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 [1–6]. Even within a given method or analytical technique, numerical instabilities could compromise findings. We instrumented a structural-connectome estimation pipeline with Monte Carlo Arithmetic [7, 8] and evaluated the stability of the derived connectomes, their features [9], and the impact on a downstream analysis [10,11]. 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. It also highlights both the potential impact of basic analytical choices and measure on the reliability of downstream analyses, and the necessity of stability evaluation 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*
***Corresponding author**: tristan.glatard@concordia.ca
†Authors contributed equally

Brain imaging relies on complex computational methods and processing pipelines. Tools are trusted to perform everything from pre-processing tasks (i.e. image reconstruction, denoising, registration, and sample imputation) to downstream modeling and statistical evaluation. While these tools undoubtedly undergo rigorous evaluation and quality assurance by developers to ensure their validity, in the absence of ground-truth this validity is often evaluated by proxy outcome statistics and agreement with previously existing literature and theory. Tests are inevitably performed on small datasets, which are known to probe several expected conditions of a tool and are executed on closely managed computational infrastructures. While this is practical from an engineering point of view, it results in a lack of biological and computational heterogeneity in tool development.

The assumption that tools perform equivalently across populations, infrastructures, and other off-target factors (i.e. the odd/evenness of voxels in an image [1]) is fundamental to their usefulness, but remains largely untested. Yet, it is possible to quantify the stability of tools with respect to a variety of these factors through both reliability evaluation and perturbation paradigms.

Measures such as Intra-Class Correlation (ICC) [12], Intra-Class Effect Decomposition (ICED) [13], and Discriminability [14] have been used to evaluate the similarity between observations within test-retest datasets (i.e. scans originating from the same versus different individuals). While some combinations of pipeline and dataset have been shown to achieve perfectly reliable estimates, for instance in functional or diffusion connectivity, this is often not the case [14, 15]. Importantly, even when this is the case, it does not necessarily imply that the estimates are of high quality (i.e. a result simply encoding numeric subject identifiers would achieve perfect reliability, but would not be biologically meaningful).

Recently, the impact of analytic choices on the reliability of results was assessed [2]. 70 different teams did their best-effort analysis of the same task-based fMRI analysis. Huge variability was observed in the findings and their confidence across research teams. This quantified the impact of scientists and their analytical decisions on the reliability of results.

On the opposite end of the spectrum, perturbation methods allow for the evaluation of the stability of a tool for

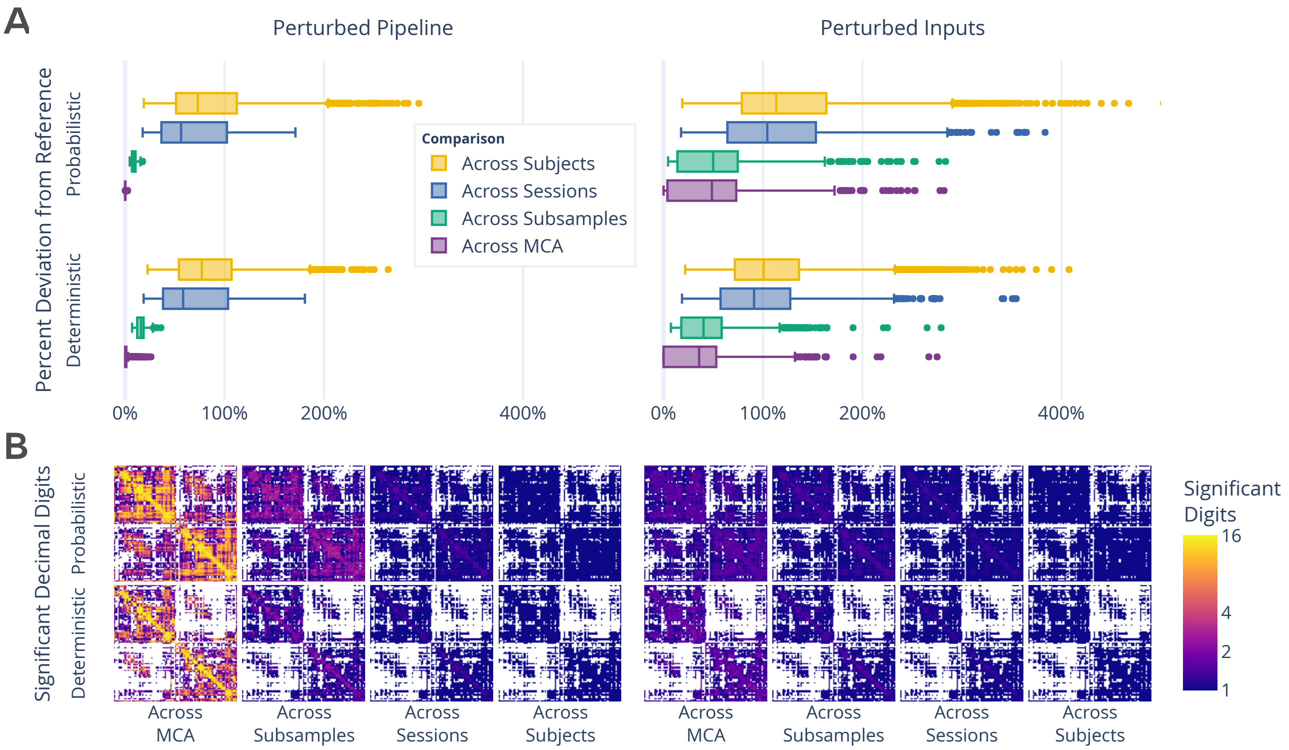

Figure 1. Exploration of perturbation-induced deviations from reference connectomes. (**A**) The absolute deviations, in the form of normalized percent deviation from reference, shown as the across MCA series relative to Across Subsample, Across Session, and Aross Subject variations. (**B**) The number 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.

datasets without repeated measures. This process involves modifying and repeating analyses across theoretically unimportant deviations, and allows for quantification of a tool's (in)stability to these changes. A "one-voxel" perturbation approach was used to study the stability of cortical surface estimation pipelines [3]. A single voxel was perturbed by 1% intensity prior to brain-surface extraction with two different tools and the resulting surfaces were compared across the original and perturbed datasets. Both for Gray and White Matter surfaces, the two tools showed significant deviations. Another perturbation approach explored the impact of the operating system on the Human Connectome Project (HCP) pipelines [16]. Structural segmentations derived using the same hardware, dataset, and pipeline package versions showed considerable differences across systems using CentOS6 and CentOS7 [4, 5].

While these previous studies shed important light on inconsistencies in results derived by commonly-used pipelines, their conclusions are subject to change depending on details irrelevant to the tool, dataset, or task being evaluated, such as which researchers are included, which voxel is perturbed, or which operating systems are used, respectively. This limits their ability to conclude on the underlying stability of a specific analytical tool evaluated. Classical numerical analysis might frame the stability of methods in terms of "conditioning", referring to the sensitivity of a system to small perturbations. However, it is difficult to directly apply it in brain imaging as the magnitude and application of an "insignificant perturbation" is unclear.

An alternative approach for studying instability, Monte Carlo Arithmetic (MCA), involves modifying mathematical operations containing floating-point numbers by adding uniform random noise simulating round-off errors at a given precision [7]. Not only can this approach be applied arbitrarily across tools, lowering the needed level of manual intervention, but also it perturbs pipeline operation by a known and customizable magnitude. A preliminary study usedMCA to evaluate the stability of a structural connectome estimation pipeline [6]. This method revealed that differences between connectomes generated from a single session of data can be as high as cross-subject level variation. While this work evaluated the variability in derived connectomes, the effect of such variability on final outcomes of interest to neuroscience is currently unclear.

The present study quantifies the numerical, analytical, and biological implications of instabilities in neuroimaging. We accomplish this by evaluating the stability of two pipelines that estimate structural connectomes, instrumented with MCA, and characterize the direct and downstream impact of observed instabilities. We perform a direct evaluation of stability by looking at the introduced variability in observed graphs and relating this to observed values of biological importance such as between subject and between session variability.

The impact on downstream analyses is first explored through curated summary statistics and estimates of their variance across perturbations. Then, we perform a neuroscience study on the derived connectomes in which we classify individuals with respect to high (> 25) or low BMI and quantify the relative performance of our model on this task across simulations.

Finally we conclude on the impact of induced instabilities on these measures, and make recommendations for future work in this area.

Graphs Vary Widely With Perturbations

Figure 1 shows the observed changes in structural connectivity induced by perturbations using two metrics: percent deviation, and number of significant digits. The deviations observed purely across MCA were displayed alongside other sources of variability, in order of magnitude, such as: subsampling, variation across sessions, and variation across subjects. In both the case of Pipeline Perturbation and Input Perturbation settings, the magnitude of percent deviation between pairs of connectomes (Figure 1A) across each category is consistent with this sorting. However, there is a much tighter bound in the distribution of differences in both the cross-MCA and cross-subsample cases with Pipeline Perturbations, where deviations rarely reach the level of session or individual variations. Connectomes generated with Input Perturbations show considerably more variability, often reaching deviations equal to or greater than those observed across individuals or sessions. While both pipelines show similar distributions in both perturbation settings, the probabilistic pipeline is more stable for cross-MCA and cross-subsample evaluations in the face of Pipeline Perturbations whereas the deterministic is more stable under Input Perturbations ($p < 0.0001$ for all; exploratory).

The number of significant digits per edge across connectomes (Figure 1B) similarly follows the expected and previously observed trend across groups. While the cross-MCA Pipeline Perturbation evaluations show nearly perfect precision (approaching 16 digits) for many edges, this evaluation uniquely shows considerable drop off in performance with the cross-subsample group (average of < 4 digits). The combination of this with results presented in Figures 1A suggests that specific edge weights are largely affected by these perturbations while macro-scale connectivity is largely unchanged. Connectomes perturbed by Input Perturbations show no more than an average of 3 significant digits across all groups. In the case of both perturbations, cross-individual significance does not exceed a single digit per edge, suggesting that groupwise average connectomes should be limited to a single digit of precision (i.e. only the order of magnitude of the edge can be relied upon).

The fact that each of these direct comparisons show distinct relationships between pipeline, perturbation mode, and data, illustrates the high dependence of stability evaluation upon specific application context.

Subject-Specific Signal is Amplified While Artifacts Are Reduced

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

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

amplification in meaningful differences while also showing a reduction in off-target differences.

Distributions of Graph Statistics Are Reliable, Individual Statistics Are Not

Connectomes are often summarized by lower-dimensional statistics more suitable for numerous analytical methods [9]. 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 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 Are Highly Variable

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.

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

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula

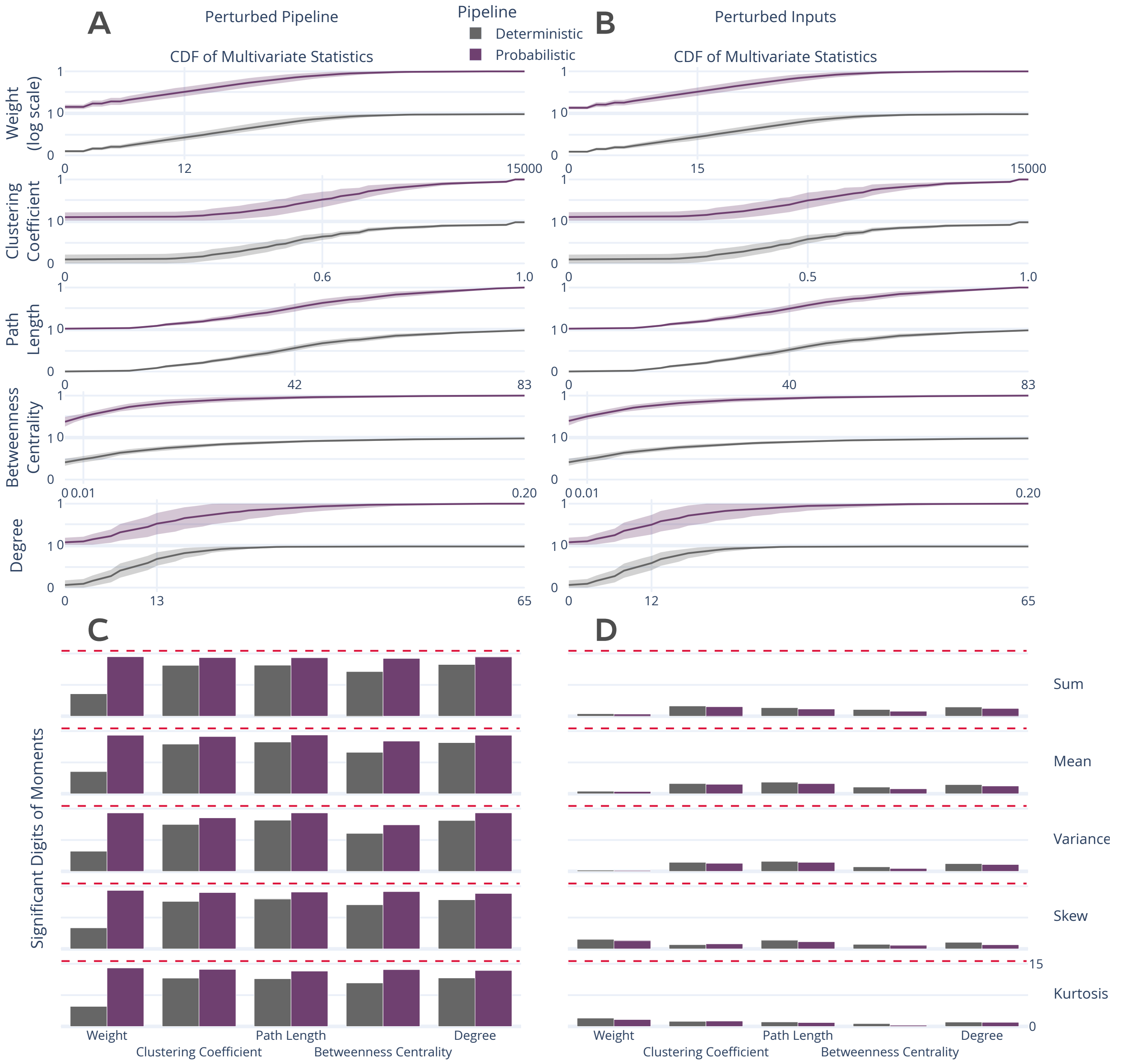

Figure 2. Distribution and stability assessment of multivariate graph statistics. (**A**, **B**) The cumulative distribution functions of multivariate statistics across all subjects and perturbation settings. There was no significant difference between the distributions in A and B. (**C**, **D**) The 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.

libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

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

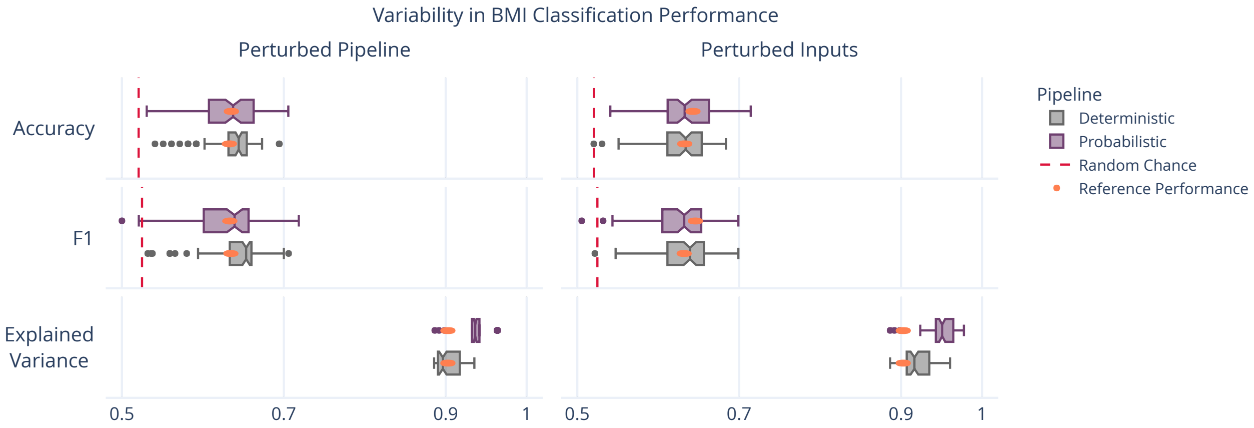

Figure 3. Observed variability in BMI classification. Training and Test sets were sampled from the MCA-generated dataset such that a single observation of each individual was present in each sampling. This sampling was performed 20 times, and each dataset was used to train a classifier with each of 2, 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.

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

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] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

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

[16] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, M. Jenkinson, and WU-Minn HCP Consortium, "The minimal preprocessing pipelines for the human connectome project," *Neuroimage*, vol. 80, pp. 105–124, Oct. 2013.

Methods

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.

Author Contributions

GK was responsible for the experimental design, data processing, analysis, interpretation, and the majority of writing. All authors contributed to the writing 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 ¡ 0.0001 for all; exploratory).

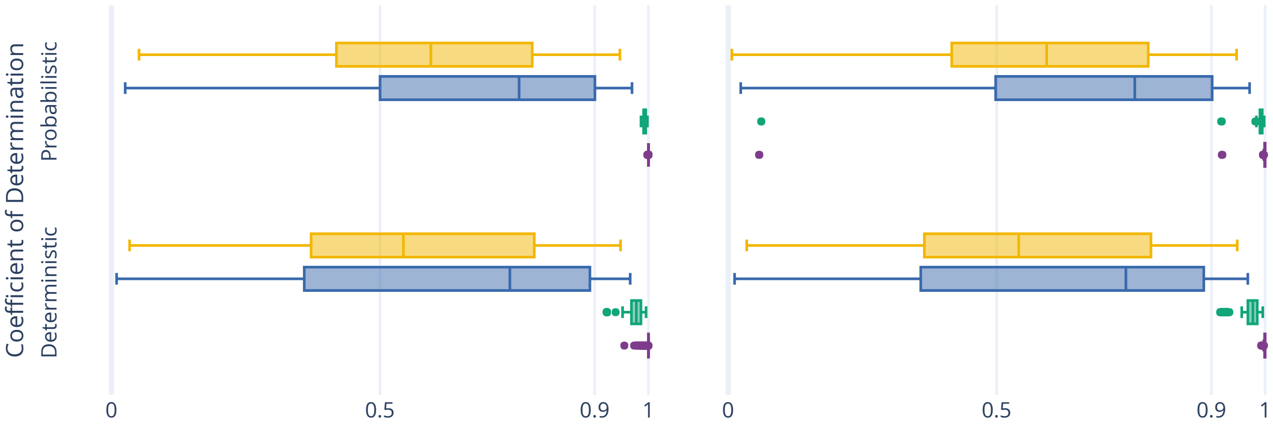

Figure S1. The correlation between perturbed connectomes and their reference.

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.

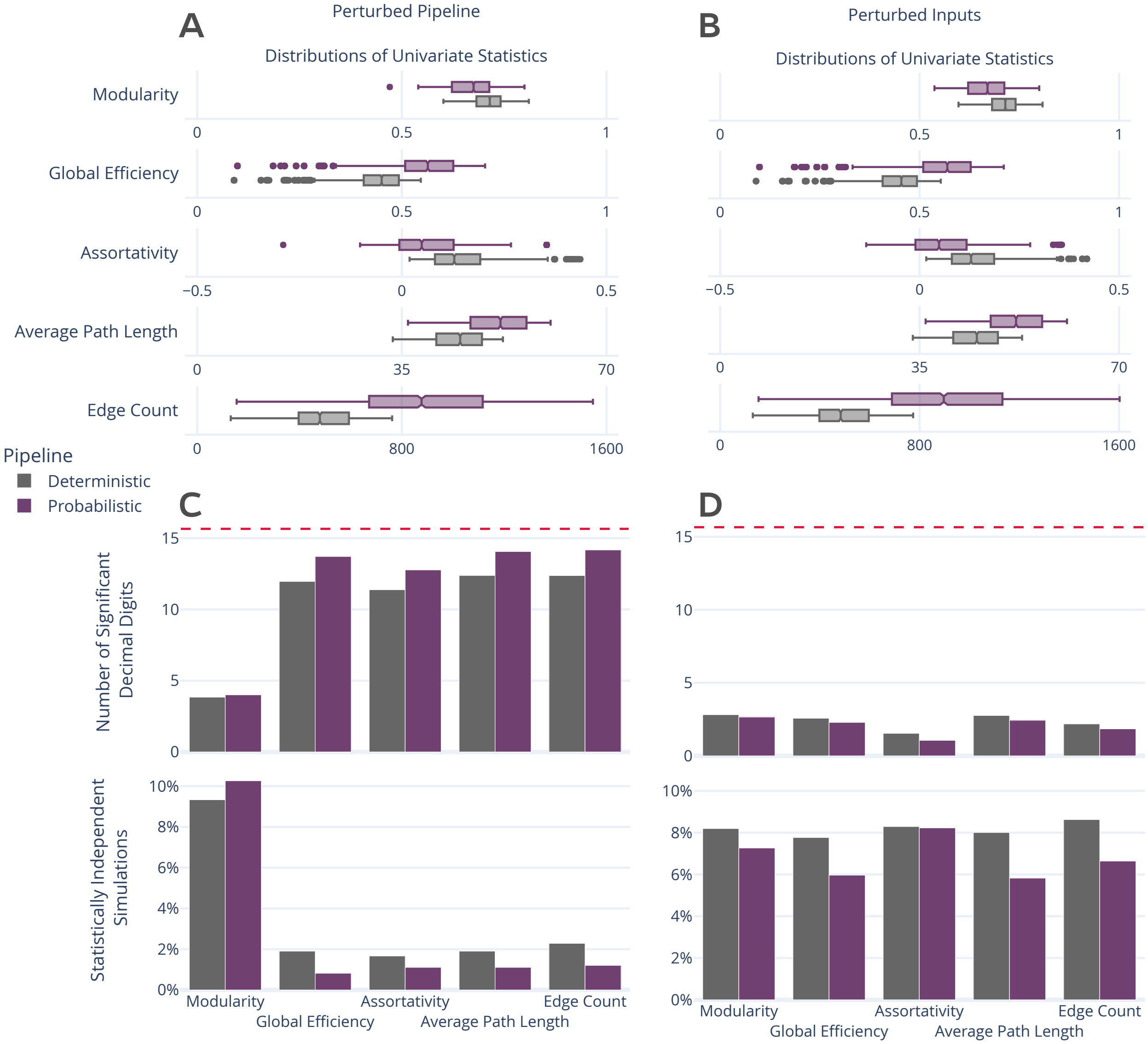

Figure S2. Distribution and stability assessment of univariate graph statistics. (**A**, **B**) The distributions of each computed univariate statistic across all subjects and perturbations for Pipeline and Input settings, respectively. There was no significant difference between the distributions in 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.