Big Data Reproducibility: Applications in Brain Imaging

Eric W. Bridgeford[1], Shangsi Wang[1], Zhi Yang[2], Zeyi Wang[1], Ting Xu[3], Cameron Craddock[3], Jayanta Dey[1], Gregory Kiar[1], William Gray-Roncal[1], Carey E. Priebe[1], Brian Caffo[1], Michael Milham[3], Xi-Nian Zuo[2,4], Consortium for Reliability and Reproduciblity, Joshua T. Vogelstein[*1]

Abstract. Reproducibility, the ability to replicate analytical findings, is a prerequisite for both scientific discovery and clinical utility. Troublingly, we are in the midst of a reproducibility crisis, in which many investigations fail to replicate. Although many believe that these failings are due to misunderstanding or misapplication of statistical inference (e.g., p-values or the dichotomization of "statistically significant"), we believe the shortcomings arise much earlier in the data science workflow, at the level of measurement, including data acquisition and reconstruction. A key to reproducibility is that multiple measurements of the same item (e.g., experimental sample or clinical participant) are similar to one another, while they are dissimilar from other items. The intra-class correlation coefficient (ICC) quantifies reproducibility in this way, but only for univariate (one dimensional) Gaussian data. In contrast, big data is multivariate (high-dimensional), non-Gaussian, and often non-Euclidean (including text, images, speech, and networks), rendering ICC inadequate. We propose a novel statistic, *discriminability*, which quantifies the degree to which individual samples are discriminable from one another, without restricting the data to be univariate, Gaussian, or even Euclidean. We then introduce the possibility of optimizing experimental design via increasing discriminability. We prove that optimizing discriminability yields an improved ability to use the data for subsequent inference tasks, without specifying the inference task *a priori*. We then apply this approach to a brain imaging dataset built by the "Consortium for Reliability and reproducibility" which consists of 28 disparate magnetic resonance imaging datasets. Optimizing discriminability improves performance on multiple subsequent inference tasks, despite that they were not considered in the optimization. We therefore suggest that designing experiments and analyses to optimize discriminability may be a crucial step in solving the reproducibility crisis.

1 Introduction

Reproducibility and reliability are central tenets in data science, whether applied to scientific discovery or clinical utility. As a rule, if results do not reproduce, we do not trust them. In statistics, reproducibility is closely related to *robustness*, which quantifies the robustness of an estimator the model misspecifications (such as additional sources of variance not accounted for in the model) [1]. Reproducibility is also related to *stability*, which generalizes robustness in certain ways, for example, whether an estimate is stable over multiple random samples of the data [2]. Engineering and operations research have been concerned with *reliability* for a long time, as they require that their products are reliable under various conditions. Very recently, the general research community became interested in these issues, as individuals began noticing and publishing failures to reproduce across fields, including neuroscience and psychology [3–5].

A number of strategies have been suggested to resolve this "reproducibility crisis." For example, the editors of "Basic and Applied Social Psychology" have banned the use of p-values [6]. Unfortunately, an analysis of the publications since banning indicates that studies after the ban tended to overstate, rather than understate, their claims, suggesting that this proposal possibly had the opposite effect [7]. More recently, the American Statistical Association released a statement recommending banning the phrase "statistically significant" for similar reasons [8].

A different strategy has been to quantify the reproducibility or reliability of ones' data after collection. This practice has been particularly popular in brain imaging, where many studies have been devoted to quantifying the reproducibility of different properties of the data [9–12]. However, this approach has severe limitations. Perhaps the most problematic aspect of this approach is clear from the popular adage, "garbage in, garbage out" [13]. If the measurements themselves are not sufficiently reproducible, then scalar summaries of the data cannot be reproducible either. This perspective, the primary of measurement, is fundamental in statistics, so much so that one of the first modern statistics textbook was R.A. Fisher's, "The Design of Experiments" [14].

[1] Johns Hopkins University, [2] Shanghai Jiaotong University, [3] Child Mind Institute, [4] Beijing Normal University, Nanning Normal University, University of Chinese Academy of Sciences. [*] Corresponding author: Joshua T. Vogelstein (jovo@jhu.edu).

Motivated by Fisher's work on experimental design, rather than recommending different post-data acquisition inferential techniques, or computing the reproducibility of data after collecting, we take a different approach. Specifically, we advocate for designing experiments to maximize reproducibility. Experimental design has a rich history, including in psychology [15] and neuroscience [16, 17]. The vast majority of work in experimental design, however, focuses on designing an experiment to answer a particular scientific question. In this big data age, however, experiments are often designed to answer many questions, including questions not even considered at the time of data acquisition. How can one even conceivably design experiments to obtain data that is particularly useful for those questions?

We propose to design experiments to optimize the discriminability of individual items (for example, participants in a study, or samples in an experiment). To do so, we introduce the discriminability statistic, which quantifies the degree to which multiple measurements of the same item are more similar to one another than they are to other items. This statistic has several advantages over existing statistics that one could potentially use to optimize experimental design. First, it is nonparametric, meaning that its validity does not depend on any parametric assumptions, such as Gaussianity. Second, it can readily be applied to multivariate Euclidean data, or even non-Euclidean data (such as images, text, speech, or networks). Third, it can be applied to any stage of the data science pipeline, from data acquisition to data wrangling to data inferences.

We provide a theoretical justification of this statistic, as well as several illustrative simulated examples demonstrating its potential value. We then demonstrate its value on a unique brain imaging dataset generated by the Consortium for Reliability and Reproducibility (CoRR) [18]. This dataset is an amalgamation of over 28 different studies, many of which were collected using different scanners, manufactured by different companies, and run by different people, using different settings. Moreover, the scanned individuals span various age ranges, sexes, and ethnicities. Nonetheless, we are interested in finding a pipeline to analyze the data such that they can be used for many different inference tasks. After evaluating nearly 200 different analysis pipelines on over 3000 scans, we determined the optimal pipeline, that is, the pipeline with the highest Discr. We then demonstrate that for every single dataset, on average, pipelines that achieve higher Discr also yield data with more information about multiple phenotypes. This is despite the fact that no phenotypic information whatsoever was incorporated into the optimal design criterion. This is in contrast with other potential design criteria, which did not exhibit this property. We therefore believe optimizing experiments to improve reproducibility, specifically by maximizing discriminability, will be useful for a wide range of disciplines and sectors. To facilitate its use, we make all of our code and data derivatives open access at https://neurodata.io/mgc.

2 Data Reproducibility Statistics

2.1 Intra-Class Correlation The intra-class correlation coefficient (ICC) is a commonly used data reproducibility statistic [19]. ICC is the fraction of the total variability that is across-item variability, that is, ICC is defined as the across-item variability divided by the within-item plus across-item variability. ICC has several severe limitations. First, it is univariate, meaning if the data are multidimensional, they must first be represented by univariate statistics, thereby discarding multivariate information. Second, ICC is based on an (overly simplistic) Gaussian assumption characterizing the data. Thus, any deviations from this assumption render the interpretation of the magnitude of ICC questionable, because non-Gaussian measurements that are highly reliable could yield quite low ICC.

2.2 Image Intra-Class Correlation (I2C2 The Image Intra-Class Correlation (I2C2) was introduced to mitigate ICC's univariate limitation [20]. Specifically, I2C2 operates on covariances matrices, rather than variances. To obtain a univariate summary of reproducibility, I2C2 operates on the trace of the covariance matrices, one of several possible strategies, similar to most multivariate analysis of variance procedures [21]. Thus, while overcoming one limitation of ICC, I2C2 still heavily leverages Gaussian assumptions of the data to justify its validity. We introduce a complementary multivariate parametric generalization of ICC which we call Principle Component Intra-Class Correlation (PICC). PICC is simply ICC computed on the the first principle component of the data. The main advantage of PICC over I2C2 is conceptual and computational simplicity; empirically, it performs approximately as well.

2.3 Non-Parametric Discriminability

Our main contribution is the introduction of a nonparametric multivariate (and non-Euclidean) data reproducibility statistic which we call Discr. Discr quantifies the degree to which multiple measurements of the same item are more similar to one another than they are to other items, without making any parametric assumptions, and without requiring the data to be univariate. Here we outline how one can compute Discr from data.

Consider n items, where each item has s measurements, resulting in $N = n \times s$ total measurements across items, then Discr can be computed as follows:

1. Compute the distance between all pairs of samples (resulting in an $N \times N$ matrix).
2. For all samples of all items, compute the fraction of times that a within-item distance is smaller than an across-item distance.
3. The Discr of the dataset is the average of the above mentioned fraction.

A high Discr indicates that within-item measurements are more similar to one another than across-item measurements. For more algorithmic details, see Algorithm 9. For formal definition of terms, see Appendix A.

3 Theoretical properties of discriminability

Under reasonably general assumptions, if within-item variability increases, predictive accuracy will subsequently decrease. Therefore, a statistic that is sensitive to within-item variance is desirable for optimal experimental design, regardless of the distribution of the data. Carmines and Zeller [22] introduces a univariate parametric framework in which predictive accuracy can be lower-bounded by a decreasing function of ICC; as a direct consequence, a strategy with a higher ICC can, on average, have higher predictive performance on subsequent inference tasks. Unfortunately, this valuable theoretical result is limited in its applicability, as it is restricted to univariate data, whereas big data analysis strategies often produce data in high dimensions. We therefore prove the following generalization of this theorem (see Appendix B for proof):

Theorem 3.1. *Under the multivariate additive noise setting,* Discr *provides a lower bound on the predictive accuracy of a subsequent classification task. Consequently, a strategy with a higher* Discr *provably provides a higher bound on predictive accuracy than a strategy with a lower* Discr*.*

Thus, Discr provides a theoretical extension of ICC to a non-parametric multivariate model, and correspondingly, motivates optimizing experiments to obtain higher Discr.

4 Empirical properties of discriminability on simulated data

4.1 Simulation settings

To develop insight into the performance of Discr, we consider four different simulation settings. Each includes between 2 and 20 items, with s measurements per item, in 2 dimensions. Figure 1A shows a two-dimensional scatterplot of each setting, and Figure 1B shows the Euclidean distance matrix between samples, ordered by item. The four settings are (see Appendix D for details):

1. **Gaussian** Each item is distributed according to a spherically symmetric Gaussian, therefore respecting the assumptions that motivate ICC.
2. **Cross** Both items have Gaussian distributions with the same mean, but they are no longer spherically symmetric and have different covariances, specifically, the different dimensions of each item have different distributions.
3. **Ball/Circle** One item is distributed in the unit ball, the other on the unit circle; Gaussian noise is added to both, but neither item is entirely characterized by a Gaussian.
4. **No Signal** Both items have the same Gaussian distribution.

4.2 Discr empirically predicts performance on subsequent inference tasks

We compare the empirical performance of Discr to PICC and I2C2 by investigating the sensitivity of each statistic to changes in within-item variability. Figure 1C shows the impact of increasing within-item variance on the four different simulation settings. For the three with predictive information, increasing variance decreases predictive accuracy (green line). As desired, Discr also decreases nearly perfectly propor-

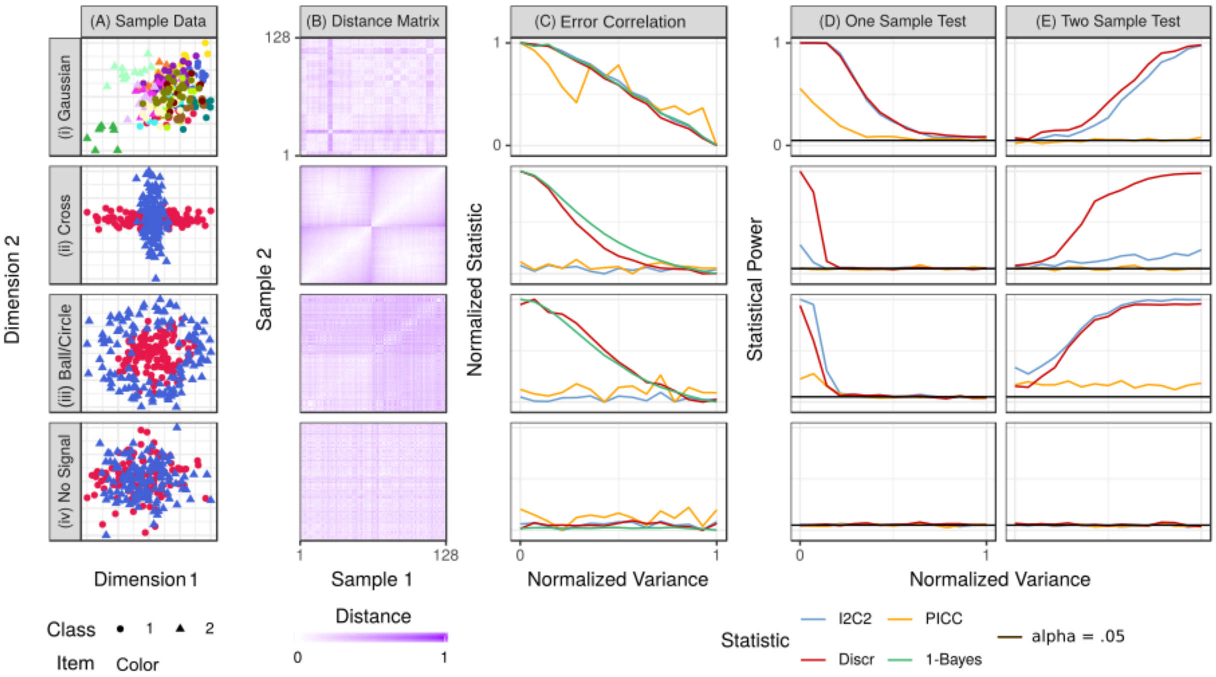

Figure 1: **Four simulations demonstrate the value of Discr for optimal experimental design**. All simulations are two-dimensional, with 128 samples, and $\alpha = 0.05$, with 500 iterations per setting. *(i) Gaussian* $K = 16$ individuals, with the item-specific and measurement-specific effects spherically Gaussian distributed. *(ii) Cross* $K = 2$ items, where each item is still Gaussian distributed, but the different items have different variances for the difference dimensions. *(iii) Annulus/Disc* $K = 2$ individuals, where one is a distributed in an annulus, and the other within the unit disc, and white Gaussian noise is added to both. *(iv) No Signal* A simulation where the two individuals have equal distribution. For each, class label is indicated by shape, and color indicates item. **(B)** The Euclidean distance matrix between samples within each simulation setting. Samples are organized by item. Simulation settings in which items are discriminable tend to have a block structure in which samples from the same item are relatively similar to one another. **(C)** A comparison of the observed statistics to $1-$ Bayes error. Only Discr correctly tracks changes in within-item variance. **(D)** One-sample test of whether data are discriminable, Discr achieves nearly as high or higher power than I2C2 and ICC for all settings and variances. **(E)** Two-sample test of which approach is more discriminable. Discr achieves highest power for all settings and variances. For all simulations, the variance is normalized (Appendix D for details).

tionally with decreasing variances. However, only in the first setting, where each item has a spherically symmetric Gaussian distribution, do I2C2 and PICC drop proportionally. Even in the second (Gaussian) setting, I2C2 and PICC are effectively uninformative about the within-item variance. And in the third (non-Gaussian) setting, they are similarly useless. This suggests that of these statistics, only Discr can serve as a satisfactory surrogate for predictive accuracy under these relatively simple settings.

4.3 A test for discriminability A prerequisite for making item-specific predictions is that items are different from one another in predictable ways, that is, are discriminable. If not, the same assay applied to the same individual on multiple trials could result in unacceptably highly variable results. Thus, prior to embarking on a machine learning search for predictive accuracy using some data, one can simply test whether the data are discriminable at all. If not, predictive accuracy will be hopeless. Letting D denote the Discr of a dataset with n items and s measurements per item, and D_0 denote the Discr of the same size dataset with zero item specific information, the formal *one-sample* hypothesis test for

$\textsc{Discr}$ is

$$\begin{aligned} H_0 &: D = D_0, \\ H_A &: D > D_0. \end{aligned} \tag{1}$$

One could replace D for $\textsc{Discr}$ with some other test statistic, such as $\texttt{PICC}$ or $\texttt{I2C2}$. We devised a permutation test to obtain a distribution of the test statistic under the null, and a corresponding p-value. To evaluate the different procedures, we compute the power of each test, that is, the probability of correctly rejecting the null when it is false (which is one minus type II error; see Appendix C.1 for details).

Figure 1D shows that $\textsc{Discr}$ achieves as high power as $\texttt{I2C2}$, and higher power than $\texttt{PICC}$, in the spherical Gaussian setting. This result demonstrates that despite the fact that $\textsc{Discr}$ does not rely on Gaussian assumptions, it still performs as well or better than parametric methods when the data satisfy these assumptions. In the cross setting, only $\textsc{Discr}$ correctly identifies that items differ from one another, despite the fact that the data are Gaussian. In the ball/disc setting, both $\textsc{Discr}$ and $\texttt{I2C2}$ perform comparably. And when there is no signal, all tests are valid, achieving power less than or equal to the critical value. Non-parametric $\textsc{Discr}$ therefore has the power of parametric approaches for data at which those assumptions are appropriate, and much higher power for other data.

4.4 A test for whether one experimental design is more discriminible than another

Given two experimental designs—which can differ either by acquisition and/or analysis details—are the measurements produced by one method more discriminable than the other? Formally, letting $D^{(1)}$ be the $\textsc{Discr}$ of the first approach, and $D^{(2)}$ be the $\textsc{Discr}$ of the second approach, we have the following *two-sample* hypothesis for discriminability:

$$\begin{aligned} H_0 &: D^{(1)} = D^{(2)}, \\ H_A &: D^{(1)} > D^{(2)}. \end{aligned} \tag{2}$$

Again, one could replace $\textsc{Discr}$ with other test statistics, and we devised a permutation test to obtain the distribution of the test statistic under the null, and p-values (see Appendix C.2 for details). Figure 1D shows $\textsc{Discr}$ achieves nearly as high or higher power than both $\texttt{I2C2}$ and $\texttt{PICC}$ for all three settings with across-item differences, and all tests are valid. The fact that $\textsc{Discr}$ achieves nearly equal or higher power than the Gaussian methods, even under Gaussian assumptions, suggests that $\textsc{Discr}$ will be a superior metric for optimal experimental design.

5 Empirical Discriminability on Real Data

5.1 Real data acquisition and analysis

Consortium for Reliability and Reproducibility (CoRR) [23] has generated functional, anatomical, and diffusion magnetic resonance imaging (dMRI) scans from >1,600 participants, often with multiple measurements, collected through 28 different studies spanning over 20 sites. Each of the sites use different scanners, technicians, and scanning protocols, thereby representing a wide variety of different acquisition settings with which one can test different analysis pipelines. Figure 2A shows the six stage sequence of analysis steps for converting the raw fMRI data into networks or connectomes, that is, estimates of the strength of connections between all pairs of brain regions. At each stage of the pipeline, we consider several different "standard" approaches, that is, approaches that have previously been proposed in the literature, typically with hundreds or thousands of citations [24]. Moreover, they have all been collected into an analysis engine, called Configurable Pipeline for the Analysis of Connectomes (C-PAC) [25]. In total, for the six stages together, we consider $2 \times 2 \times 2 \times 2 \times 4 \times 3 = 192$ different analysis pipelines. Because each stage is nonlinear, it is possible that the best sequence of choices is not equivalent to the best choices on their own. For this reason, publications that evaluate a given stage using any metric, could result in misleading conclusions if one is searching for the best sequence of steps. The dMRI connectomes were acquired via 48 analysis pipelines using the Neurodata MRI Graphs ($\texttt{ndmg}$) pipeline [26]. Appendix E provides specific details for both fMRI and dMRI analysis, as well as the options attempted.

5.2 Different analysis strategies yield widely disparate discriminabilities Figure 2B shows the analysis strategy has a large impact on the Discr of the resulting fMRI connectomes. Each column shows one of 64 different analysis strategies, ordered by how significantly different they are from the pipeline with greatest Discr (averaged over all datasets, tested using the above two-sample test). Interestingly, pipelines with worse average Discr also tend to have higher variance across datasets. The best pipeline, FNNNCP, uses FSL registration, no frequency filtering, no scrubbing, no global signal regression, CC200 parcellation, and converts edges weights to ranks. The majority of the strategies $(51/64 \approx 80\%)$ show significantly worse Discr than the optimal strategy at $\alpha = 0.05$ (Discr 1-sample test). In other words, several standard procedures for analyzing these data *reduce* Discr on average, calling into question whether they should be used in resting state fMRI connectomics.

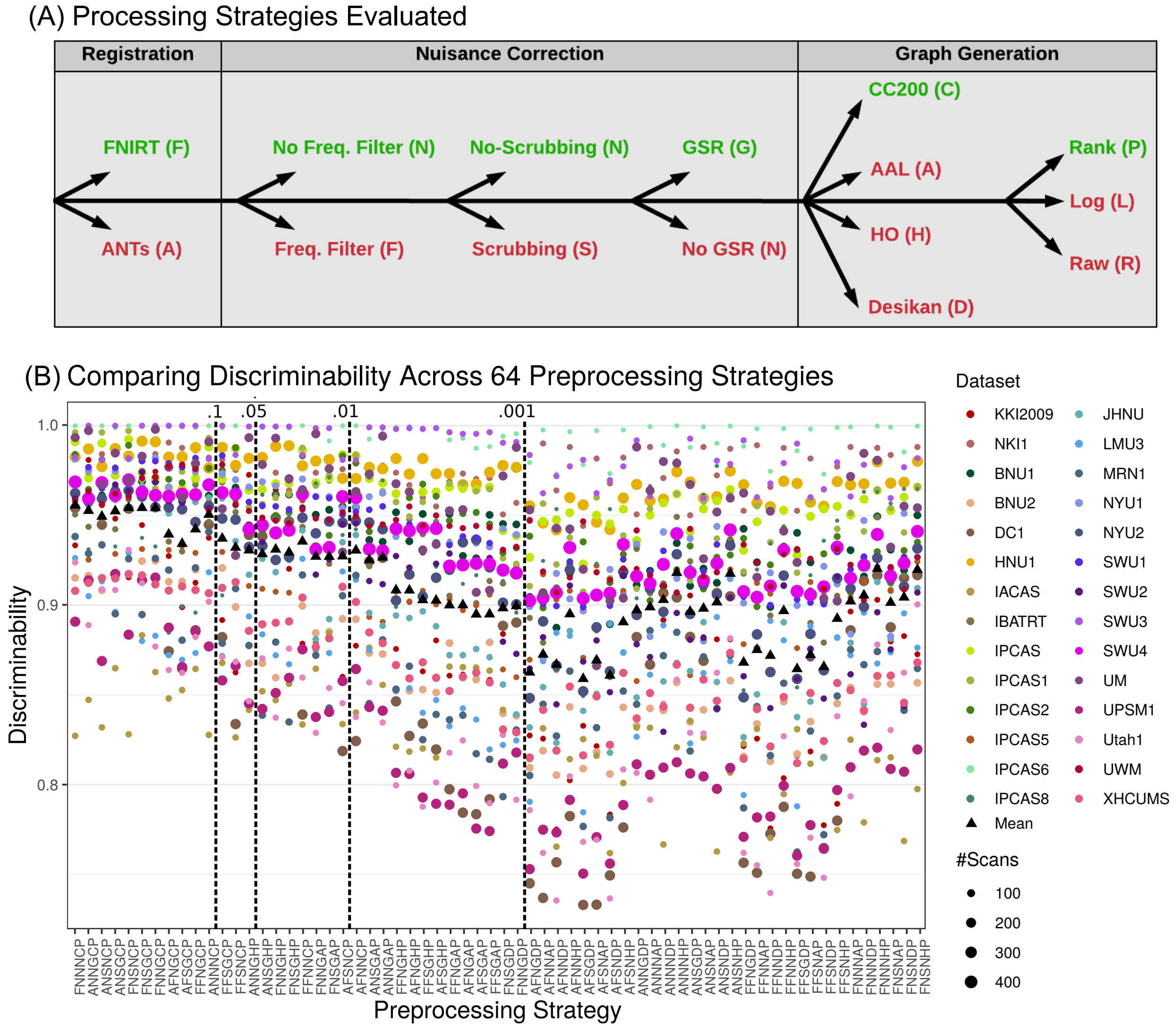

Figure 2: **Different analysis strategies yield widely disparate discriminabilities. (A)** a schematic illustrating the analysis options for the 192 fMRI pipelines under consideration (described in Appendix E). The optimal choices are green. **(B)** Discr of fMRI Connectomes analyzed 64 ways. Functional correlation matrices are estimated from 28 multi-session studies from the CoRR dataset using each pipeline. The analysis strategy codes are assigned sequentially according to the abbreviations listed for each step in **(A)**. The mean Discr per pipeline is a weighted sum of its discriminabilities across datasets. Each pipeline is compared to the optimal pipeline with the highest mean Discr, FNNNCP, using the above two-sample hypothesis test. The remaining strategies are arranged according to p-value, indicated in the top row.

5.3 Discr identifies which acquisition and analysis decision are most important for improving performance While the above analysis provides evidence for which sequence of analysis steps is best, it does not provide information about which choices individually have the largest impact on overall Discr. To do so, it is inadequate to simply fix a pipeline and only swap out algorithms for a single stage, as such an analysis will only provide information about that fixed pipeline. Therefore, we evaluate each choice in the context of all 192 considered pipelines in Figure 3A. If one were to independently select the best option for each analysis stage (FNNGCP), although it is not exactly the same as the pipeline with highest Discr (FNNNCP), it is also not significantly worse (Discr 2-sample test, p-value ≈ 0.14). Moreover, except for scrubbing, each stage has a significant impact on Discr after correction for multiple hypotheses (Wilcoxon signed-rank statistic, p-values all < 0.001).

Another choice is whether to estimate connectomes using functional or diffusion MRI. Whereas both data acquisition strategies have known problems [27], the Discr of the two experimental modalities has not been directly compared. Using four datasets from CoRR that acquired both fMRI and dMRI on the same subjects, and have quite similar demographic profiles, we tested whether fMRI or dMRI derived connectomes were more discriminable. For three of the four datasets, dMRI connectomes were more discriminable. This is not particularly surprising, given the susceptibility of fMRI data to changes in state rather than trait (e.g., amount of caffeine prior to scan [25]).

The above results motivate investigating which aspects of the dMRI analysis strategy were most effective. We focus on two criteria: how to scale the weights of connections, and how many regions of interest (ROIs) to use. Figure 3C.i shows that the log transform tends to yield more discriminable connectomes, though not significantly so (Wilcoxon signed-rank statistic, p-value≈ 0.40). Both rank and log transform significantly exceed raw edge weights (Wilcoxon signed-rank statistic, p-value< 0.001). Figure 3C.ii shows that parcellations with larger numbers of ROIs tend to have larger discriminabilities. Unfortunately, most parcellations with semantic labels (e.g., visual cortex) have hundreds not thousands of parcels. This result therefore motivates the development of more refined semantic labels.

5.4 Optimizing Discr improves downstream inference performance We next examined the relationship between the Discr of each pipeline, and the amount of information it preserves about two properties of interest: sex and age. Based on the simulations above, we expect that analysis pipelines with higher Discr will yield connectomes with more information about covariates. Indeed, Figure 4 shows that, for every single dataset (28 in total), a pipeline with higher Discr tends to preserve more information about both covariates. The amount of information is quantified by the effect size of the multiscale graph correlation statistic MGC [28, 29], a statistic that quantifies the magnitude of association for both linear and nonlinear dependence structures. In contrast, if one were to use either PICC or I2C2 to select the optimal pipeline, for many datasets, subsequent predictive performance would degrade. These results are highly statistically significant: the slopes of effect size versus Discr across datasets are positive for both age and sex (Fisher's corrected [30] t-test, p-value < 0.001 for both), but of the others, only PICC versus age has a significantly positive slope, with a trivially small effect size (0.004).

6 Discussion We propose the use of the Discr as a simple and intuitive measure for experimental design featuring multiple measurements. Numerous efforts have established the value of *quantifying* reliability, repeatability, and replicability (or stability) using parametric measures such as ICC and I2C2. However, they have not been used to optimize reproducibility—that is, they are only used post-hoc to determine reproducibility, not used as criteria for searching over the design space—nor have non-parametric multivariate generalizations of these statistics been available. We derive one-sample (goodness-of-fit) and two-sample (equality) tests for Discr, and demonstrate via theory and simulation that Discr provides numerous advantages over existing techniques across a range of simulated settings. Our neuroimaging use-case exemplifies the utility of these features of the Discr framework for optimal experimental design.

Discr provides a number of connections with related statistical algorithms worth further consideration. Discr is related to energy statistics [31], in which the statistic is a function of distances between

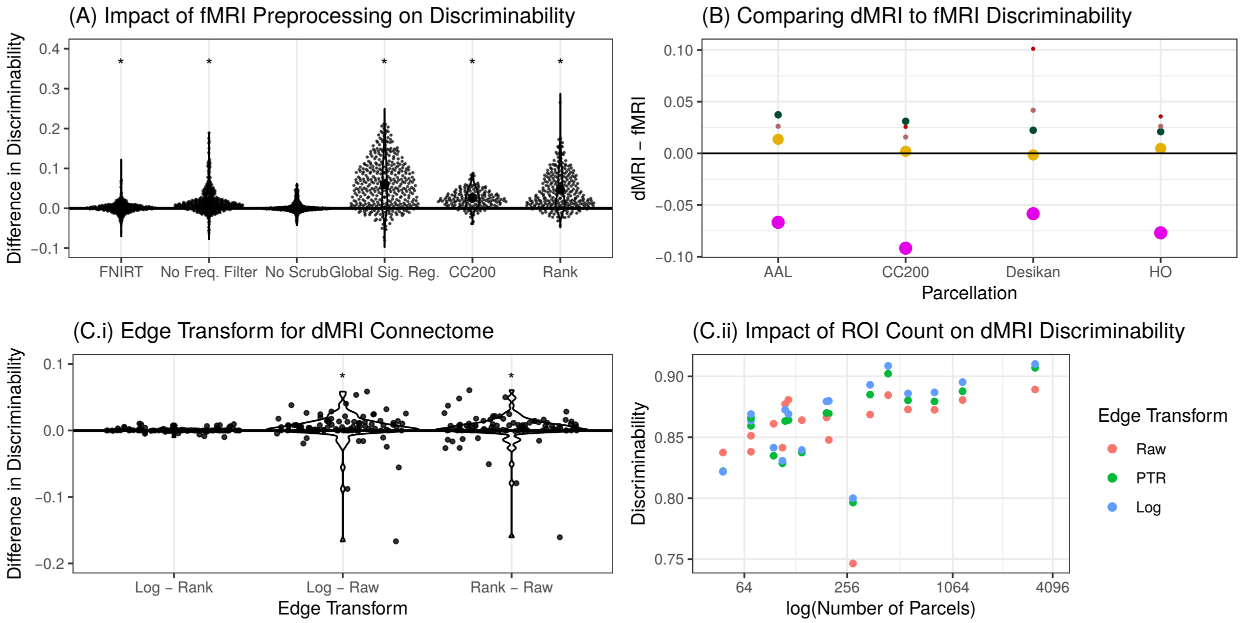

Figure 3: **Parsing the relative impact on Discr of various acquisition and analytic choices**. **(A)** The pipelines are aggregated for a particular analysis step, with pairwise comparisons with the remaining analysis options held fixed. The beeswarm plot shows the difference between the overall best performing option and the second best option for each stage (mean in bigger black dot); the x-axis label indicates the best performing strategy. The best strategies are FNIRT, no frequency filtering, no scrubbing, global signal regression, the CC200 parcellation, and ranks edge transformation. A Wilcoxon signed-rank test is used to determine whether the mean for the best strategy exceeds the second best strategy: a * indicates that the p-value is at most 0.001 after Bonferroni correction. Of the best options, only no scrubbing is *not* significantly better than alternative strategies. Note that the options that perform marginally the best are not significantly different than the best performing strategy overall, as shown in Figure 2. **(B)** A comparison of the discriminabilities for the 4 datasets with both fMRI and dMRI connectomes. dMRI connectomes tend to be more discriminable, in 14 of 20 total comparisons. **(C.i)** Comparing raw edge weights (Raw), ranking (Rank), and log-transforming the edge-weights (Log) for the diffusion connectomes, the Log and Rank transformed edge-weights tend to show higher Discr than Raw. **(C.ii)** As the number of ROIs increases, the Discr tends to increase.

observations [32]. Energy statistics provide approaches for goodness-of-fit (one-sample) and equality testing (two-sample), and multi-sample testing [33]. Similar to Discr, energy statistics make relatively few assumptions. However, energy statistics requires a large number of measurements per item, which is often unsuitable for biological data where we frequently have only a small number of repeated measurements. Discr is most closely related to multiscale generalized correlation (MGC) [28, 29], which combines energy statistics with nearest neighbors, as does Discr.

While Discr provides experimental design guidance for big data, other considerations may play a role in a final determination. For example, the connectomes analyzed here are *resting-state*, as opposed to *task-based* fMRI connectomes. Recent literature suggests that the global signal in a rs-fMRI scan may be a nuisance variable for task-based approaches [34, 35]. Thus, while Discr is an effective tool for experimental design, knowledge of the techniques in conjunction with the inference task is still a necessary component of any investigation.

It is important to emphasize that Discr, as well the related statistics, are neither necessary, nor sufficient, for a measurement to be practically useful. For example, categorical covariates, such as sex, are often meaningful in an analysis, but not discriminable. Human fingerprints are discriminable, but typically not biologically useful. In addition, none of the statistics studied here are immune to sample characteristics, thus interpreting results across studies deserves careful scrutiny. For example, having a sample with variable ages will increase the inter-subject dissimilarity of any metric dependent on age

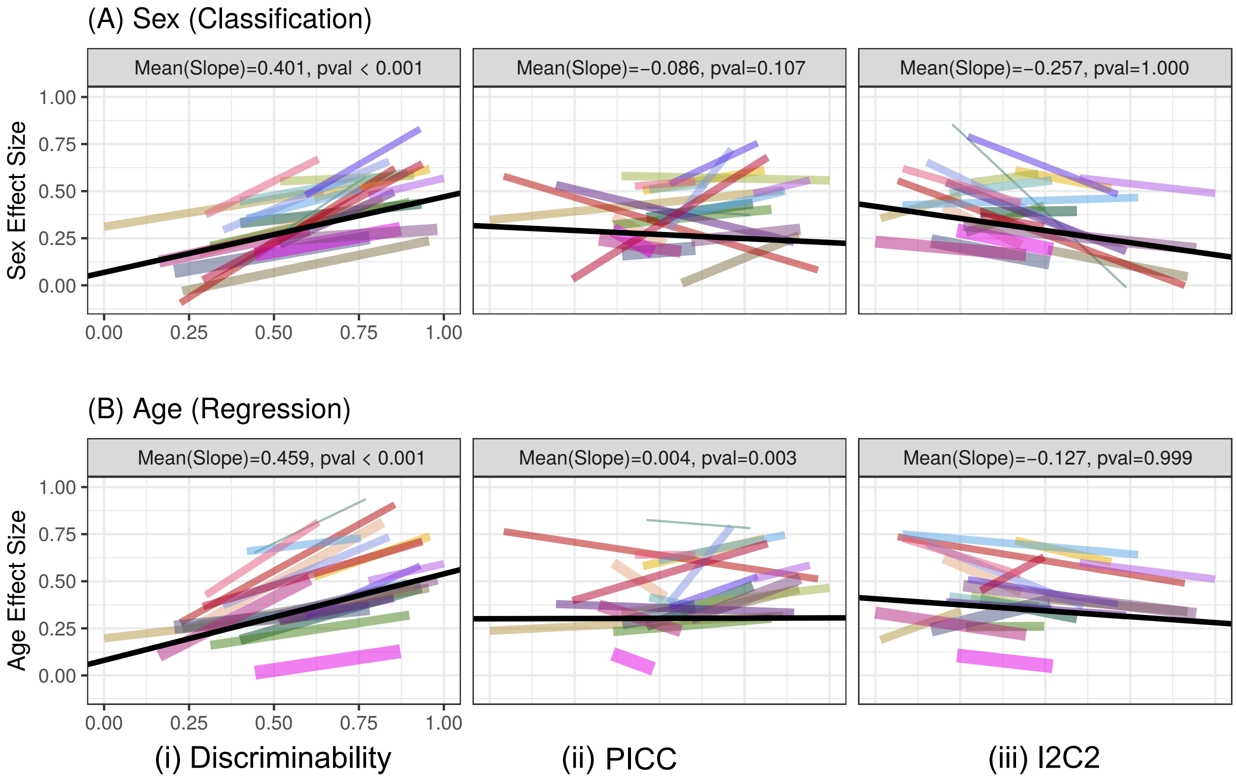

Figure 4: **Optimizing Discr improves downstream inference performance**. Using the connectomes from the 64 pipelines with raw edge-weights, we examine the relationship between connectomes vs **(A)** sex and **(B)** age. The columns evaluate difference approaches for computing effect size, including **(i)** Discr, **(ii)** PICC, and **(iii)** I2C2. Each panel shows effect size (*x axis*) versus MGC (*y axis*). Both the *x* and *y* axes are normalized by the minimum and maximum statistic. For each study, the effect size is regressed onto . Color and line width correspond to the study and number of scans, respectively (see Figure 2B). The solid black line is the weighted mean over all studies. Discr is the only statistic in which *all* slopes exceed zero. Moreover, we find that the corrected p-value [30] is significant across datasets for both covariates (med. p-value $< .001$). This indicates that pipelines with higher Discr correspond to larger effect sizes for the covariate of interest, and that this relationship is stronger for Discr than other statistics. Appendix E.2 details the methodologies employed.

(such as the connectome). With these caveats in mind, Discr remains as a key experimental design consideration a wide variety of settings.

Due to the high volume of open-access data with informative downstream inferential covariates, as well as the large number of open-source libraries for analyzing data, the connectomics use-case provided herein serves to illustrate how Discr can be used to facilitate experimental design, and mitigate reproducibility issues. We envision that Discr will find substantial applicability across disciplines and sectors beyond brain imaging, such as genomics, pharmaceutical research, and many other aspects of big data science and industry. To this end, we provide open-source implementations of Discr for both Python and R [36, 37]. Code for reproducing all the figures in this manuscript is available at https://neurodata.io/mgc.

Acknowledgements This work was partially supported by the National Science Foundation NEURONEX award DMS-1707298, and the Defense Advanced Research Projects Agency's (DARPA) SIMPLEX program through SPAWAR contract N66001-15-C-4041. X-NZ receives funding supports by the National Basic Research (973) Program (2015CB351702), the Natural Science Foundation of China (81471740, 81220108014), the China - Netherlands CAS-NWO Programme (153111KYSB20160020), Beijing Municipal Science and Tech Commission (Z161100002616023, Z171100000117012), the Major

Project of National Social Science Foundation of China (14ZDB161), the National R&D Infrastructure and Facility Development Program of China, Fundamental Science Data Sharing Platform (DKA2017-12-02-21), and Guangxi BaGui Scholarship (201621).

References

[1] Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*. Wiley, 2 edition edition, February 2009.

[2] Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, September 2013.

[3] John P A Ioannidis. Why most published research findings are false. *PLoS Med.*, 2(8):e124, August 2005.

[4] Monya Baker. Over half of psychology studies fail reproducibility test. *Nature Online*, August 2015.

[5] Prasad Patil, Roger D Peng, and Jeffrey T Leek. What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspect. Psychol. Sci.*, 11(4):539–544, July 2016.

[6] David Trafimow and Michael Marks. Editorial. *Basic Appl. Soc. Psych.*, 37(1):1–2, January 2015.

[7] Ronald D Fricker, Katherine Burke, Xiaoyan Han, and William H Woodall. Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban. *Am. Stat.*, 73 (sup1):374–384, March 2019.

[8] Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. Moving to a World Beyond "p < 0.05". *Am. Stat.*, 73(sup1):1–19, March 2019.

[9] Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John C S Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, Antao Chen, Bing Chen, Jiangtao Chen, Xu Chen, Stanley J Colcombe, William Courtney, R Cameron Craddock, Adriana Di Martino, Hao-Ming Dong, Xiaolan Fu, Qiyong Gong, Krzysztof J Gorgolewski, Ying Han, Ye He, Yong He, Erica Ho, Avram Holmes, Xiao-Hui Hou, Jeremy Huckins, Tianzi Jiang, Yi Jiang, William Kelley, Clare Kelly, Margaret King, Stephen M LaConte, Janet E Lainhart, Xu Lei, Hui-Jie Li, Kaiming Li, Kuncheng Li, Qixiang Lin, Dongqiang Liu, Jia Liu, Xun Liu, Yijun Liu, Guangming Lu, Jie Lu, Beatriz Luna, Jing Luo, Daniel Lurie, Ying Mao, Daniel S Margulies, Andrew R Mayer, Thomas Meindl, Mary E Meyerand, Weizhi Nan, Jared A Nielsen, David O'Connor, David Paulsen, Vivek Prabhakaran, Zhigang Qi, Jiang Qiu, Chunhong Shao, Zarrar Shehzad, Weijun Tang, Arno Villringer, Huiling Wang, Kai Wang, Dongtao Wei, Gao-Xia Wei, Xu-Chu Weng, Xuehai Wu, Ting Xu, Ning Yang, Zhi Yang, Yu-Feng Zang, Lei Zhang, Qinglin Zhang, Zhe Zhang, Zhiqiang Zhang, Ke Zhao, Zonglei Zhen, Yuan Zhou, Xing-Ting Zhu, and Michael P Milham. An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci Data*, 1: 140049, December 2014.

[10] David O'Connor, Natan Vega Potler, Meagan Kovacs, Ting Xu, Lei Ai, John Pellman, Tamara Vanderwal, Lucas C Parra, Samantha Cohen, Satrajit Ghosh, Jasmine Escalera, Natalie Grant-Villegas, Yael Osman, Anastasia Bui, R Cameron Craddock, and Michael P Milham. The Healthy Brain Network Serial Scanning Initiative: a resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions. *Gigascience*, 6(2):1–14, February 2017.

[11] Xi-Nian Zuo, Ting Xu, and Michael Peter Milham. Harnessing reliability for neuroscience research. *Nat Hum Behav*, 3(8):768–771, August 2019.

[12] Aki Nikolaidis, Anibal Solon Heinsfeld, Ting Xu, Pierre Bellec, Joshua Vogelstein, and Michael Milham. Bagging Improves Reproducibility of Functional Parcellation of the Human Brain. July 2019.

[13] David J Hand. *Measurement: A Very Short Introduction*. Oxford University Press, 1 edition edition, 2016.

[14] Ronald A Fisher. *The Design of Experiments*. Macmillan Pub Co, 1935.

[15] R E Kirk. Experimental Design. In Irving Weiner, editor, *Handbook of Psychology, Second Edition*, volume 12, page 115. John Wiley & Sons, Inc., Hoboken, NJ, USA, September 2012.

[16] Anders M Dale. Optimal experimental design for event-related fmri. *Human brain mapping*, 8(2-3): 109–114, 1999.

[17] Liam Paninski. Asymptotic theory of information-theoretic experimental design. *Neural Comput.*, 17(7):1480–1507, July 2005.

[18] Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John CS Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1:140049, 2014.

[19] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[20] H Shou, A Eloyan, S Lee, V Zipunnikov, AN Crainiceanu, MB Nebel, B Caffo, MA Lindquist, and CM Crainiceanu. Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (i2c2). *Cognitive, Affective, & Behavioral Neuroscience*, 13(4):714–724, 2013.

[21] Carl J Huberty and Stephen Olejnik. *Applied MANOVA and Discriminant Analysis*. John Wiley & Sons, May 2006.

[22] Edward G Carmines and Richard A Zeller. *Reliability and Validity Assessment*. SAGE Publications, November 1979.

[23] Xi-Nian Zuo, Clare Kelly, Jonathan S Adelstein, Donald F Klein, F Xavier Castellanos, and Michael P Milham. Reliable intrinsic connectivity networks: test–retest evaluation using ica and dual regression approach. *Neuroimage*, 49(3):2163–2177, 2010.

[24] Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10): 4734–4739, 2010.

[25] S Sikka, B Cheung, R Khanuja, S Ghosh, C Yan, Q Li, J Vogelstein, R Burns, S Colcombe, C Craddock, et al. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). In *5th INCF Congress of Neuroinformatics, Munich, Germany*, volume 10, 2014.

[26] Gregory Kiar, Eric Bridgeford, Will Gray Roncal, Consortium for Reliability (CoRR), Reproducibility, Vikram Chandrashekhar, Disa Mhembere, Sephira Ryman, Xi-Nian Zuo, Daniel S Marguiles, R Cameron Craddock, Carey E Priebe, Rex Jung, Vince Calhoun, Brian Caffo, Randal Burns, Michael P Milham, and Joshua Vogelstein. A High-Throughput Pipeline Identifies Robust Connectomes But Troublesome Variability. *bioRxiv*, page 188706, apr 2018. doi: 10.1101/188706. URL https://www.biorxiv.org/content/early/2018/04/24/188706.

[27] Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S Ghosh, Chaogan Yan, Qingyang Li, Daniel Lurie, Joshua Vogelstein, Randal Burns, Stanley Colcombe, Maarten Mennes, Clare Kelly, Adriana Di Martino, Francisco X. Castellanos, and Michael Milham. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC). *Frontiers in Neuroimformatics*, July 2013.

[28] Cencheng Shen, Carey E Priebe, and Joshua T Vogelstein. From Distance Correlation to Multiscale Generalized Correlation. *Journal of American Statistical Association*, October 2017. URL http://arxiv.org/abs/1710.09768.

[29] Joshua T Vogelstein, Eric W Bridgeford, Qing Wang, Carey E Priebe, Mauro Maggioni, and Cencheng Shen. Discovering and deciphering relationships across disparate data modalities. *Elife*, 8, January 2019. URL http://dx.doi.org/10.7554/eLife.41690.

[30] Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[31] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *J. Stat. Plan. Inference*, 143(8):1249–1272, August 2013.

[32] Maria L Rizzo and Gábor J Székely. Energy distance. *WIREs Comput Stat*, 8(1):27–38, January 2016.

[33] Maria L Rizzo, Gábor J Székely, et al. Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010.

[34] Kevin Murphy and Michael D Fox. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage*, 154:169–173, July 2017.

[35] Thomas T Liu, Alican Nalci, and Maryam Falahpour. The global signal in fMRI: Nuisance or information? *Neuroimage*, 150:213–229, April 2017.

[36] Sambit Panda, Satish Palaniappan, Junhao Xiong, Ananya Swaminathan, Sandhya Ramachandran, Eric W Bridgeford, Cencheng Shen, and Joshua T Vogelstein. mgcpy: A comprehensive high dimensional independence testing python package. July 2019.

[37] Eric Bridgeford, Censheng Shen, Shangsi Wang, and Joshua T. Vogelstein. Multiscale generalized correlation, May 2018. URL https://doi.org/10.5281/zenodo.1246967.

[38] Zeyi Wang, Eric W Bridgeford, Joshua T Vogelstein, and et al Caffo, Brian. Statistical analysis of data reproducibility measures.

[39] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[40] REAC Paley and A Zygmund. On some series of functions,(3). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 190–205. Cambridge Univ Press, 1932.

[41] Pierre A Devijver and Josef Kittler. *Pattern recognition: A statistical approach*. Prentice hall, 1982.

[42] Patrick E Meyers, Ganesh C Arvapalli, Sandhya C Ramachandran, Paige F Frank, Allison D Lemmer, Eric W Bridgeford, and Joshua T Vogelstein. Standardizing human brain parcellations. *Biorxiv*, October 2019.

[43] Stephen M Smith et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23 Suppl 1:S208–19, jan 2004. ISSN 1053-8119. URL http://www.ncbi.nlm.nih.gov/pubmed/15501092.

[44] Mark W Woolrich et al. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45(1 Suppl):S173–86, mar 2009. ISSN 1095-9572. URL http://www.sciencedirect.com/science/article/pii/S1053811908012044.

[45] Mark Jenkinson et al. FSL. *NeuroImage*, 62(2):782–90, aug 2012. ISSN 1095-9572. URL http://www.ncbi.nlm.nih.gov/pubmed/21979382.

[46] John Mazziotta et al. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association*, 8(5):401–430, 2001.

[47] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian Nimmo-Smith. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8, 2014.

[48] Eleftherios Garyfallidis, Matthew Brett, Marta Morgado Correia, Guy B Williams, and Ian Nimmo-Smith. Quickbundles, a method for tractography simplification. *Frontiers in neuroscience*, 6:175, 2012.

[49] Disa Mhembere, William Gray Roncal, Daniel Sussman, Carey E Priebe, Rex Jung, Sephira Ryman, R Jacob Vogelstein, Joshua T Vogelstein, and Randal Burns. Computing scalable multivariate glocal invariants of large (brain-) graphs. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 297–300. IEEE, 2013.

[50] Nathalie Tzourio-Mazoyer et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1): 273–289, 2002.

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

[52] Nikos Makris, Jill M Goldstein, David Kennedy, Steven M Hodge, Verne S Caviness, Stephen V Faraone, Ming T Tsuang, and Larry J Seidman. Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia research*, 83(2):155–171, 2006.

[53] JL Lancaster. The Talairach Daemon, a database server for Talairach atlas labels. *NeuroImage*, 1997. ISSN 1053-8119.

[54] R Cameron Craddock, Saad Jbabdi, Chao-Gan Yan, Joshua T Vogelstein, F Xavier Castellanos, Adriana Di Martino, Clare Kelly, Keith Heberlein, Stan Colcombe, and Michael P Milham. Imaging

human connectomes at the macroscale. *Nat. Methods*, 10(6):524–539, June 2013. URL http://dx.doi.org/10.1038/nmeth.2482.

[55] Chandra S Sripada et al. Lag in maturation of the brain‘s intrinsic functional architecture in attention-deficit/hyperactivity disorder. *Proceedings of the National Academy of Sciences*, 111 (39):14259–14264, 2014.

[56] Daniel Kessler et al. Modality-spanning deficits in attention-deficit/hyperactivity disorder in functional networks, gray matter, and white matter. *The Journal of Neuroscience*, 34(50):16555–16566, 2014.

[57] Rahul S Desikan et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.01.021.

[58] Cencheng Shen and Joshua T Vogelstein. Decision Forests Induce Characteristic Kernels. November 2018. URL http://arxiv.org/abs/1812.00029.

[59] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Appendix A. Population and Sample Discriminability.

Suppose that $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$ represents a physical property of interest for a particular item i. In a biological context, for instance, an item could be a participant in a study, and the property of interest could be the individual's true brain network, or connectome. We cannot directly observe the physical property, but rather, we must first measure $\boldsymbol{\theta}_i$ and then "wrangle" it. Call the measurement function, $f \in \mathcal{F}$ for a family of possible measurement functions $\mathcal{F}$ That is, $f : \boldsymbol{\Theta} \to \boldsymbol{\mathcal{W}}$. So, measurements of $\boldsymbol{\theta}_i$ are observed as $f(\boldsymbol{\theta}_i) = \boldsymbol{w}_i$. However, $\boldsymbol{w}_i$ may be a noisy, with measurement artefacts. Alternately, $\boldsymbol{w}_i$ might not be the property of interest, for example, if the property is a network, perhaps $\boldsymbol{w}_i$ is a multivariate time-series, from which we can estimate a network. We therefore have another function, $g \in \mathcal{G} : \boldsymbol{\mathcal{W}} \to \boldsymbol{\mathcal{X}}$, which represents the data wrangling procedure to take the measurement and produce an informative derivative (for instance, confound removal). The family of possible data wrangling procedures to produce the informative derivative is $\mathcal{G}$. In this fashion, the output of interest is $\boldsymbol{x}_i = g(f(\boldsymbol{\theta}_i))$.

The goal of experimental design is to choose an f and g that yield high-quality and useful inferences, that is, that yield $\boldsymbol{x}$'s that we can use for various inferential purposes. When we have repeated measurements of the same items, we can use those samples to our advantage. Given $\boldsymbol{x}_i^j$, which is the jth measurement of sample i, we would expect $\boldsymbol{x}_i^j$ to be more similar to $\boldsymbol{x}_i^{j'}$ (another measurement of the same item), than to any measurement of a different item $\boldsymbol{x}_{i'}^{j''}$. Formally, let $\delta : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}} \to [0, \infty)$ be a distance metric, we define the population discriminability:

$$D_{\delta,f,g} = \mathbb{P}\Big(\delta(\boldsymbol{x}_i^j, \boldsymbol{x}_i^{j'}) < \delta(\boldsymbol{x}_i^j, \boldsymbol{x}_{i'}^{j''})\Big)$$

That is, "population discriminability" D represents the average probability that the *within-item distance* $\delta(\boldsymbol{x}_i^j, \boldsymbol{x}_i^{j'})$ is less than the *between-item distance* $\delta(\boldsymbol{x}_i^j, \boldsymbol{x}_{i'}^{j''})$. Discriminability depends on the choice of distance δ, as well as the measurement protocal f and the analysis choices g.

The population discriminability represents a property of the distribution of $\boldsymbol{\theta}_i$. In real data since we do not observe the true distribution, we instead rely on the sample discriminability. Suppose a dataset consists of $i \in \{1, \ldots, n\}$ items, where each item i has J_i repeat measurements. The sample discriminability is defined:

$$\texttt{Discr}\Big\{\boldsymbol{x}_i^j\Big\}_{j\in[J_i], i\in[n]} = \frac{\sum_{i\in[n]}\sum_{j\in[J_i]}\sum_{j'\neq j}\sum_{i'\neq i}\sum_{j''\in[J_{i'}]}\mathbb{I}\Big\{\delta(\boldsymbol{x}_i^j, \boldsymbol{x}_i^{j'}) < \delta(\boldsymbol{x}_i^j, \boldsymbol{x}_{i'}^{j''})\Big\}}{\sum_{i\in[n]}\sum_{j\in[J_i]}\sum_{j'\neq j}\sum_{i'\neq i}\sum_{j''\in[J_{i'}]} 1}.$$

It can be shown [38] that the under the multivariate additive noise model; that is, $\boldsymbol{x}_i^j = \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i^j$ where $\boldsymbol{\epsilon}_i^j \overset{ind}{\sim} f_\epsilon$, $\text{var}\Big(\boldsymbol{\epsilon}_i^j\Big) < \infty$, and $\mathbb{E}\Big[\boldsymbol{\epsilon}_i^j\Big] = \boldsymbol{c}$, that the sample discriminability, Discr is both a consistent and unbiased estimator for population discriminability.

Appendix B. Discriminability Provides an Informative Bound for Inference.

During experimental design, the extent of subsequent inference tasks may be unknown. A natural question may be, what are the implications of the selection of a discriminable experimental design?

Formally, assume the task of interest is binary classification: that is, $\mathcal{Y} = \{0, 1\}$, and we seek a classifier $h\colon \mathcal{X} \to \mathcal{Y}$. The goal of experimental design in this context is to choose the options (f^*, g^*) that will minimize the classification loss:

$$(f^*, g^*) = \underset{(f,g)\in\mathcal{F}\times\mathcal{G}}{\text{argmin}}\ \mathbb{P}(h(f(g(\boldsymbol{\theta}))) \neq y)$$

For a fixed (f, g), the minimal prediction error is achieved by the Bayes classifier [39]:

$$h^*_{f,g}(\boldsymbol{\theta}_i) \triangleq \underset{y\in\{0,1\}}{\text{argmax}}\,\mathbb{P}\big(y_i = y\big|f(g(\boldsymbol{\theta}_i))\big),$$

and let $L^*_{f,g}$ denote Bayes error, that is, the error achieved by $h^*_{f,g}$.

Theorem B.1. *Assume the multivariate gaussian additive noise setting; that is:*

$$\boldsymbol{x}_i^j = \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i^j \tag{3}$$

where $\boldsymbol{\theta}_i \overset{iid}{\sim} \mathbb{P}_v$, $var(\boldsymbol{\theta}_i) = \boldsymbol{\Sigma}_\theta$, $\boldsymbol{\epsilon}_i^j \overset{iid}{\sim} \mathbb{P}_\epsilon$, *and* $var\Big(\boldsymbol{\epsilon}_i^j\Big) = \boldsymbol{\Sigma}_\epsilon$ *with* $\mathbb{E}\Big[\boldsymbol{\epsilon}_i^j\Big] = \boldsymbol{c}$. *There exists a decreasing function* $\gamma(\cdot)$ *which depends only on* $\boldsymbol{\theta}$ *and* y *s.t.:*

$$L^*_{f,g} \leq \gamma(D_{f,g})$$

That is, the Bayes error can, in fact, be upper bounded by a decreasing function of discriminability, as shown in the proof below. As a direct consequence of this theorem, we see:

Corollary B.1. *Assume* (f_1, g_1) *and* (f_2, g_2) *are two analysis strategies, and suppose that* $D_{f_1,g_1} > D_{f_2,g_2}$. *Then:*

$$L^*_{f_1,g_1} \leq L^*_{f_2,g_2}.$$

In other words, the Bayes error achieved by strategy (f_1, g_2) can, in fact, be upper bounded by the Bayes error achieveable by strategy (f_2, g_2). Consequently, under the described setting, the pipeline that achieves a higher Discr can facilitate improved inference than competing strategies, despite the fact that the task is unknown during data acquisition and analysis. Complementarily, note that if we were to instead consider the predictive accuracy $1 - L^*_{f,g}$, we can obtain a similar result to obtain a lower bound on the predictive accuracy via an increasing function of Discr. That is, in the context of the corollary, a more discriminable pipeline will tend to have a higher accuracy on an arbitrary predictive task.

Proof of Theorem (B.1).

Consider the additive noise setting, that is $\boldsymbol{x}_i^j = \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i^j$,

$$\begin{aligned}
D &= \mathbb{P}\big(\delta_{i,t,t'} < \delta_{i,i',t,t''}\big)\\
&= \mathbb{P}(\|\boldsymbol{x}_i^j - \boldsymbol{x}_i^{j'}\| < \|\boldsymbol{x}_i^j - \boldsymbol{x}_{i'}^{j''}\|)\\
&= \mathbb{P}(\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| < \|\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i^j - \boldsymbol{\theta}_{i'} - \boldsymbol{\epsilon}_{i'}^{j''}\|)\\
&\leq \mathbb{P}(\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| < \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\| + \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|)\\
&= \mathbb{P}(\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\| < \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|)\\
&= \frac{1}{2}\mathbb{P}(\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\| < \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|\|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\| < 0) +\\
&\quad \frac{1}{2}\mathbb{P}(\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\| < \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|\|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\| > 0)\\
&= \frac{1}{2} + \frac{1}{2}\mathbb{P}(\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\| < \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|\|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\| > 0)\\
&= \frac{1}{2} + \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big| < \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|)\\
&= 1 - \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big| > \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|).
\end{aligned}$$

To bound the probability above, we bound the $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|$ and $\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big|$ separately. We start with the first term

$$\mathbb{E}(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|^2) = \mathbb{E}(\boldsymbol{\theta}_i^T\boldsymbol{\theta}_i + \boldsymbol{\theta}_{i'}^T\boldsymbol{\theta}_{i'} - 2\boldsymbol{\theta}_i^T\boldsymbol{\theta}_{i'}) = 2\sigma_2^2.$$

Here, $\sigma_2^2 = \mathrm{tr}(\boldsymbol{\Sigma}_\theta)$ is the trace of covariance matrix of $\boldsymbol{\theta}_i$. We can apply Markov's Inequality for any $t > 0$:

(4) $$\mathbb{P}(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\| < t) \geq 1 - \frac{2\sigma_2^2}{t^2}.$$

Let $\sigma_1^2 = \mathrm{tr}(\boldsymbol{\Sigma}_\epsilon)$ denote the trace of covariance matrix of ϵ_i^j, and let a and b be two constants satisfy

$$\mathbb{E}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big|^2) \geq a^2\sigma_1^2,$$
$$\frac{\mathbb{E}^2(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big|^2)}{\mathbb{E}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big|)^4} \geq b$$

Furthermore, we let $t^2 = \sqrt{2}a\sigma_1\sigma_2$, and let

$$\theta = \frac{t^2}{\mathbb{E}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big|^2)} \leq \frac{\sqrt{2}a\sigma_1\sigma_2}{a^2\sigma_1^2} = \frac{\sqrt{2}\sigma_2}{a\sigma_1}.$$

If $a^2\sigma_1^2 \geq 2\sigma_2^2$, then $\theta \leq 1$. According to the Paley-Zygmund Inequality [40], that is,

$$\mathbb{P}(Z > \theta\mathbb{E}[Z]) \geq (1-\theta)^2\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}$$

for all $0 \leq \theta \leq 1$ and $Z \geq 0$, we can plug in the θ above to achieve

$$\mathbb{P}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big|^2 > t^2) \geq b\left(1 - \frac{t^2}{a^2\sigma_1^2}\right)^2 = b\left(1 - \frac{\sqrt{2}\sigma_2}{a\sigma_1}\right)^2.$$

Also plug in the t^2 for the inequality 4, we have

$$\mathbb{P}(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|^2 < t^2) \geq 1 - \frac{2\sigma_2^2}{t^2} = 1 - \frac{\sqrt{2}\sigma_2}{a\sigma_1}.$$

Understand the fact that $\boldsymbol{\theta}$'s and $\boldsymbol{\epsilon}$'s are independent, we can combine the two inequalities

$$\begin{aligned}
D &= \mathbb{P}(\delta_{i,t,t'} < \delta_{i,i',t,t''}) \\
&= \mathbb{P}(\|\boldsymbol{x}_i^j - \boldsymbol{x}_i^{j'}\| < \|\boldsymbol{x}_i^j - \boldsymbol{x}_{i'}^{j''}\|) \\
&\leq 1 - \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big| > \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|) \\
&\leq 1 - \frac{1}{2}\mathbb{P}(\big|\|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_i^{j'}\| - \|\boldsymbol{\epsilon}_i^j - \boldsymbol{\epsilon}_{i'}^{j''}\|\big|^2 > t^2)P(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}\|^2 < t^2) \\
&\leq 1 - \frac{1}{2}b\left(1 - \frac{\sqrt{2}\sigma_2}{a\sigma_1}\right)^3
\end{aligned}$$

Note that the resulted bound holds true even if $a^2\sigma_1^2 < 2\sigma_2^2$, as the right hand side becomes greater than 1. So we can have a bound on $\frac{\sigma_2}{\sigma_1}$,

(5) $$\frac{\sigma_2}{\sigma_1} \geq \frac{a}{\sqrt{2}}\left(1 - \left(\frac{2-2D}{b}\right)^{1/3}\right)$$

To obtain a bound on Bayes error, we apply Devijver and Kittler's result [41], which is

$$L \leq \frac{2\pi_0\pi_1}{1 + \pi_0\pi_1\Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\mu}}.$$

Here, π_0 and π_1 are prior probabilities for two classes. $\Delta\mu$ is the difference between means of two classes. Since $\boldsymbol{\epsilon}$ is assumed to be independent of $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\Delta\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x}|\boldsymbol{y}=0) - \mathbb{E}(\boldsymbol{x}|\boldsymbol{y}=1) = \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{y}=0) - \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{y}=1).$$

$\boldsymbol{\Sigma}$ is the weighted covariance matrix of $\boldsymbol{x}$,

$$\begin{aligned}\boldsymbol{\Sigma} &= \pi_0 \mathsf{Var}(\boldsymbol{x}|\boldsymbol{y}=0) + \pi_1 \mathsf{Var}(\boldsymbol{x}|\boldsymbol{y}=1)\\ &= \pi_0 \mathsf{Var}(\boldsymbol{\theta}|\boldsymbol{y}=0) + \pi_1 \mathsf{Var}(\boldsymbol{\theta}|\boldsymbol{y}=1) + \mathsf{Var}(\boldsymbol{\epsilon}).\end{aligned}$$

If we further assume $\mathsf{Var}(\boldsymbol{\epsilon}) = \sigma_1^2\boldsymbol{\Sigma}'$ where the trace of $\boldsymbol{\Sigma}'$ is 1, then inequality 5 implies $\sigma_1^2 \leq \sigma_{1*}^2$, where

$$\sigma_{1*} = \frac{\sqrt{2}\sigma_2}{a(1-(\frac{2-2D}{b})^{1/3})}.$$

Hence, $\boldsymbol{\Sigma} \leq \boldsymbol{\Sigma}_*$ where

$$\boldsymbol{\Sigma}_* = \pi_0 \mathsf{Var}(\boldsymbol{\theta}|\boldsymbol{y}=0) + \pi_1 \mathsf{Var}(\boldsymbol{\theta}|\boldsymbol{y}=1) + \sigma_{1*}^2\boldsymbol{\Sigma}'.$$

Therefore, $\boldsymbol{\Sigma}^{-1} \geq \boldsymbol{\Sigma}_*^{-1}$, and we have

$$L < \frac{2\pi_0\pi_1}{1+\pi_0\pi_1\Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\mu}} < \frac{2\pi_0\pi_1}{1+\pi_0\pi_1\Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}_*^{-1}\Delta\boldsymbol{\mu}}.$$

□

Appendix C. Hypothesis Testing.

C.1 One-Sample Test Recall the one-sample hypothesis test, shown in Equation (1). We approximate the distribution of $\hat{D}$ under the null through a permutation approach. The item labels of our N samples are first permutated randomly, and $\hat{D}_{0,N}$ is computed each time given the observed data $\boldsymbol{X}$ and the permuted labels. For a level α significance test, we compare $\hat{D}$ to the $(1-\alpha)$ quantile $\mathcal{Q}_{1-\alpha}$ of the empirical null distribution $\hat{D}_{0,N}$, and reject the null hypothesis if $\hat{D}_N < \mathcal{Q}_{1-\alpha}$. This approach provides higher power than the former approach, under similar assumptions.

Note that the permutation-based approach requires r computations of the sample Discr. The total computational complexity is then $\mathcal{O}\big(N^2 \max(p, rs)\big)$. This approach is only linear in the number of desired repetitions, and therefore is sensible for most settings in which the sample Discr can itself be computed. Moreover, we can greatly speed this computation up through parallelization. With T cores, the computational complexity is instead $\mathcal{O}\big(N^2 \max(p, \frac{r}{T}s)\big)$, as shown in Algorithm 9. We extend this one-sample test to both PICC and I2C2 to provide a robust p-value associated with both statistics of interest. Note that the permutation approach can be generalized to any statistic quantifying repeatability based on repeated measurements.

Algorithm 1 Discr One-Sample Permutation Test

```
Input: (1) $\left\{\boldsymbol{x}_i^j\right\}_{j\in[J_i],i\in[n]}$ $n$ items of data, each featuring $J_i$ measurements.
       (2) $r$ an integer for the number of permutations.

Output: $p \in [0,1]$ the $p$-value associated with the test.
function $p = $ oneSampleTest($\{\boldsymbol{x}_i^j\}_{j\in[J_i],i\in[n]}, r$)
    $d_a = \texttt{Discr}\left\{\boldsymbol{x}_i^j\right\}_{j\in[J_i],i\in[n]}$        ▷ compute observed sample discriminability
    ▷ Note that this for-loop can be parallelized over $T$ cores, as the loops are independent
    for $i$ in $1,\ldots,r$ do
        $\pi = \texttt{Shuffle}(n, \{J_i\}_{i=1}^n)$        ▷ a random shuffling of the measurements
        $d_i = \texttt{Discr}\left\{\boldsymbol{x}_{\pi(i,j)}\right\}_{j\in[J_i],i\in[n]}$        ▷ Compute Discr with random order of sample ids
    end for
    $p = \frac{1}{r+1}\left(\sum_{i=1}^r \mathbb{I}_{\{d_a \geq d_i\}} + 1\right)$ ▷ $p$-value is fraction of times observed is more extreme than under null
    return $p$
end function
```

Figure 5: Discr **One-Sample Test Overview.** Our implementation of the permutation test for the one-sample test of the hypothesis given in Equation (1) requires $\mathcal{O}\left(N^2 \max\left(p, \frac{r}{T}s\right)\right)$ time, where r is the number of permutations and T is the number of cores available for the permutation test. The `Shuffle` function is the function which rearranges all of the data within the dataset, without regard to item nor measurement index. The output provides a new measurement index for each item i and measurement j.

C.2 Two-Sample Test We implement two-sample testing using a permutation approach, similar to the one-sample testing. First, compute the observed difference in Discr between two design choices. The null distribution of the difference in Discr is constructed by first taking random convex combinations of the observed data from each of the two methods choices (the "randomly combined datasets"). Discr is computed for each of the two randomly combined datasets for each permutation. Finally, for each permutation, the all pairs of observed differences in Discr is computed. Finally, the observed statistic is compared with the differences under the null of the randomly combined datasets. The p-value is the fraction of times that the observed statistic is more extreme than the null. Note that we can use this approach for both one and two-tailed hypotheses for an experimental design having higher Discr, lower Discr, and equal Discr relative a second approach; we implement all three in the software implementation of the two-sample test. The Algorithm for the two-sample test is shown in Figure 6, with the alternative hypothesis as specified in Equation (2). The computational complexity is then $\mathcal{O}\big(\frac{r}{T}N^2\max(p, \max_i(s_i))\big)$. Note that for each permutation, the limiting step is the computation of the Discr in $\mathcal{O}\big(N^2\max(p, s)\big)$. This is then offset through parallelization over T cores in the implementation. We extend this two-sample test to both PICC and I2C2 to provide a robust p-value associated with both statistics of interest, for similar reasons to the above. Again, this permutation approach can be generalized to any statistic quantifying repeatability based on repeated measurements.

Algorithm 2 Discr Two-Sample Test

```
Input: (1) $\left\{ \boldsymbol{x}_i^j \right\}_{j \in [J_i], i \in [n]}$ $n$ items of data, each featuring $J_i$ measurements, from the first sample.
    (2) $\left\{ \boldsymbol{z}_i^j \right\}_{j \in [J_i], i \in [n]}$ $n$ the observed data, from the second sample.
    (3) $r$ an integer for the number of permutations.

Output: $p \in [0, 1]$ the $p$-value associated with the test.
function $p$ =TWOSAMPLETEST($\{\boldsymbol{x}_i^j\}_{j \in [J_i], i \in [n]}, \{\boldsymbol{z}_i^j\}_{j \in [J_i], i \in [n]}, r$)
    $\hat{D}^{(1)} = \texttt{Discr}\left\{ \boldsymbol{x}_i^j \right\}_{j \in [J_i], i \in [n]}$    ▷ The Discr of the first sample.
    $\hat{D}^{(2)} = \texttt{Discr}\left\{ \boldsymbol{z}_i^j \right\}_{j \in [J_i], i \in [n]}$    ▷ The Discr of the second sample.
    $d_a = \hat{D}^{(1)} - \hat{D}^{(2)}$    ▷ The observed difference in Discr between samples $1$ and $2$.
        ▷ The for-loop below can be parallelized over $T$ cores, as each loop is an independent
    for $i$ in $1 : r$ do
            ▷ Generate a synthetic null dataset for each of the $2$ samples, using a convex combination of the elements of each sample
        for $k$ in $1 : 2$ do
            $\pi = \texttt{Shuffle}(n, \{J_i\}_{i=1}^n)$    ▷ a random shuffle of the measurements
            $\psi = \texttt{Shuffle}(n, \{J_i\}_{i=1}^n)$
            $\lambda_i^j \overset{iid}{\sim} \text{Unif}(0, 1)$    ▷ for $j = 1, \ldots, n$, where $\boldsymbol{\Lambda} = (\lambda_j)_{j=1}^n$
            $\boldsymbol{u}_i^j = \lambda_i^j \boldsymbol{x}_{\pi(i,j)} + (1 - \lambda_i^j) \boldsymbol{z}_{\psi(i,j)}$    ▷ Convex combination of random elements from each sample
            $d_i^{(k)} = \texttt{Discr}\left\{ \boldsymbol{u}_i^j \right\}_{j \in [J_i], i \in [n]}$    ▷ Compute Discr of the convexly combined elements
        end for
    end for
        ▷ Compute all pairs differences in Discr using the convexly-combined samples
    for $i$ in $1, \ldots, r - 1$ do
        for $j$ in $i + 1, \ldots, r$ do
            $d_n \leftarrow c\left( d_n, d_{n,i}^{(1)} - d_{n,j}^{(2)}, d_{n,j}^{(2)} - d_{n,i}^{(1)} \right)$    ▷ Null distribution of the difference
        end for
    end for
        ▷ $p$-value is fraction of times that observed Discr is more extreme than synthetic datasets
    $p = \frac{2}{r(r-1)+1} \left( \sum_{i=1}^{|d_n|} \mathbb{I}_{\{d_a \leq d_{n,i}\}} + 1 \right)$
    return $p$
end function
```

Figure 6: Discr **Two-Sample Test Overview** Our implementation of the permutation test for the hypothesis given in Equation (2) requires $\mathcal{O}\left(\frac{r}{T} N^2 \max(p, s)\right)$ time, where r is the number of permutations and T is the number of cores available for the permutation test. Above, the only alternative considered is that $H_A : d_a > 0$; our code-based implementation provides strategies for $H_A : d_a < 0$ and $H_A : d_a = 0$ as well.

Appendix D. Simulations.

The following simulations were constructed, where $\sigma_{min}, \sigma_{max}$ are the variance ranges, and settings were run at 15 intervals in $[\sigma_{min}, \sigma_{max}]$ for 500 repetitions per setting. For a simulation setting with variance σ, the variance is reported as the normalized variance, $\bar{\sigma} = \frac{\sigma - \sigma_{min}}{\sigma_{max} - \sigma_{min}}$. Dimensionality is 2, the number of items is K, and the total number of measurements across all items is 128. Typically, i indicates the individual identifier, and j the measurement index. Notationally, in the below descriptions, we adopt the convention that $\boldsymbol{z}_i^j$ obeys the true distribution for a single observation j of item i, and $\boldsymbol{x}_i^j$ incorporates the controlled error term $\boldsymbol{\epsilon}_i^j$, which is the term which is varied the simulation. Further, measurements are simulated from each item randomly with probability $\pi_i = \frac{1}{K}$; that is, the number of measurements for each item n_i is s.t. $n_1, \ldots, n_K \overset{d}{\sim} \mathrm{Multinom}\big(\frac{1}{K}, \ldots, \frac{1}{K}, 128\big)$.

D.1 One Sample Testing and Bayes Error

1. No Signal: $K = 2$ items, where the true distributions for class 1 and class 2 are the same.
   - $\boldsymbol{z}_i^j \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $i = 1, \ldots, 2$, $t = 1, \ldots, 64$. Note: $\boldsymbol{0} \in \mathbb{R}^2$ is $\boldsymbol{0}$, and likewise for $\boldsymbol{I}$
   - $\boldsymbol{\epsilon}_i^j \overset{iid}{\sim} \mathcal{N}\big(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 20]$
   - $\boldsymbol{x}_i^j = \boldsymbol{z}_i^j + \boldsymbol{\epsilon}_i^j \overset{iid}{\sim} \mathcal{N}\big(\boldsymbol{0}, (1 + \sigma^2)\boldsymbol{I}\big)$
2. Cross: $K = 2$ items, where the true distributions for class 1 and class 2 are orthogonal.
   - $\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 0.1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 2 \end{bmatrix}$
   - $\boldsymbol{z}_i^j \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, $i = 1, 2$
   - $\boldsymbol{\epsilon}_i^j \overset{iid}{\sim} \mathcal{N}\big(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 20]$
   - $\boldsymbol{x}_i^j = \boldsymbol{z}_i^j + \boldsymbol{\epsilon}_i^j$
3. Gaussian: $K = 16$ items, where the true distributions are each gaussian.
   - $\boldsymbol{\mu}_i \overset{iid}{\sim} \pi_1 \mathcal{N}(\boldsymbol{0}, 4\boldsymbol{I})$, $i = 1, \ldots, 16$
   - $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$
   - $\boldsymbol{z}_i^j \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$
   - $\boldsymbol{\epsilon}_i^j \overset{iid}{\sim} \mathcal{N}\big(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 20]$
   - $\boldsymbol{x}_i^j = \boldsymbol{z}_i^j + \boldsymbol{\epsilon}_i^j$
4. Ball/Circle: $K = 2$ items, where 1 item is uniformly distributed on the unit ball with gaussian error, and the second item is uniformly distributed on the unit sphere with gaussian error.
   - $\boldsymbol{z}_{1t} \overset{iid}{\sim} \mathbb{B}(r = 1) + \mathcal{N}(\boldsymbol{0}, 0.1\boldsymbol{I})$ samples uniformly on unit ball of radius 2 with Gaussian error
   - $\boldsymbol{z}_{2t} \overset{iid}{\sim} \mathbb{S}(r = 1.5) + \mathcal{N}(\boldsymbol{0}, 0.1\boldsymbol{I})$ samples uniformly on unit sphere of radius 2 with Gaussian error
   - $\boldsymbol{\epsilon}_i^j \overset{iid}{\sim} \mathcal{N}\big(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 10]$
   - $\boldsymbol{x}_i^j = \boldsymbol{z}_i^j + \boldsymbol{\epsilon}_i^j$

Bayes error was estimated by simulating $n = 10{,}000$ points according to the above simulation settings, and approximating the Bayes error through numerical integration. The classification labels for $K = 2$ simulations were consistent with the individual labels, and for the $K = 16$, the first class consists of the 8 distributions whose means were leftmost, and the rest of the distributions were the other class.

D.2 Two Sample Testing Items are sampled with the same true distributions $\boldsymbol{z}_i^j$ as before, with the following augmentation:

$$\boldsymbol{x}_{i,k}^j = \begin{cases} \boldsymbol{z}_i^j & k = 1 \\ \boldsymbol{z}_i^j + \boldsymbol{\epsilon}_i^j & k = 2 \end{cases}$$

That is, the observed data $\boldsymbol{x}_{i,k}^j$ for item i, observation j, and sample $k \in [2]$ is such that the first sample is distributed according to the true item distribution, and the second sample is distributed according to the true item distribution with an added noise term, where $\boldsymbol{\epsilon}_i^j \overset{iid}{\sim} \mathcal{N}\big(\mathbf{0}, \sigma^2 \boldsymbol{I}\big)$:

1. No Signal: $K = 2$
   - $\sigma \in [0, 10]$
2. Cross: $K = 2$
   - $\sigma \in [0, 2]$
3. Gaussian: $K = 16$
   - $\sigma \in [0, 2]$
4. Ball/Circle: $K = 2$
   - $\sigma \in [0, 10]$

By construction, one would anticipate Discr of the first sample to exceed that of the second sample, as the second sample has additional error. Therefore, the natural hypothesis is:

$$H_0 : D^{(1)} = D^{(2)}, \qquad H_A : D^{(1)} > D^{(2)}$$

Appendix E. Connectomics Application.

E.1 Data Acquisition and Analysis

fMRI Analysis Pipelines The fMRI connectomes were acquired as follows. Motion correction is performed via `mcflirt` to estimate the 6 motion parameters (x, y, z translation and rotations). Registration is performed by first performing a cross-modality registration from the functional to the anatomical MRI using `flirt-bbr`, followed by registration to the anatomical template using either (1) FSL-`fnirt` or (2) ANTs-`SyN`, two techniques for non-linear registration. Frequency filtering was performed by either (1) not frequency filtering, or (2) bandpass filtering signal outside of the $[.01, .1]$ Hz range. Volumes were either (1) not scrubbed, or (2) scrubbed if motion exceeded 0.5 mm, in which case the preceding volume and succeeding two volumes were removed. Global signal regression was either (1) not performed, or (2) performed by removing the global mean signal across all voxels in the functional timeseries. Moreover, across all analysis pipelines, the top 5 principal components (`compcor`), Friston 24 parameters, and a quadratic polynomial were fit and regressed from the functional timeseries. Finally, the voxelwise timeseries were spatially downsampled using (1) the CC200 parcellation, (2) the AAL parcellation, (3) the Harvard-Oxford parcellation, or (4) the Desikan-Killany parcellation. Graphs were estimated by (1) computing the rank of the raw absolute correlations, (2) log-transforming the raw absolute correlations, or (3) computing the raw absolute correlation between pairs of regions of interest in each parcellation. No mean centering was performed for functional connectivity estimates. Specific data analysis instructions for deployment in `AWS` can be found in the https://neurodata.io/m2g. All data analysis was performed in the `AWS` cloud using `CPAC` version `3.9.2` [27]. All parcellations are available in `neuroparc` human brain atlases [42].

dMRI Analysis Pipelines The dMRI connectomes were acquired as follows. The dMRI scans were corrected for eddy currents using FSL's `eddy-correct` [43]. FSL's "standard" linear registration pipeline was used to register the sMRI and dMRI images to the MNI152 atlas [43–46]. A tensor model is fit using DiPy [47] to obtain an estimated tensor at each voxel. A deterministic tractography algorithm is applied using DiPy's EuDX [47, 48] to obtain streamlines, which indicate the voxels connected by an axonal fiber tract. Graphs are formed by contracting voxels into graph vertices depending on spatial [49], anatomical [50–53], or functional [54–57] similarity. Given a parcellation with vertices V and a corresponding mapping $P(v_i)$ indicating the voxels within a region i, we contract our fiber streamlines as follows. $w(v_i, v_j) = \sum_{u \in P(v_i)} \sum_{w \in P(v_j)} \mathbb{I}\left\{F_{u,w}\right\}$ where $F_{u,w}$ is true if a fiber tract exists between voxels u and w, and false if there is no fiber tract between voxels u and w. The specific parcellations leveraged are detailed in Kiar et al. [26], consisting of parcellations defined in the MNI152 space [50–57]. All parcellations are available in `neuroparc` human brain atlases [42].

E.2 Effect Size Investigation In this investigation, we are interested in learning how maximization based on the observed notion of reliability correlates with real performance on a downstream inference task. Recalling Corollary (B.1), we explore the implications of this corollary in a large neuroimaging dataset provided by the Consortium for Reliability and Reproducibility [18], and demonstrate that selection of the experimental design via Discr, in fact, facilitates improved downstream inference on both a regression and classification task. This provides strong motivation for leveraging the Discr for experimental design.

Ideally, for a particular summary statistic, a high value will generally correlate with a positive effect size. For datasets $i = 1, \ldots, M$ where M is the total number of datasets, an analysis strategy $j = 1, \ldots, 192$ for 192 total analysis strategies, and $k = 1, \ldots, 3$ are our summary statistics of interest (Discr, PICC, and I2C2), we fit the standard linear regression model $Y = \beta X + \epsilon$, where we model the effect size Y estimated by MGC [58] via a linear relationship with X, the observed sample statistic for approach k (Discr, PICC, or I2C2), with coefficient β. Note that the interpretation of β is the expected change in the effect size Y due to a single unit change in the observed sample statistic X. Both Y and X are uniformly normalized across all strategies within a single dataset to facilitate intuitive comparison across methods. For each summary statistic k, we pose the following hypothesis:

$$H_0 : \beta = 0; \quad H_A : \beta > 0$$

Acceptance of the alternative hypothesis would have the interpretation that an increase in the observed sample statistic X would tend to correspond to an increase in the observed effect size Y, and the relevant test is the one-way t-test. Acceptance of the alternative hypothesis against the null provides evidence that an increase in the sample statistic corresponds to an increase in the observed effect size, where the neither of the responses (age, sex) were known or considered at the time the data were analyzed nor the sample statistics were computed. This provides evidence that the statistic is informative for experimental design within the context of this investigation. Model fitting for this investigation is conducted using the lm package in the R programming language [59].

E.3 Dataset Descriptions

Useful Data Links All relevant analysis scripts and data for figure reproduction in this manuscript made publicly available, and can be found at https://neurodata.io/mgc.

Dataset	Manuf.	Model	TE (ms)	TR (ms)	STC	#Timepts	#Sub	#Ses	#Scans	Discr
KKI2009	NA	NA	NA	NA	NA	NA	21	1	42	0.93
NKI24	Siemens	TrioTim	30	645	inter.	900	24	2	47	0.98
BNU1	Siemens	TrioTim	30	2000	inter.	200	50	2	100	0.97
BNU2	Siemens	TrioTim	30	variable	inter.	variable	50	2	100	0.92
DC1	Philips	NaN	35	2500	inter.	120	114	4	244	0.95
HNU1	GE	MR750	30	2000	inter.	300	30	10	300	0.98
IACAS	GE	Signa	30	2000	inter.	240	28	3	59	0.83
IBATRT	Siemens	TrioTim	30	1750	seq.	220	36	2	50	0.95
IPCAS	NA	NA	NA	NA	NA	NA	78	2	156	0.99
IPCAS1	Siemens	TrioTim	30	2000	inter.	205	30	2	60	1.00
IPCAS2	Siemens	TrioTim	30	2500	inter.	212	35	2	70	0.98
IPCAS5	Siemens	TrioTim	30	2000	inter.	170	22	2	44	0.96
IPCAS6	Siemens	TrioTim	30	2500	inter.	242	2	15	30	1.00
IPCAS8	Siemens	TrioTim	30	2000	inter.	240	13	2	26	0.96
JHNU	Siemens	TrioTim	30	2000	inter.	250	30	2	60	0.96
LMU3	Siemens	TrioTim	30	3000	inter.	120	25	2	50	0.93
MRN1	NA	NA	NA	NA	NA	NA	53	2	88	0.94
NYU1	Siemens	Allegra	25	2000	NaN	197	25	3	75	0.98
NYU2	Siemens	Allegra	15	2000	inter.	180	187	3	252	0.96
SWU1	Siemens	TrioTim	30	2000	inter.	240	20	3	59	0.97
SWU2	Siemens	TrioTim	30	2000	inter.	300	27	2	54	0.96
SWU3	Siemens	TrioTim	30	2000	inter.	242	24	2	48	0.98
SWU4	Siemens	TrioTim	30	2000	inter.	242	235	2	467	0.97
UM	Siemens	TrioTim	30	2000	seq.	150	80	2	160	0.99
UPSM1	Siemens	TrioTim	29	1500	seq.	200	100	3	230	0.89
Utah1	Siemens	TrioTim	28	2000	inter.	240	26	2	52	0.92
UWM	GE	MR750	25	2600	inter.	231	25	2	50	0.96
XHCUMS	Siemens	TrioTim	30	3000	inter.	124	24	5	120	0.91

Figure 7: **fMRI Dataset Descriptions**. In the above table, STC corresponds to slice timing correction. Rows with NA entries do not have available metadata associated with the scanning protocol. The sample discriminabilities correspond to the Discr of the best performing pipeline overall, FNNNCP.

Dataset	Manuf.	Model	TE (ms)	TR (ms)	#Dir	bval $\frac{s}{mm^2}$	#Sub	#Ses	#Scans	Discr
BNU1	Siemens	TrioTim	89	8000	30	1000	57	2	113	1.00
HNU1	GE	MR750	Min	8600	33	1000	30	10	300	0.99
KKI2009	NA	NA	NA	NA	NA	NA	21	2	42	1.00
NKI24	Siemens	TrioTim	95	2400	137	1500	20	2	40	1.00
SWU4	Siemens	TrioTim	NaN	NaN	93	1000	227	2	454	0.88

Figure 8: **dMRI Dataset Descriptions**. In the above table, Dir corresponds to the number of diffusion directions. Rows with NA entries do not have available metadata associated with the scanning protocol. The sample discriminabilities correspond to the Discr of the pipeline with the CPAC200 parcellation and the log-transformed edges.