Numerical Instabilities in Analytical Pipelines Compromise the Reliability of Network Neuroscience

Gregory Kiar[1], Yohan Chatelain[2], Pablo de Oliveira Castro[3], Eric Petit[4], Ariel Rokem[5], Gaël Varoquaux[6], Bratislav Misic[1], Tristan Glatard[2†], Alan C. Evans[1†]

Abstract

The analysis of brain-imaging data requires complex and often non-linear transformations to support findings on brain function or pathologies. And yet, recent work has shown that variability in the choices that one makes when analyzing data can lead to quantitatively and qualitatively different results, endangering the trust in conclusions[1–4]. Even within a given method or analytical technique, numerical instabilities could compromise findings[5–8]. We instrumented a structural-connectome estimation pipeline with Monte Carlo Arithmetic[9,10], a technique to introduce random noise in floating-point computations, and evaluated the stability of the derived connectomes, their features[11,12], and the impact on a downstream analysis[13,14]. The stability of results was found to be highly dependent upon which features of the connectomes were evaluated, and ranged from perfectly stable (i.e. no observed variability across executions) to highly unstable (i.e. the results contained no trustworthy significant information). The extreme range and variability in results presented here could severely hamper our understanding of brain function in brain-imaging studies. However, it also highlights potential paths forward, such leveraging this variance to reduce bias in estimates of brain connectivity. This paper demonstrates that stability evaluations are necessary as a core component of typical analytical workflows.

Keywords

Stability — Reproducibility — Network Neuroscience — Neuroimaging

[1] *Montréal Neurological Institute, McGill University, Montréal, QC, Canada*
[2] *Department of Computer Science and Software Engineering, Concordia University, Montréal, QC, Canada*
[3] *Department of Computer Science, Université of Versailles, Versailles, France*
[4] *Exascale Computing Lab, Intel, Paris, France*
[5] *Department of Psychology and eScience Institute, University of Washington, Seattle, WA, USA*
[6] *Parietal project-team, INRIA Saclay-ile de France, France*
†Authors contributed equally

The modelling of brain networks, called connectomics, has shaped our understanding of the structure and function of the brain across a variety of organisms and scales over the last decade[12,15–19]. In humans, these wiring diagrams are obtained *in vivo* through Magnetic Resonance Imaging (MRI), and show promise towards identifying biomarkers of disease. This can not only improve understanding of so-called "connectopathies", such as Alzhiemer's Disease and Schizophrenia, but potentially pave the way for therapeutics[20–24].

However, the analysis of brain imaging data relies on complex computational methods and software pipelines. Tools are trusted to perform everything from pre-processing tasks to downstream statistical evaluation. While these tools undoubtedly undergo rigorous evaluation on bespoke datasets, in the absence of ground-truth this is often evaluated through measures of reliability[25–28], proxy outcome statistics, or agreement with existing theory. Importantly, this means that tools are not necessarily of known or consistent quality, and it is not uncommon that equivalent experiments may lead to diverging conclusions[2,6–8]. While many scientific disciplines suffer from a lack of reproducibility[29], this was recently explored in brain imaging by a 70 team consortium which performed equivalent analyses and found widely inconsistent results[2].

The present study approached evaluating reproducibility from a systemic perspective in which a series brain imaging studies were numerically perturbed and the biological implications of the observed instabilities were quantified. We accomplished this through the use of Monte Carlo Arithmetic (MCA)[9], a technique which enables characterization of the sensitivity of a system to small perturbations. We explored the impact of perturbations through the direct comparision of structural connectomes, the consistency of their features, and their eventual application in a neuroscience study. Finally we conclude on the consequences of the observed instabilities and make recommendations for future work in this area.

Graphs Vary Widely With Perturbations

A subset of the Nathan Kline Institute Rockland Sample dataset[30] was randomly selected to contain 25 individuals with two sessions of imaging data, each of which was subsampled into two components, resulting in four collections per in-

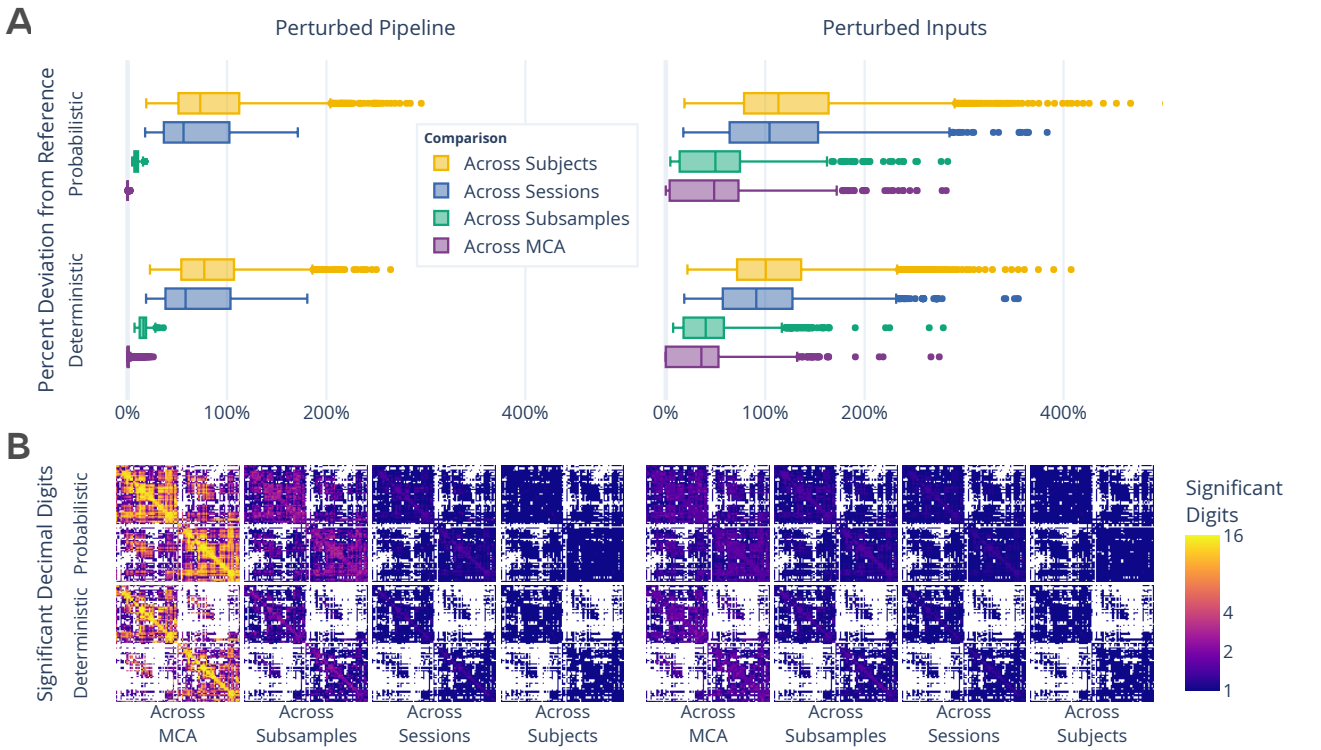

Figure 1. Exploration of perturbation-induced deviations from reference connectomes. (**A**) The absolute deviations, in the form of normalized percent deviation from reference, shown as the across MCA series relative to Across Subsample, Across Session, and Aross Subject variations. (**B**) The number of significant decimal digits in each set of connectomes as obtained after evaluating the effect of perturbations. In the case of 16, values can be fully relied upon, whereas in the case of 1 only the first digit of a value can be trusted. Pipeline- and Input-perturbations are shown on the left and right, respectively.

dividual. Structural connectomes were generated with canonical deterministic and probabilistic pipelines[31,32] which were instrumented with MCA, mimicking computational noise at either the inputs or throughout the pipelines[5,10]. The pipelines were sampled 20 times per collection and once without perturbations, resulting in a total of $4,200$ connectomes.

The stability of connectomes was evaluated through the deviation from reference and the number of significant digits (Figure 1). The comparisons were grouped according to differences across simulations, subsampling of data, sessions of acquisition, or subjects. While the similarity of connectomes decreases as the collections become more distinct, connectomes generated with Input Perturbations show considerable variability, often reaching deviations equal to or greater than those observed across individuals or sessions (Figure 1A; right). This finding suggests that instabilities inherent to these pipelines may mask session or individual differences, limiting the trustworthiness of derived connectomes for these analyses. While both pipelines show similar performance, the probabilistic pipeline is more stable in the face of Pipeline Perturbations whereas the deterministic is more stable under Input Perturbations ($p < 0.0001$ for all; exploratory). An evaluation of the stability of graph correlations can in Supplemental Section S1.

The number of significant digits per edge across connectomes (Figure 1B) similarly decreases across groups. While the cross-MCA comparison of connectomes generated Pipeline Perturbations show nearly perfect precision for many edges (approaching the maximum of 15.7 digits for 64-bit data), this evaluation uniquely shows considerable drop off in performance across data subsampling (average of < 4 digits). Input Perturbations show no more than an average of 3 significant digits across all groups. Significance across individuals did not exceed a single digit per edge in any case, indicating that only the magnitude of edges in groupwise average connectomes can be trusted. The combination of these results with those presented in Figure 1A suggests that while specific edge weights are largely affected by instabilities, macro-scale network topology is stable.

Subject-Specific Signal is Amplified While Artifacts Are Reduced

- explain discriminability as a measure of reliability
- explain hypothesis 1, result, and impact (MCA improves reliability)
- explain hypothesis 2, result, and impact (MCA reduces session-noise)

Table 1. The impact of instabilities evaluated through the separability of the dataset based on simulation, subsample, session, and subject (reported as mean ± standard deviation Discriminability). While a perfectly separable dataset would be represented by a score of 1.0, the chance performance is 1/the number of classes. In the case of Hypothesis 1, the evaluation of similarity across individuals, the chance performance is 0.04. In the case of Hypotheses 2 and 3, the evaluation of similarity across sessions or subsamples, respectively, the chance performance is 0.5. The alternative hypothesis, indicating significant separation across classes, is accepted for all experiments, with $p < 0.005$.

				Reference Execution		**Perturbed Pipeline**		**Perturbed Inputs**	
Exp.	Subj.	Sess.	Samp.	Det.	Prob.	Det.	Prob.	Det.	Prob.
1.1	All	All	1	0.64 ± 0.00	0.65 ± 0.00	0.82 ± 0.00	0.82 ± 0.00	0.77 ± 0.00	0.75 ± 0.00
1.2	All	1	All	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.02	0.90 ± 0.02
1.3	All	1	1			1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.02	0.90 ± 0.02
2.4	1	All	All	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.88 ± 0.12	0.85 ± 0.12
2.5	1	All	1			1.00 ± 0.00	1.00 ± 0.00	0.89 ± 0.11	0.84 ± 0.12
3.6	1	1	All			0.99 ± 0.03	1.00 ± 0.00	0.71 ± 0.07	0.61 ± 0.05

- explain hypothesis 3, result, and impact (MCA reduces subsample-noise)
- summarize gross impact (the reliability of brain imaging datasets is improved through the adoption of MCA)

Beyond direct difference on the connectomes, we evaluated the discriminability between groups, to characterize quantitatively the impact of perturbations on the separability of the dataset (Table 1). For hypothesis 1, which explores the separability of the dataset with respect to participant labels, an ideal dataset would have a discriminability score of 1.0. In experiment 1.1, that which mimics a typical test-retest scenario, we observe that the dataset is separable with a discriminability score of 0.64 or greater in each of these experiments (all statistically significant, $p < 0.005$). Both MCA instrumentations significantly increase the discriminability, and therefore reliability, of the dataset in this experiment ($p < 0.001$ for all). This result impactfully suggests the utility of both MCA methods for synthesizing more robust and reliable individual estimates of connectivity.

Experiment 1.1 is unsurprisingly the least discriminable test of this hypothesis, as experiments 1.2 and 1.3 rely on the same session of data, either distinguished by downsampling or perturbations, respectively. Input Perturbations lead to a decrease in the separability of individuals in these experiments, but the reliability still exceeds the cross-session case.

Hypothesis 2 considers the separability of connectomes based upon session, rather than subject. In this case, performance was computed within-individual and aggregated. An ideal test-retest dataset – one where there is no difference between two observations of an individual – would have a discriminability score of 0.5 in these experiments.

The discriminability in both the unperturbed and Perturbed Pipeline settings is 1.0, meaning that there is a significant difference between sessions of data. However, while still significant relative to chance, Input Perturbations lead to considerably lower discriminability, reducing the difference between distinct sessions of data. This suggests that Input Perturbations reduce session-dependent bias.

Finally, experiment 3.6 evaluates the separability of samples based on subsampling. Similarly to the previous, the performance is once again computed through analyzing individual sessions of data and aggregating across sessions and individuals, with an ideal score of 0.5. While this experiment could not be evaluated using reference executions, the Pipeline Perturbation performance showed near perfect separation between direction subsamples whereas Input Perturbations considerably lower this separability towards chance, similar to as was observed in experiments 2.4 and 2.5. This is further evidence which suggests that the Input Perturbations may have an application in obtaining more robust estimates of individual connectivity, as across each experiment it shows an amplification in meaningful differences while also showing a reduction in off-target differences.

Distributions of Graph Statistics Are Reliable, Individual Statistics Are Not

- define measures and cite their common use
- define comparison of distributions and discuss implication (distributions are stable)
- define comparison of moments (individual statistics are unstable)
- summarize impact (group-wise distributions should be relied upon rather than individual differences)

Connectomes are often summarized by lower-dimensional statistics more suitable for numerous analytical methods[12]. Figure 2 explores the stability of these graph-theoretical metrics computed from the perturbed graphs, including weight, clustering coefficient, path length, betweenness centrality, and degree. Due to the variable length of the edgewise statistics, cumulative density functions for each statistic were evaluated over a fixed range and the mean density and associated standard error were computed for each bin (Figures 2A and

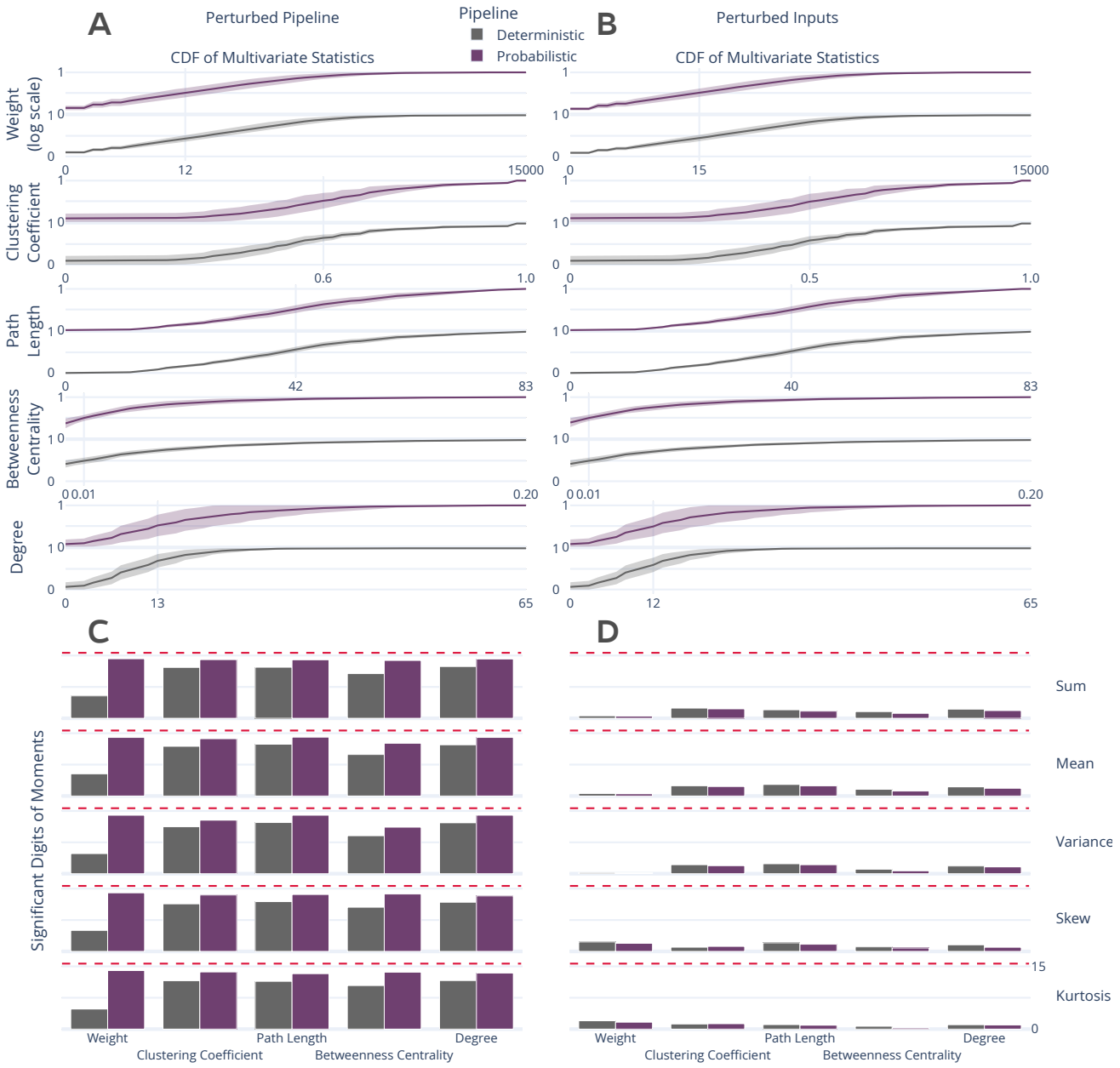

Figure 2. Distribution and stability assessment of multivariate graph statistics. (**A**, **B**) The cumulative distribution functions of multivariate statistics across all subjects and perturbation settings. There was no significant difference between the distributions in A and B. (**C**, **D**) The number of significant digits in the first five moments of each statistic across perturbations. The dashed red line refers to the maximum possible number of significant digits.

2B), with the distributions' minimum, median, and maximum values denoted on each x-axis. There was no significant difference in distributions observed for each statistic across the two perturbation settings. The first 5 moments of these statistics within individuals as observed with Pipeline Perturbations (Figure 2C) were stable with more than 10 significant digits with the exception of edge weight when using the deterministic pipeline. In the case of all statistics, the probabilistic pipeline was more stable than the deterministic pipeline ($p < 0.0001$; exploratory). In stark contrast, these moments were highly unstable in the face of Input Perturbations (Figure 2D), in which no measure had more than 5 significant digits of information, and several moment and statistic pairs had less than a single significant digit, such as the variance in edge weight or the kurtosis of betweenness centrality. In general, there was not a strong relationship between the order of the moment and its stability. A similar analysis was performed for univariate statistics in Supplemental Section S2.

The large discrepancy between the stability of individual estimates in these settings versus the similarity of aggregated

CDFs suggests that while individual estimates are unstable, the comparison between aggregates or groups may be considered much more reliable.

The Strength of Brain-Behaviour Relationships is Eroded

- define classification task and cite justification
- state reference performance and distribution of perturbed performances
- summarize impact (the strength of biological signal is highly unstable and erodes confidence in models of brain behaviour relationships)

While the variability of explicit features of connectomes was summarized above, these networks are commonly-used as inputs to machine learning models. Here, connectomes were projected into a low dimensional space using PCA and then input a logistic regression classifier (Figure 3). The number of principal components was selected as the minimum number of components required to capture 90% of the variance in the reference set; this resulted in 20 components. Using the reference performance, i.e. that using only unperturbed graphs (Figure 3; orange overlay), the classification accuracies were 0.635 and 0.628, and the F1 scores were 0.636 and 0.630 for data derived using the deterministic and probabilistic pipelines, respectively, with the average explained variance at 90% in both cases. The random chance performance for these evaluations measures were 0.521 and 0.519, respectively (Figure 3; dashed red line).

When performing this analysis using sampled instances of the perturbed dataset across both pipelines and perturbation methods, the portion of explained variance in the sample with 20 components ranged from 0.886 -– 0.978. The classification accuracy ranged from 0.520 – 0.716 and the F1 score ranged from 0.510 -– 0.725. These results range from at or below random chance performance, to considerable accuracy that outperforms that obtained using the reference dataset.

Discussion

This paper explores the impact of perturbing a structural connectome estimation pipeline with small amounts of noise, on the order of machine error. These impacts were explored through three distinct comparisons of the derived connectomes: comparison of networks directly, comparison of commonly used network features, and comparison of the performance of a downstream analytical task. In each analysis we found considerable variability across the perturbed executions. Perhaps more notably, we found that the variability was itself highly variable across data, perturbation, feature, and measure. In each exploration we found that the stability of results ranged from nearly perfectly trustworthy (i.e. no variation or fully significant) to untrustworthy often to the level of only two or three significant digits of information, and occasionally with no significant digits at all.

The adoption of Monte Carlo Arithmetic for this application allowed for not only the characterization of concrete measures of stability within individual samples, such as the number of significant digits in edges of a connectome or a derived statistic, but allowed these values to be situated with respect to other biologically meaningful sources of variability such as cross-session or cross-individual variation. Importantly, the evaluation of the stability of graphs in the cross-direction subsampling setting would not have been possible without either a realistic noise model or a simulation technique similar to MCA in which multiple results are derived from a single sample of input data.

While both MCA approaches identified instabilities in the tested pipelines, in the case of the Test-ReTest analysis, the adoption of MCA improved the discriminability, and thus the reliability of the results. In the case of traditional cross-individual comparison, both perturbation methods improved the reproducibility of the dataset significantly, increasing the separability of individuals from 0.64– to 0.82 and 0.77 for the deterministic pipeline for Pipeline and Input Perturbation methods, respectively, and similarly for the probabilistic pipeline. This suggests that MCA may be a cost effective and context-agnostic method for data augmentation.

Conversely to the increase in reliability across individuals, in the case when discrimination between classes was not desired (i.e. cross-session and within-session cross-subsample comparisons), Input Perturbations in particular led to the reduced separability of the dataset. For this discrimination task, a well-conditioned system would provide results that would closely resemble one another, such that the separability of the classes is low. The fact that this was not the case suggests that the system (i.e. tool and data combination) was unstable, leading to results which were distinct without biologically-meaningful variability. The repeated generation of connectomes with Input Perturbed MCA in this experiment reduces this separability, however, showing that fixed biases in this system were partially responsible for the separability. The discriminability of the cross-direction group in particular shifted dramatically towards an ideal random-chance value. This result highlights that MCA can be used not only to identify instabilities within pipelines but also as a technique for the generation of distributions of plausible results which may be aggregated or considered en masse.

The important difference of this final discriminability experiment compared to the previous ones is that there is no variability in data acquisition or preprocessing that contribute to the observed differences. This means that the separability observed in this experiment is only due to the modeling pipelines. While data and processing tools are inseparable when performing a stability evaluation, the tiered experiment approach performed here allows for the characterization of this relationship at different scales and of independent pipeline components.

One emergent theme across all analyses was that Input Perturbations led to considerably less stable than Pipeline

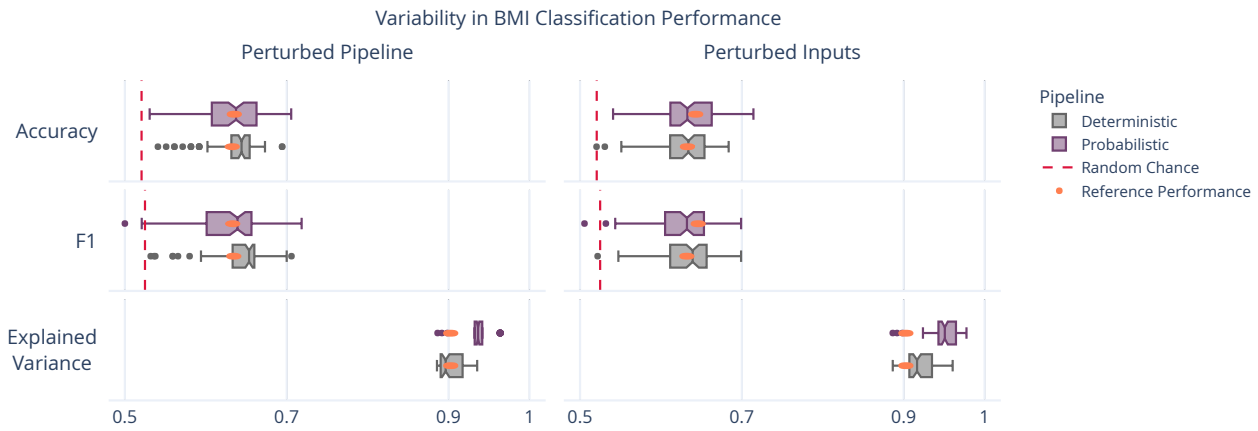

Figure 3. Observed variability in BMI classification. Training and Test sets were sampled from the MCA-generated dataset such that a single observation of each individual was present in each sampling. This sampling was performed 20 times, and each dataset was used to train a classifier with each of 2, 5, 10, and N-fold cross validation, and the shown metrics are the average across each of these training paradigms. The dashed red lines indicate random-chance performance, and the orange dots show the performance using the reference executions.

Perturbations, confirming observations in [10]. However, the distributions of connectomes and their statistics were not distinct from one another. This is strong support for the group-wise evaluation of graph statistics over the use of individual estimates for applications in brain imaging.

The difference in stability between the two instrumentations opens the door for many avenues for future work. The two implementations of MCA were determined, for practical reasons, as the boundary between software libraries; Pipeline Perturbations included a much more complete instrumentation, resulting in a higher density of perturbations. We hypothesize that the increased density of Pipeline Perturbations allow for the law of large numbers to mask critically unstable points within the pipeline, whereas errors introduced in the Input Perturbation implementation are allowed to propagate longer before correction. This question is non-trivially answered as understanding the true density of "coverage" of an arbitrary pipeline is impossible without direct assessment. Once coverage is evaluated for a given tool, instrumentations perturbing different degrees of the pipeline have potential to shed light on the relationship between the magnitude, frequency, and density of perturbations on the observed stability, as well as reduce the computational burden of MCA.

Another observation which became apparent across the analyses was that variability changed significantly based on measure or resolution. For instance, graphs generated with Input Perturbations were perceived to be significantly more similar to one another when compared with correlation compared to percent deviation or significant digits measures. Importantly, this does not imply correlation as a superior statistic or means for the comparison of graphs, but just that it is more robust, or less sensitive, under these circumstances. If an initial evaluation of these graphs were performed solely using correlation, it would perhaps lead the researcher to believe that their measures are perfectly stable. However, if their downstream analysis depended on one of the statistics demonstrated through Figures 2 and 3 to be highly unstable, this assumption would be incorrect. What this suggests is rather that the stability of any measures intended to be used in an analysis should be quantified, as the stability of one measure cannot be perfectly inferred from another.

A Limit in Detection of Individual Differences

The finding that individually derived network statistics were unreliable in the Input Perturbations setting despite resulting in no aggregated change has important implications on applications of connectomics. This finding bounds the success of studying individual differences, a central objective in brain imaging (Dubois and Adolphs 2016). More specifically, the relationships found between pheno/genotypic data and these statistics or connectomes in general will be limited by the reliability of the statistics themselves. While this has obvious implications in machine-learning applications such as the classification task studied above, this limitation extends to hypothesis testing as well. Though the result of an individual comparison in a hypothesis test will have a reported false-positive rate, the accuracy of this rate is dependent on the reliability of the sample used. For instance, if MCA were performed and each single observation of a highly unstable sample were used in a hypothesis test, the true false positive rate would be a combination of the variability in the results of that hypothesis test and the reported rate under the parameters of the model. Without a repeated-measure setting such as MCA, it is impossible to empirically estimate the reliability of the samples used in hypothesis testing, meaning that the reliability of accepted hypotheses is unknown, regardless of the reported false positive rate. In fact, it is a virtual certainty that

the true false positive rate for a given hypothesis exceeds that reported by the terminal test simply as a result of numerical instabilities.

Eroding the Biological Plausibility of Derived Connectomes

One important insight uniquely provided by the classification task performed above is with regards to the biological plausibility of derived connectomes. While previous analyses explored the variability of the connectomes or their features, with the exception of comparisons to cross-subject variability these values were not situated with respect to meaningful biological differences. The results shown here demonstrate that our MCA-based perturbations not only distort results, but do so while retaining meaningful signal rather than merely degrading our estimates. This analysis allows us to demonstrate considerable variability in classification performance, in which we observe the reference execution to be approximately centered in the distribution of results. Critically, while MCA perturbations are data-agnostic, this result is strong evidence in favour of the quality of derived connectomes observed using MCA, suggesting that it can be used effectively in conjunction with, or in lieu of, context-aware noise models, as is often the case in brain imaging.

Shortcomings and Future Questions

A limitation of the approach presented in this paper is the difficulty of instrumentation for arbitrary libraries with MCA. This is non-trivial for many tools due to a dependence on recompilation under specific settings. For this reason all modeling pipelines tested here were constructed using Dipy (Garyfallidis et al. 2014), a canonical Python library which provides accessible implementations of commonly-used algorithms in tractography. Pre-processing was not perturbed in these experiments. Other work has shown that linear registration is sensitive to minor perturbations, a prerequisite and core piece of many elements of pre-processing such as motion correction and alignment [8]. In practice, this suggests that the instabilities observed here would be compounded with those introduced in pre-processing.

Additionally, the analyses performed in this paper were using a single dataset and a single set of similar analysis pipelines. While the conclusions stated here have been observed in other modalities and tools used in neuroimaging through other means, the nature of this work makes it impossible to know how reliable other tools or datasets are when perturbed and evaluated similarly. Extending this work to both typical fMRI and structural MRI workflows is of interest, and the topic of future projects.

This paper also does not address methodological flexibility or compare this to the observed instability. As was demonstrated by NARPS in a task-based fMRI analysis setting [26], there exists a nearly boundless space of possible combinations of tools scientists could choose to compose their pipelines. As a result, research groups often end up using unique processing pipelines and infrastructures, which may lead to divergent conclusions even in the face of modeling identical datasets. The evaluation performed here does not compare observed the impact of MCA-induced instabilities to variability of this sort. In fact, these studies explore instability at the opposite ends of the analysis spectrum: from human introduced variability in the conceptualization and realization of an analysis workflow on the one hand, down to the unavoidable error implicit in the digital representation of data. It is of extreme interest to combine these approaches and explore the interaction of these scientific degrees of freedom with effects from software implementations, libraries, and parametric choices.

Finally, it is important to state explicitly that the work presented here does not invalidate the analytical pipelines but merely state that many studies are accompanied by an unknown degree of uncertainty due to machine-introduced errors. The desired outcome of this paper is to motivate a shift in scientific computing – particularly in neuroimaging – towards a paradigm which values the explicit evaluation of the trustworthiness of claims alongside the claims themselves.

Notes

You could mention that instabilities affect not only the topology, but the geometry of reconstructed networks.

It means that we are, as a field, overconfident in numerical stability, and that in a typical setting this would lead to wrong conclusions. Or am I over-stating it? Because if that's the case, it needs to be unequivocally stated here. And in the abstract. And maybe also in the title of the article. this amount of numerical noise is to be expected in typical settings, not only in MCA. Is that correct? And that has implications for inferences in the typical setting.

References

[1] D. Glen, P. Taylor, J. Seidlitz, M. Glen, C. Liu, P. Molfese, and R. Reynolds, "Through thick and thin: Measuring thickness in MRI with AFNI," *Annual Meeting of the Organization for Human Brain Mapping*, 2018.

[2] R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock *et al.*, "Variability in the analysis of a single neuroimaging dataset by many teams," *Nature*, pp. 1–7, 2020.

[3] C. M. Bennett, M. B. Miller, and G. L. Wolford, "Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction," *Neuroimage*, vol. 47, no. Suppl 1, p. S125, 2009.

[4] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates," *Proceedings of the national academy of sciences*, vol. 113, no. 28, pp. 7900–7905, 2016.

[5] G. Kiar, P. de Oliveira Castro, P. Rioux, E. Petit, S. T. Brown, A. C. Evans, and T. Glatard, "Comparing perturbation models for evaluating stability of neuroimaging pipelines," p. 109434202092623, 2020.

[6] A. Salari, G. Kiar, L. Lewis, A. C. Evans, and T. Glatard, "File-based localization of numerical perturbations in data analysis pipelines," *arXiv preprint arXiv:2006.04684*, 2020.

[7] L. B. Lewis, C. Y. Lepage, N. Khalili-Mahani, M. Omidyeganeh, S. Jeon, P. Bermudez, A. Zijdenbos, R. Vincent, R. Adalat, and A. C. Evans, "Robustness and reliability of cortical surface reconstruction in CIVET and FreeSurfer," *Annual Meeting of the Organization for Human Brain Mapping*, 2017.

[8] T. Glatard, L. B. Lewis, R. Ferreira da Silva, R. Adalat, N. Beck, C. Lepage, P. Rioux, M.-E. Rousseau, T. Sherif, E. Deelman, N. Khalili-Mahani, and A. C. Evans, "Reproducibility of neuroimaging analyses across operating systems," *Front. Neuroinform.*, vol. 9, p. 12, Apr. 2015.

[9] D. S. Parker, *Monte Carlo Arithmetic: exploiting randomness in floating-point arithmetic*. University of California (Los Angeles). Computer Science Department, 1997.

[10] C. Denis, P. de Oliveira Castro, and E. Petit, "Verificarlo: Checking floating point accuracy through monte carlo arithmetic," 2016.

[11] R. F. Betzel, A. Griffa, P. Hagmann, and B. Misic, "Distance-dependent consistency thresholds for generating group-representative structural brain networks," *bioRxiv*, 2018.

[12] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, Sep. 2010.

[13] B.-Y. Park, J. Seo, J. Yi, and H. Park, "Structural and functional brain connectivity of people with obesity and prediction of body mass index using connectivity," *PLoS One*, vol. 10, no. 11, p. e0141376, Nov. 2015.

[14] A. Gupta, E. A. Mayer, C. P. Sanmiguel, J. D. Van Horn, D. Woodworth, B. M. Ellingson, C. Fling, A. Love, K. Tillisch, and J. S. Labus, "Patterns of brain structural connectivity differentiate normal weight from overweight subjects," *Neuroimage Clin*, vol. 7, pp. 506–517, Jan. 2015.

[15] T. E. Behrens and O. Sporns, "Human connectomics," *Current opinion in neurobiology*, vol. 22, no. 1, pp. 144–153, 2012.

[16] M. Xia, Q. Lin, Y. Bi, and Y. He, "Connectomic insights into topologically centralized network edges and relevant motifs in the human brain," *Frontiers in human neuroscience*, vol. 10, p. 158, 2016.

[17] J. L. Morgan and J. W. Lichtman, "Why not connectomics?" *Nature methods*, vol. 10, no. 6, p. 494, 2013.

[18] M. P. Van den Heuvel, E. T. Bullmore, and O. Sporns, "Comparative connectomics," *Trends in cognitive sciences*, vol. 20, no. 5, pp. 345–361, 2016.

[19] J. Dubois and R. Adolphs, "Building a science of individual differences from fMRI," *Trends Cogn. Sci.*, vol. 20, no. 6, pp. 425–443, Jun. 2016.

[20] A. Fornito and E. T. Bullmore, "Connectomics: a new paradigm for understanding brain disease," *European Neuropsychopharmacology*, vol. 25, no. 5, pp. 733–748, 2015.

[21] G. Deco and M. L. Kringelbach, "Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders," *Neuron*, vol. 84, no. 5, pp. 892–905, 2014.

[22] T. Xie and Y. He, "Mapping the alzheimer's brain with connectomics," *Frontiers in psychiatry*, vol. 2, p. 77, 2012.

[23] M. Filippi, M. P. van den Heuvel, A. Fornito, Y. He, H. E. H. Pol, F. Agosta, G. Comi, and M. A. Rocca, "Assessment of system dysfunction in the brain through mri-based connectomics," *The Lancet Neurology*, vol. 12, no. 12, pp. 1189–1199, 2013.

[24] M. P. Van Den Heuvel and A. Fornito, "Brain networks in schizophrenia," *Neuropsychology review*, vol. 24, no. 1, pp. 32–48, 2014.

[25] J. J. Bartko, "The intraclass correlation coefficient as a measure of reliability," *Psychol. Rep.*, vol. 19, no. 1, pp. 3–11, Aug. 1966.

[26] A. M. Brandmaier, E. Wenger, N. C. Bodammer, S. Kühn, N. Raz, and U. Lindenberger, "Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED)," *Elife*, vol. 7, Jul. 2018.

[27] E. W. Bridgeford, S. Wang, Z. Yang, Z. Wang, T. Xu, C. Craddock, J. Dey, G. Kiar, W. Gray-Roncal, C. Coulantoni *et al.*, "Eliminating accidental deviations to minimize generalization error: applications in connectomics and genomics," *bioRxiv*, p. 802629, 2020.

[28] G. Kiar, E. Bridgeford, W. G. Roncal, V. Chandrashekhar, and others, "A High-Throughput pipeline identifies robust connectomes but troublesome variability," *bioRxiv*, 2018.

[29] M. Baker, "1,500 scientists lift the lid on reproducibility," 2016.

[30] K. B. Nooner, S. J. Colcombe, R. H. Tobe, M. Mennes, M. M. Benedict, A. L. Moreno, L. J. Panek, S. Brown, S. T. Zavitz, Q. Li, S. Sikka, D. Gutman, S. Bangaru, R. T. Schlachter, S. M. Kamiel, A. R. Anwar, C. M. Hinz, M. S. Kaplan, A. B. Rachlin, S. Adelsberg, B. Cheung, R. Khanuja, C. Yan, C. C. Craddock, V. Calhoun, W. Courtney, M. King, D. Wood, C. L. Cox, A. M. C. Kelly, A. Di Martino, E. Petkova, P. T. Reiss, N. Duan, D. Thomsen, B. Biswal, B. Coffey, M. J. Hoptman, D. C. Javitt, N. Pomara, J. J. Sidtis, H. S. Koplewicz, F. X. Castellanos, B. L. Leventhal, and M. P. Milham, "The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry," *Front. Neurosci.*, vol. 6, p. 152, Oct. 2012.

[31] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. van der Walt, M. Descoteaux, I. Nimmo-Smith, and Dipy Contributors, "Dipy, a library for the analysis of diffusion MRI data," *Front. Neuroinform.*, vol. 8, p. 8, Feb. 2014.

[32] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams, and I. Nimmo-Smith, "QuickBundles, a method for tractography simplification," *Front. Neurosci.*, vol. 6, p. 175, Dec. 2012.

Methods

$$\cos^3\theta = \frac{1}{4}\cos\theta + \frac{3}{4}\cos 3\theta \tag{1}$$

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

1. First item in a list
2. Second item in a list
3. Third item in a list

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

- First item in a list
- Second item in a list
- Third item in a list

Author Contributions

GK was responsible for the experimental design, data processing, analysis, interpretation, and the majority of writing. All authors contributed to the revision of the manuscript. YC, POC, and EP were responsible for MCA tool development and software testing. AR, GV, and BM contributed to experimental design and interpretation. TG contributed to experimental

design, analysis, and interpretation. TG and ACE were responsible for supervising and supporting all contributions made by GK. The authors declare no competing interests for this work.

Acknowledgments

This research was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (award no. CGSD3-519497-2018). This work was also supported in part by funding provided by Brain Canada, in partnership with Health Canada, for the Canadian Open Neuroscience Platform initiative.

Additional Information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Tristan Glatard at tristan.glatard@concordia.ca.

S1. Graph Correlation

The correlations between observed graphs (Figure 1B) across each grouping follow the same trend to percent deviation. However, notably different from percent deviation, there is no significant difference in the correlations between Pipeline or Input instrumentations. By this measure, the probabilistic pipeline is more stable in all cross-MCA and cross-directions except for the combination of Input Perturbation and cross-MCA ($p \not< 0.0001$ for all; exploratory).

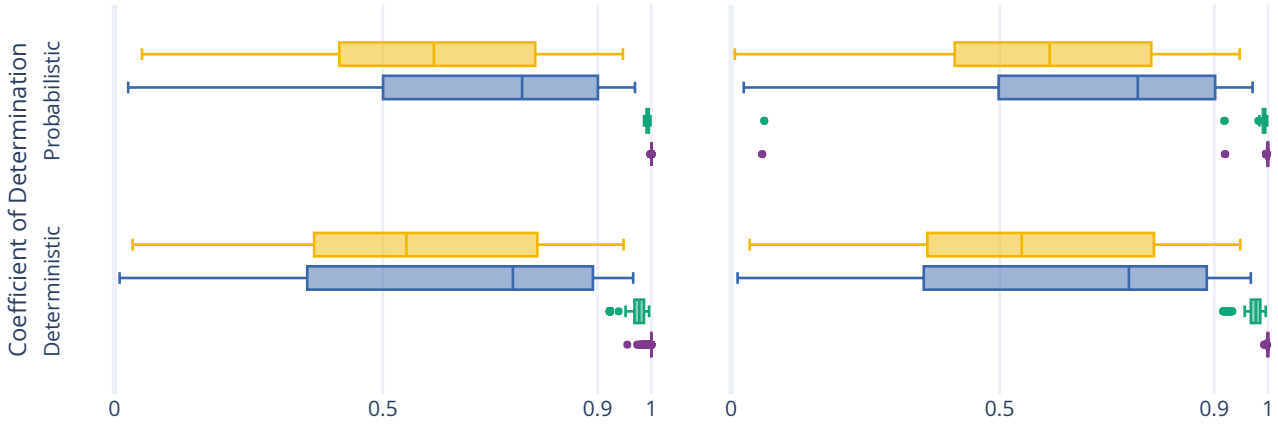

Figure S1. The correlation between perturbed connectomes and their reference.

S2. Univariate Graph Statistics

Figure 2 explores the stability of these graph-theoretical metrics computed from the perturbed graphs, including modularity, global efficiency, assortativity, average path length, and edge count. When aggregated across individuals and perturbations, the distributions of these statistics (Figures 2A and 2B) show no significant differences between perturbation methods for either deterministic or probabilistic pipelines. However, when quantifying the stability of these measures across connectomes derived from a single session of data, the two perturbation methods show considerable differences. The number of significant digits in univariate statistics for Pipeline Perturbation instrumented connectome generation exceeded 11 digits for all measures except modularity, which contained more than 4 significant digits of information (Figure 2C). When detecting outliers from the distributions of observed statistics for a given session, the false positive rate (using a threshold of $p = 0.05$) was approximately 2% for all statistics with the exception of modularity which again was less stable with an approximately 10% false positive rate. The probabilistic pipeline is significantly more stable than the deterministic pipeline ($p ¡ 0.0001$; exploratory) for all features except modularity. When similarly evaluating these features from connectomes generated in the Input Perturbation setting, no statistic was stable with more than 3 significant digits or a false positive rate lower than nearly 6% (Figure 2D). The deterministic pipeline was more stable than the probabilistic pipeline in this setting ($p ¡ 0.0001$; exploratory).

Two notable differences between the two perturbation methods are, first, the uniformity in the stability of the statistics, and second, the dramatic decline in stability of individual statistics in the Input Perturbation setting despite the consistency in the overall distribution of values. It is unclear at present if the discrepancy between the stability of modularity in the Pipeline Perturbation context versus the other statistics suggests the implementation of this measure is the source of instability or if it is implicit to the measure itself. The dramatic decline in the stability of features derived from Input Perturbed graphs despite no difference in their overall distribution both shows that while individual estimates may be unstable the comparison between aggregates or groups may be considered much more reliable.

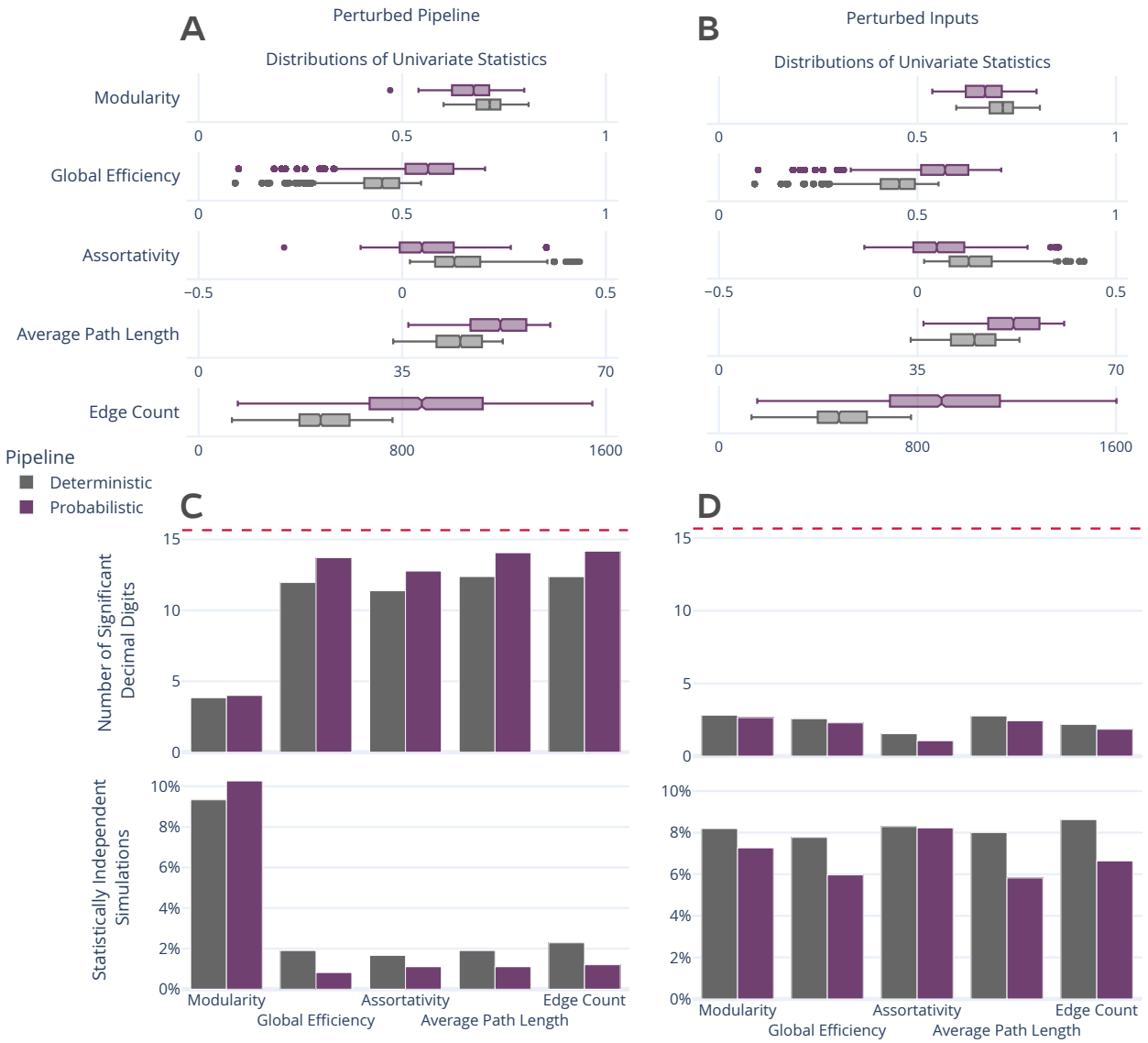

Figure S2. Distribution and stability assessment of univariate graph statistics. (**A**, **B**) The distributions of each computed univariate statistic across all subjects and perturbations for Pipeline and Input settings, respectively. There was no significant difference between the distributions in A and B. (**C**, **D**; top) The number of significant decimal digits in each statistic across perturbations, averaged across individuals. The dashed red line refers to the maximum possible number of significant digits. (**C**, **D**; bottom) The percentage of connectomes which were deemed significantly different ($p < 0.05$) from the others obtained for an individual.