Numerical Instabilities in Analytical Pipelines Lead to Large and Meaningful Variability in Brain Networks

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

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[1–3]. Even within a given method or analytical technique, numerical instabilities could compromise findings[4–7]. We instrumented a structural-connectome estimation pipeline with Monte Carlo Arithmetic[8,9], a technique to introduce random noise in floating-point computations, and evaluated the stability of the derived connectomes, their features[10,11], and the impact on a downstream analysis[12,13]. 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). While the extreme range and variability in results presented here could severely hamper our understanding of brain organization in brain-imaging studies, it also leads to an increase in the reliability of datasets. This paper highlights the potential of leveraging the induced variance in estimates of brain connectivity to reduce the bias in networks alongside increasing the robustness of their applications in the detection or classification of individual differences. This paper demonstrates that stability evaluations are necessary for understanding error and bias inherent to scientific computing, including but not limited to neuroimaging, and that they should be a 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[11,14–18]. 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 Alzheimer's Disease and Schizophrenia, but potentially pave the way for therapeutics[19–23].

However, the analysis of brain imaging data relies on complex computational methods and software. 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[24–27], 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[1,5–7]. While many scientific disciplines suffer from a lack of reproducibility[28], this was recently explored in brain imaging by a 70 team consortium which performed equivalent analyses and found widely inconsistent results[1], and it is likely that software instabilities played a role.

The present study approached evaluating reproducibility from a computational perspective in which a series of brain imaging studies were numerically perturbed such that the plausibility of results was not affected, and the biological implications of the observed instabilities were quantified. We accomplished this through the use of Monte Carlo Arithmetic (MCA)[8], 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 and opportunities afforded by the observed instabilities and make recommendations for the roles stability analyses may play towards increasing the reliability of brain imaging research.

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[29] 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[30,31] which were instrumented with MCA, replicating computational noise at either the inputs or throughout the pipelines[4,9]. 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. 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 with 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). In addition, input perturbations show no more than an average of 3 significant digits across all groups, demonstrating a significant limitation in the reliability independent edge weights. Significance across individuals did not exceed a single digit per edge in any case, indicating that only the magnitude of edges in naively computed groupwise average connectomes can be trusted. The combination of these results

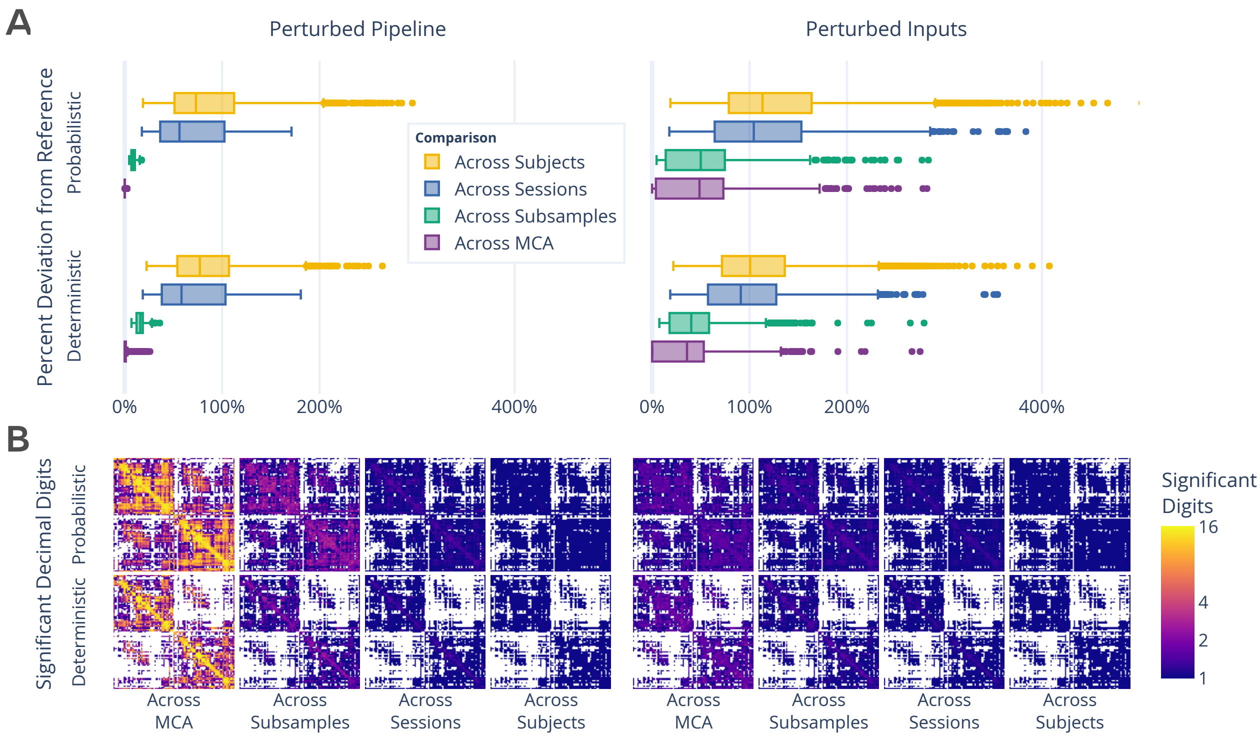

Figure 1. Exploration of perturbation-induced deviations from reference connectomes. (**A**) The absolute deviations, in the form of normalized percent deviation from reference, shown as the across MCA series relative to Across Subsample, Across Session, and Aross Subject variations. (**B**) The number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

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[26] in which the similarity of samples across multiple measurements were compared to distinct samples in the dataset (Table 1, with additional experiments and explanation in Supplemental Section S2). The ability to separate connectomes across subjects (Hypothesis 1) is an essential prerequisite for the application of brain imaging towards identifying individual differences[18]. In testing hypothesis 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 ($p < 0.005$ 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, it is similarly reliant on network similarity across equivalent acquisitions (Hypothesis 2). In this case, connectomes were grouped based upon session, rather than subject, and the ability to distinguish one session

Table 1. The impact of instabilities as evaluated through the separability of the dataset based on individual (or subject) differences, session, and subsample. The performance is reported as mean Discriminability. While a perfectly separable dataset would be represented by a score of 1.0, the chance performance, indicating minimal separability, is 1/the number of classes. H_3 could not be tested using the reference executions due to too few possible comparisons. The alternative hypothesis, indicating significant separation, was accepted for all experiments, with $p < 0.005$.

			Reference Execution		Perturbed Pipeline		Perturbed Inputs	
Comparison	**Chance**	**Target**	Det.	Prob.	Det.	Prob.	Det.	Prob.
H_1: Across Subjects	0.04	1.0	0.64	0.65	0.82	0.82	0.77	0.75
H_2: Across Sessions	0.5	0.5	1.00	1.00	1.00	1.00	0.88	0.85
H_3: Across Subsamples	0.5	0.5			0.99	1.00	0.71	0.61

from another was computed within-individual and aggregated. Both the unperturbed and pipeline perturbation settings perfectly preserved differences between cross-sectional sessions with a score of 1.0 ($p < 0.005$; optimal score: 0.5; chance: 0.5), indicating a dominant session-dependent signal for all individuals despite no intended biological differences. 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 (Hypothesis 3). 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.

Across all cases, the induced perturbations showed an amplification of meaningful biological signal alongside a reduction of off-target signal. This result appears strikingly like a manifestation of the well-known bias-variance tradeoff[32] in machine learning, a concept which observes a decrease in bias as variance is favoured by a model. In particular, this highlights that numerical perturbations can be used to not only evaluate the stability of pipelines, but that the induced variance may be leveraged for the interpretation as a robust distributions of possible results.

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

Exploring the stability of topological features of connectomes is relevant for typical analyses, as low dimensional features are often more suitable than full connectomes for many analytical methods in practice[11]. 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[10] was explored in Figure 2. The cumulative den-

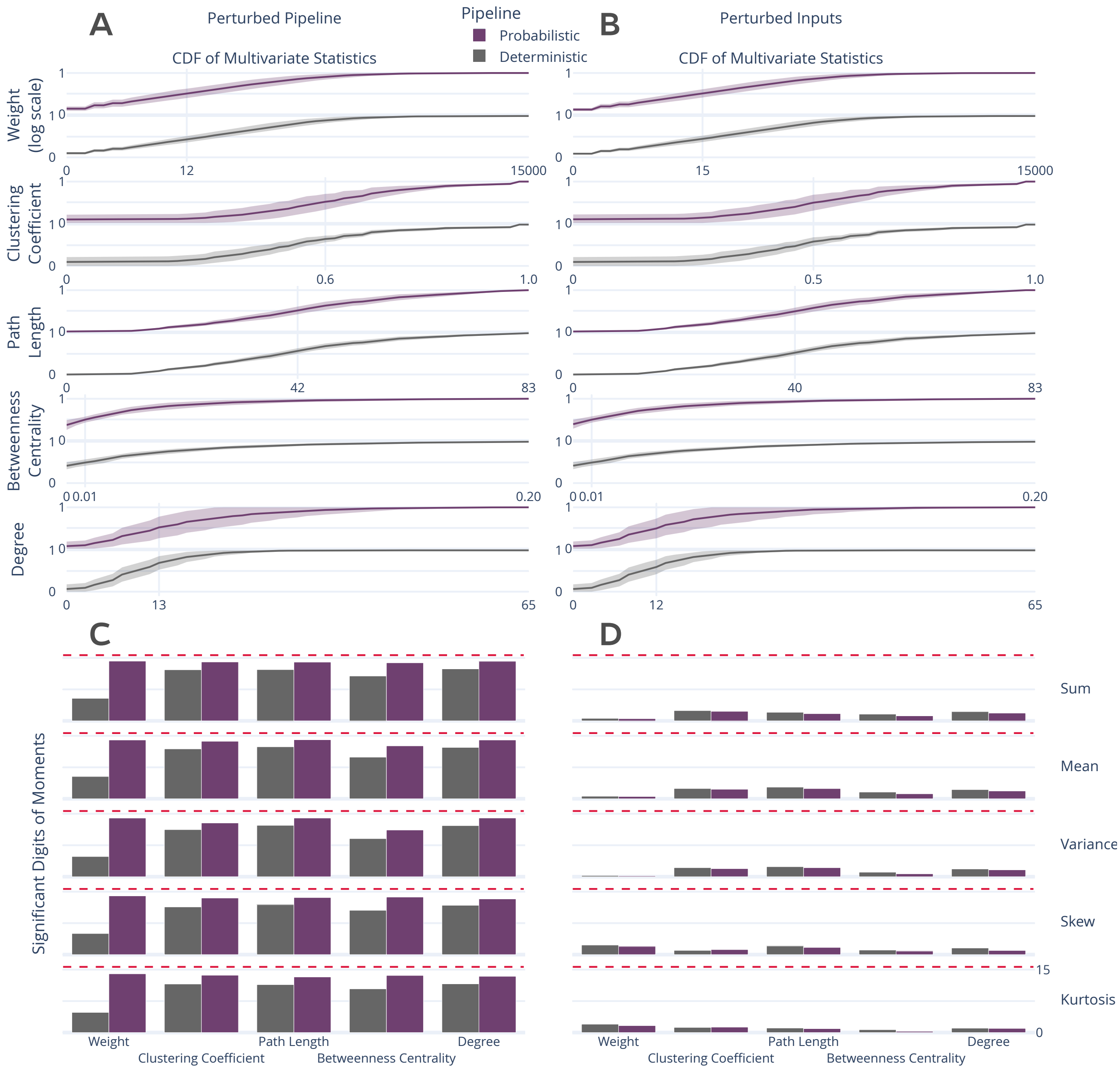

Figure 2. Distribution and stability assessment of multivariate graph statistics. (**A**, **B**) The cumulative distribution functions of multivariate statistics across all subjects and perturbation settings. There was no significant difference between the distributions in A and B. (**C**, **D**) The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

sity 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 between the distributions for each feature across the two perturbation settings, suggesting that the topological features summarized by these multivariate features are 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 less than a single significant digit, indicating a complete lack of reliability. This dramatic degradation in stability for individual measures strongly suggests that these features may be unreliable as individual biomarkers when derived from a single pipeline evaluation, though their reliability may be increased when studying their distributions across perturbations. A similar analysis was performed for univariate statistics and can be found in Supplemental Section S3.

Uncertainty in Brain-Phenotype 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-phenotype relationships[18]. To explore the stability of these analyses, we modelled the relationship between high- or low- Body Mass Index (BMI) groups and brain connectivity[12,13], 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 perturbed models varied from 0.520 – 0.716 and 0.510 – 0.725, respectively, ranging from at or below random performance to outperforming performance on the reference dataset. This large variability illustrates a previously uncharacterized margin of uncertainty in the modelling of this relationship, and limits confidence in reported accuracy scores on singly processed datasets. The portion of explained variance in these samples ranged from 88.6% –– 97.8%, similar to the reference, 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-phenotype relationships is not possible, but rather it sheds light on impactful uncertainty that must be accounted for in this process, and supports the use of ensemble modeling techniques.

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 trustworthy 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 as it is currently performed. In particular, this bounds the success of studying individual differences, a central objective in brain imaging[18], 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 aggregation of networks, either across perturbations for an individual or across groups, over the use of individual estimates.

Underestimated False Positive Rates While the instability of brain networks was used here to demonstrate the limitations of modelling brain-phenotype 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

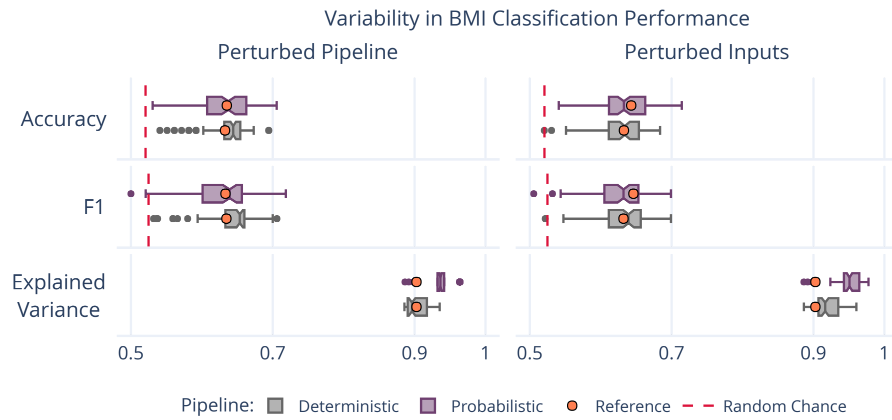

Figure 3. Variability in BMI classification across the sampling of an MCA-perturbed dataset. The dashed red lines indicate random-chance performance, and the orange dots show the performance using the reference executions.

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 uncertainty inherent to derived data is compounded with traditional arguments limiting the trustworthiness of claims[33], and hampers the ability of researchers to evaluate the quality of results. 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. The induced variability in derived brain networks may be leveraged to estimate aggregate connectomes with lower bias than any single independent observation, leading to learned relationships that are more generalizable and ultimately more useful.

Cost-Effective Data Augmentation The evaluation of reliability in brain imaging has historically relied upon the expensive collection of repeated measurements choreographed by massive cross-institutional consortia[34,35]. 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 absence of a realistic noise model or simulation technique similar to MCA, the evaluation of network stability across data subsampling would not have been possible.

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 pre-processing such as motion correction and alignment, is sensitive to minor perturbations[7]. It is likely that the instabilities across the entire processing workflow would be compounded with one another, resulting in even greater variability. While the analyses performed in this paper evaluated a single dataset and set of pipelines, extending this work to other modalities and analyses 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[1]. 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 presence of unknown error-bars associated with experimental findings limits the impact of results due to increased uncertainty. The desired outcome of this paper is to motivate a shift in scientific computing – both in neuroimaging and more broadly – towards a paradigm which favours the explicit evaluation of the trustworthiness of claims alongside the claims themselves.

Methods

Dataset

The Nathan Kline Institute Rockland Sample (NKI-RS)[29] dataset contains high-fidelity imaging and phenotypic data from over 1,000 individuals spread across the lifespan. A subset of this dataset was chosen for each experiment to both match sample sizes presented in the original analyses and to minimize the computational burden of performing MCA. The selected subset comprises 100 individuals ranging in age from 6 – 79 with a mean of 36.8 (original: 6 – 81, mean 37.8), 60% female (original: 60%), with 52% having a BMI over 25 (original: 54%).

Each selected individual had at least a single session of both structural T1-weighted (MPRAGE) and diffusion-weighted (DWI) MR imaging data. DWI data was acquired with 137 diffusion directions; more information regarding the acquisition of this dataset can be found in the NKI-RS data release[29].

In addition to the 100 sessions mentioned above, 25 individuals had a second session to be used in a test-retest analysis. Two additional copies of the data for these individuals were generated, including only the odd or even diffusion directions (64 + 9 B0 volumes = 73 in either case). This allowed for an extra level of stability evaluation to be performed between the levels of MCA and session-level variation.

In total, the dataset is composed of 100 downsampled sessions of data originating from 50 acquisitions and 25 individuals for in depth stability analysis, and an additional 100 sessions of full-resolution data from 100 individuals for subsequent analyses.

Processing

The dataset was preprocessed using a standard FSL[36] workflow consisting of eddy-current correction and alignment. The MNI152 atlas[37] was aligned to each session of data, and the resulting transformation was applied to the DKT parcellation[38]. Downsampling the diffusion data took place after preprocessing was performed on full-resolution sessions, ensuring that an additional confound was not introduced in this process when comparing between downsampled sessions. The preprocessing described here was performed once without MCA, and thus is not being evaluated.

Structural connectomes were generated from preprocessed data using two canonical pipelines from Dipy[30]: deterministic and probabilistic. In the deterministic pipeline, a constant solid angle model was used to estimate tensors at each voxel and streamlines were then generated using the EuDX algorithm[31]. In the probabilistic pipeline, a constrained spherical deconvolution model was fit at each voxel and streamlines were generated by iteratively sampling the resulting fiber orientation distributions. In both cases tracking occurred with 8 seeds per 3D voxel and edges were added to the graph based on the location of terminal nodes with weight determined by fiber count.

The random state of the probabilistic pipeline was fixed for all analyses. Fixing this random seed allowed for explicit attribution of observed variability to Monte Carlo simulations rather than internal state of the algorithm.

Perturbations

All connectomes were generated with one reference execution where no perturbation was introduced in the processing. For all other executions, all floating point operations were instrumented with Monte Carlo Arithmetic (MCA)[8] through Verificarlo[9]. MCA simulates the distribution of errors implicit to all instrumented floating point operations (flop). This rounding is performed on a value x at precision t by:

$$inexact(x) = x + 2^{e_x - t}\xi \tag{1}$$

where e_x is the exponent value of x and ξ is a uniform random variable in the range $(-\frac{1}{2}, \frac{1}{2})$. MCA can be introduced in two places for each flop: before or after evaluation. Performing MCA on the inputs of an operation limits its precision, while performing MCA on the output of an operation highlights round-off errors that may be introduced. The former is referred to as Precision Bounding (PB) and the latter is called Random Rounding (RR).

Using MCA, the execution of a pipeline may be performed many times to produce a distribution of results. Studying the distribution of these results can then lead to insights on the stability of the instrumented tools or functions. To this end, a complete software stack was instrumented with MCA and is made available on GitHub at `https://github.com/gkiar/fuzzy`.

Both the RR and PB variants of MCA were used independently for all experiments. As was presented in[4], both the degree of instrumentation (i.e. number of affected libraries) and the perturbation mode have an effect on the distribution of observed results. For this work, the RR-MCA was applied across the bulk of the relevant libraries and is referred to as Pipeline Perturbation. In this case the bulk of numerical operations were affected by MCA.

Conversely, the case in which PB-MCA was applied across the operations in a small subset of libraries is here referred to as Input Perturbation. In this case, the inputs to operations within the instrumented libraries (namely, Python and Cython) were perturbed, resulting in less frequent, data-centric perturbations. Alongside the stated theoretical differences, Input Perturbation is considerably less computationally expensive than Pipeline Perturbation.

All perturbations targeted the least-significant-bit for all data ($t = 24$ and $t = 53$ in float32 and float64, respectively[9]). Simulations were performed 20 times for each pipeline execution. A detailed motivation for the number of simulations can be found in[39].

Evaluation

The magnitude and importance of instabilities in pipelines can be considered at a number of analytical levels, namely: the induced variability of derivatives directly, the resulting downstream impact on summary statistics or features, or the ultimate change in analyses or findings. We explore the nature and severity of instabilities through each of these lenses. Unless otherwise stated, all p-values were computed using Wilcoxon signed-rank tests.

Direct Evaluation of the Graphs

The differences between simulated graphs was measured directly through both a direct variance quantification and a comparison to other sources of variance such as individual- and session-level differences.

Quantification of Variability Graphs, in the form of adjacency matrices, were compared to one another using three metrics: normalized percent deviation, Pearson correlation, and edgewise significant digits. The normalized percent deviation measure, defined in[4], scales the norm of the difference between a simulated graph and the reference execution (that without intentional perturbation) with respect to the norm of the reference graph. The purpose of this comparison is to provide insight on the scale of differences in observed graphs relative to the original signal intensity. A Pearson correlation coefficient[40] was computed in complement to normalized percent deviation to identify the consistency of structure and not just intensity between observed graphs.

Finally, the estimated number of significant digits, s', for each edge in the graph is calculated as:

$$s' = -log_{10}\frac{\sigma}{|\mu|} \tag{2}$$

where μ and σ are the mean and unbiased estimator of standard deviation across graphs, respectively. The upper bound on significant digits is 15.7 for 64-bit floating point data.

The percent deviation, correlation, and number of significant digits were each calculated within a single session of data, thereby removing any subject- and session-effects and providing a direct measure of the tool-introduced variability across perturbations. A distribution was formed by aggregating these individual results.

Class-based Variability Evaluation To gain a concrete understanding of the significance of observed variations we explore the separability of our results with respect to understood sources of variability, such as subject-, session-, and pipeline-level effects. This can be probed through Discriminability[26], a technique similar to ICC[24] which relies on the mean of a ranked distribution of distances between observations belonging to a defined set of classes. The discriminability statistic is formalized as follows:

$$Disc. = Pr(\|g_{ij} - g_{ij'}\| \leq \|g_{ij} - g_{i'j'}\|) \tag{3}$$

where g_{ij} is a graph belonging to class i that was measured at observation j, where $i \neq i'$ and $j \neq j'$.

Discriminability can then be read as the probability that an observation belonging to a given class will be more similar to other observations within that class than observations of a different class. It is a measure of reproducibility, and is discussed in detail in[26]. This definition allows for the exploration of deviations across arbitrarily defined classes which in practice can be any of those listed above. We combine this statistic with permutation testing to test hypotheses on whether differences between classes are statistically significant in each of these settings.

With this in mind, three hypotheses were defined. For each setting, we state the alternate hypotheses, the variable(s) which were used to determine class membership, and the remaining variables which may be sampled when obtaining multiple observations. Each hypothesis was tested independently for each pipeline and perturbation mode, and in every case where it was possible the hypotheses were tested using the reference executions alongside using MCA.

H_{A1}: <u>Individuals are distinct from one another</u>
Class definition: *Subject ID*
Comparisons: *Session (1 subsample)*, *Subsample (1 session)*, *MCA (1 subsample, 1 session)*

H_{A2}: Sessions within an individual are distinct
Class definition: *Session ID* | *Subject ID*
Comparisons: *Subsample*, *MCA (1 subsample)*

H_{A3}: Subsamples are distinct
Class definition: *Subsample* | *Subject ID, Session ID*
Comparisons: *MCA*

As a result, we tested 3 hypotheses across 6 MCA experiments and 3 reference experiments on 2 pipelines and 2 perturbation modes, resulting in a total of 30 distinct tests.

Evaluating Graph-Theoretical Metrics

While connectomes may be used directly for some analyses, it is common practice to summarize them with structural measures, which can then be used as lower-dimensional proxies of connectivity in so-called graph-theoretical studies[11]. We explored the stability of several commonly-used univariate (graphwise) and multivariate (nodewise or edgewise) features. The features computed and subsequent methods for comparison in this section were selected to closely match those computed in[10].

Univariate Differences For each univariate statistic (edge count, mean clustering coefficient, global efficiency, modularity of the largest connected component, assortativity, and mean path length) a distribution of values across all perturbations within subjects was observed. A Z-score was computed for each sample with respect to the distribution of feature values within an individual, and the proportion of "classically significant" Z-scores, i.e. corresponding to $p < 0.05$, was reported and aggregated across all subjects. The number of significant digits contained within an estimate derived from a single subject were calculated and aggregated.

Multivariate Differences In the case of both nodewise (degree distribution, clustering coefficient, betweenness centrality) and edgewise (weight distribution, connection length) features, the cumulative density functions of their distributions were evaluated over a fixed range and subsequently aggregated across individuals. The number of significant digits for each moment of these distributions (sum, mean, variance, skew, and kurtosis) were calculated across observations within a sample and aggregated.

Evaluating A Brain-Phenotype Analysis

Though each of the above approaches explores the instability of derived connectomes and their features, many modern studies employ modeling or machine-learning approaches, for instance to learn brain-phenotype relationships or identify differences across groups. We carried out one such study and explored the instability of its results with respect to the upstream variability of connectomes characterized in the previous sections. We performed the modeling task with a single sampled connectome per individual and repeated this sampling and modelling 20 times. We report the model performance for each sampling of the dataset and summarize its variance.

BMI Classification Structural changes have been linked to obesity in adolescents and adults[41]. We classified normal-weight and overweight individuals from their structural networks (using for overweight a cutoff of BMI > 25[13]). We reduced the dimensionality of the connectomes through principal component analysis (PCA), and provided the first N-components to a logistic regression classifier for predicting BMI class membership, similar to methods shown in[12,13]. The number of components was selected as the minimum set which explained $> 90\%$ of the variance when averaged across the training set for each fold within the cross validation of the original graphs; this resulted in a feature of 20 components. We trained the model using k-fold cross validation, with $k = 2, 5, 10$, and N (equivalent to leave-one-out; LOO).

Data & Code Provenance

The unprocessed dataset is available through The Consortium of Reliability and Reproducibility (`http://fcon_1000.projects.nitrc.org/indi/enhanced/`), including both the imaging data as well as phenotypic data which may be obtained upon submission and compliance with a Data Usage Agreement. The connectomes generated through simulations have been bundled and stored permanently (https://doi.org/10.5281/zenodo.4041549), and are made available through The Canadian Open Neuroscience Platform (https://portal.conp.ca/search, search term "Kiar").

All software developed for processing or evaluation is publicly available on GitHub at https://github.com/gkpapers/2020ImpactOfInstability. Experiments were launched using Boutiques[42] and Clowdr[43] in Compute Canada's HPC cluster environment. MCA instrumentation was achieved through Verificarlo[9] available on Github at https://github.com/verificarlo/verificarlo. A set of MCA instrumented software containers is available on Github at https://github.com/gkiar/fuzzy.

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. Correspondence and requests for materials should be addressed to Tristan Glatard at tristan.glatard@concordia.ca.

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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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," *The International Journal of High Performance Computing Applications*, 2020.

[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] 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.

[7] 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.

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

[9] C. Denis, P. de Oliveira Castro, and E. Petit, "Verificarlo: Checking floating point accuracy through monte carlo arithmetic," *2016 IEEE 23nd Symposium on Computer Arithmetic (ARITH)*, 2016.

[10] R. F. Betzel, A. Griffa, P. Hagmann, and B. Mišić, "Distance-dependent consensus thresholds for generating group-representative structural brain networks," *Network neuroscience*, vol. 3, no. 2, pp. 475–496, 2019.

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

[12] 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.

[13] 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, "Pat-

terns of brain structural connectivity differentiate normal weight from overweight subjects," *Neuroimage Clin*, vol. 7, pp. 506–517, Jan. 2015.

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

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

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

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

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[33] J. P. Ioannidis, "Why most published research findings are false," *PLoS medicine*, vol. 2, no. 8, p. e124, 2005.

[34] 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.

[35] 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.

[36] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *Neuroimage*, vol. 62, no. 2, pp. 782–790, Aug. 2012.

[37] J. L. Lancaster, D. Tordesillas-Gutiérrez, M. Martinez, F. Salinas, A. Evans, K. Zilles, J. C. Mazziotta, and P. T. Fox, "Bias between mni and talairach coordinates analyzed using the icbm-152 brain template," *Human brain mapping*, vol. 28, no. 11, pp. 1194–1205, 2007.

[38] A. Klein and J. Tourville, "101 labeled brain images and a consistent human cortical labeling protocol," *Front. Neurosci.*, vol. 6, p. 171, Dec. 2012.

[39] D. Sohier, P. De Oliveira Castro, F. Févotte, B. Lathuilière, E. Petit, and O. Jamond, "Confidence intervals for stochastic arithmetic," Jul. 2018.

[40] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, I. Cohen, Y. Huang, J. Chen, and J. Benesty, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–4.

[41] C. A. Raji, A. J. Ho, N. N. Parikshak, J. T. Becker, O. L. Lopez, L. H. Kuller, X. Hua, A. D. Leow, A. W. Toga, and P. M. Thompson, "Brain structure and obesity," *Hum. Brain Mapp.*, vol. 31, no. 3, pp. 353–364, Mar. 2010.

[42] T. Glatard, G. Kiar, T. Aumentado-Armstrong, N. Beck, P. Bellec, R. Bernard, A. Bonnet, S. T. Brown, S. Camarasu-Pop, F. Cervenansky, S. Das, R. Ferreira da Silva, G. Flandin, P. Girard, K. J. Gorgolewski, C. R. G. Guttmann, V. Hayot-Sasson, P.-O. Quirion, P. Rioux, M.-É. Rousseau, and A. C. Evans, "Boutiques: a flexible framework to inte-

grate command-line applications in computing platforms," *Gigascience*, vol. 7, no. 5, May 2018.

[43] G. Kiar, S. T. Brown, T. Glatard, and A. C. Evans, "A serverless tool for platform agnostic computational experiment management," *Front. Neuroinform.*, vol. 13, p. 12, Mar. 2019.

[44] H. Huang and M. Ding, "Linking functional connectivity and structural connectivity quantitatively: a comparison of methods," *Brain connectivity*, vol. 6, no. 2, pp. 99–108, 2016.

S1. Graph Correlation

The correlations between observed graphs (Figure S1) across each grouping follow the same trend to as percent deviation, as shown in Figure 1. 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).

The marked lack in drop-off of performance across these settings, inconsistent with the measures show in Figure 1 is due to the nature of the measure and the graphs. Given that structural graphs are sparse and contain considerable numbers of zero-weighted edges, the presence or absense of an edge dominated the correlation measure where it was less impactful for the others. For this reason and others[44], correlation is not a commonly used measure in the context of structural connectivity.

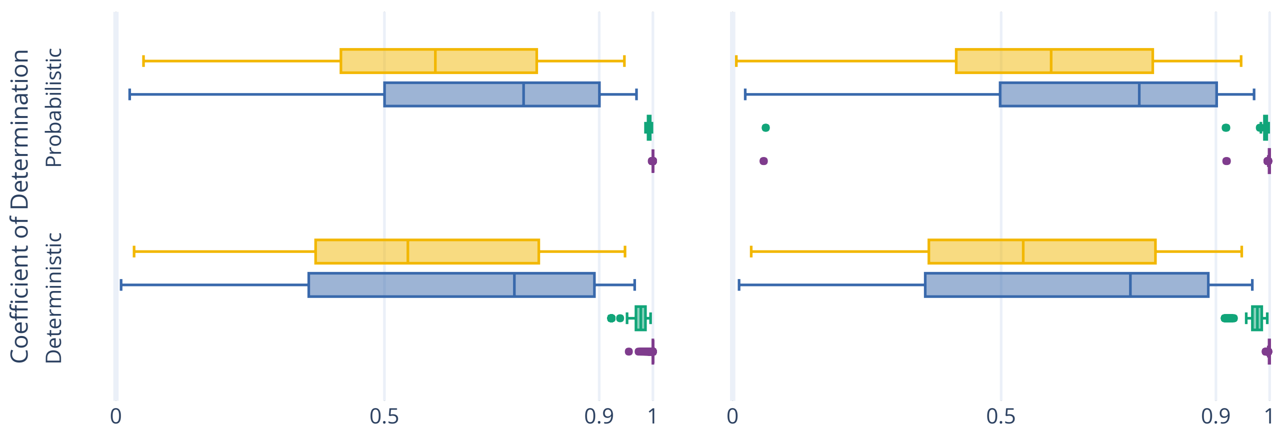

Figure S1. The correlation between perturbed connectomes and their reference.

S2. Complete Discriminability Analysis

Table S1. The complete results from the Discriminability analysis, with results reported as mean $\pm$ standard deviation Discriminability. As was the case in the condensed table, the alternative hypothesis, indicating significant separation across groups, was 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

The complete discriminability analysis includes comparisons across more axes of variability than the condensed version. The reduction in the main body was such that only axes which would be relevant for a typical analysis were presented. Here, each of Hypothesis 1, testing the difference across subjects, and 2, testing the difference across sessions, were accompanied with additional comparisons to those shown in the main body.

Subject Variation Alongside experiment 1.1, that which mimicked a typical test-retest scenario, experiments 1.2 and 1.3 could be considered a test-retest with a handicap, given a single aqcuisition per individual was compared either across subsamples or simulations, respectively. For this reason, it is unsurprising that the dataset achieved considerably higher discriminability scores.

Session Variation Similar to subject variation, the session variation was also modelled across either both or a single subsample. In both of these cases the performance was similar, and the finding that input perturbation reduced the off-target signal was consistent.

S3. Univariate Graph Statistics

Figure S2 explores the stability of univariate 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 S2A and S22B) showed 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 S2C). 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 S2D). 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; this finding is consistent with that presented for multivariate statistics.

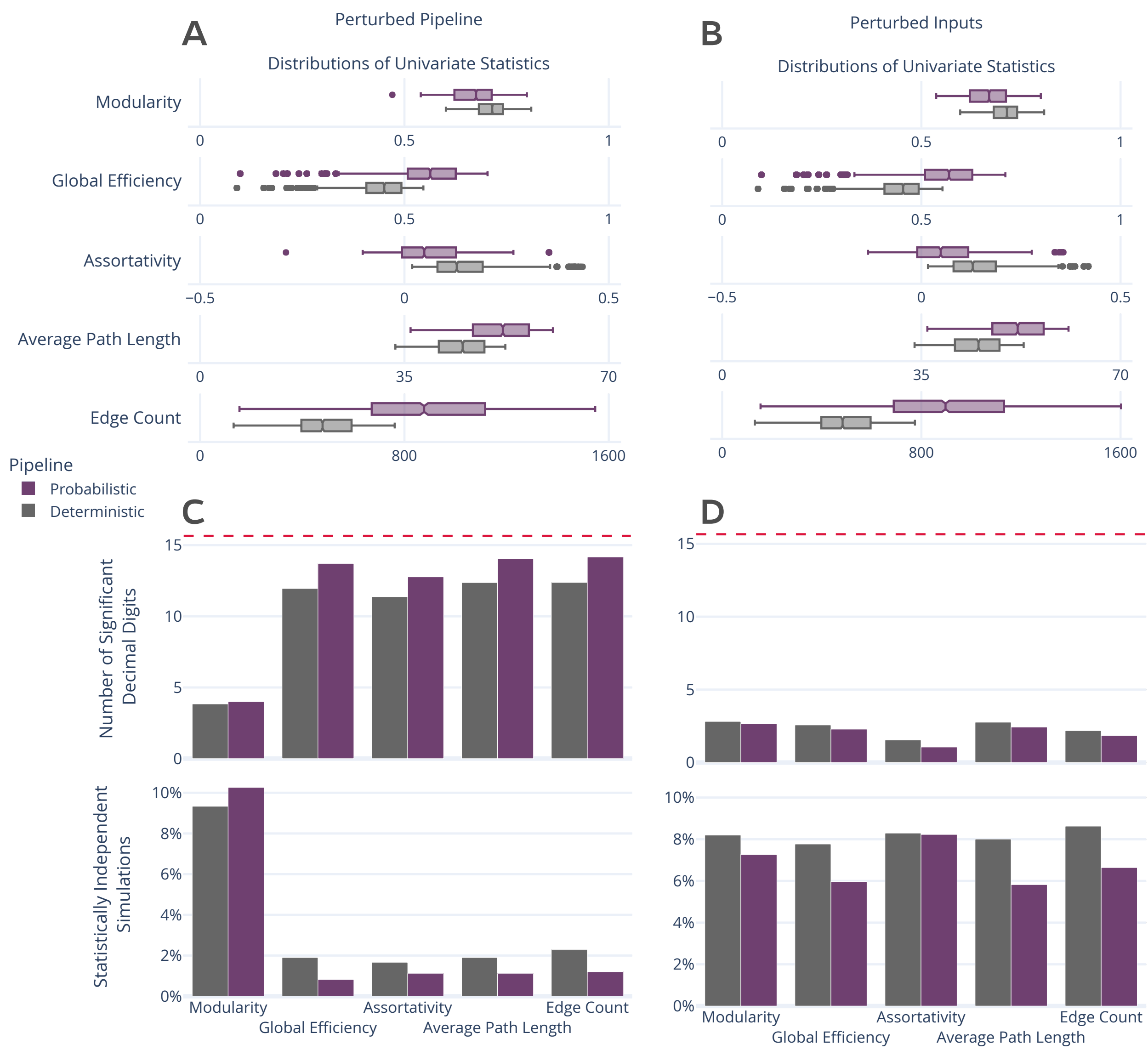

Figure S2. Distribution and stability assessment of univariate graph statistics. (**A**, **B**) The distributions of each computed univariate statistic across all subjects and perturbations for Pipeline and Input settings, respectively. There was no significant difference between the distributions in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