A Comprehensive Cloud Framework for Accurate and Reliable Human Connectome Estimation and Meganalysis

Gregory Kiar, Eric W. Bridgeford, Vikram Chandrashekhar, Disa Mhembere, Randal Burns, William R. Gray Roncal*, Joshua T. Vogelstein*
1 Johns Hopkins University & 2 Child Mind Institute

Abstract

The connectivity of the human brain is fundamental to understanding the principles of cognitive function, and the mechanisms by which it can go awry. To that extent, tools for estimating human brain networks are required for single subject, group level, and cross-study analyses. We have developed an open-source, cloud-enabled, turn-key pipeline that operates on (groups of) raw diffusion and structure magnetic resonance imaging data, estimating brain networks (connectomes) across 24 different spatial scales, with quality assurance visualizations at each stage of processing. Running a harmonized analysis on 10 different datasets comprising 2,295 subjects and 2,861 scans reveals that the connectomes across datasets are similar on coarse scales, but quantitatively different on fine scales. Our framework therefore illustrates that while general principles of human brain organization may be preserved across experiments, obtaining reliable p-values and clinical biomarkers from connectomics will require further harmonization efforts.

1 Introduction

Neuroimaging methods such as magnetic resonance imaging (MRI) are becoming increasingly more accessible and available in clinical and research populations. Specifically, recent advances in Diffusion Weighted MRI (DWI) provide high contrast for the connective tissue of the brain (i.e. white matter), enabling the study of structural networks within the brain (connectomes). As such, DWI data is being collected at an unprecedented rate, including both healthy and diseased populations [1–3]. These datasets provide us with a unique opportunity to discover the principles of connectome coding, as well as potentially identifying clinically useful biomarkers.

To fully capitalize on these data requires processing pipelines that satisfy a number of desiderata. First, pipelines should yield *accurate* and *reliable* estimates of data derivatives with each processing stage. A pipeline's accuracy can be evaluated by comparing it to known neuroscience; its reliability can be assessed with repeated measurements (such as test-retest data). The requirement for accurate and reliable estimates follow from desiring accurate and reliable inferences on the basis of the estimated connectomes; if the connectomes are either inaccurate or unreliable, it is unlikely that the subsequent inferences will be. The pipelines should exhibit these properties across datasets, this includes datasets collected using different acquisition parameters, and from different institutions. This *robustness* to dataset variability facilitates comparing results across datasets, a requirement for high-confidence in scientific or medical studies. Moreover, the pipeline should be able to be run on different platforms, with minimal installation and configuration energy. This *usability* criterion ensures that the pipelines can be run by different analysts using different hardware resources, which may be required for privacy reasons. To date, a number of DWI processing pipelines have been proposed, none of which satisfy each of these desiderata however.

We present NDMG, an accurate, reliable, robust turn-key solution for structural connectomes estimation that can be deployed at scale either in the cloud or locally for cross-study analysis. Leveraging existing tools such as FSL [4–6], Dipy [7], the MNI152 brain atlas [8] and others, NDMG is a one-click pipeline that lowers the barrier to entry for connectomics. By virtue of harmonized processing, the NDMG pipeline enables scientific "meganalysis" in which data from multiple studies can be pooled, opening the door for more highly generalizable statistical analyses of the structure of the human brain. We ran NDMG on 10 datasets comprising 2,295 subjects with 2,861 scans, for each gener-

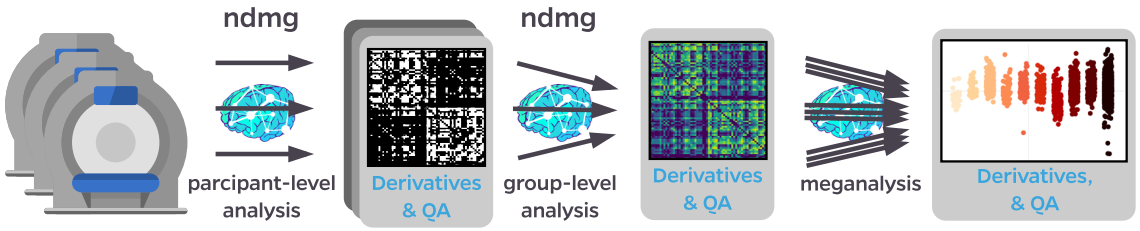

Figure 1: **ndmg usage workflow**. The NDMG pipeline enables accurate, reliable, robust, and usable analysis at the participant-, group-, and meganalysis levels. Raw diffusion and structural scans are provided to the participant-level analysis, which in turn outputs derivatives including connectomes at many different parcellation scales. By organizing the data according to the BIDS specification, group-level analysis is automatically performed, and pooled across datasets for meganalysis.

ating connectomes at 24 different scales, for a total of nearly 70,000 estimated connectomes, all of which are now publicly available. This is the largest yet database of connectomes [9], and the largest meganalysis of connectomics [10]. In addition to demonstrating that NDMG is accurate, reliable, robust, and usable, our results indicate that previously documented qualitative properties of connectomes are preserved across datasets. Yet, quantitatively, even upon harmonizing the connectome estimating, there are significant quantitative differences across datasets. This suggests that further work is required to utilize connectomes to produce accurate and reliable p-values or clinical biomarkers across datasets.

2 Results

The NDMG pipeline enables three tiers of analysis: subject-level, group-level, and meganalysis (see Figure 1). The Brain Imaging Data Structure (BIDS) is a recently proposed specification for organizing multi-scan, multi-subject, multi-modality datasets [11; 12]. Each session of data, consisting of a structural scan (T1w/MPRAGE), a diffusion scan (DWI), and the diffusion parameters files (b-values, b-vectors), can then be used as inputs to generate a connectome.

Subject-level analysis performs several transformations from raw diffusion images towards generating connectomes, producing plots and derivatives for each intermediate stage. Once subject-level analysis has completed for a cohort of data, NDMG group-level analysis can be performed on the generated graphs. At this stage, summary statistics are computed for all graphs generated within the dataset. These statistics are then plotted for additional quality assurance by the scientist, and saved for rigorous quantitative evaluation. Since NDMG processes all subjects and groups identically, it enables the pooling of data across cohorts, that is, a "meganalysis." Meganalysis expands the sample population to potentially improve statistical power, confirm reliability, and ultimately the scientific impact of their findings. All of the derived graphs and intermediate derivatives have been made publicly available on our Amazon S3 bucket, `mrneurodata`, and can also be accessed through `http://m2g.io`.

2.1 Subject-Level Analysis

The subject-level of NDMG has been developed by leveraging and interfacing existing tools, including FSL [4–6], Dipy [7], the MNI152 atlas [8], and a variety of parcellations defined in the MNI152 space [13–19]. All algorithms which required hyper-parameter selection were initially set to the suggested parameters for each tool, and tuned to improve the quality, reliability, and robustness of the results.

Conceptually, this pipeline can be broken up into four key components: (1) registration, (2) tensor estimation, (3) tractography, and (4) graph generation (see Figure 2). The NDMG pipeline has been validated through reliability on multiple measurement datasets (including test-retest). Below we provide a brief description of each step, Appendix A provides further details. Subject-level analysis in NDMG takes approximately 1-hour to complete using 1 CPU core

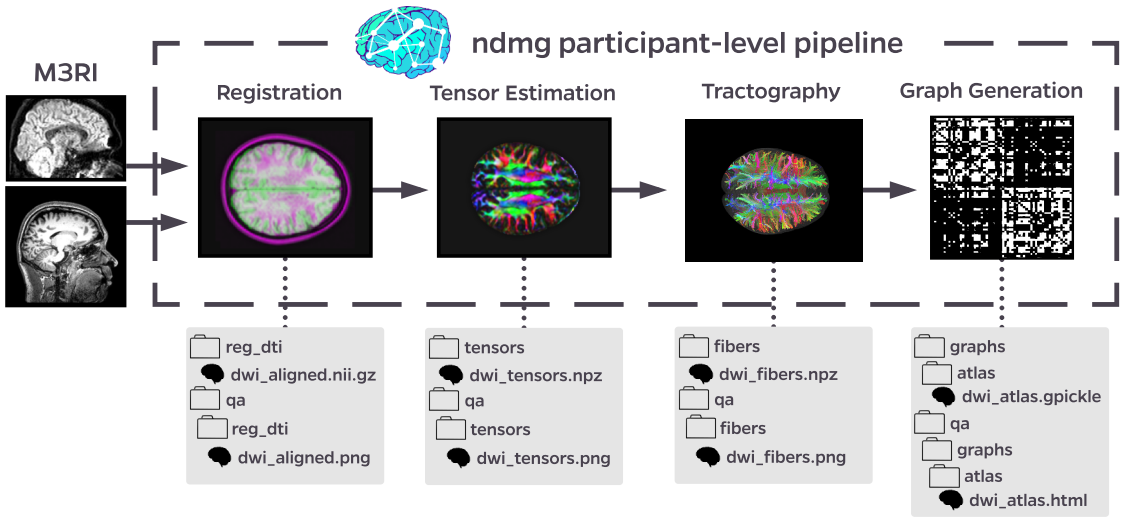

Figure 2: **ndmg subject-level analysis**. The subject-level of the NDMG pipeline transforms raw diffusion weighted MRI data into structural connectomes. Here, NDMG consists of four main steps: registration, tensor estimation, tractography, and graph generation . At each stage, NDMG produces both data derivatives and quality assurance figures of the derivative, as illustrated.

and 12 GB of RAM.

A crucial component of the design of NDMG was to include quality assurance (QA) figures for each stage to the pipeline, to enable users to easily detect whether or not the pipeline is producing accurate results. Snapshots of these QA figures are shown Figure 2.

The subject-level analysis was run on 2,295 subjectss including 2,861 scans; each generating connectomes across each of the 24 parcellations in NDMG, resulting in 68,664 total brain-graphs.

Registration NDMG leverages FSL [4–6] for a series of linear registrations. Taking as input the minimally preprocessed DWI and T1W images, the end result is the DWI volumes aligned to the MNI152 atlas [8]. The registration pipeline implemented is "standard" when working with diffusion data and FSL's tools. The QA figure produced at this stage is cross-sectional images at different depths in the three canonical planes (sagittal, coronal, and axial) of an overlay of the DWI image and the MNI152 atlas.

Tensor Estimation A voxelwise tensor image from the DWI image stack using a simple 6-component tensor model from Dipy [7]. The aligned diffusion volumes and b-values/b-vectors files are transformed into a 6-dimensional tensor volume, a single dimension for each component of the resulting tensors. A fractional anisotropy map of the tensors is provided for QA, again using multiple depths in each of the three canonical image planes.

Tractography Streamlines are generated from the tensors using Dipy's EuDX [20], a deterministic tractography algorithm closely related to FACT [21]. Each voxel within the tensor image, confined to the boundary of the brain mask, is used as a seed-point in EuDX and fibers are produced and then pruned based on their length. NDMG provides a QA plot visualizing a subset of the generated streamlines within a mask of the MNI152 brain so that the user can verify that their structure resembles that of the fractional anisotropy map generated in the previous step.

Graph Generation Connectomes are created by tracing fibers through pre-defined parcellations. As fibers are traced, an undirected edge is added to the graph for every pair of regions along the path, where the weight of the edge is the cumulative number of fibers between two regions. NDMG includes neuroanatomically delineated parcellations,

Table 1: **Processed public M3R datasets.** The derivatives from each dataset processed with NDMG are publicly available at `http://m2g.io`. Multiple measurement datasets were evaluated using discriminability ("Discr" below), where 1 indicates perfectly discriminable connectomes. The pooled discriminability is the computed by pooling the datasets and ignoring group labels. Age is reported as the dataset mean $\pm$ standard deviation; "Rep's" is the number of scans per subject.

Dataset	Scanner	# Dirs	Age (yrs)	% Male	# Subj's	Rep's	Total Scans	Discr
BNU1 [2]	Siemens	30	23.0 ± 2.3	53	57	2	114	0.984
BNU3 [2]	Siemens	64	22.5 ± 2.1	50	48	1	47	-
HNU1 [2]	GE	33	24.4 ± 2.3	50	30	10	300	0.993
KKI2009 [22]	Philips	33	31.8 ± 9.4	52	21	2	42	1.0
MRN1313	-	70	-	-	1313	1	1299	-
NKI1 [2]	Siemens	137	34.4 ± 12.8	0	24	2	40	0.984
NKI-ENH [23]	Siemens	137	42.5 ± 19.6	40	198	1	198	-
SWU4 [2]	-	93	20.0 ± 1.3	51	235	2	454	0.884
Templeton114	Siemens	70	21.8 ± 3.0	58	114	1	114	-
Templeton255	Siemens	150	-	-	255	1	253	-
Pooled					**2295**		**2861**	**0.979**

such as the HarvardOxford cortical and sub-cortical atlases [16], JHU [15], Talairach [17], Desikan [13], and AAL [14] atlases, algorithmically delineated parcellations, such as slab907 [18], Slab1068 [19], CC200 [24], and 16 downsampled (DS) parcellations [25] ranging from 70 to 72,783 nodes that we developed. QA for the graph generation step includes a heatmap of the adjacency matrix, as well as several univariate and multivariate graph statistics: betweenness centrality, clustering coefficient, hemisphere-separated degree sequence, edge weight, eigenvalues of the graph laplacian, locality statistic-1, and the number of non-zero edges [25]. The hemisphere-separated degree sequence we developed to indicate, for each vertex, its ipsilateral degree and its contralateral degree, which we found quite useful for QA. Appendix A.4 includes definitions and implementation details for each of the statistics.

2.2 Group-Level Analysis

Once connectomes have been generated for a dataset, NDMG group-level analysis computes and plots multiscale graph summary statistics as well as reliability statistics. We ran the NDMG group analysis on 10 different datasets, listed in Table 1.

Graph Summary Statistics Each subject's connectome can be summarized by a set of graph statistics, as described above. For QA purposes, we visualize each session's summary statistics overlaid on one another. For example, Figure 3 demonstrates that each graph from the BNU3 dataset using the Desikan atlas has relatively similar values for the statistics. It is clear from both the degree plot and the mean connectome plot that the DWI connectomes tend to have more connections within a hemisphere than across a hemisphere, as expected. Appendix A.5 illustrates how NDMG also computes the average value for each univariate and multivariate statistic for each atlas, and demonstrates similarities across scales, indicating that that the basic structure of the connectomes is preserved across different atlases.

Reliability Group level results from NDMG that include repeated measurements are quantitatively assessed using a statistic called discriminability [26]. 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:

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

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 dataset, each connectome is more alike to connectomes from the same subject than to others. Table 1 lists the discriminability score of each dataset with repeated measurements; NDMG achieves a discriminability score of nearly 0.99 or greater on most

datasets, with a the lowest scoring nearly 0.9.

2.3 Meganalysis

Many sources of variability contribute to the observed summary statistics, including subject, measurement, population, and analysis. By virtue of harmonizing the analysis across subjects, we are able assess the remaining degrees of variability due to measurement and population specific effects. Although population level effects are expected, for example, when comparing two different populations with different demographics, for inferences based on neuroimaging to be valid, variability across measurements must be relatively small.

Coarse Grained Similarities Across Groups Figure 4 shows the mean connectome computed from each dataset, as well as the weighted mega-mean and mega-standard deviation connectomes combining all datasets. These means have very similar structures and intensity profiles. For example, each connectome seems to contain a higher number of ipsi-lateral connections than contra-lateral connections. Moreover, the ipsi-lateral connections form more consistent across groups than are the contra-lateral connections. Finally, the ipsi-lateral connectivity within left (nodes 1-35) and right (nodes 36-70) hemispheres, respectively, are very similar in structure.

To test each of these conjectures, we assume that each groups' connectome is a sample from a random graph model. We compute the variance across datasets for each connectome edge. Indeed, the variances of the ipsi-lateral connections are stochastically smaller than those of the contra-lateral connections.

Fine-Grained Difference Across Groups with Implications Although at a coarse resolution each group of connectomes exhibited similar properties, the above analysis is insufficient to determine the extent of "batch effects"— sources of variabil-

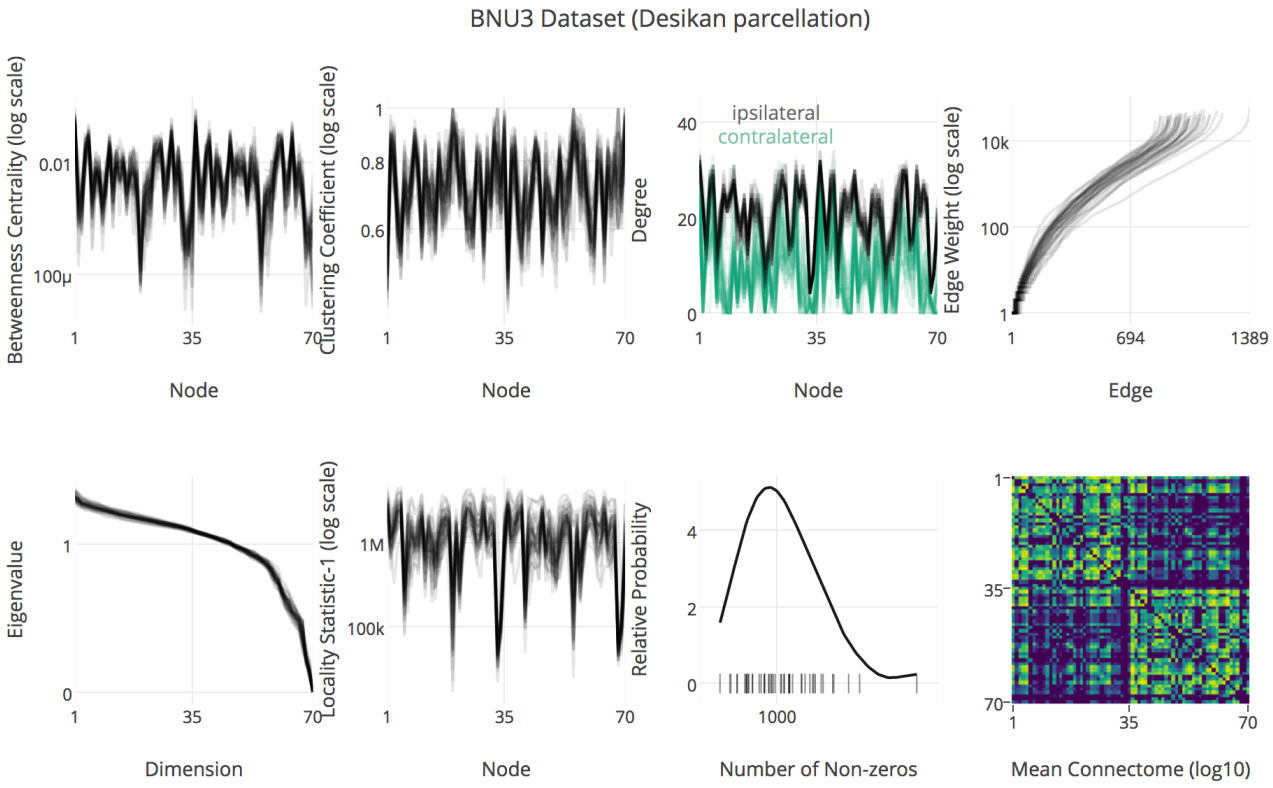

Figure 3: **Graph summary statistics**. NDMG computes and plots of connectome-specific summary statistics after estimating graphs for each session, providing immediate quality assurance for the group.

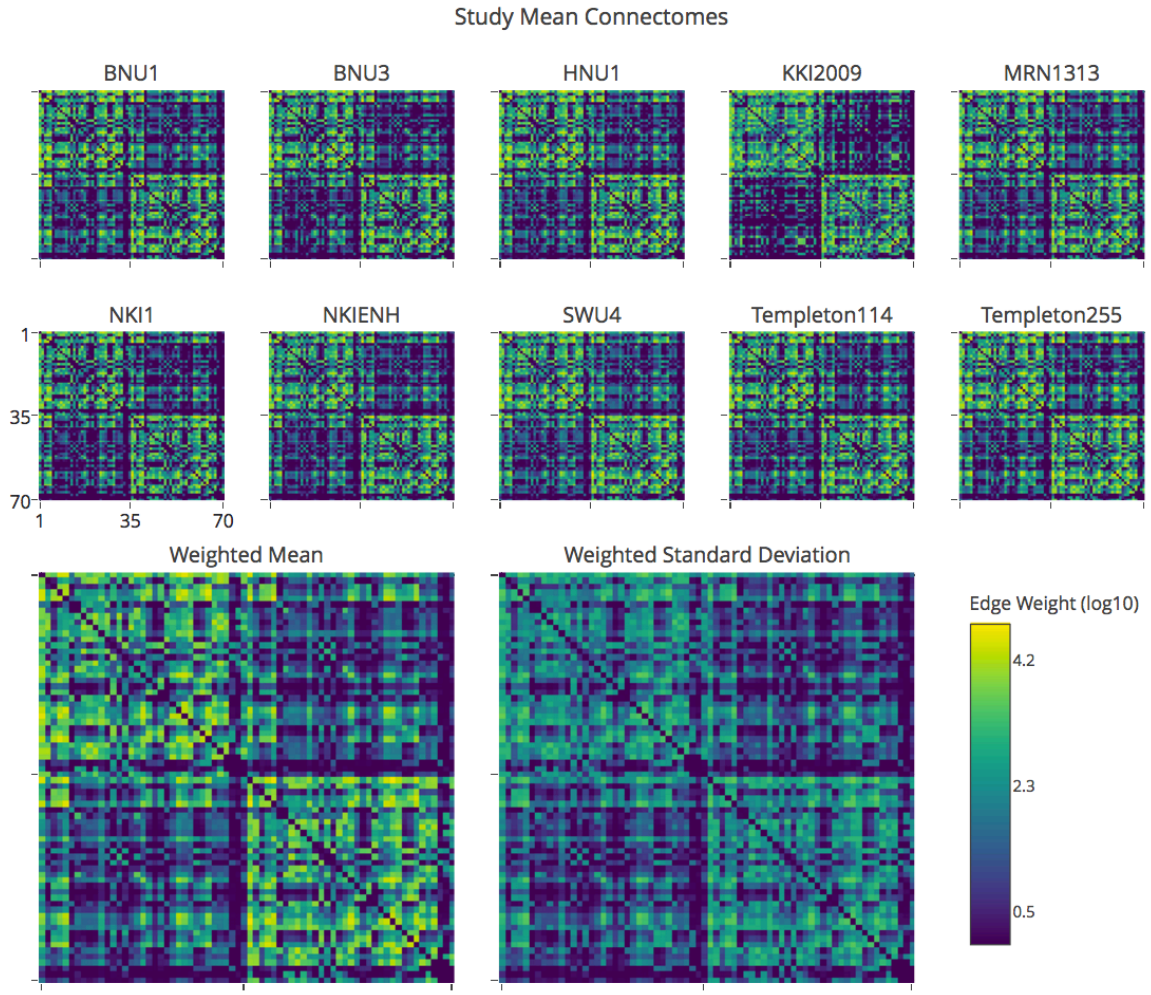

Figure 4: **Multi-Study Mean Connectomes**. Dataset-mean connectomes and a combined mean-of-mean and standard deviation-of-mean connectomes were computed from the Desikan labeled graphs produced by NDMG, resulting in the largest known mean connectome to-date, consisting of 2,861 sessions. Datasets appear qualitatively similar to one another, with minor deviations particularly visible in the contra-lateral regions of the graphs. As expected, ipsi-lateral connectivity is consistently more dense than contra-lateral connectivity. Similarly, the standard deviation connectome, which highlights edges that are more highly variable, shows higher ipsi-lateral variance. This suggests that not only are these connections more likely to occur, but they have a higher variance, as well.

ity such as scanner, acquisition sequence, and operator, which are not 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. Whereas we desire high subject-level discriminability indicating that subject variability is larger than other sources of variability within a group, here we desire low group-level discriminability, indicating that group variability is smaller than biological

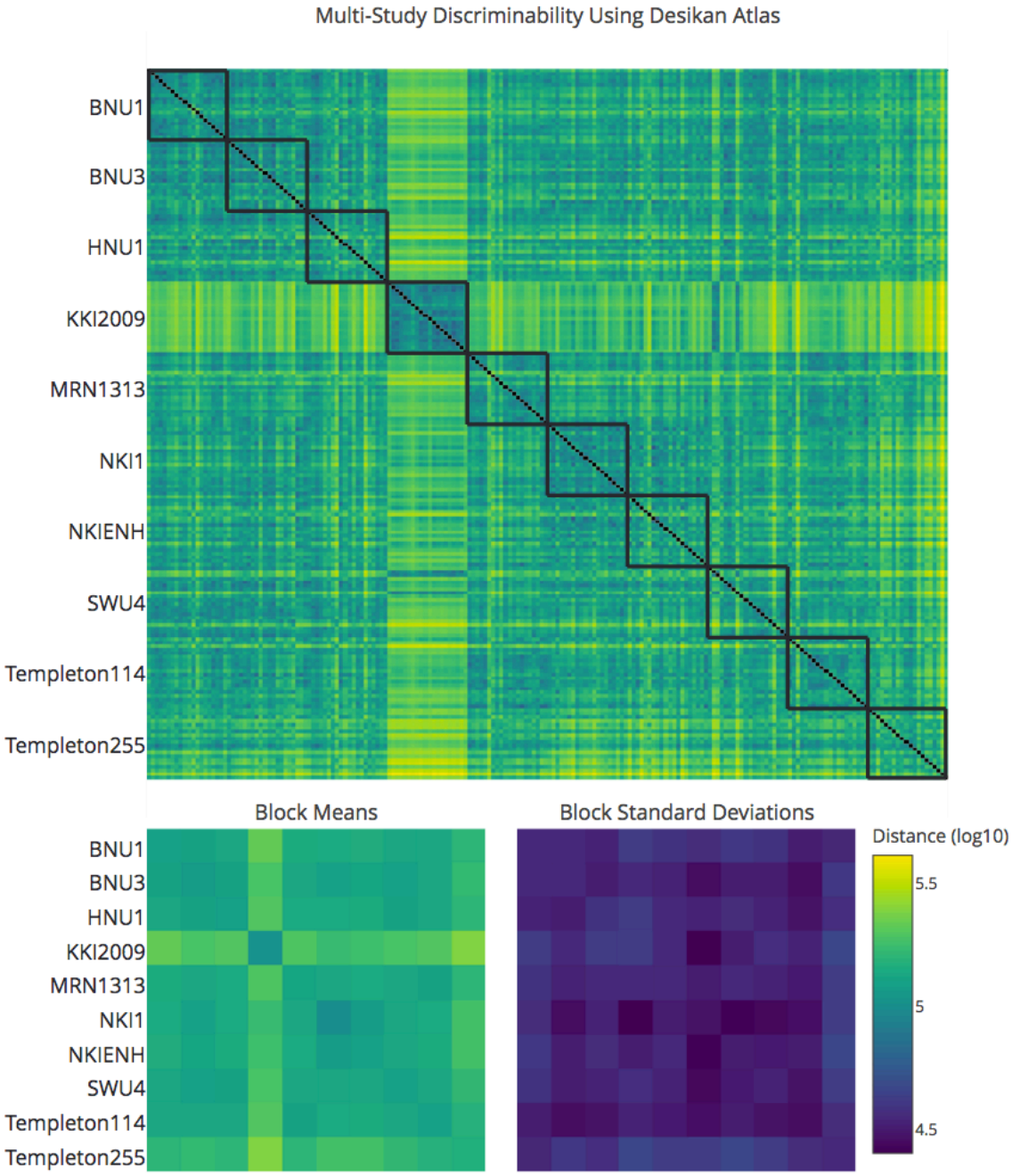

Figure 5: **Prevalence of batch effects**. Discriminability was computed across the ten processed datasets using the dataset-id as the class label. If no significant difference between datasets exists, the discriminability score would not be significantly different from chance, a score of 0.363. Here, the discriminability score was 0.632, which is significant with a p-value of less than 0.0001 when performing a permutation test, suggesting that there is significant dataset specific signal in the graphs.

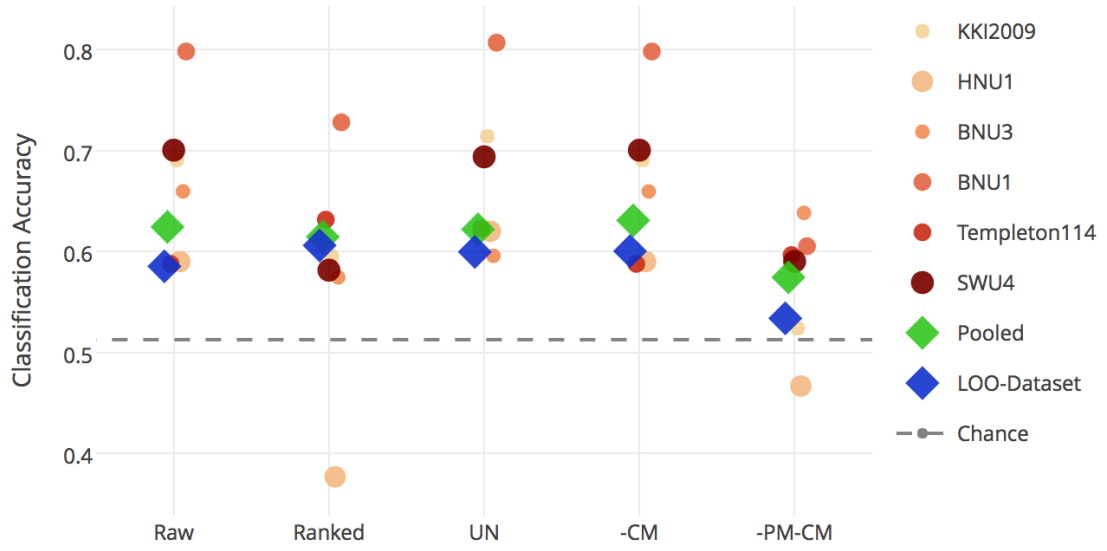

Figure 6: **Connectome Sex Classification**. Using K-Nearest Neighbours classification, several cross-validation attempts were made on "raw" and further post-processed connectomes to improve accuracy in sex classification. The attempted normalizations were: ranking edges, unit normalization, dataset with subtracting cohort means, and dataset with subtracting cohort and population means. The cross-validations methods were leave-out-one (LOO): subject within each dataset, subject for pooled dataset, dataset for pooled dataset. If the batch effects were insignificant, then we would expect the two pooled-dataset methods to have equivalent performance. We notice that LOO-subject performs greater in all normalization strategies, indicating that the batch effect has considerable impact on downstream classification.

variability. Chance discriminability, indicating that connectomes were all 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 0.363, a significant difference as the p < 0.0001 level (see [26] for details). Figure 5 shows the average discriminability both within and across groups. KKI2009 sticks out as being an outlier group. KKI2009 was acquired on a Philips scanner, whereas the other groups predominantly used Siemens scanners and, and a single group used GE (HNU1). Removing either or both groups that used non-Siemens scanner did not meaningfully change discriminability (0.626 with p<0.0001, and 0.627 with p<0.0001, for removing KKI2009, and both KKI2009 and HNU1, respectively). 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, however, 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 not be possible. Rather, it is the relative size of the batch effect as compared to the effect of the signal of interest that regulates the impact of batch effects. To determine the implications of the batch effect on 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 [27]. FIgure 6 depicts the leave-one-out (LOO) subject out classification error for each of the six DWI datasets with sex information; the accuracy ranges from around 60% 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 sample size, and therefore improve accuracy. However, the pooled data exhibited poor accuracy, with LOO-dataset performing even worse than many of the individual datasets.

It is possible that this batch effect could be mediated by some normalization scheme. 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. None of the normalization schemes operationally improved performance on the pooled data. These results collectively suggests that the batch effect is large, and signals found in one dataset were idiosyncratic to that dataset, rather than representing true neurobiological signals.

3 Discussion

The NDMG pipeline is a reliable tool for structural connectome estimation with a low barrier to entry for neuroscientists, 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 [28; 29]. 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 and makes comparing graphs trivial across studies, avoiding 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 in a principled meganalysis using the open source data derivatives generated here.

Previous efforts have developed pipelines for DWI data. For example, PANDAS [30] and CMTK [31] are flexible pipelines enabling users to select hyper-parameters for their dataset, a useful feature, but they do not provide a reference pipeline that is optimized for any particular criteria across datasets. MRCAP [32] and MIGRAINE [33] provide reference pipelines, but are difficult to deploy, and also lacked vetting across datasets.

Other efforts have focused on multi-site data. Specifically, [34] used fMRI-derived connectomes from the ABIDE dataset demonstrating an impressive ability to minimized batch effects. Unfortunately, most ABIDE datasets lack DWI data, so a similar strategy for NDMG is not currently possible. Additionally, a variety of studies propose methods for data harmonization upon either minimally preprocessed or raw MRI data [35–37] which could be explored within the context of the NDMG pipeline.

We integrated NDMG with a number of different computing platforms, including OpenNeuro[1], CBRAIN[2], and Amazon Web Services[3], as well as provide a Docker image so that it can be run from the web or locally on disparate computational configurations such as laptops and institutional clusters with ease.

Author Information GK[1] wrote the code, designed the experiments, ran the analysis, and wrote the paper; EB[1] did all fMRI analysis; EB & VC[1] assisted with all aspects of the manuscript; DM[1] wrote the large graph analysis, advised by RB[1]. WGR[1] designed the python package based on a preliminary implementation that he wrote, and co-advised GK, JTV[1,2] oversaw everything and is the corresponding author: ⟨jovo@jhu.edu⟩.

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] J. A. Nielsen, B. A. Zielinski, P. T. Fletcher, A. L. Alexander, N. Lange, E. D. Bigler, J. E. Lainhart, and J. S. Anderson, "Multisite functional connectivity mri classification of autism: Abide results," 2013.

[2] X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos *et al.*, "An open science resource for establishing reliability and reproducibility in functional connectomics," *Scientific data*, vol. 1, p. 140049, 2014.

[3] S. Das, A. P. Zijdenbos, D. Vins, J. Harlap, and A. C. Evans, "Loris: a web-based data management system for multi-center studies," *Frontiers in neuroinformatics*, vol. 5, p. 37, 2012.

[4] S. M. Smith *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL." *NeuroImage*, vol. 23 Suppl 1, pp. S208–19, jan 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15501092

[5] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL." *NeuroImage*, vol. 45, no. 1 Suppl, pp. S173–86, mar 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811908012044

[6] M. Jenkinson *et al.*, "FSL." *NeuroImage*, vol. 62, no. 2, pp. 782–90, aug 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21979382

[7] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. Van Der Walt, M. Descoteaux, and I. Nimmo-Smith, "Dipy, a library for the analysis of diffusion mri data," *Frontiers in neuroinformatics*, vol. 8, p. 8, 2014.

[8] J. Mazziotta *et al.*, "A four-dimensional probabilistic atlas of the human brain," *Journal of the American Medical Informatics Association*, vol. 8, no. 5, pp. 401–430, 2001.

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

[10] G. Varoquaux and R. C. Craddock, "Learning and comparing functional connectomes across subjects," *NeuroImage*, vol. 80, pp. 405–415, 2013.

[11] K. Gorgolewski, T. Auer, V. Calhoun, C. Craddock, S. Das, E. Duff, G. Flandin, S. Ghosh, T. Glatard, Y. Halchenko *et al.*, "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments."

[12] K. J. Gorgolewski, F. Alfaro-Almagro, T. Auer, P. Bellec, M. Capotă, M. M. Chakravarty, N. W. Churchill, A. L. Cohen, R. C. Craddock, G. A. Devenyi, A. Eklund, O. Esteban, G. Flandin, S. S. Ghosh, J. S. Guntupalli, M. Jenkinson, A. Keshavan, G. Kiar, F. Liem, P. R. Raamana, D. Raffelt, C. J. Steele, P.-O. Quirion, R. E. Smith, S. C. Strother, G. Varoquaux, Y. Wang, T. Yarkoni, and R. A. Poldrack, "Bids apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods," *PLOS Computational Biology*, vol. 13, no. 3, pp. 1–16, 03 2017. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pcbi.1005209

[13] R. 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.

[14] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

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

[16] N. Makris, J. M. Goldstein, D. Kennedy, S. M. Hodge, V. S. Caviness, S. V. Faraone, M. T. Tsuang, and L. J. Seidman, "Decreased volume of left and total anterior insular lobule in schizophrenia," *Schizophrenia research*, vol. 83, no. 2, pp. 155–171, 2006.

[17] J. Lancaster, "The Talairach Daemon, a database server for Talairach atlas labels," *NeuroImage*, 1997.

[18] C. 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*, vol. 111, no. 39, pp. 14 259–14 264, 2014.

[19] D. Kessler *et al.*, "Modality-spanning deficits in attention-deficit/hyperactivity disorder in functional networks, gray matter, and white matter," *The Journal of Neuroscience*, vol. 34, no. 50, pp. 16 555–16 566, 2014.

[20] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams, and I. Nimmo-Smith, "Quickbundles, a method for tractography simplification," *Frontiers in neuroscience*, vol. 6, p. 175, 2012.

[21] S. Mori *et al.*, "Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging," *Annals of neurology*, vol. 45, no. 2, pp. 265–269, 1999.

[22] B. A. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. L. Lim, J. A. Farrell, J. A. Bogovic, J. Hua, M. Chen, S. Jarso *et al.*, "Multi-parametric neuroimaging reproducibility: a 3-t resource study," *Neuroimage*, vol. 54, no. 4, pp. 2854–2866, 2011.

[23] K. B. Nooner, S. Colcombe, R. Tobe, M. Mennes, M. Benedict, A. Moreno, L. Panek, S. Brown, S. Zavitz, Q. Li *et al.*, "The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry," *Frontiers in neuroscience*, vol. 6, p. 152, 2012.

[24] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fmri atlas generated via spatially constrained spectral clustering," *Human brain mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[25] D. Mhembere, W. G. Roncal, D. Sussman, C. E. Priebe, R. Jung, S. Ryman, R. J. Vogelstein, J. T. Vogelstein, and R. Burns, "Computing scalable multivariate glocal invariants of large (brain-) graphs," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 297–300.

[26] S. Wang, Z. Yang, X.-N. Zuo, M. Milham, C. Craddock, G. Kiar, W. R. Gray Roncal, E. Bridgeford, CORR, C. E. Preibe, and J. T. Vogelstein, "Optimal decisions for discovery science via maximizing discriminability: Applications in neuroimaging," Tech. Rep., 2017.

[27] J. T. Vogelstein, W. G. Roncal, R. J. Vogelstein, and C. E. Priebe, "Graph classification using signal-subgraphs: Applications in statistical connectomics," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1539–1551, 2013.

[28] J. P. Owen, E. Ziv, P. Bukshpun, N. Pojman, M. Wakahiro, J. I. Berman, T. P. Roberts, E. J. Friedman, E. H. Sherr, and P. Mukherjee, "Test–retest reliability of computational network measurements derived from the structural connectome of the human brain," *Brain connectivity*, vol. 3, no. 2, pp. 160–176, 2013.

[29] D. Petrov, A. Ivanov, J. Faskowitz, B. Gutman, D. Moyer, J. Villalon, N. Jahanshad, and P. Thompson, "Evaluating 35 methods to generate structural connectomes using pairwise classification," *arXiv preprint arXiv:1706.06031*, 2017.

[30] Z. Cui *et al.*, "PANDA: a pipeline toolbox for analyzing brain diffusion images." *Frontiers in human neuroscience*, vol. 7, p. 42, jan 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2013.00042/abstract

[31] A. Daducci *et al.*, "The connectome mapper: an open-source processing pipeline to map connectomes with MRI." *PloS one*, vol. 7, no. 12, p. e48121, jan 2012. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0048121

[32] W. R. Gray, J. A. Bogovic, J. T. Vogelstein, B. A. Landman, J. L. Prince, and R. J. Vogelstein, "Magnetic resonance connectome automated pipeline," *IEEE Pulse*, vol. 3, no. 2, pp. 42–48, 2011.

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

[34] A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux, "Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example," *NeuroImage*, vol. 147, pp. 736–745, 2017.

[35] H. Mirzaalian, L. Ning, P. Savadjiev, O. Pasternak, S. Bouix, O. Michailovich, S. Karmacharya, G. Grant, C. E. Marx, R. A. Morey *et al.*, "Multi-site harmonization of diffusion mri data in a registration framework," *Brain Imaging and Behavior*, pp. 1–12, 2017.

[36] J.-P. Fortin, D. Parker, B. Tunc, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur *et al.*, "Harmonization of multi-site diffusion tensor imaging data," *bioRxiv*, p. 116541, 2017.

[37] E. Olivetti, S. Greiner, and P. Avesani, "Adhd diagnosis from multiple data sources with batch effects," *Frontiers in systems neuroscience*, vol. 6, p. 70, 2012.

[38] J.-D. Tournier, F. Calamante, D. G. Gadian, and A. Connelly, "Direct estimation of the fiber orientation density function from diffusion-weighted mri data using spherical deconvolution," *NeuroImage*, vol. 23, no. 3, pp. 1176–1185, 2004.

[39] D. S. Tuch, T. G. Reese, M. R. Wiegell, and V. J. Wedeen, "Diffusion mri of complex neural architecture," *Neuron*, vol. 40, no. 5, pp. 885–895, 2003.

Appendix A Processing Pipeline

Here we take a deep-dive into each of the modules of the NDMG 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 A.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, a desired feature of the pipeline was that it could be run on non-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, as non-linear methods were found to have higher variability across datasets while simultaneously increasing the resource and time requirements of the pipeline (not shown).

As is seen in Figure 7B1, 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 as 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.

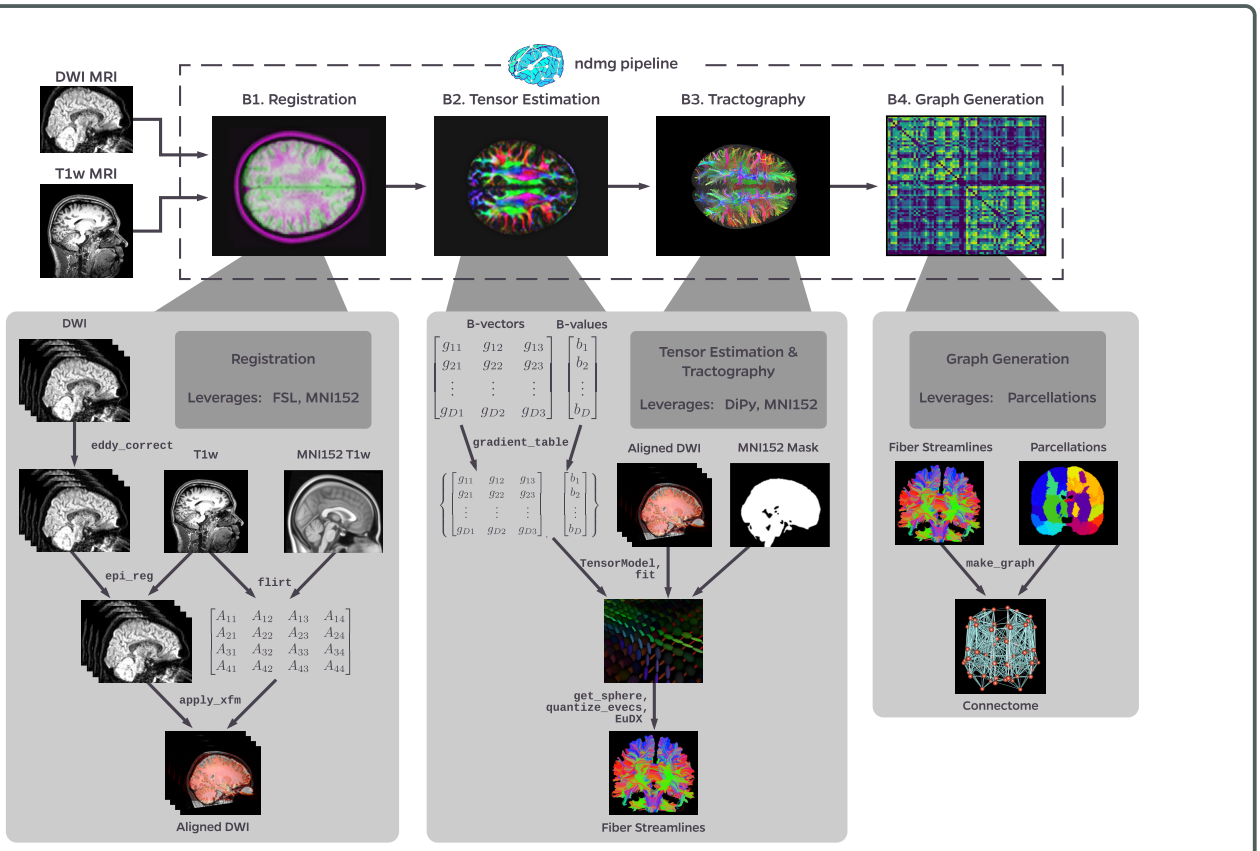

Figure 7: **ndmg detailed pipeline**. The NDMG pipeline consists of 4 main steps: Registration (**B1**), Tensor Estimation (**B2**), Tractography (**B3**), and Graph Generation (**B4**). Each of these sections leverages pubicly available tools and data to robustly produce the desired derivative of each step. Alongside derivative production, NDMG produces QA figures at each stage, as can be seen in **B1-4**, that enable qualitative evaluation of the pipeline's performance.

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, as this enables 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 3 slices of the first B0 volume of the aligned DWI image overlaid on the MNI152 template in the 3 principle coordinate planes, providing 9 plots in total which enable qualitative assessment of the quality of alignment.

Appendix A.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 [7]. 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 [38; 39]. Because NDMG is designed to run robustly on 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[4], computes a 6-component tensor for each voxel in the image, reducing the DWI image stack to a single 6-dimensional image which 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 9 panels as above.

Appendix A.3 Tractography

In keeping with the theme of computationally efficient and robust methods, NDMG uses DiPy's deterministic tractography algorithm, `EuDX` [20]. 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, allowing 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 A.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.

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.

Table 2: **Graph statistics**. Each of the graph statistics computed by NDMG.

Statistic	Operates On	Implementation
Betweenness Centrality	Binarized Graph	NetworkX
Clustering Coefficient	Binarized Graph	NetworkX
Degree Sequence	Binarized Graph	NetworkX
Edge Weight Sequence	Weighted Graph	NetworkX
Eigen Values	Weighted Graph	NetworkX and Numpy
Locality Statistic-1	Weighted Graph	ndmg and NetworkX
Number of Non-Zero Edges	Binarized Graph	NetworkX
Cohort Mean Connectome	Weighted Graph	Numpy

Appendix A.5 Group-Level Multi-Scale Analysis

Figure 8 shows the group-level summary statistics of connectomes belonging to same dataset over 13 parcellations ranging from 48 nodes up to 500 nodes; an additional 11 parcellations with up to over 70,000 nodes are not shown here for clarity. For each parcellation, vertex statistics are scaled/normalized by number of vertices in the parcellation and smoothed as described in **??** 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, 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 [25]. However, graphs from the smaller DS parcellations look less similar to those from the neuroanatomically defined parcellations (JHU [15], Desikan [13], HarvardOxford [16], CC200 [24]). This suggests that the neuroanatomically defined parcellations are more similar to one another than they are to the downsampled parcellations.

Appendix A.6 Multi-Site Analysis

Figure 9 shows a variety of uni– and multi-variate statistics of the average connectome from each of the datasets enumerated in Table 1 using the Desikan parcellation. Each dataset largely appears to have similar trends across each of the statistics shown.

Notes

[1] `https://openneuro.org/`
[2] https://portal.cbrain.mcgill.ca
[3] http://scienceinthe.cloud/
[4] `http://nipy.org/dipy/examples_built/reconst_dti.html`

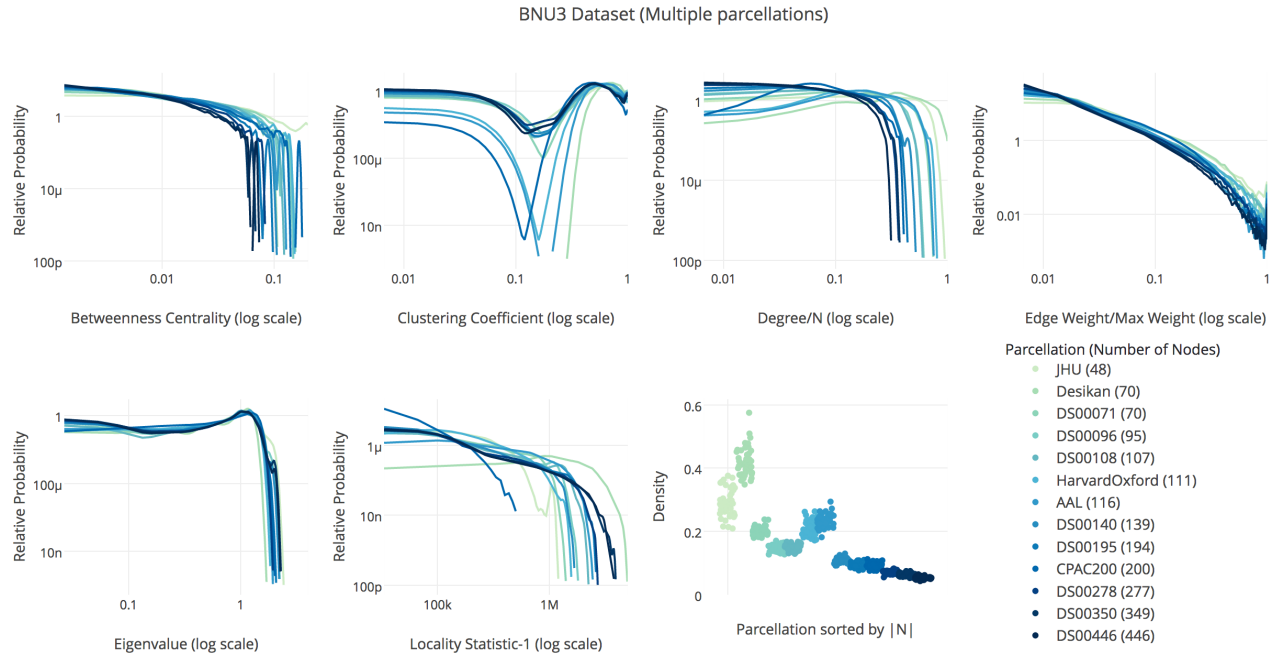

Figure 8: **Multi-scale graph analysis**. NDMG produces connectomes at a variety of scales, enabling investigation of graph properties between parcellation schemes. We can observe that the statistics are qualitatively similar in shape across scales, however, they are quantitatively significantly different. 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.

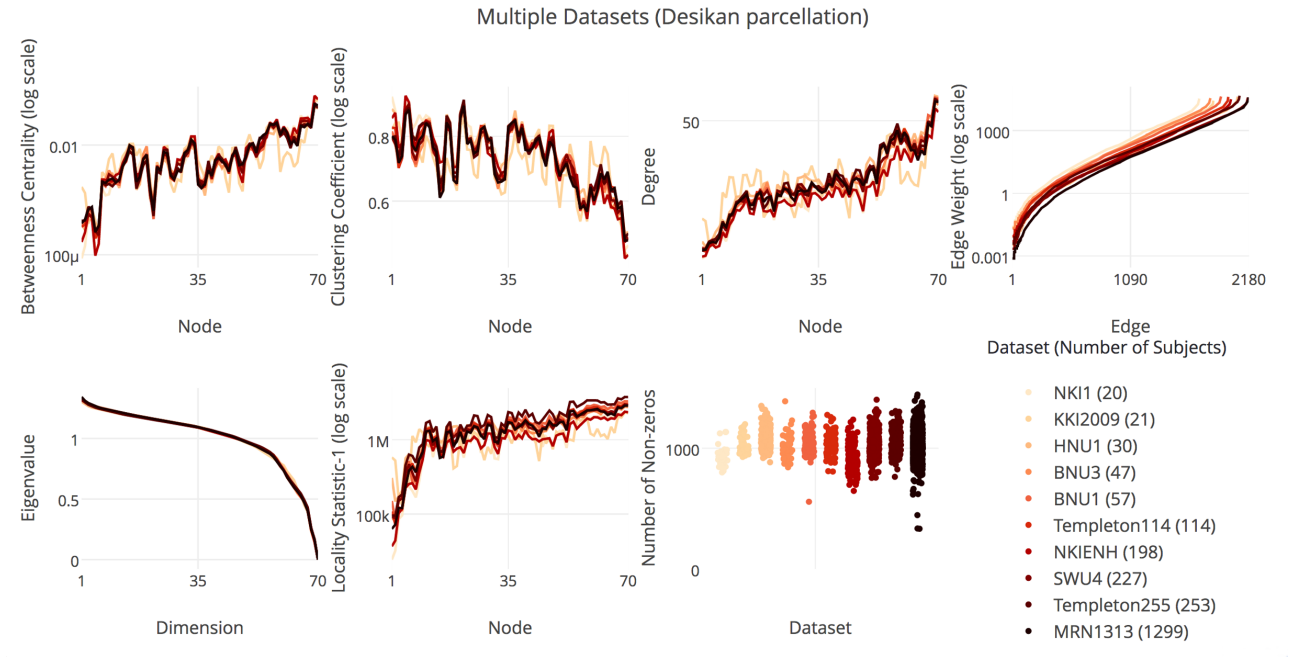

Figure 9: **Multi-site graph analysis**. Average connectomes from ten datasets processed with NDMG are qualitatively compared by way of their summary statistics on the Desikan parcellation. The Desikan atlas used in NDMG has been modified to include two additional regions, one per hemisphere, which fills 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. The interactive version of this plot, available through `http://m2g.io`, provides hover-text of the region names.