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

Prior to exploring the analytic impact of instabilities, a direct understanding of the induced variability was required. A subset of the Nathan Kline Institute Rockland Sample (NKIRS) 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 individual. Structural connectomes were generated with canonical deterministic and probabilistic pipelines[31,32] which were instrumented with MCA, replicatingcomputational noise at either the inputs or throughout the pipelines[5,10]. The pipelines

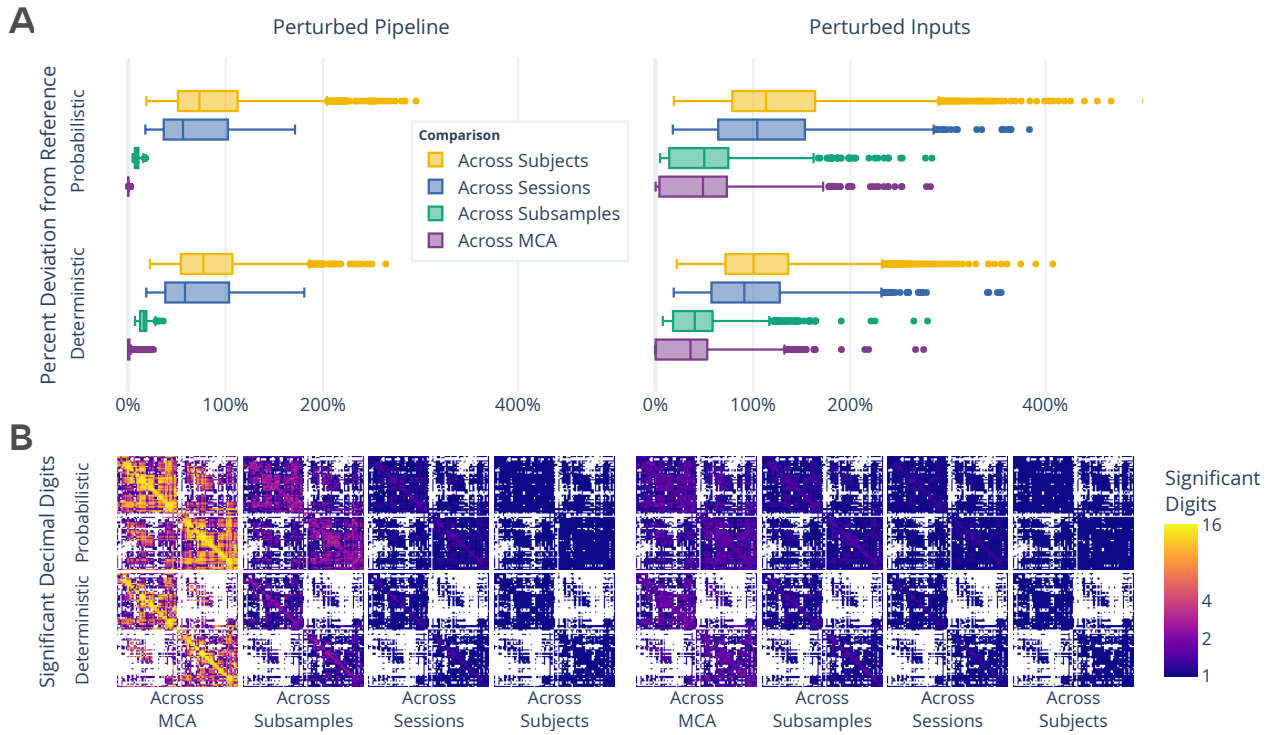

Figure 1. Exploration of perturbation-induced deviations from reference connectomes. (**A**) The absolute deviations, in the form of normalized percent deviation from reference, shown as the across MCA series relative to Across Subsample, Across Session, and Aross Subject variations. (**B**) The number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

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. While both pipelines show similar performance, the probabilistic pipeline was more stable in the face of pipeline perturbations whereas the deterministic was more stable to input perturbations ($p < 0.0001$ for all; exploratory). The stability of correlations can be found 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 Off-Target Biases Are Reduced

We assessed the reproducibility of the dataset through mimicking and extending a typical test-retest experiment[27] in which the similarity of samples across multiple measurements were compared to distinct samples in the dataset (Table 1). The ability to separate connectomes across subjects (Experiments 1.1, 1.2, and 1.3) is an essential prerequisite for the application of brain imaging towards identifying individual differences[19]. In experiment 1.1, we observe that the dataset is separable with a score of 0.64 and 0.65 ($p < 0.001$; optimal score: 1.0; chance: 0.04) without any instrumentation. However, we can see that inducing instabilities through MCA improves the reliability of the dataset to over 0.75 in each case ($p < 0.001$ for all), significantly higher than without instrumentation or greater in

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 reliable 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 groups, 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

($p < 0.005$ for all). This result impactfully suggests the utility of perturbation methods for synthesizing robust and reliable individual estimates of connectivity, serving as a cost effective and context-agnostic method for dataset augmentation.

While the separability of individuals is essential for the identification of brain networks, this modelling is similarly reliant on network similarity across equivalent acquisitions (Experiments 2.4, 2.5). In this case, connectomes were grouped based upon session, rather than subject, and the ability to distinguish one session from another was computed within-individual and aggregated. Both the unperturbed and pipeline perturbation settings perfectly preserved differences between cross-sectional session with a score of 1.0 ($p < 0.005$; optimal score: 0.5; chance: 0.05). However, while still significant relative to chance (score: 0.85 and 0.88; $p < 0.005$ for both), input perturbations lead to significantly lower separability of the dataset ($p < 0.005$ for all). This reduction of the difference between sessions of data within individuals suggests that increased variance caused by input perturbations reduces the impact of non-biological acquisition-dependent bias inherent in the brain graphs.

Though the previous sets of experiments inextricably evaluate the interaction between the dataset and tool, the use of subsampling allowed for characterizing the separability of networks sampled from within a single acquisition (Experiment 3.6). While this experiment could not be evaluated using reference executions, the executions performed with pipeline perturbations showed near perfect separation between subsamples, with scores of 0.99 and 1.0 ($p < 0.005$; optimal: 0.5; chance: 0.5). Given that there is no variability in data acquisition or preprocessing that contributes to this reliable identification of scans, the separability observed in this experiment may only be due to instability or bias inherent to the pipelines. The high variability introduced through input perturbations considerably lowered the reliability towards chance (score: 0.71 and 0.61; $p < 0.005$ for all), further supporting this as an effective method for obtaining lower-bias estimates of individual connectivity.

In all cases the induced perturbations showed an amplification of meaningful biological signal alongside a reduction of off-target bias across all experiments. This result highlights that stability evaluation can be used not only to identify instabilities and variance within pipelines, but that the observed variance may be leveraged for the generation of robust distributions of results.

Distributions of Graph Statistics Are Reliable, But Individual Statistics Are Not

Exploring she stability of topological features of connectomes approaches that of the stability of analyses, as these features are often more suitable than full connectomes for many analytical methods in practice[12]. A separate subset of the NKIRS dataset was randomly selected to contain a single non-subsampled session for 100 individuals, and connectomes were generated as above.

The stability of several commonly-used multivariate graph features[11] was explored in Figure 2. The cumulative density of the features was computed within individuals and the mean density and associated standard error were computed for across individuals (Figures 2A and 2B). There was no significant difference beetween the distributions for each feature across the two perturbation settings, suggesting that the topological features summarized by these multivariate features is robust across both perturbation modes.

In addition to the comparison of distributions, the stability of the first 5 moments of these features was evaluated (Figures 2C and 2D). In the face of pipeline perturbations, the feature-moments were stable with more than 10 significant digits with the exception of edge weight when using the deterministic pipeline, though the probabilistic pipeline was more stable for all comparisons ($p < 0.0001$; exploratory). In stark contrast, input perturbations led to highly unstable feature-moments (Figure 2D), such that none contained more than 5 significant digits of information and several contained than a single significant digit, indicating a complete lack of reliability. This dramatic degradation in stability for individual

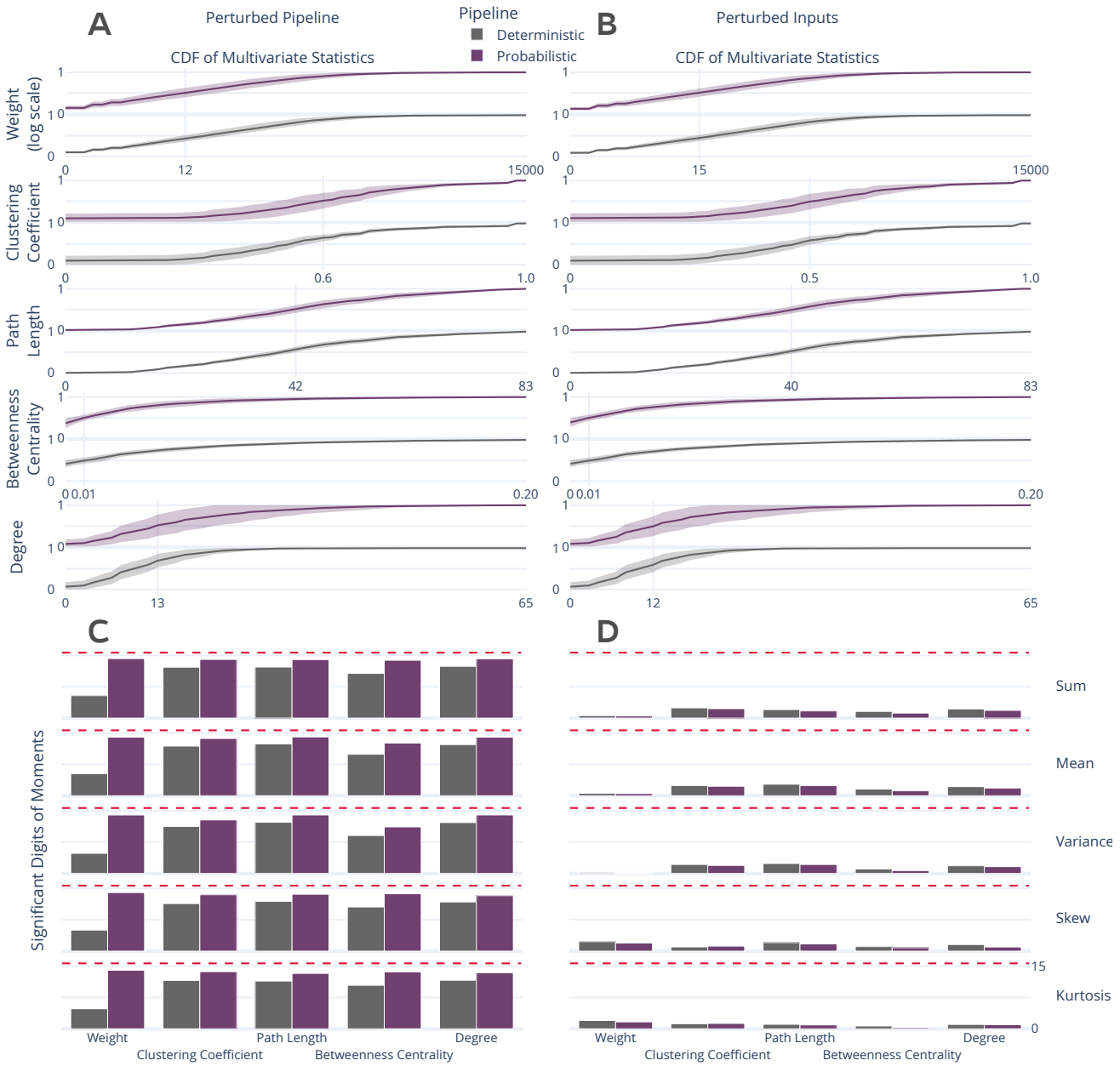

Figure 2. Distribution and stability assessment of multivariate graph statistics. (**A**, **B**) The cumulative distribution functions of multivariate statistics across all subjects and perturbation settings. There was no significant difference between the distributions in A and B. (**C**, **D**) The number of significant digits in the first 5 five moments of each statistic across perturbations. The dashed red line refers to the maximum possible number of significant digits.

measures, combined with the stability in their cross-subject distributions, strongly suggests that while these features may be unreliable as individual biomarkers, they may be used to robustly describe network topologies at a group-level. A similar analysis was performed for univariate statistics and can be found in Supplemental Section S2.

Wide Margins in Brain-Behaviour Relationships

While the variability of connectomes and their features was summarized above, networks are commonly-used as inputs to machine learning models tasked with learning brain-behaviour relationships[19]. To explore the stability of these analyses, we modelled the relationship between BMI and brain connectivity[13,14], using standard dimensionality reduction and classification tools, and compared this to reference and random performance (Figure 3).

The analysis was perturbed through distinct samplings of the dataset across both pipelines and perturbation methods. The accuracy and F1 score for the models varied from 0.520 — 0.716 and 0.510 -– 0.725, respectively, ranging from at or

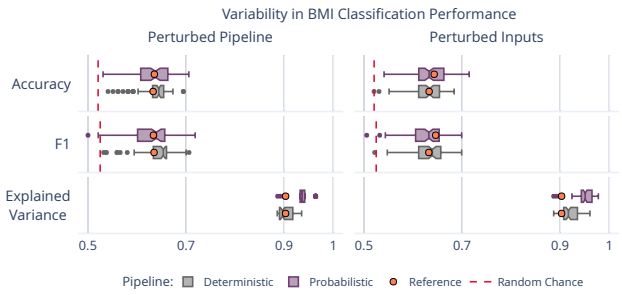

Figure 3. Variability in BMI classification across the sampling of an MCA-perturbed dataset. The dashed red lines indicate random-chance performance, and the orange dots show the performance using the reference executions.

below random performance to outperforming the reference performance. This large variability illustrates a previously uncharacterized margin of uncertainty in the modelling of this relationship, and erodes confidence in reported accuracy scores on singly processed datasets. The portion of explained variance in these samples ranged from 88.6% -– 97.8%, closely surrounding the reference dataset, suggesting that the range in performance was not due to a gain or loss of meaningful signal, but rather the reduction of bias towards specific outcome. Importantly, this finding does not suggest that modelling brain-behaviour relationships is not possible, but rather it sheds light on impactful uncertainty that must be accounted for in this process.

Discussion

The perturbation of structural connectome estimation pipelines with small amounts of noise, on the order of machine error, led to considerable variability in derived brain graphs. Across all analyses the stability of results ranged from nearly perfectly trustworthy (i.e. no variation) to completely unreliable (i.e. containing no significant digits of information). Given that the magnitude of introduced numerical noise is to be expected in typical settings, this finding has potentially significant implications for inferences in brain imaging. In particular, this bounds the success of studying individual differences, a central objective in brain imaging[19], given that the quality of relationships between phenotypic data and brain networks will be limited by the stability of the connectomes themselves. This issue was accentuated through the crucial finding that individually derived network features were unreliable despite there being no significant difference in their aggregated distributions. This finding is not damning for the study of brain networks as a whole, but rather is strong support for the groupwise evaluation of networks over the use of individual estimates.

Trouble with Over Confidence While the instability of brain networks was used here to demonstrate the limitations of modelling brain-behaviour relationships in the context of machine learning, this limitation extends to classical hypothesis testing, as well. Though performing individual comparisons in a hypothesis testing framework will be accompanied by reported false positive rates, the accuracy of these rates is critically dependent upon the reliability of the samples used. In reality, the true false positive rate for a test would be a combination of the reported confidence and the underlying variability in the results, a typically unknown quantity.

When performing these experiments outside of a repeated-measure context, such as that afforded here through MCA, it is impossible to empirically estimate the reliability of samples. This means that the reliability of accepted hypotheses is also 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 the reported value simply as a result of numerical instabilities. This overconfidence in the numerical stability of our analyses limits the ability of researchers to evaluate the quality of results, and ultimately progress science. The accompaniment of brain imaging experiments with direct evaluations of their stability, as was done here, would allow researchers to simultaneously improve the numerical stability of their analyses and accurately gauge confidence in them. Furthermore, the induced variability in derived brain networks may be leveraged to shift the bias-variance tradeoff such that learned relationships are more generalizable and ultimately the utility of such relationships is increased.

Cost-Effective Data Augmentation The evaluation of reliability in brain imaging has historically relied upon the expensive collection of repeated measurements choreographed by the massive cross-institutional consortia[33,34]. The finding that perturbing experiments using MCA both increased the reliability of the dataset and decreased off-target differences across acquisitions opens the door for a promising paradigm shift. Given that MCA is data-agnostic, this technique could be used effectively in conjunction with, or in lieu of, realistic noise models to augment existing datasets. While this of course would not replace the need for repeated measurements when exploring the effect of data collection paradigm or study longitudinal progressions of development or disease, it could be used in conjunction with these efforts to increase the reliability of each distinct sample within a dataset. In contexts where repeated measurements are collected to increase the fidelity of the dataset, MCA could potentially be employed to increase the reliability of the dataset and save millions of dollars on data collection. This technique also opens the door for the characterization of reliability across axes which have been traditionally inaccessible. For instance, in the absense of a realistic noise model, the evaluation of network stability across data subsampling would not have been possible without a simulation technique similar to MCA.

Shortcomings and Future Questions Given the complexity of recompiling complex software libraries, pre-processing was not perturbed in these experiments. Other work has shown that linear registration, a core piece of many elements of preprocessing such as motion correction and alignment, is sensitive to minor perturbations[8]. It is likely that the instabilities

across the entire processing workflow would be compounded with one another, resulting in even greater instability. While the analyses performed in this paper evaluated a single dataset and set of pipelines, extending this work to other modalities and workflows is of interest for future projects.

This paper does not explore methodological flexibility or compare this to numerical instability. Recently, the nearly boundless space of analysis pipelines and their impact on outcomes in brain imaging has been clearly demonstrated[2]. The approach taken in these studies complement one another and explore instability at the opposite ends of the spectrum, with human variability in the construction of an analysis workflow on one end and the unavoidable error implicit in the digital representation of data on the other. 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 analytical pipelines used in brain imaging, but merely sheds light on the fact 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.

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 *et al.*, "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.

[33] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[34] 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, no. 1, pp. 1–13, 2014.

Methods

Unedited past here. Will copy-paste methods more-or-less as-is from the previous draft.

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

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 ¡ 0.0001 for all; exploratory).

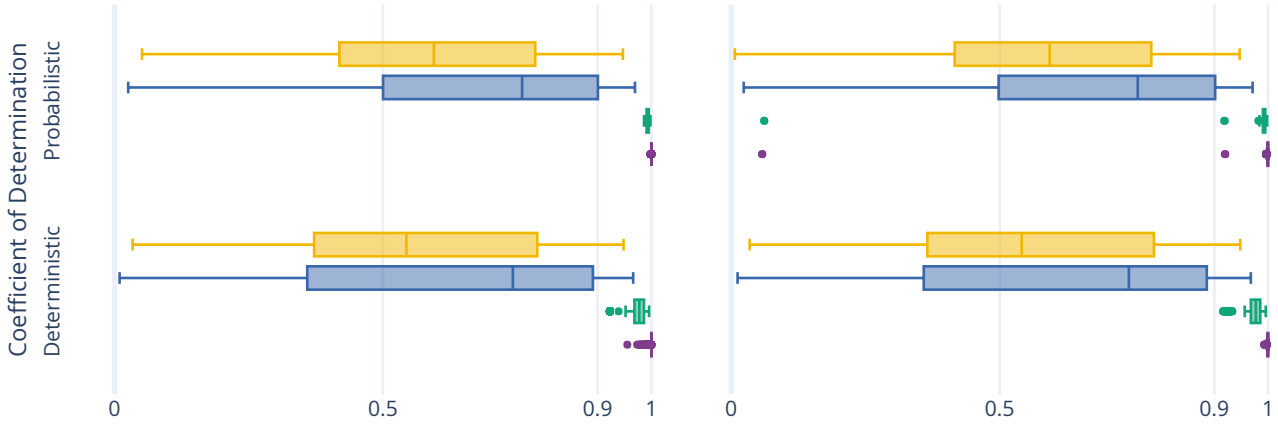

Figure S1. The correlation between perturbed connectomes and their reference.

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.

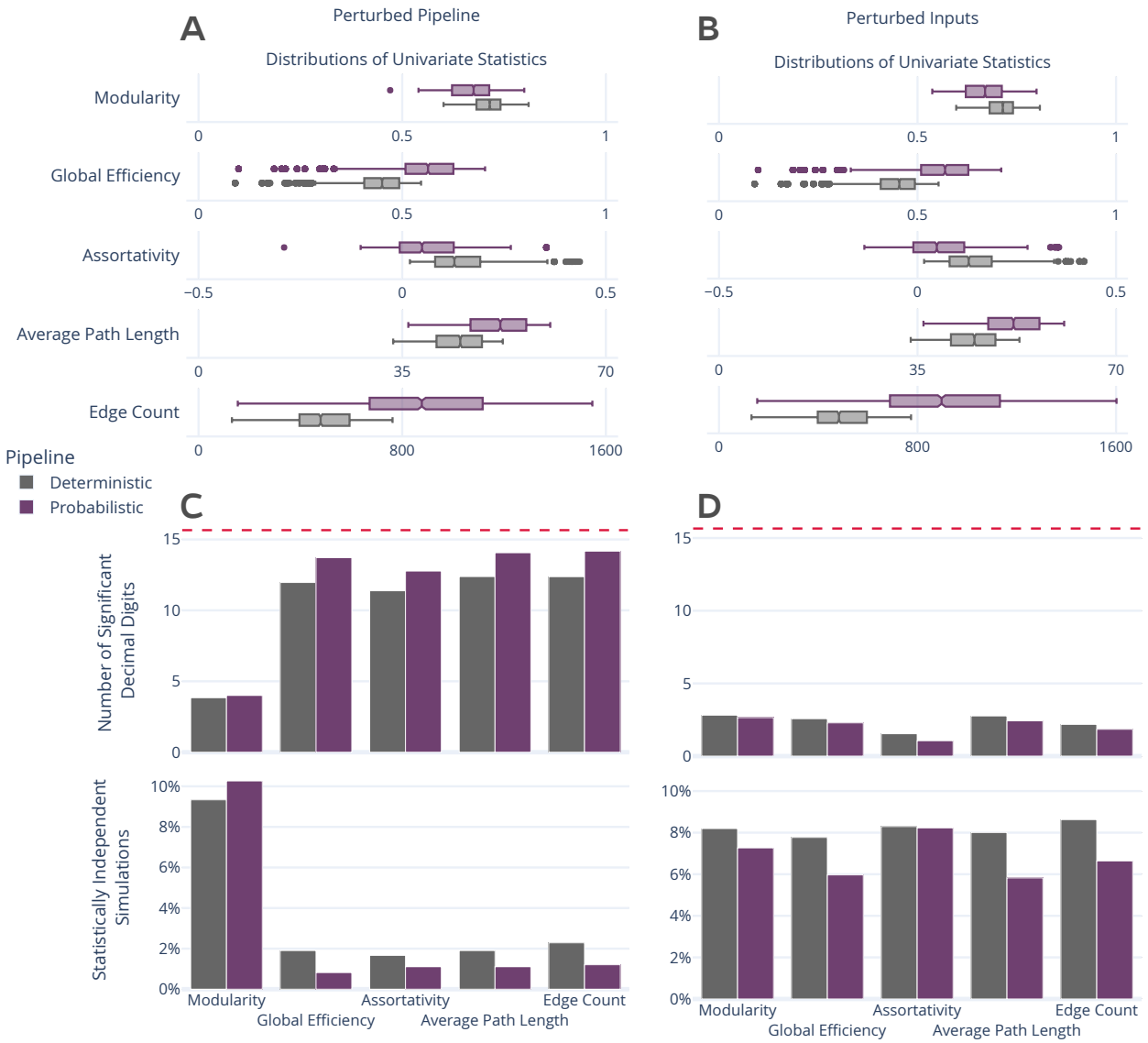

Figure S2. Distribution and stability assessment of univariate graph statistics. (**A**, **B**) The distributions of each computed univariate statistic across all subjects and perturbations for Pipeline and Input settings, respectively. There was no significant difference between the distributions in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