axial projection of the brain mask with the fibers contained. This allows

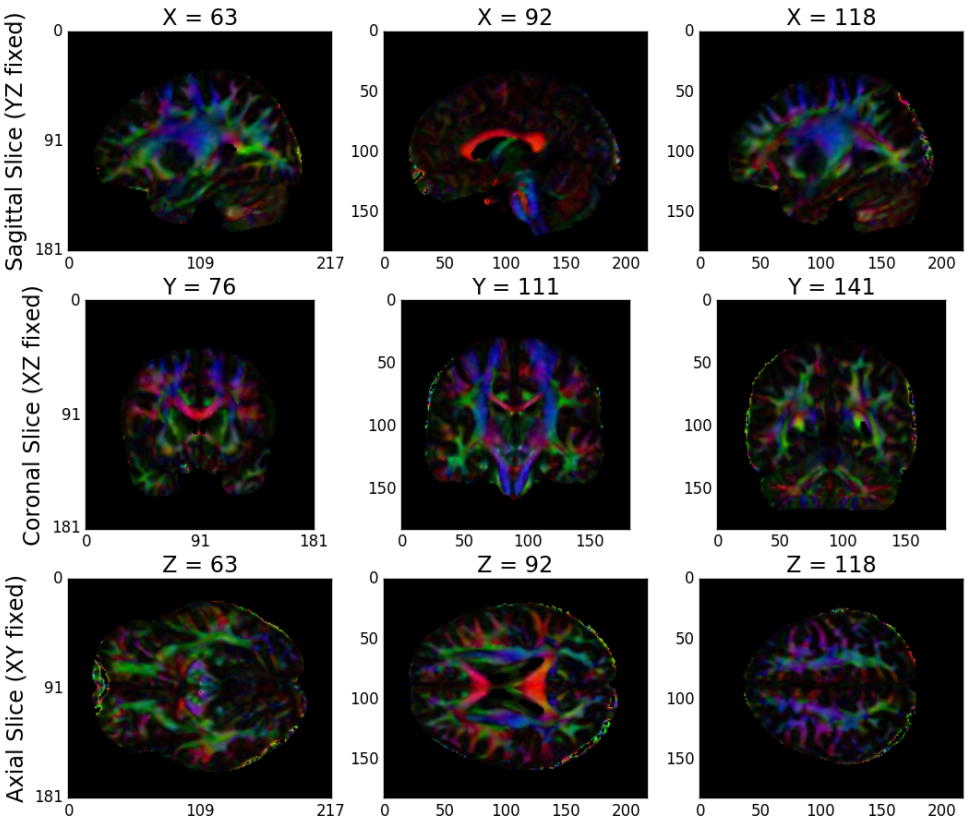


Figure 8: NDMG-d **Tensor Estimation QAX**. NDMG-d produces tensor QAX showing the voxelwise determininstic tensor model fit to the aligned DWI sequence.

the user to verify, for example, that streamlines are following expected patterns within the brain and do not leave the boundary of the mask.

Appendix B.4 Graph Estimation

NDMG uses the fiber streamlines to generate connectomes across multiple parcellations. The connectomes generated are graph objects, with nodes in the graph representing regions of interest (ROIs), and edges representing connectivity via fibers. An undirected edge is added to the graph for each pair of ROIs a given streamline passes through. Edges are undirected because DWI data lacks direction information. Edge weight is the number of streamlines which pass through a given pair of regions. NDMG uses 24 parcellations, including all standard public DWI parcellations known by the authors. Users may run NDMG using any additional parcellation defined in MNI152 space simply by providing access to it on the command-line. To package an additional parcellation with NDMG, please contact the maintainers.

Appendix C Functional Pipeline

Here we take a deep-dive into each of the modules of the NDMG-f pipeline. We will explain algorithm and parameter choices that were implemented at each step, and the justification for why they were used over alternatives.

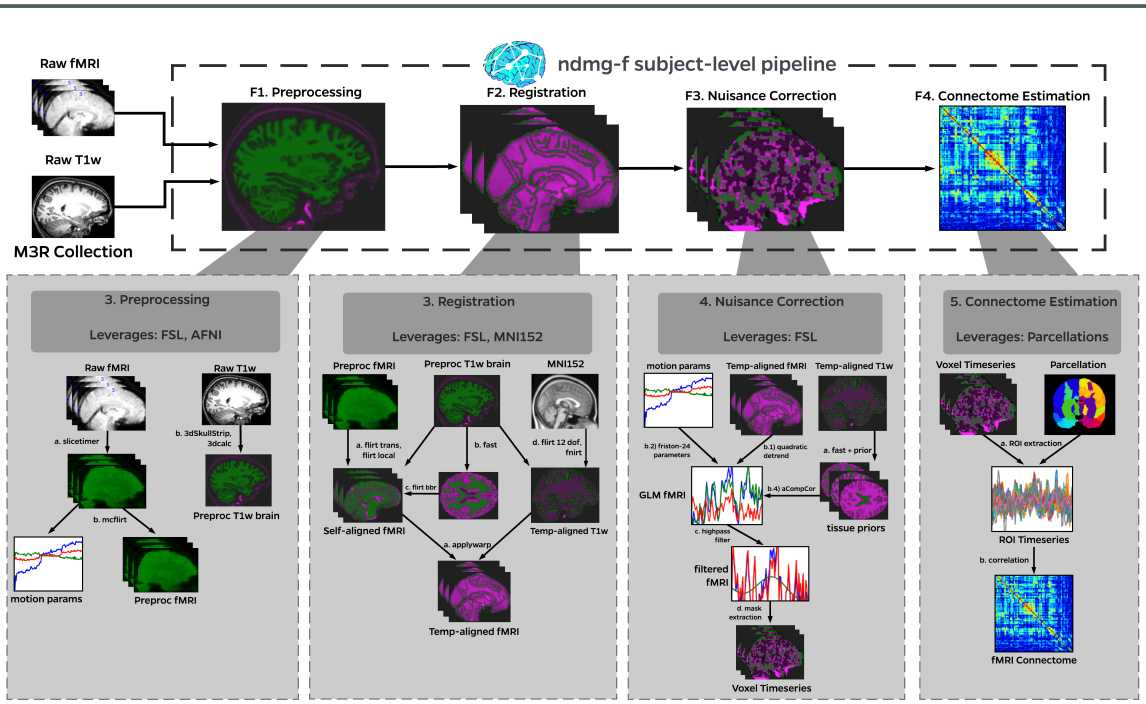


Figure 9: NDMG-f **detailed pipeline**. The NDMG-f pipeline consists of 4 main steps: Preprocessing (**F1**), Registration (**F2**), Nuisance Correction (**F3**), and Graph Generation (**F4**). Each of these sections leverages pubicly available tools and data to robustly produce the desired derivative of each step. Alongside derivative production, NDMG-f produces QA figures at each stage, as can be seen in **F1-4**, that enable qualitative evaluation of the pipeline's performance.

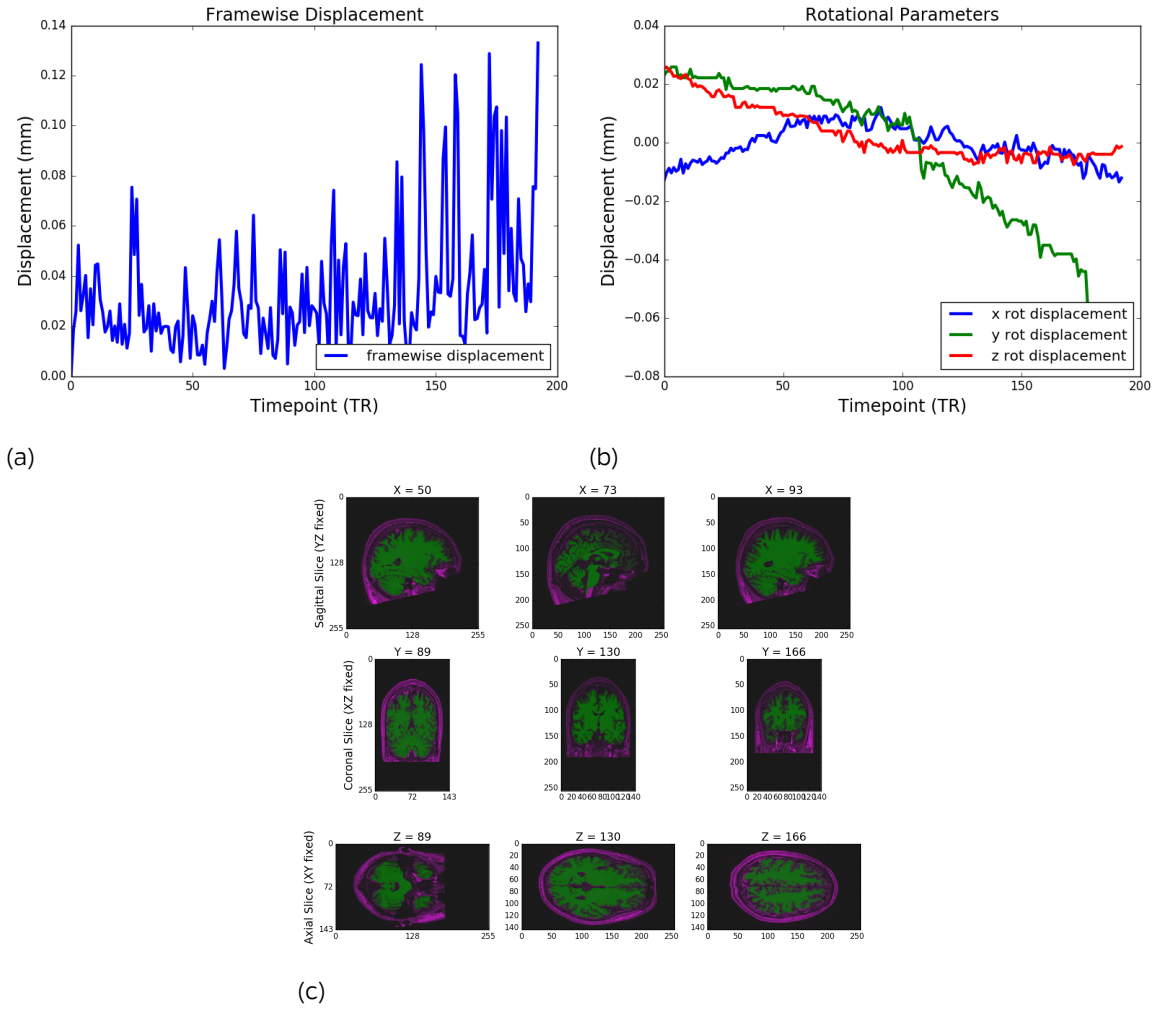


Figure 10: NDMG-f **Preprocessing QAX**. For preprocessing, NDMG-f produces QAX figures showing the framewise displacement per timestep (10a), the rotational (10b) and translational motion parameters, plots of the raw and corrected brain, and the success of the brain extraction process (10c).

Appendix C.1 Preprocessing

Slice Timing Correction To collect an individual 4D EPI sequence, a 3D volume is constructed as a combination of individual 2D slices. The 2D slices are collected incrementally; that is, we collect each 2D slice for approximately 10 milliseconds, and the entire 3D volume is complete in about 30 2D slices. This gives a repetition time, or TR, for the volume of 1 to 3 seconds, depending on the scanner. In response to a stimulus, we expect a typical brain response on the order of 16 seconds. This means that during the course of a single volume being collected, a different amount of BOLD may be present present in the first slice than when the last slice is actually collected [74], yet the results are recorded as a single timepoint. The data are essentially a "sliding snapshot" over the course of one TR; observations are not all at a fixed point in time.

NDMG-f accounts for slice timing aliasing by accepting a user-provided acquisition sequence (the order in which slices are collected for a single time point). Given the acquisition sequence of the 2D slices, we can compute the TR shift of each 2D slice using the TR information from the header of the brain image. A slice that occurs first in a TR will have a shift of 0, while a slice that occurs at the end of a TR will have a shift of 1. A slice that occurs exactly in the middle of a TR has a shift of 0.5. For each voxel in an individual slice, interpolation is used to re-center our observations to all have a TR shift of 0.5. Slicetiming is accompmlished using the slicetimer utility provided by FSL [74].

Motion Correction During an fMRI session, participants sit in a small, cramped scanner, often for 5 to 10 minutes. During the course of a study, it is fairly common for participants to move their heads, even if only small amounts. Small shifts will lead to a person's head being in different spatial positions at each timestep [23], which hampers our efforts to standardize the spatial properties of each subject's brain down the line through registration. This is because registrations are performed by estimating the registration on the first volume [37; 63; 75], after which the estimated transformation is applied across the temporal dimension. This means that if each 3D volume is not aligned spatially, inconsistencies in registration will generally decrease functional connectome quality [70].

Fortunately, given that the subject's brain is the same scaling for each 3D volume (the brain shape itself is not changing in time), a 6 degree of freedom rigid affine transformation for each 3D volume (1 translational and 1 rotational parameter per x, y, z direction the subject could move his/her head) using the mean fMRI slice as the reference. Motion correction is implemented using the mcflirt utility [35], which is a simplification of FSL's FLIRT registration tool.

Anatomical Preprocessing To preprocess the anatomical t1w image, AFNI's 3dSkullstrip [1] is used. 3dSkullstrip provides modifications to the BET algorithm [36] to make it more robust without hyperparameters. Note that 3dSkullstrip renormalizes intensities, so to regain the original intensities, the result is binarized and fed as a step function (essentially making it a mask) through 3dcalc and multiplied voxelwise with the original image, yielding the original image intensities of the brain and excluding the regions determined to be skull.

Appendix C.2 Registration

Self Registration To register our input fMRI to our reference T1w image, a 3 degree of freedom (DOF) affine transformation is estimated with x, y, and z translational parameters with FSL's FLIRT [34] using the $sch3Dtrans_3dof$ schedule file provided as part of the FSL package. This centers the functional brain on the T1w brain optimally and serves to improve the initialization of registrations in later steps. Next, a locally-optimized transformation from the fMRI brain to the T1w brain is estimated. Again, this transformation is heavily robust, and has its hyperparameters tuned to focus on local features of the input (fMRI) and reference (T1w) spaces in the $simple3d.sch$ FLIRT schedule file. This schedule file is chosen due to the input fMRI potentially having a narrow field of view, resolution constraints, or tearing that will perform poorly using a more global alignment.

Next, a third alignment is estimated using the locally-aligned fMRI to the structural T1w image using the bbr cost function provided by FSL [33]. In functional MRI, the white-matter/gray-matter border is fairly apparent as

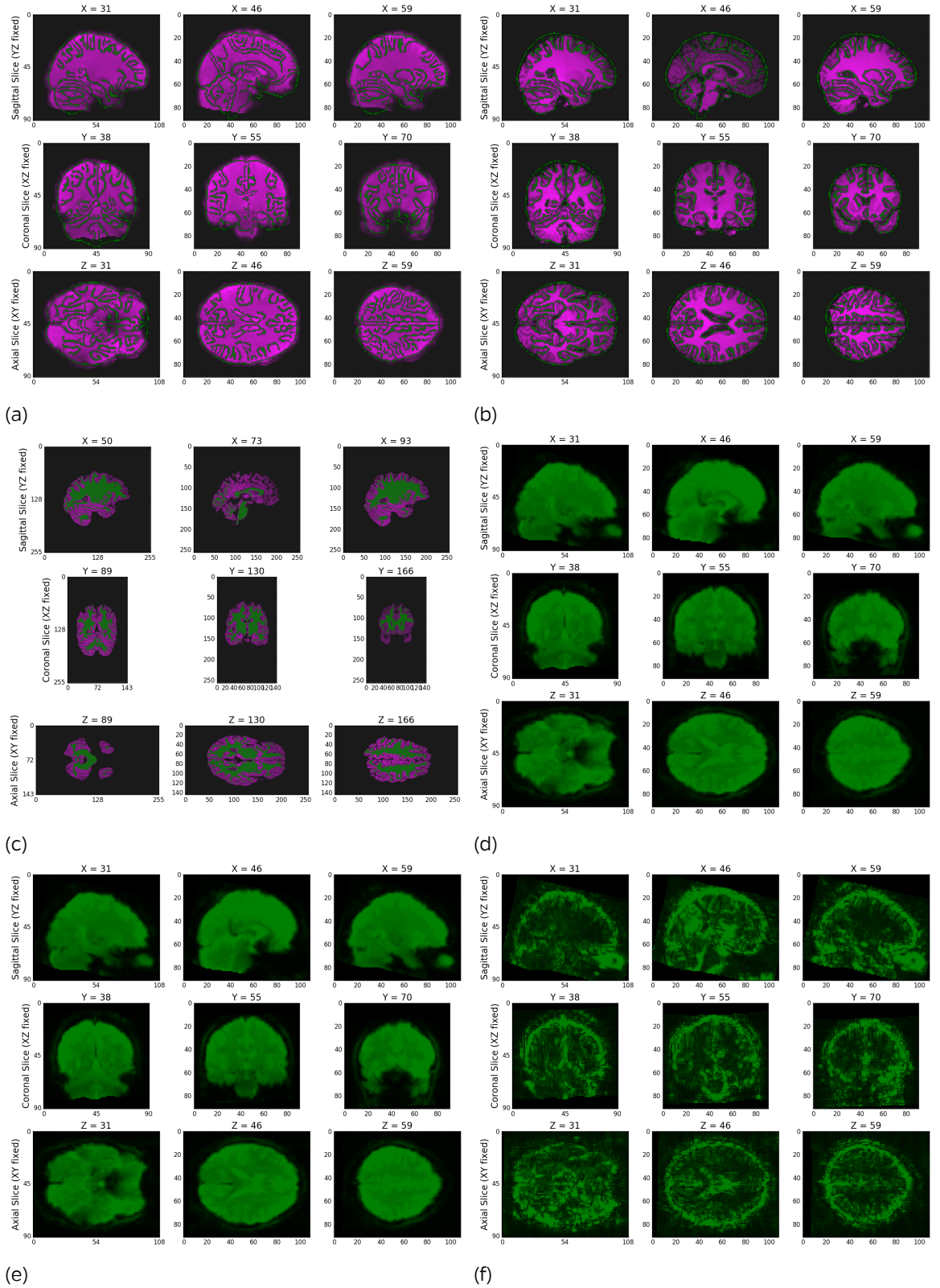


Figure 11: NDMG-f **Registration QAX**. For registration, NDMG-f produces summary figures showing the preprocessed epi image overlaid on the t1w, the registered epi image overlaid on the template (11a), the registered t1w image overlaid on the template (11b), the white-matter mask used in FLIRT-bbr (11c), the voxelwise mean intensity (11d), the voxelwise signal-to-noise ratio (11e), and the voxelwise contrast-to-noise ratio (11f).

the gray-matter generally shows higher intensities than the white-matter regions. Leveraging this observation, the white-matter boundary can be aligned between the fMRI and the T1w scan with high accuracy [33]. A 6 DOF transformation from the fMRI to the T1w image is estimated, and the T1w is then segmented to produce a white-matter mask using FSL's FAST algorithm [77]. FLIRT is used with the bbr (boundary-based registration) cost-function to align the boundaries of the white-matter in the fMRI and T1w scans optimally. This provides a high-quality alignment for intra-modal registration from an EPI to a T1w image.

Template Registration A template brain represents the anatomical average brain of the sampling of subjects it is collected over. This anatomically average brain theoretically represents the average brain we will find during our external investigations, allowing minimal alignment on average. For FNGS, we assume that users will be using the MNI152 template [30].

A gentle linear transformation of the T1w brain to the template atlas is estimated using the local-optimisation schedule file from before. We use this local-optimisation registration as the starting point for a more extensive 12-DOF global FLIRT alignment than the self-registration case. Given that the template brain will theoretically be less similar than simple translations, rotations and scalings can provide, a non-linear registration is estimated from the T1w to the template space. This is accomplished using FSL's FNIRT algorithm [3], with hyper-parameter tuning specific for the MNI152 template. This non-linear transformation is applied first to the T1w image. The non-linear transformation is then combined with the result of the self-alignment step and applied to the functional volume. Applying the transformation only once prevents unnecessary fixed-precision multiplies, as each application of a transformation will lead to a loss of information due to the fixed number of bits each voxel intensity is stored as.

Appendix C.3 Nuisance Correction

General Linear Model Over the course of an fMRI scanning session, many sources of noise arise that must be corrected for in order to make quality data inferences. The scanner heats up during a scanning session (producing a high strength magnetic field for sessions lasting up to ten or more minutes, which in turn produces an enormous amount of heat). As the scanner heats, the signal recorded tends to drift (first demonstrated by [62] who showed that a heated scanner detected "brain activity" in cadavers). This drift has been shown to be approximately quadratic [66], so a second-degree polynomial regressor is created.

While spatial motion correction removes the visual impact of head motion, spurious signal artifacts remain present. These artifacts can be characterized by the position of the brain in the scanner and the prior positions of the brain in the scanner, as first shown by Friston et al. [23]. This history relationship can be effectively captured by the current volume and the preceding volume, as well as their squares, so 24 regressors are estimated where we have 4 regressors (1 current frame, 1 shifted frame, 1 squared-current frame, 1 square-shifted frame) for each of our 6 (x, y, z translation and rotation) motion regressors. These regressors are known as the Friston 24 parameter regressors.

Finally, it has been shown by [4] that the fMRI signal is corrupted by physiological noise, from physiological functions such as blood flow or vessel dilation. The physiological confounds present in our functional data have been shown to be effectively captured by the top 5 principal components from the white-matter and lateral-ventricle signal [4], [10]. We estimate csf and white-matter masks using the FAST algorithm, [77] with priors obtained from the MNI152 parcellation [46]. This estimated mask is eroded by 2 voxels on all sides to avoid any potential signal distortion from the gray-matter signal, since gray-matter signal is expected to correlate with our stimuli. Any signal bleeding into the white-matter voxels (since the gray-matter/white-matter boundary has a slight bleed-over region) that could get removed by our PCA would be disastrous on our downstream inferences.

The regressors are incorporated into the design matrix X of the general linear model (GLM) shown in (2). For our n voxels, the t timestep BOLD signal, we can decompose $T_{raw} \in \mathbb{R}^{t \times n}$ as:

$$T_{raw} = X\beta + \epsilon \tag{2}$$

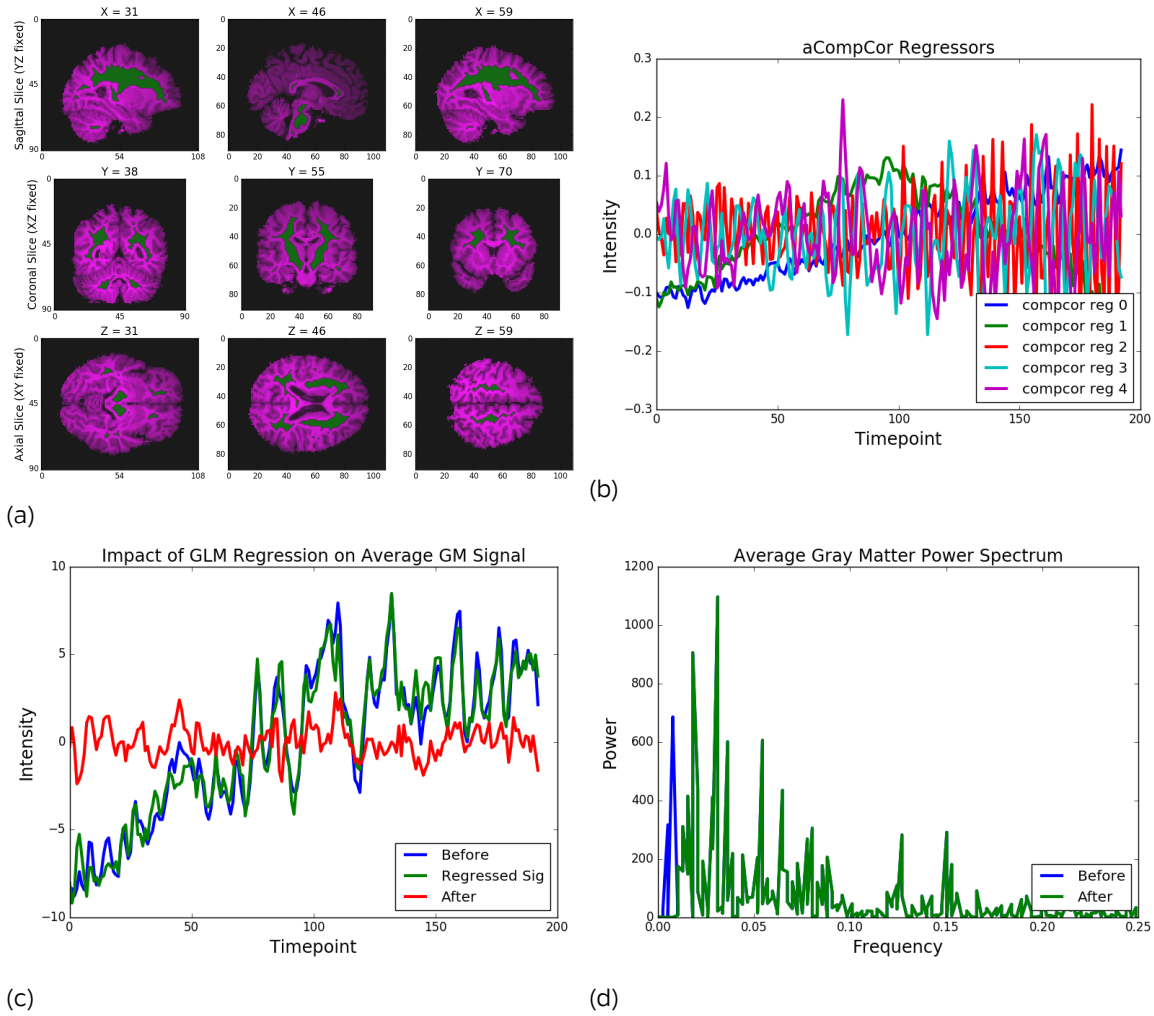


Figure 12: NDMG-f **Nuisance Correction QAX**. For the nuisance correction step, the NDMG-f pipeline produces quality control figures focusing on the regressors being accounted for. We visualize the white matter, gray matter, cerebro-spinal fluid, and eroded white matter mask (shown in 12a). For the General Linear Model, we look at the regressors estimated from the eroded white matter and cerebro-spinal timemseries by aCompCor (shown in 12b), the regressors estimated by the Friston 24 parameter model, and the average gray matter signal before and after regression (shown in 12c). For frequency filtering, we show the average gray matter signal before and after frequency filtering, and the average gray matter power spectrum before and after filtering (shown in 12d).

where $\epsilon \in \mathbb{R}^{t \times n}$ is the timeseries that cannot be accounted for in the regressor design matrix $X \in \mathbb{R}^{t \times r}$ with coefficients $\beta \in \mathbb{R}^{r \times n}$. Minimizing the squared-error loss of T with respect to $X\beta$ will provide a best-estimate of the coefficients $\hat{\beta}$ of the regressors in X, where the desired, regressor-corrected timeseries is just ϵ since we know that ϵ are the components of our raw signal that cannot be fit by our design regressors [66]. This can solve using the least-squares solution to minimize the squared-error loss function:

$$\hat{\beta} = (X^T X)^{-1} X^T T_{raw}$$

Using the estimate $\hat{\beta}$ for our regressors X, the GLM-corrected timeseries is:

$$\epsilon = T_{raw} - X\hat{\beta}$$

Low Frequency Drift Removal Using the GLM-corrected timeseries, low-frequency drift that may still be present in the functional volume can then be removed . Any physiological response due to a stimulus will have a period of around 16 seconds (or a frequency of 0.063 Hz) and will not exceed the period of any stimuli present, as has been shown in [44]. Using this information, sinusoidal fourier modes with frequencies lower than most brain stimuli are removed. Conservatively, we set a threshold of 0.01 Hz for highpass-filtering out low-frequency noise (this should not remove task-dependent signal as long as our task has a period less than about 100 seconds).

T1 Effect Removal During the fMRI session, the first few volumes may appear to have brighter intensities as the T1 effects are not fully saturated. External attempts to remove the T1 effect include component-correction and global mean normalization of each slice, both of which have been shown to potentially remove brain signal [8]. To account for this, the first 15 seconds of the fMRI sequence are discarded, which tends to account for the majority of contrast-dependent issues.

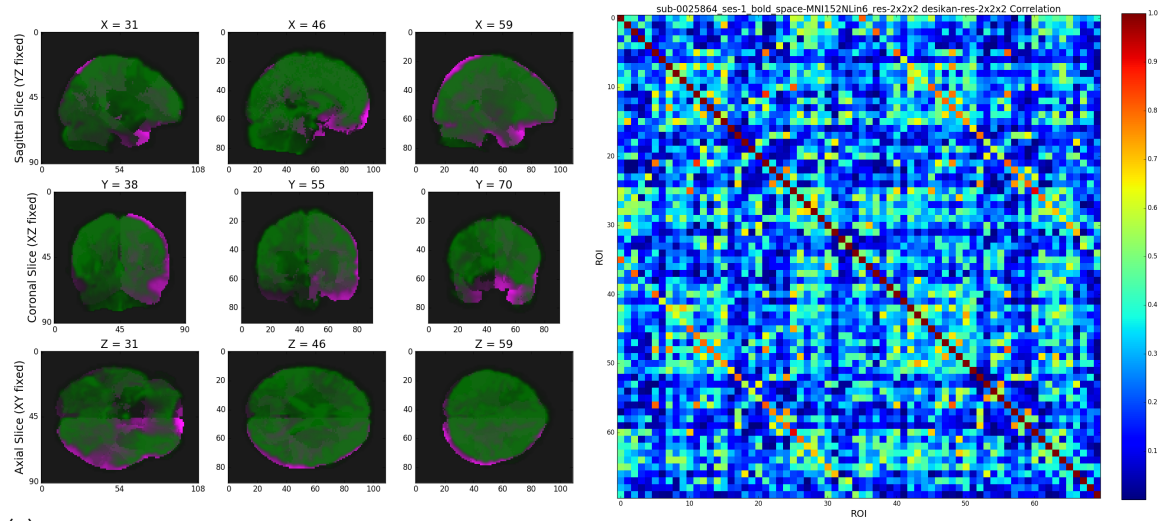
Appendix C.4 Graph Estimation

It is often of neurological significance to think of the brain as segmented into regions of neurons performing a similar task based on their spatial layout. To accomplish this, neuroscientists parcellate the registered brains into Regions of Interest (ROIs), whereby individual voxels are labelled on the template atlas. The voxels parcellated into a given are averaged spatially to yield an ROI timeseries. This gives us a significantly downsampled representation of the brain, often yielding in excess of 100 fold compression and also reducing voxel-by-voxel noise that may be present. The downsampled timeseries can then be analyzed using traditional statistical methods with greater robustness and computational efficiency [59].

Finally, the ROI timeseries can be used to estimate a functional connectome. For our connectome, we want each edge to represent a relationship between two regions of the brain. Intuitively, we can use our functional timeseries to show similarity between regions, whereby the simplest comparison is in terms of the correlation of functional activity. Performing the pairwise correlation between each pair of regions yields us a simplified graphical representation of our original brain. We then rank the edges in our graph from lowest weight to largest, and define the ranks as the edge-weights of our functional connectome [73].

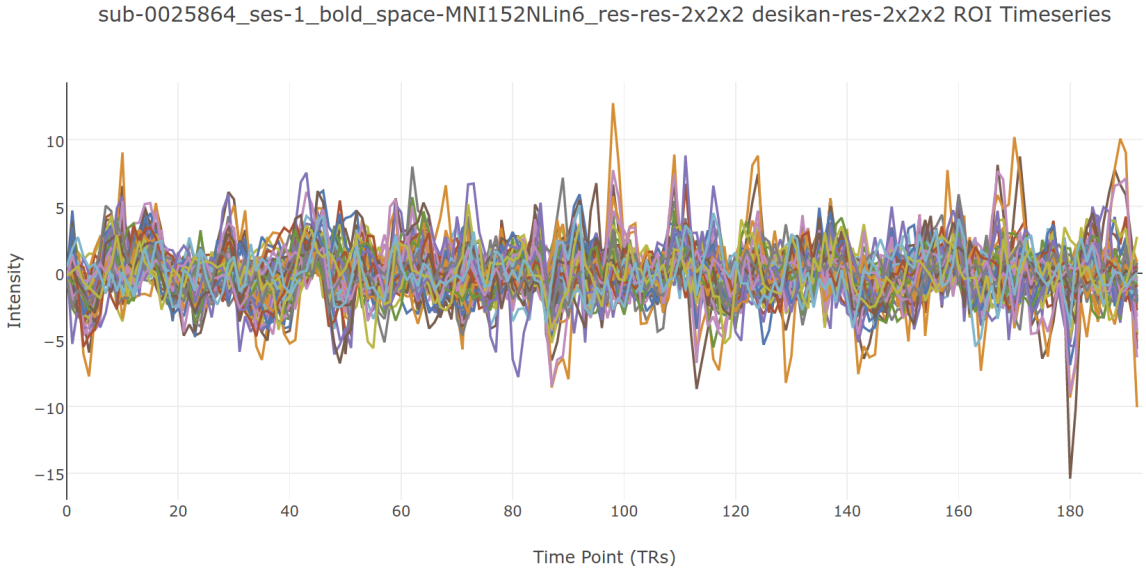
Appendix D Multi-scale Multi-Connectome Analysis

NDMG computes eight node- or edge-wise statistics of each connectome. Each illustrates a non-parametric graph property. The graph statistics are primarily computed with NetworkX and Numpy, and all implementations for NDMG live within the `graph_qa` module. Below, for each statistic we provide a link to the code/documentation of the statistic as it was implemented.



(a)

(b)



0

(c)

Figure 13: NDMG-f **Graph Generation QAX**. For graph generation, we visualize the fMRI sequence shown in green over the parcellation shown in purple (13a), the T1w image in green over the parcellation shown in purple, the correlation matrix (13b), and the average timeseries per ROI (13c).

Table 3: **Graph statistics**. Each of the graph statistics computed by NDMG. The binarized graphs for NDMG-d were formed by thresholding the non-zero edges. The binarized graphs for NDMG-f were formed by thresholding edges with correlations greater than 0.1, which was identified in [73] as having the highest discriminability for functional connectomes.

Statistic	NDMG-d	NDMG-f	**Implementation**
Betweenness Centrality	Binarized Graph	Binarized Graph	NetworkX
Clustering Coefficient	Binarized Graph	Binarized Graph	NetworkX
Degree Sequence	Binarized Graph	Weighted Graph	NetworkX
Edge Weight Sequence	Binarized Graph	–	NetworkX
Eigen Values	Graph Laplacian	Graph Laplacian	NetworkX and Numpy
Locality Statistic-1	Binarized Graph	Weighted Graph	ndmg and NetworkX
Number of Non-Zero Edges	Binarized Graph	Binarized Graph	NetworkX
Path Length	–	Weighted Graph	NetworkX
Cohort Mean Connectome	Weighted Graph	Weighted Graph	Numpy

Appendix D.1 Group-Level Multi-Scale Analysis

Figure 14 shows the group-level summary statistics of diffusion connectomes belonging to same dataset over 13 parcellations ranging from 48 nodes up to 500 nodes; for clarity, an additional 11 parcellations with up to over 70,000 nodes are not shown here. Similarly, Figure **??** shows the group-level summary statistics of functional connectomes belonging to the same dataset over 5 parcellations ranging from 52 to 200 nodes. For each parcellation, vertex statistics are scaled/normalized by number of vertices in the parcellation and smoothed as described in Appendix B.4 for comparison purposes. For most of the statistics, the "shape" of the distributions are relatively similar across scales, though their actual magnitudes can vary somewhat dramatically. In particular, in Figure 14, graphs from the the downsampled block-atlases (DS) appear to be scaled versions of one another, as may be expected because they are related to one-another by a region-growing function [47]. However, graphs from the smaller DS parcellations look less similar to those from the neuroanatomically defined parcellations (JHU [54], Desikan [18], HarvardOxford [45], CC200 [12]). This suggests that the neuroanatomically defined parcellations are more similar to one another than they are to the downsampled parcellations.

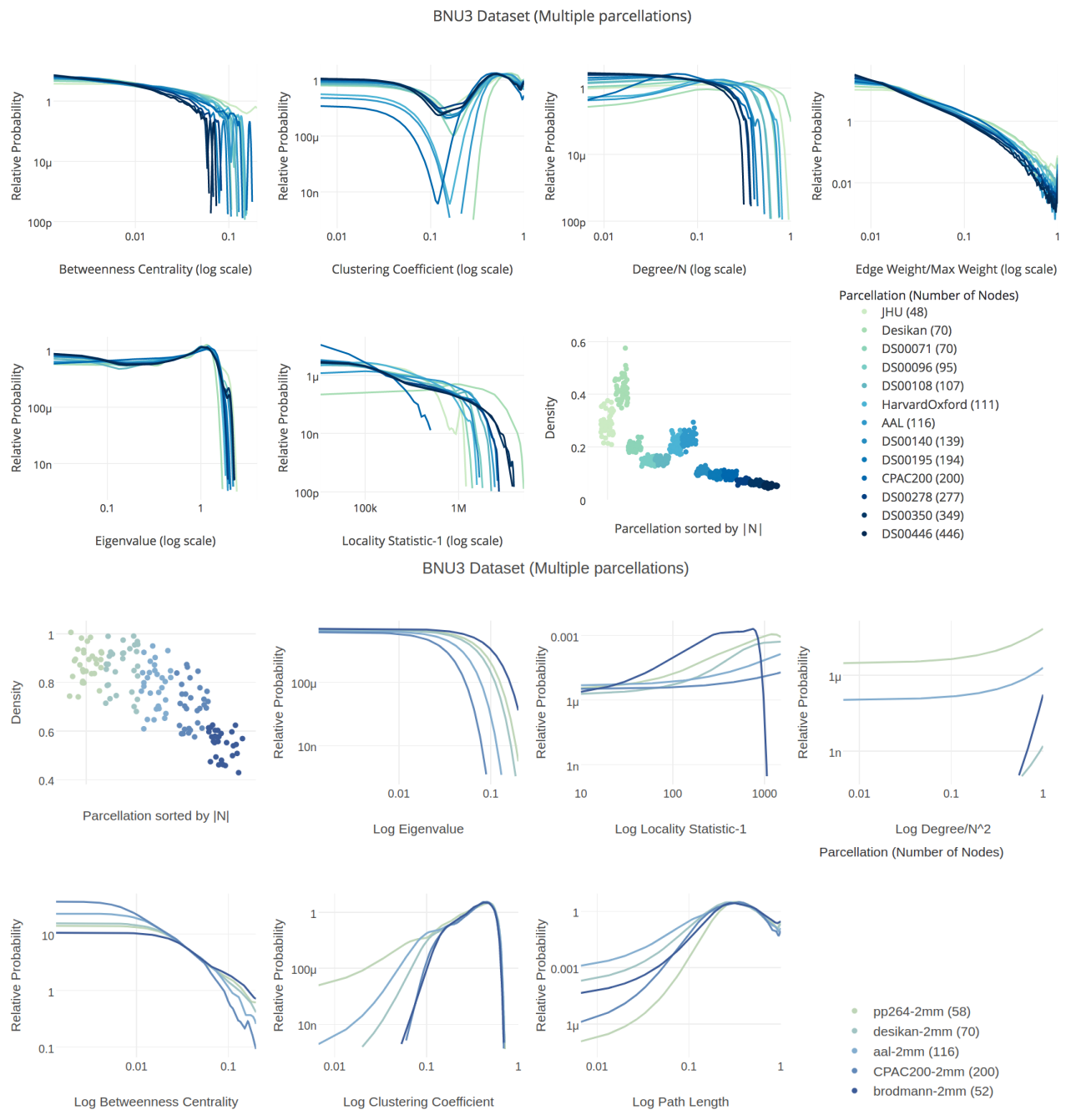


Figure 14: **Multi-Scale Connectome Analysis.** NDMG produces connectomes at a variety of scales, enabling investigation of graph properties between parcellation schemes. We can observe that the statistics are qualitatively similar in shape across scales, however, they are quantitatively significantly different, for both the diffusion (top) and functional (bottom) connectomes. This suggests that claims made or analyses performed on a given scale may not hold when applied to another scale. This is impactful, as the choice of parcellation has significant bearing on the results of a scientific study.

Appendix D.2 Multi-Site Analysis

Figure 15 shows a variety of uni- and multi-variate statistics of the average diffusion connectome from each of the datasets enumerated in Table 2 with diffusion data using the Desikan parcellation. Figure **??** shows the same statistics computed on the average functional connectome from each of the datasets enumerated in Table 2 with functional data using the Desikan parcellation. In both the diffusion and functional connectomes, each dataset largely appears to have similar trends across each of the statistics shown.

Appendix E Statistical Connectomics using a Structured Independent Edge Model

Appendix E.1 Task

Given:

- n samples of graphs, $G_1 = \{(g_i)\}_{i=1}^{n}$ from one population, and m samples of graphs, $G_2 = \{(g_i)\}_{i=1}^{m}$.
- A graph, $g_i \in G_j$, where $g_i = (E, V, w)$ for $N = |V|$ regions of interest and $w(v_i, v_j) = w_{ij}$.
- a partitioning of the edges into E_1 and E_2, where $E_1 \cup E_2 = E$ and $E_1 \cap E_2 = \emptyset$.

a. Does the connectivity for the edges E_1 exceed those of E_2 within a particular modality?

b. Does the difference in connectivity for the edges E_1 and E_2 of one modality exceed that of another modality?

Appendix E.2 Statistical Model

Assume we have a random variable A that can be characterized by the Stochastic Block Model with parameters G, B:

$$A \sim SBM(G, B)$$

where G is a grouping of the N vertices in our graph into C communities V_i where $\bigcup_{i=1}^{C} V_i = V$, and $V_i \cap V_j = \emptyset$ for all $i \neq j$. B represents the parameters for within and between group edge probabilities. Assume that the number of edges in each subgraph are binomially distributed with the parameter p, we can estimate the number of edges for each group with the pmf (noting that in our case, we are given n and k a priori):

$$f_B(p|n,k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Then the likelihood function is of the form:

$$L(p|n,k) = \prod_{k=0}^{n} f_B(n,k|p) = \prod_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

$$log(L(p|n,k)) = \sum_{k=0}^{n} \log\left(\binom{n}{k}\right) + k\log(p) + (n-k)\log(1-p)$$

Maximizing with respect to p:

$$\frac{\delta log(L(p|n,k))}{\delta p} = \sum_{k=0}^{n} \frac{k}{p} - \frac{n-k}{1-p} = 0$$
$$\frac{k}{p} = \frac{n-k}{1-p}$$
$$\hat{p} = \mathbb{E}[p] = \frac{k}{n}$$

to get the variance term, we note that $\hat{p} = \frac{k}{n}$, so then $Var(p) = Var\left(\frac{k}{n}\right) = \frac{1}{n^2}Var(k)$. The binomial distribution can be thought of as an aggregation of n independent bernoulli trials with probability p; that is, $X_i \overset{iid}{\sim} Bern(p)$ where $\mathbb{E}[X_i] = p$. Given that the variance of independent events sum, we can expand:

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} E\left[X_i^2\right] - E\left[X_i\right]^2$$
$$\mathbb{E}\left[X_i^2\right] = 0^2(1-p) + 1^2(p) = p$$
$$Var(k) = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[X_i\right]^2$$
$$= np(1-p)$$

Then:

$$Var(\hat{p}) = \frac{1}{n^2}Var(k) = \frac{\hat{p}(1-\hat{p})}{n}$$

where p is the probability of a given edge, k are the number of connected edges, and n is the number of possible edges. We can therefore define an estimator of B, $\hat{B}$ where connections between community V_l and V_m can be modelled iid:

$$\hat{B}_{lm} = \mathcal{N}(\mu_{lm}, \sigma_{lm})$$

where $\hat{\mu}_{lm} = \frac{1}{|C_l \times C_m|} \sum_{(i,j) \in E(C_l \times C_m)} A_{ij}$, and $\hat{\sigma}^2_{lm} = \frac{\hat{\mu}_{lm}(1-\hat{\mu}_{lm})}{|C_l \times C_m|}$.

Assuming our edges are iid, we can generalize the above model very simply by instead of considering our vertices to exist in communities, placing our edges into two communities E_1 and E_2, where $E_1 \cup E_2 = E$ and $E_1 \cap E_2 = $. In a 2 edge-community model, we can simplify:

$$A \sim SIEM(G, B)$$

where G is a grouping of our N^2 possible edges into C communities E_i where $\bigcup_{i=1}^{C} E_i = E$, and $E_i \cap E_j = \emptyset$ for all $i \neq j$. B represents the parameters for within and between group edge probabilities.

Then we can define an estimator for B as follows:

$$\hat{B} \sim \mathcal{N}(\mu_B, \Sigma_B)$$

where:

$$\mu_B^{(k)} = p_k = \frac{1}{|E_k|} \sum_{(i,j)\in E_k} M_{ij}$$
$$\sigma_B^{(k)} = \frac{p_k(1-p_k)}{|E_k|}$$

In a 2-community case (as studied here):

$$\hat{\mu}_B = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$
$$\hat{\Sigma}_B = \begin{bmatrix} \frac{p_1(1-p_1)}{|E_1|} & 0 \\ 0 & \frac{p_2(1-p_2)}{|E_2|} \end{bmatrix} = \begin{bmatrix} \sigma_{p_1} & 0 \\ 0 & \sigma_{p_2} \end{bmatrix}$$

where p_j represents the probability of an edge in the j^{th} edge-community, and σ_j the variance of edges in that particular edge-community. Then, given a connectome as an adjacency matrix $M \in \{0,1\}^{N\times N}$ with N vertices, we can compute estimators as follows:

$$E_1 = \{(i,j) : \text{edge } (i,j) \in E_1\}$$
$$E_2 = \{(i,j) : \text{edge } (i,j) \in E_2\}$$
$$\hat{p}_1 = \frac{1}{|E_1|} \sum_{(i,j)\in E_1} M_{ij}$$
$$\hat{p}_2 = \frac{1}{|E_2|} \sum_{(i,j)\in E_2} M_{ij}$$
$$\sigma_{\hat{p}_1} = s_1 = \frac{p_1(1-p_1)}{|E_1|}$$
$$\sigma_{\hat{p}_2} = s_2 = \frac{p_2(1-p_2)}{|E_2|}$$

Then we have $\delta = |p_1 - p_2|$ representing the difference in connectivity from E_1 to E_2.

Appendix E.3 Test Statistic

[Welch's T-Test](https://en.wikipedia.org/wiki/Welch

$$T = \frac{\bar{\delta}_1 - \bar{\delta}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and the degrees of freedom can be calculated as follows:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}}$$

where $\nu_1 = n_1 - 1,\ \nu_2 = n_2 - 1$.
We can then use a one-sided test given T, ν to get a $p-$ value.

Appendix E.4 P-Value

1) We can compute a p-value of falsely rejecting the null hypothesis by simply finding the area:

$$p = \int_{-T_{observed}}^{\infty} p(x, df)dx = 1 - \int_{-\infty}^{T_{observed}} p(x, df)dx$$

where $p(x, df)$ is the pdf for the T distribution with degrees of freedom df.

Appendix E.5 Statistical Power

1) The statistical power can be computed as the inverse of the probability of making a Type II (β) error. A type II error can be defined as follows:

$$\beta = \mathbb{P}(\text{reject } H_A \text{ in favor of } H_0 \mid H_A \text{ is true}) = \mathbb{P}(T_{observed} > T_{critical})$$

where $T_{critical}$ is the test-statistic at the given level of significance α specified by our test. To compute the power, we will compute the rejection cutoff for the test-statistic, and then simulate data under the alternative hypothesis, and see how many times we would reject the null hypothesis in our simulated data.

We examine the above model on collections of functional and diffusion connectomes, as well as the megameans of each modality, with respect to:

a. ipsi vs. contra-lateral edge communities

b. bilateral vs. non-bilateral edge communities

Notes

[1] `http://nipy.org/dipy/examples_bcouilt/reconst_dti.html`

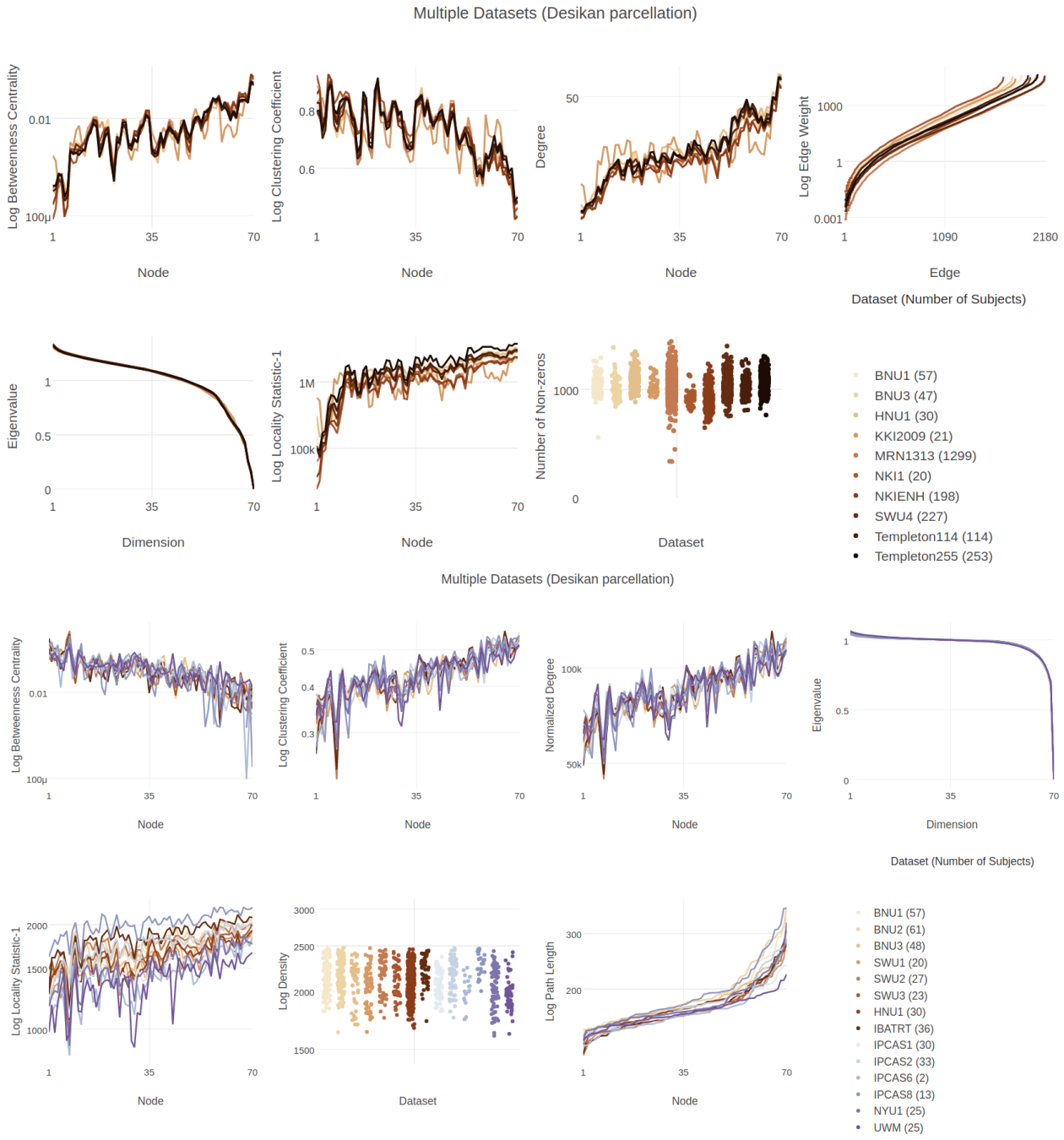


Figure 15: **Multi-Site Connectome Analysis**. Average connectomes from ten diffusion datasets processed with NDMG-d are qualitatively compared by way of their summary statistics on the Desikan parcellation in the top figure. The Desikan atlas used in NDMG has been modified to include two additional regions, one per hemisphere, which fills in a hole in the parcellation near the corpus callosum. The nodes in this plot have been sorted such that the degree sequence of the left hemisphere (Desikan nodes 1-35) of the BNU1 dataset is monotonically non-decreasing, and that corresponding left-right nodes are next to one another. On the bottom, we repeat the same analysis on the functional connectomes. Like the statistics computed in for the diffusion connectomes, the statistics are again qualitatively similar but quantitatively disparate. This suggests that claims made or analyses performed on a given scale may not hold when applied to another scale. Again, we see that parcellation choice has an impact on the implications of a study.