Optimal Experimental Design for Big Data: Applications in Brain Imaging

Eric W. Bridgeford[1], Shangsi Wang[2], Zhi Yang[3], Zeyi Wang[4], Ting Xu[5], Cameron Craddock[6], Gregory Kiar[7], William Gray-Roncal[8], Carey E. Priebe[9], Brian Caffo[10], Michael Milham[11], Xi-Nian Zuo[12], Consortium for Reliability and Reproduciblity , and Joshua T. Vogelstein[13]

Abstract. The cost of data collection and processing is becoming prohibitively expensive for many research groups across disciplines, a problem that is exacerbated by the dependence of ever larger sample sizes to obtain reliable inferences for increasingly subtle questions. And yet, as more data is available and open access, more researchers desire to analyze it for different questions, often including previously unforeseen questions. To further increase sample sizes, existing datasets are often amalgamated. These *reference datasets*—datasets that serve to answer many disparate questions for different individuals—are increasingly common and important. *Reference pipelines* efficiently and flexibly analyze on all the datasets. How can one optimally design these reference datasets and pipelines to yield derivative data that are simultaneously useful for many different tasks? We propose an approach to experimental design that leverages multiple measurements for each distinct item (for example, an individual). The key insight is that each measurement of the same item should be more similar to other measurements of that item, as compared to measurements of any other item. In other words, we seek to optimally *discriminate* one item from another. We formalize the notion of discriminability, and introduce both a non-parameteric and parametric statistic to quantify the discriminability of potentially multivariate or non-Euclidean datasets. With this notion, one can make optimal decisions—either with regard to acquisition or analysis of data—by maximizing discriminability. Crucially, this optimization can be performed in the absence of any task-specific (or supervised) information. We show that optimizing decisions with respect to discriminability yields improved performance on subsequent inference tasks. We apply this strategy to a brain imaging dataset built by the "Consortium for Reliability and Reproducability" which consists of 24 disparate magnetic resonance imaging datasets, each with up to hundreds of individuals that were imaged multiple times. We show that by optimizing pipelines with respect to discriminability, we improve performance on multiple subsequent inference tasks, even though discriminability does not consider the tasks whatsoever.

1 Introduction

As the size of data increases, scientists face two questions: (i) in what manner should data be collected, and (ii) what strategies should be used to process the data. When the data will be used for multiple different inference tasks, there is a conflict: if one optimizes for a single inference task, information required for other inference tasks could be lost. This problem is exacerbated when the data will be used for *unknown* future inference tasks. In such scenarios, how can one make decisions that yield high-quality inferences for many subsequent tasks? In other words, which experimental and analytical properties of the measurements should one optimize?

One goal would be to maximize aspects of measurement validity, such as, the degree to which the measurements corresponds to what it is purporting to measure. However, often aspects of measurement validity can not be observed directly [10, 41]. Instead, researchers often leverage a related concept of statistical unbiasedness; an estimate is unbiased if its expected value is identical to the true value. Unbiasedness can be seen as a kind of validity. However, unbiasedness often comes at a cost

[1]Department of Biostatistics, Johns Hopkins University (ebridge2@jhu.edu)
[2]Department of Applied Mathematics and Statistics, Johns Hopkins University
[3]CAS Institute of Psychology, Shanghai Jiaotong University
[4]Department of Biostatistics, Johns Hopkins University
[5]Child Mind Institute
[6]Department of Diagnostic Medicine, Dell Medical School, University of Texas at Austin
[7]Department of Biomedical Engineering, McGill University
[8]Johns Hopkins University Applied Physics Laboratory
[9]Department of Applied Mathematics and Statistics, Institute for Computational Medicine, Johns Hopkins University
[10]Department of Biostatistics, Johns Hopkins University
[11]Childmind Institute
[12]CAS Institute of Psychology, Beijing Normal University, Nanning Normal University, University of Chinese Academy of Sciences
[13]Corresponding author; Department of Biomedical Engineering, Institute for Computational Medicine, Kavli Neuroscience Discovery Institute, Johns Hopkins University (jovo@jhu.edu).

in variance. To give a simple example, a broken clock is not valid, in that its measurement of time does not correspond accurately to the true time with high likelihood (only twice per day). Yet, it has zero variance. Conversely, unbiased estimates can often be improved upon by introducing bias to decrease variance [35].

To complicate matters, in scientific measurement, some sources of variability can be of interest, such as veridical biological heterogeneity. In contrast, many sources of variability in scientific measurement are a nuisance, such as measurement noise. Thus, a natural quantity to optimize would be a function that preserves biological variability while mitigating extraneous variability. If one has acquired multiple measurements per item (e.g., an individual), then the intra-class correlation coefficient (`ICC`) is a possible quantity to optimize. `ICC` is a statistic based on a simplified model of variability; `ICC` decomposes all sources of variability into either within-item (assumed to be measurement noise) and across-item (assumed to be veridical heterogeneity) variability. The `ICC` is then the fraction of the total variability that is across-subject variability. `ICC` is bounded between 0 and 1, and therefore provides an index that can be naturally compared across datasets. `ICC` is therefore a useful quantity to optimize in experimental design. However, optimizing `ICC` has, to our knowledge, not previously been proposed, perhaps because it requires acquiring multiple measurements per item. This is despite the fact that `ICC` is the de facto standard metric for evaluating the reliability of an experiment.

That said, `ICC` has several other limitations if one were to use it to optimize experimental design. First, it is a univariate measure, meaning if the data are multivariate, they must first be represented by univariate statistics, thereby discarding the multivariate information. Second, `ICC` is based on a particular model of the data: Gaussianity. Once Gaussianity of the data does not hold, interpretation of the magnitude of `ICC` is no longer as straightforward, as non-Gaussian measurements that are highly reliable could yield quite low `ICC`.

We therefore generalize `ICC` in two ways. First, we introduce a multivariate parametric generalization, `ICCoPCA`, in which we compute the first principle component of the data, and then compute `ICC` of that representation. Second, we introduce a multivariate nonparametric generalizeation, replacing the variance computation with a rank-based distance computation. We refer to as the *discriminability* statistic. For both generalizations, we introduce a permutation procedure to obtain both one-sample and multi-sample test statistics and p-values. The multi-sample testing allows us to formally compare experiments for the study of repeatability and reliability. We provide an extensive simulation benchmark to illustrate the value of using these two statistics for optimal experimental design.

The motivation of this work is a reference brain imaging dataset generated by the Consortium for Reliability and Reproducibility (CoRR) [40]. This dataset is an amalgamation of over 30 different datasets, many of which were collected using different scanners, manufactured by different companies, 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 reference pipeline to process the data such that they can be used for many different inference tasks. After evaluating nearly 200 different pipelines on over 3000 scans, we determined the optimal pipeline, that is, the pipeline with the highest discriminability. We then demonstrate that for every single dataset, on average, as one makes the pipeline achieve higher discriminability, the amount of information retained about various phenotypes increases. 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 in general, much less ubiquitously. We therefore believe this approach to optimal experimental design 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 Results

2.1 Discriminability Discriminability is a non-parametric statistic of a joint distribution in a hierarchical model, that can be used to differentiate between classes of items (or individuals). Consider n items, where each item has s measurements, resulting in $N = n \times s$ total measurements across items.

Discriminability is 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 subjects, compute the fraction of times that a within-item distance is smaller than an across-item distance (resulting in $N \cdot (s-1)$ numbers between 0 and 1).
3. The *discriminability* of the dataset is the average of the above mentioned fraction across items (resulting in a single number between 0 and 1).

A high discriminability indicates that within-item measurements are more similar to one another than across-item measurements. For more algorithmic details, see Algorithm A.1. For formal definition of terms, including the population variant of discriminability, see Appendix C.

2.2 Properties of Discriminability

Simulation Settings To develop insight into the performance of the discriminability, we develop several benchmark simulations, both within and beyond the theoretical guarantees provided by discriminability and other methods. For four different benchmark problems, we sample 10 measurements from between 2 and 20 items in 2 dimensions. Figure 2.1A shows a two-dimensional scatterplot of the benchmark problem set, and Figure 2.1B shows the distance matrix between samples, ordered by item. The performance of discriminability, and competing methods, is analyzed in the context of the following questions relevant to investigating the effectiveness of a processing pipeline using the discriminability one and two-sample frameworks, along with the classification task, are described in Appendix A.1 and A.2. Three simulations contain signal for both one and two sample scenarios, in which a true relationship is present in the data. A fourth simulation contains no signal, in which the samples of interest have an identical distribution.

Discriminability Optimizes the Bound on an Unspecified Performance Task When choosing an optimal reference method, an important consideration is whether a method choice will facilitate known or unknown inference. Formally, a reference method facilitates the measurement of a true property of interest for each individual i, where the true property is discriminable within the population. We are interested in the prediction of a phenotypic property property of each subject i. How can we choose a reference method that facilitates measurements that will improve our prediction of the phenotypic property?

As discussed by Wang et al. [36], we are interested in bounding the minimum (best) possible predictive error among all possible prediction rules. Assuming only that each measurement and phenotypic pair represent independent and identically distributed draws from the population (that is, our individuals are a random sample) and that the measurement noise is bounded and addative with respect to the true property of interest, the the predictive error can be upper bounded by a decreasing function of discriminability. This has the implication that a higher discriminability provides a lower bound on the minimum predictive error of *any* inference task. An immediately consequence of this is that by choosing a more discriminable pipeline, we are more likely to see improved inference performance on downstream prediction tasks. To our knowledge, this work introduces the concept of using multiple measurements to formally bound the predictive error on the downstream inference task regardless of whether the task is known *a priori*, providing theoretical motivation to maximize the discriminability for subsequent inference.

In Figure 2.1C, we compare the statistics of interest (normalized between 0 and 1) to $1-$ the bayes error. In all of the signal relationships, discriminability correlates highly with $1-$ the Bayes Error, whereas `I2C2` and `ICCoPCA` only display this property when the data follows the `MANOVA` model. This demonstrates that in the simulation settings considered, discriminability provides an useful criterion for selection of the particular simulation setting that minimizes the Bayes Error (and maximizes $1-$ the Bayes Error), despite the fact that the classification task was unknown *a priori*.

Discriminability Uncovers the Dependence of Measurements between Subjects To what extent are the within-subject measurements more similar than the between-subject measurements? We consider the

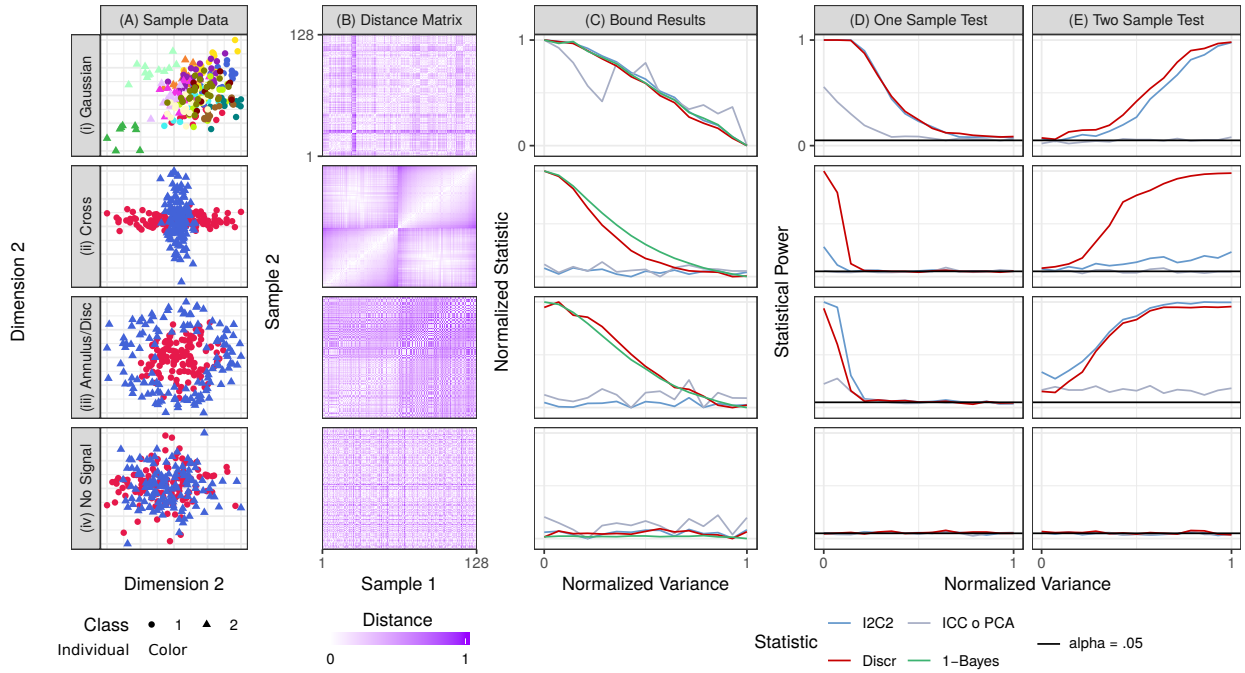

Figure 2.1: Four simulations demonstrating the flexibility of the discriminability framework across a range of class relationships. For all simulations, the number of dimensions $d = 2$, the number of samples $n = 128$, and $\alpha = .05$, with 500 iterations per setting. *(i)* **Gaussian** $K = 16$ individuals, with the subject-specific and measurement-specific effects Gaussian distributed. This simulation follow the model specified by `MANOVA`, and when projected to a single dimension, follow the model specified by `ANOVA`. *(ii)* **Cross** $K = 2$ individuals, where each class is still Gaussian distributed, but the covariance structuring is orthogonal between the two classes. *(iii)* **Annulus/Disc** $K = 2$ individuals, where one is a distributed in an annulus, and the other within the unit sphere. *(iv)* **No Signal** A simulation where the two individuals have equal true distribution, and there should be no detectable difference between the individuals in either one or two sample testing. **(B)** A plot of the distance matrix between samples within each simulation setting. Samples are organized by individual label, so the near-diagonal blocks of smaller distance indicate that samples are relatively discriminable across individuals. **(C)** A comparison of the observed statistics to $1-$ the bayes error, where the classification task is as indicated by the shape of each point in **(A)**. **(D)** One-sample test of equality using the discriminability framework as described in Appendix A.1, to test whether the observed statistic is observed by random chance. **(E)** Two-sample test of whether the observed statistic for one dataset exceeds that of another using the discriminability procedure described in Appendix A.2.

following hypothesis:

$$H_0 : D = D_0;\ H_A : D > D_0 \tag{2.1}$$

where D_0 is the discriminability that we would observe if the measurements are not discriminable within the population. A test of this hypothesis is known as a one-sample test, or a test of goodness-of-fit. The typical criterion for evaluating a statistical test is the statistical power, or the probability that a test correctly rejects the null at a given type-one error level. To test this hypothesis, we determine the null distribution of the sample discriminability statistic $\hat{D}$ using a permutation approach. As seen in Appendix A.1, this approach constructs the approximate null distribution of the sample discriminability by repeatedly permuting the measurement labels and computing the discriminability $\hat{D}_0$, and comparing the observed sample discriminability $\hat{D}$ with the given labels to those computed with the randomly permuted labels $\hat{D}_0$. This affords a test that balances statistical power without substantially sacrificing computational efficiency, as the permutations can be efficiently parallelized. We extend this goodness of fit test to both `ICCoPCA` and `I2C2`, to obtain a robust estimate of a p-value associated with the relative fit of the observed reference statistic. As shown in Figure 2.1D, sample discriminability uncovers the

relationship present in each of the different simulation approaches. Discriminability provides comparable power to I2C2 for both the Gaussian and Annulus/Disc simulation scenario, but is the only statistic that provides meaningful power in the Cross simulation. In 2.1D.iv, where no relationship is present, all tests demonstrate the property of testing validity; that is, the tests accept the alternative hypothesis at a rate equal to the cutoff threshold $\alpha = .05$.

Discriminability for Optimal Design Given two approaches for obtaining a given dataset—which can differ either by experimental protocols and/or processing pipelines—are the measurements produced by one approach more discriminable than the other? Formally, let $\hat{D}^{(1)}$ be the discriminability of the first reference method, and $\hat{D}^{(2)}$ be the discriminability of the second reference method. For instance, given a dataset processed by two difference reference methods, we may be interested in whether one dataset is more discriminable than the other. We consider the following hypothesis:

$$H_0 : D^{(1)} = D^{(2)}; \quad H_A : D^{(1)} > D^{(2)} \tag{2.2}$$

A test of this hypothesis is known as a two-sample test, or a test of equality. Again, we formally test this hypothesis using a permutation test. As shown in Appendix A.2, for each permutation, we construct synthetic null datasets for each reference dataset by taking random convex combinations of the measurements. We compute the all pairs differences in discriminability of the synthetic null datasets $\hat{D}_0^{(1)} - \hat{D}_0^{(2)}$ to form the approximate null distribution of the observed difference in the sample discriminability between the two reference methods $\hat{D}^{(1)} - \hat{D}^{(2)}$ (the test statistic), and compare the test statistic to its approximate null distribution. Again, we can distribute the permutations across the number of available threads for computational efficiency. As before, we extend this test to both ICCoPCA and I2C2 to obtain a robust estimate of a p-value associated with the relative fit of the observed reference statistic. As shown in Figure 2.1D, sample discriminability again shows high power across all signal relationships. Discriminability provides comparable power to I2C2 for both the Gaussian and Annulus/Disc simulation scenario, but again is the only statistic that provides meaningful power in the Cross simulation. In 2.1E.iv, all statistics again display testing validity.

2.3 Mega-Analysis of Statistical Connectomics Dataset with Discriminability Discriminability provides an intuitive, straightforward approach for comparison of reference pipelines. Below, we provide a thorough investigation of a rich neuroimaging dataset, provided by the Consortium for Reliability and Reproducibility (CoRR).

Real Data Collection and Processing The CoRR Dataset [39] provides functional MRI (fMRI) and diffusion MRI (dMRI) scans from > 1600 participants, often with multiple measurements, collected through 33 different cross-site studies, and for 4 of the studies, multi-modal analysis (both fMRI and dMRI scans collected for the same participant from repeated trials). Most of the studies adhere to slightly disparate scanning protocols, facilitating the opportunity to investigate discriminability as a function of both data measurement and preprocessing techniques. An exhaustive list of the collection and pre-processing techniques employed is in Figure 2.2A, and the related manuscripts [1, 29]. The fMRI and dMRI scans were processed to acquire brain graphs, or connectomes. All fMRI connectomes from datasets with repeated measurements were acquired via 192 different preprocessing pipelines using the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [29]. Figure 2.2A summarizes the different preprocessing strategies attempted for fMRI connectome acquisition. The dMRI connectomes were acquired via 48 preprocessing pipelines using the Neurodata MRI Graphs (ndmg) pipeline [13]. Appendix C provides specific details for both fMRI and dMRI preprocessing, as well as the options attempted.

Different Processing Strategies Yield Widely Disparate Discriminabilities First, we investigate the relationship between discriminability and fMRI preprocessing pipeline choice. As shown in Figure 2.2, preprocessing strategy has a prominent impact on the downstream discriminability of the resulting fMRI connectomes. Particularly, note that both the weighted-mean sample discriminability and the per-dataset variance in the discriminability shift markedly from the lower-performing strategies (right) to the better performing strategies (left). The pipelines are all compared to the pipeline with the highest

weighted-mean sample discriminability, FNNNCP (FSL registration, no frequency filtering, no scrubbing, no global signal regression, CC200 parcellation, ranked edges), using the hypothesis posed in Equation (2.2) via the two-sample (equality) test, to investigate whether the best pipeline provides a significant improvement in discriminability over each compared strategy. The majority of the strategies ($51/64 = 79.6\%$) show significantly worse discriminability than the optimal strategy at $\alpha = .05$. This highlights that choice of preprocessing pipeline has a major, and often significant, impact on the downstream discriminability.

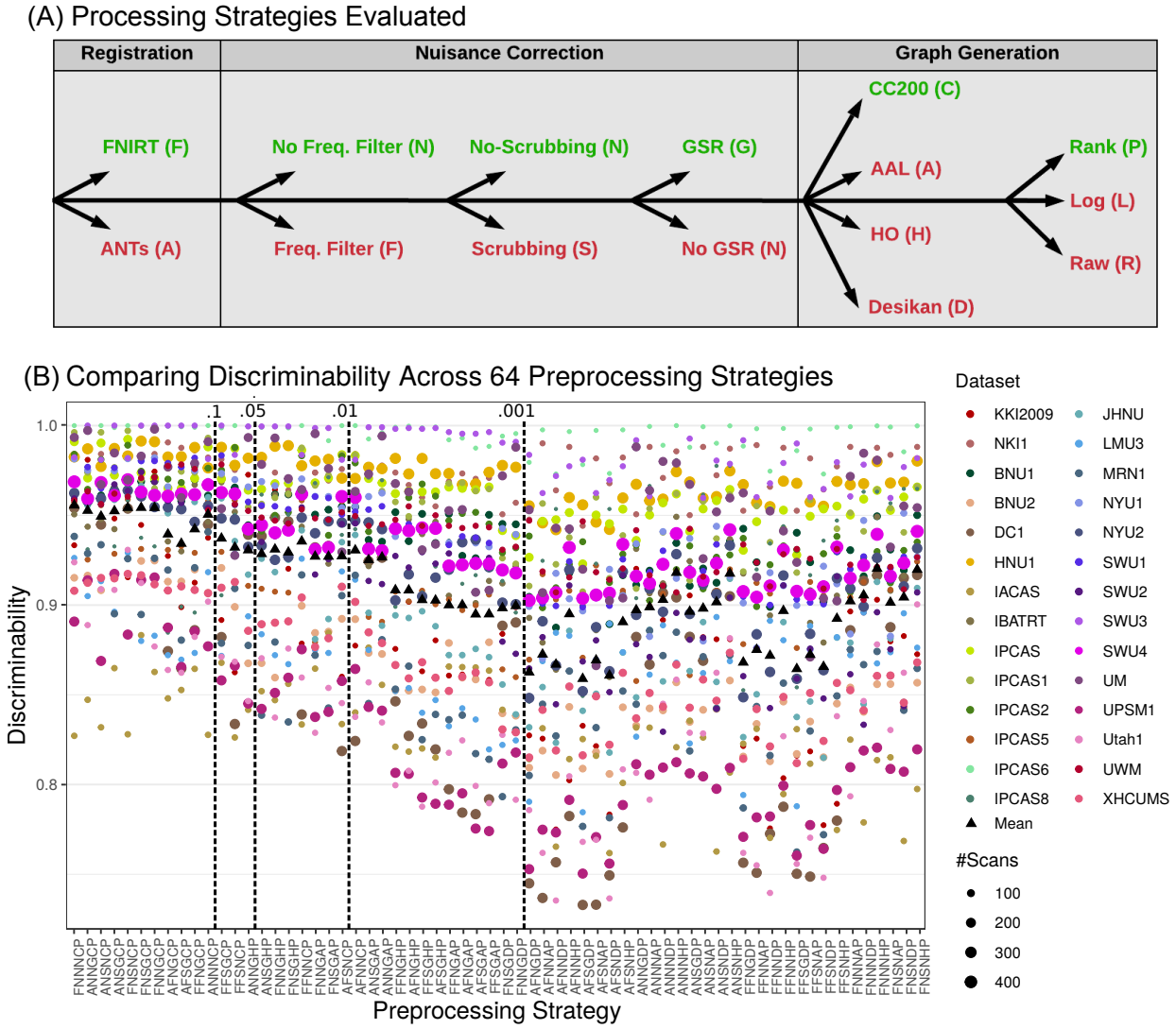

Figure 2.2: **(A)** a schematic tree illustrating the preprocessing options attempted, and where they fall in the fMRI preprocessing pipeline. Each option is detailed in Appendix C. The optimal options marginally are shown in green, and are discussed in detail in Figure 2.3. **(B)** Discriminability of fMRI Connectomes Processed 64 ways. Functional correlation matrices are acquired from 28 multi-session studies from the CoRR dataset across 4 parcellations. The preprocessing strategy codes are assigned sequentially according to the abbreviations listed for each step in **(A)**. The edges are ranked according to the pairwise correlation. The per-study discriminability is estimated and weighted according to the number of subjects for the particular study to estimate the mean discriminability for each strategy. Each pipeline is compared to the optimal pipeline with the highest mean discriminability, FNNNCP, using the two-sample (equality) test of the hypothesis in Equation (2.2) to determine whether the mean discriminability of the optimal processing strategy exceeds that of other strategies. The remaining strategies are arranged according to p-value, indicated in the top row.

Second, we investigate the impact of individual pre-processing options on the downstream dis-

criminability for fMRI data. We begin by visualizing the discriminability marginalized for each pre-processing step in Figure 2.3. Figure 2.3A investigates the impact of different rs-fMRI preprocessing strategies, showing the difference between the choice that shows the best average discriminability and the other possible options. We find that if one were to independently select the best option for each pre-processing stage (FNNGCP), it not be significantly worse than the pipeline with the highest discriminability FNNNCP (p-value = $.14$). Moreover, for each step in the pre-processing pipeline, we compare the option with the highest mean discriminability to the option with the second highest mean discriminability using the Wilcoxon Signed-Rank Test. We find that FNIRT, no frequency filtering, global signal regression, the CC200 parcellation, and ranked edge-transform each provide a significant increase in average discriminability over the alternative strategies after correction for multiple hypotheses (p-values all $< .001$).

Third, we investigate different analysis choices for dMRI data, Figures 2.3C.i and 2.3C.ii show the impact of different dMRI preprocessing strategies. We find that the log-transformed and rank-transformed edges perform relatively similarly, while both greatly outperform the raw connectome edge weights. Moreover, the number of parcels within the parcellation typically provides an improvement in discriminability, regardless of the edge-transform attempted.

Optimal Discriminability Provides Improved Downstream Inference In this experiment, we seek to understand the dependent relationship between preprocessing approach and downstream inference. Does seeking reference methods with a higher discriminability tend to improve the inferential capacity of the data? In Figure 2.4, we examine the dependence between the pre-processed connectomes from each of the 64 pipelines (using the raw connectome edge-weights) and a covariance of interest (either age, a regression covariate, or sex, a classification covariate). We determine the nature of the dependence using MGC [28, 34], a generalization of the distance correlation that enhances finite-sample statistical power for the identification of potentially non-linear dependences in the data. Under the hypothesis posed by MGC, a larger statistic corresponds to a greater effect size in the processed graphs from the covariate of interest.

To assess whether optimizing reference method selection using the discriminability preserves the dependence, we regress the MGC statistic onto each of our reference statistics (discriminability, ICC$\circ$PCA, ANOVA$\circ$PCA, I2C2) for each dataset. We find that discriminability is the only reference statistic in which all of the slopes exceed zero for each dataset across both covariates of interest. To test this observation, we consider a one-tailed null hypothesis that the slope is ≤ 0 against the alternative that the slope exceeds zero. Formally testing this hypothesis we find that, unlike the other reference methods, the p-value for discriminability is significant across both tasks after Fisher's correction [7] (median p-value $< .001$ for both sex and age). For each regression line, this has the interpretation that increasing values of discriminability tend to correspond to a larger effect size due to the covariate of interest. This example captures the intuitive notion that reference methods that are more discriminable lead to more substantial effect sizes in an unspecified downstream inference task.

3 Discussion

We propose the use of the sample discriminability, a simple and intuitive measure of the replicability of a reference approach featuring multiple measurements. Numerous efforts have established the empirical value of maintaining a notion of intra-class stability [6, 23, 37], with little theoretical justification for the importance of such approaches. Under a relatively general model, we prove that discriminability provides a lower bound on the predictive accuracy for any downstream inference task, known or unknown. This provides clear motivation for the sample discriminability in reference method selection, in which only a subset of potential tasks may be known at any given time, instilling a harmony between theory and practice for reference method selection. We derive one-sample (goodness-of-fit) and two-sample (equality) tests for the statistical comparison of collection and analysis pipelines in terms of their discriminability, and demonstrate via simulation that discriminability provides numerous advantages over existing techniques across a range of benchmarks both within and outside our theoretical setting. Our neuroimaging use-case exemplifies the utility of these features of the discriminability

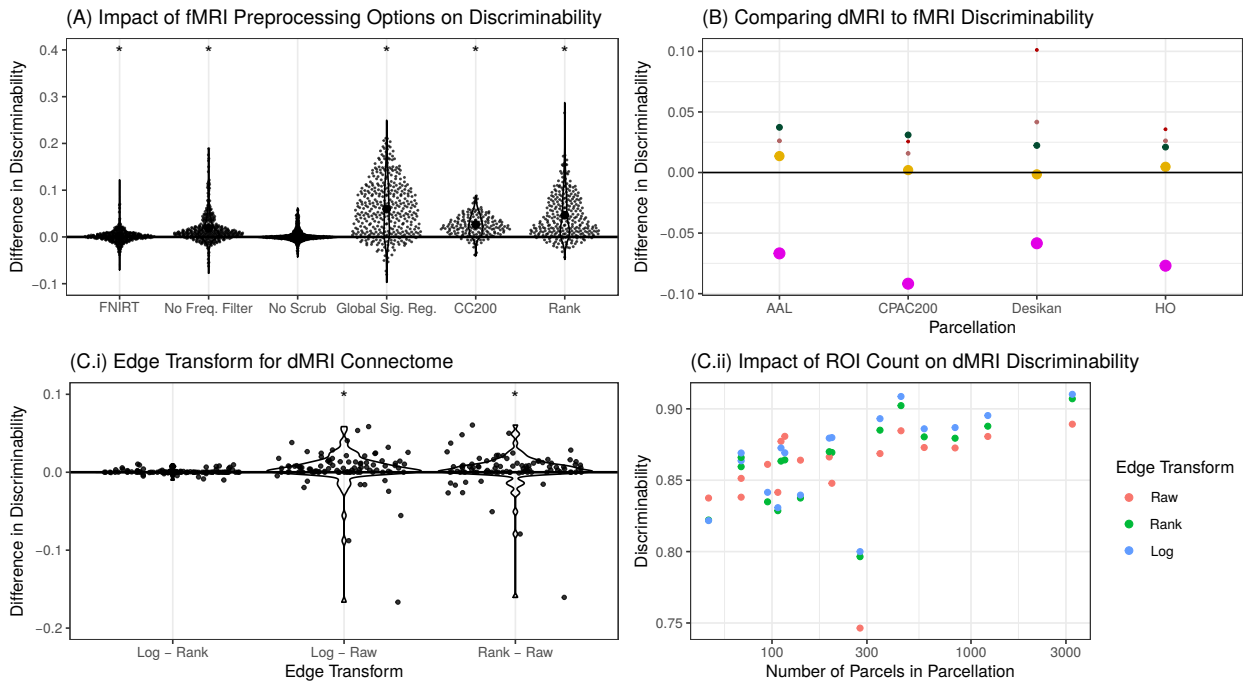

Figure 2.3: Differences in discriminability with choices of pre-processing options. **(A)** Choice of functional pre-processing options. The pipelines are aggregated for a particular preprocessing step, with pairwise comparisons with the remaining preprocessing options held fixed. The difference between the overall best performing option and the second best option at each combination of fixed strategies is shown, with the x axis indicating the best performing strategy. The mean and a reference line at 0 are shown. 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 $.001$ after Bonferroni correction for the number of hypothesis tests attempted, showing that the strategy provides a significant improvement over the competing strategy. 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 2.2. **(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)** Using 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 the highest discriminabilities. **(C.ii)** As the number of ROIs increases, the discriminability tends to increase.

framework for optimal reference selection.

Discriminability provides a number of connections with related statistical algorithms worth further consideration. Discriminability is related to energy statistics [32], in which the statistic is a function of distances between observations [25]. Energy statistics provide approaches for goodness-of-fit (one-sample) and equality testing (two-sample), for which discriminability has demonstrated utility. Distance Components `DISCO` provides a measure of dispersion of successive observations for each subject [26]. Similar to discriminability, `DISCO` makes relatively general assumptions, only requiring the observations to lie in a space with a known distance measure. However, `DISCO` requires a large number of measurements per subject, which is often unsuitable for biological data where we frequently have only a small number of repeated trials per subject. Moreover, discriminability provides similar intuition to the multi-scale generalized correlation (`MGC`) [28, 34], a procedure to discover dependencies between disparate properties of data. Like `MGC`, discriminability uses distance methods in conjunction with analysis of the nearest neighbors to a given observation to determine the relationship between data.

Moreover, in a complementary manuscript, we explore the theoretical and empirical impact of us-

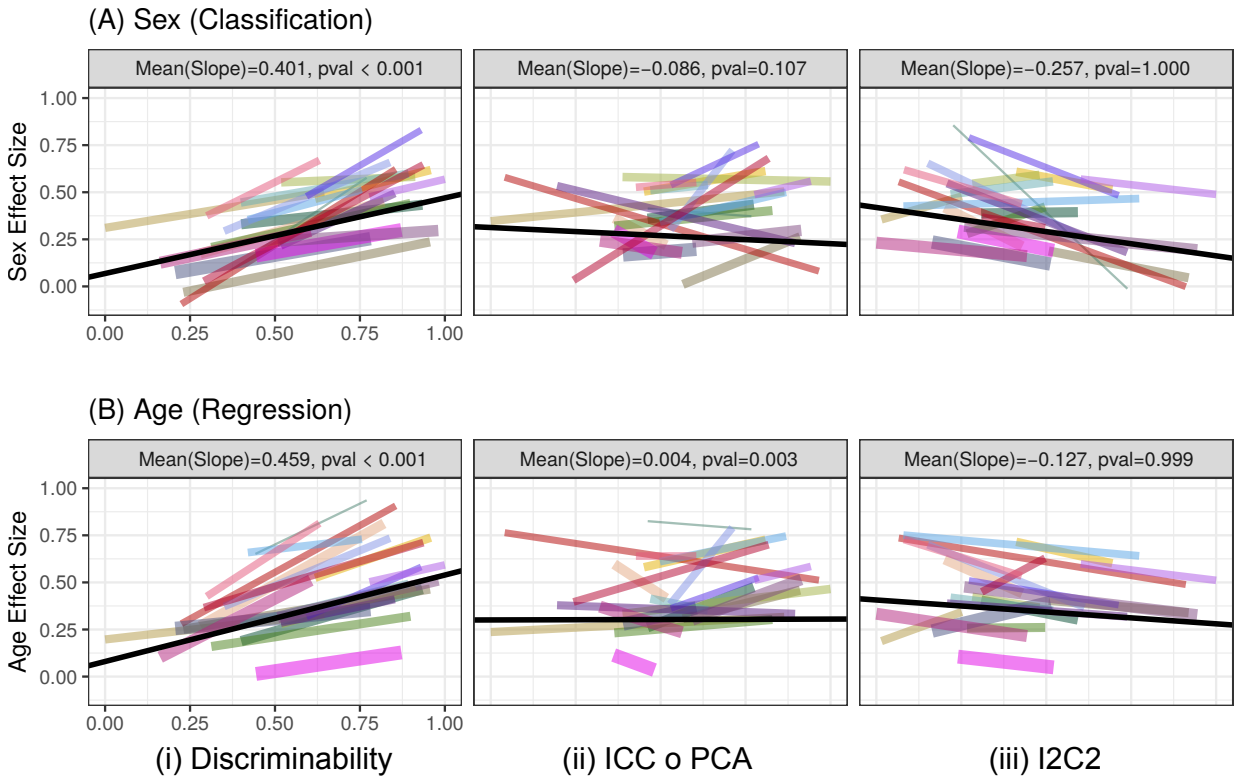

Figure 2.4: **Selection of Methods via Discriminability Increases Effect Size from Covariates**. Using the preprocessed connectomes from the 64 pipelines with raw edge-weights, we examine the dependence between **(A)** connectomes and sex, and **(B)** connectomes and age. Color and line width correspond to the Dataset and number of scans, respectively, from Figure 2.2B, and the weighted mean is shown by the solid black line. A reference statistic is computed (*x axis*) for each pipeline per dataset, along with the MGC statistic between the graphs and the covariate of interest (*y axis*) to measure the effect size in the connectomes due to the covariate. Both the *x* and *y* axes are normalized by the minimum and maximum statistic to facilitate comparison between the reference statistics. For each dataset, the effect size is regressed onto the reference statistic. Sample discriminability is the only reference statistic in which all slopes exceed zero. Moreover, we find that the corrected p-value [7] is significant across datasets for both covariates (med. p-value $< .001$). This indicates that pipelines with higher sample discriminability correspond to larger effect sizes for the covariate of interest, and that this relationship is stronger for sample discriminability than other reference statistics. Appendix C.2 details the methodologies employed.

ing related statistics to the discriminability, each leveraging a similar approach of using distance-based ranking for reference method identification. The theoretical and applied components of this work focus on understanding discriminability in the context of a model featuring additive gaussian noise, and distances are computed via the Euclidean distance. This framework holds in a fairly general statistical setting including various levels of model misspecification, as shown in [36]. A natural future investigation is to explore the impact of selection of appropriate distance metrics with discriminability. For example, in a high-dimensional scenario, an aptly chosen kernel may facilitate markedly improved performance both computationally and empirically. Moreover, different distance metrics may provide improved empirical performance under alternative models. Further, researchers may be interested in selection of the optimal distance metric for a downstream task using the discriminability two-sample test. In the scenario in which multiple experiments are conducted, we emphasize the importance of proper correction for multiple hypotheses. Additionally, we present a generalization of the ICC and ANOVA to multidimensional reference data by projecting onto the direction of maximal variance. Wang et al. [36] provide theoretical and empirical scenarios in which this approach provides advantages or disadvantages to

the discriminability and related statistics.

While to our knowledge discriminability provides the only direct framework for reference method selection, the researcher must still make informed considerations of the reference method with the potential downstream inference tasks of interest. For instance, the connectomes collected herein are resting-state fMRI connectomes; that is, the rs-fMRI scans were performed while a subject was sitting in a scanner unprompted. Recent literature has shown that while the global signal in a rs-fMRI scan may be a nuisance variable [15, 19] that can be regressed out (through Global Signal Regression, or GSR), the approach mathematically introduces artificial anticorrelations between different subnetworks in the resulting connectome [19, 20]. Negatively correlated subnetworks may be artificially augmented by the GSR procedure [20], and therefore downstream inference tasks focusing on the interpretation or analysis of anticorrelated networks may lose validity. To this end, we emphasize that while discriminability serves as an effective tool for comparison of reference methods, knowledge of the employed techniques in conjunction with the inference task is still a necessary component of an investigation.

On this note, it is important to emphasize that discriminability, as well the related metrics, 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 not biologically useful. In addition, none of the measures studied herein are immune to sample characteristics and thus care must be taken to interpret them across studies. For example, having a sample with variable ages will increase the inter-subject dissimilarity or variance of any metric dependent on age (such as the connectome). However, with these caveats in mind, discriminability remains as a key component of the practical utility of a measurement in a wide variety of settings.

Due to the high volume of available open-access data with informative downstream inferential covariates and pre-processing resources facilitating comparison of disparate reference methods, the connectomics use-case provided herein serves as an informative example of how discriminability can be used to facilitate reference method selection. We envision that discriminability will find substantial applicability across disciplines and sectors even beyond brain imaging, such as genomics, pharmaceutical research, and many other aspects of big-data science. To this end, we provide open-source implementations of discriminability for both `python` and `R` [2, 22]. Code for reproducing all the figures in this manuscript is available in the neurodata/r-mgc repository.

Acknowledgements This work was partially supported by the National Science Foundation award DMS-1707298, and the Defense Advanced Research Projects Agency's (DARPA) SIMPLEX program through SPAWAR contract N66001-15-C-4041. Xi-Nian Zuo receives funding supports by the National Basic Research (973) Program (2015CB351702), the Natural Science Foundation of China (81471740, 81220108014), Beijing Municipal Science and Tech Commission (Z161100002616023, Z171100000117012), the China - Netherlands CAS-NWO Programme (153111KYSB20160020), the Major Project of National Social Science Foundation of China (14ZDB161), the National RD Infrastructure and Facility Development Program of China, Fundamental Science Data Sharing Platform (DKA2017-12-02-21), and Guangxi BaGui Scholarship (201621).

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] Emily S Finn, Dustin Scheinost, Daniel M Finn, Xilin Shen, Xenophon Papademetris, and R Todd Constable. Can brain state be manipulated to emphasize individual differences in functional connectivity? *Neuroimage*, 160:140–151, October 2017.

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Kevin Murphy, Rasmus M Birn, Daniel A Handwerker, Tyler B Jones, and Peter A Bandettini. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage*, 44(3):893–905, February 2009.

[21] Kenichi Oishi et al. *MRI atlas of human white matter*. Academic Press, 2010.
[22] 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.
[23] Usama Pervaiz, Diego Vidaurre, Mark W Woolrich, and Stephen M Smith. Optimising network modelling methods for fMRI. August 2019.
[24] 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.
[25] Maria L Rizzo and Gábor J Székely. Energy distance. *WIREs Comput Stat*, 8(1):27–38, January 2016.
[26] 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.
[27] Cencheng Shen and Joshua T Vogelstein. Decision Forests Induce Characteristic Kernels. November 2018. URL http://arxiv.org/abs/1812.00029.
[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] 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.
[30] 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.
[31] Chandra S Sripada et al. Lag in maturation of the brainâĂŹs intrinsic functional architecture in attention-deficit/hyperactivity disorder. *Proceedings of the National Academy of Sciences*, 111 (39):14259–14264, 2014.
[32] 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.
[33] 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.
[34] 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.
[35] Charles Stein W. James. Estimation with quadratic loss. *Fourth Berkeley Symposium*, 1961.
[36] Zeyi Wang, Eric W Bridgeford, Joshua T Vogelstein, and et al Caffo, Brian. Statistical analysis of data reproducibility measures.
[37] Zeyi Wang, Haris Sair, Ciprian Crainiceanu, Martin Lindquist, Bennett A Landman, Susan Resnick, Joshua T Vogelstein, and Brian Caffo. On statistical tests of functional connectome fingerprinting. October 2018.
[38] 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.
[39] 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.
[40] 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.

[41] Xi-Nian Zuo, Ting Xu, and Michael P Milham. Harnessing reliability for neuroscience research. *Nature Human Behaviour*, 3:768–771, August 2019.

Appendix A. Hypothesis Testing.

A.1 One-Sample Test Recall the one-sample hypothesis test, shown in Equation (2.1). To construct a formal test using the sample discriminability, we can use two approaches. First, recall that under the assumption that $s_i = s_{i'}$ for all i, i', that we obtain a variance bound of $\hat{D}$, as shown in Wang et al. [36]. Using this bound on the variance, we can directly obtain a $(1-\alpha)$ confidence interval for $\hat{D}$, where we reject the null hypothesis if 0.5 does not lie within the confidence interval. Note that for the analytical one-sample test, the most complex part is computation of the discriminability itself. Then the computational complexity of the analytical one-sample test is $\mathcal{O}\big(N^2 \max(p, \max_i(s_i))\big)$, which is the same as the computational complexity of discriminability. While simple, this approach has several drawbacks. Particularly, this approach provides only a loose bound on the variance, yielding low test power.

Instead, we can approximate the distribution of $\hat{D}$ under the null through a permutation approach. We repeatedly permute the subject labels of our N samples, and compute $\hat{D}_{0,N}$ each time given 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 discriminability. The total computational complexity is then $\mathcal{O}\big(N^2 \max(p, rs)\big)$. We note that while more computationally costly than the analytical one-sample test, this approach is only linear in the number of desired repetitions, and therefore is sensible for most settings in which the sample discriminability can itself be computed. Moreover, we can greatly speed this computation up through parallelization. With T cores, the computational complexity is instead $\mathcal{O}\big(N^2 \max\big(p, \frac{r}{T}s\big)\big)$, as shown in Algorithm A.1. We extend this one-sample test to both `ICCoPCA` and `I2C2` to provide a robust p-value associated with both reference statistics of interest. In the event that the model is correctly specified by the `ANOVA` or `MANOVA` model, the permutation approach will produce a p-value that converges to the value obtained analytically through the assumptions made by the `ANOVA` or `MANOVA` model, respectively. In the event that the model is improperly specified by the `ANOVA` or `MANOVA` model but still maintain the notion of independence specified under the discriminability framework, the analytic approach will produce an invalid p-value, as the assumed null distribution of the test statistic will be inaccurate. This may lead to over or under estimation of the observed effect size. However, the permutation approach will provide a proper estimation of the null distribution of the test statistic.

A.2 Two-Sample Test Similar to the one-sample test, the two-sample test can be implemented through analytical or permutation approaches. Like the one-sample test, the two-sample test affords low power in the analytical derivation, and must place restrictive assumptions to achieve an analytical confidence interval. This confidence interval can be simply inverted to compute a p-value associated with the observed test statistic.

For the permutation-based approach, we begin by computing the observed difference in discriminability between two reference method choices. We begin to construct the null distribution of the difference in discriminability by first taking random convex combinations of the observed data from each of the two reference method choices (the "randomly combined datasets"). We compute the discriminability of each of the two randomly combined datasets for each permutation. Finally, for each permutation, we compute all pairs of observed differences in discriminability. We then compare the observed statistic 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 a pipeline having higher discriminability,

Algorithm A.1 Discriminability One-Sample Permutation Test

$\boldsymbol{X} \in \mathbb{R}^{n \times d}$ a data matrix with n samples in d dimensions.
$\boldsymbol{Y} \in \{1, \ldots, K\}^n$ sample ids for each sample.
$r \in \mathbb{Z}_+$ the number of permutations.
$\delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ a distance function. $p \in [0, 1]$ the p-value associated with the test. $D_{ij} = \delta(x_i, x_j)$ ▷ *compute distance matrix $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ $d_a = Discriminability(\boldsymbol{X}, \boldsymbol{Y})$ ▷ *compute observed discriminability ▷ Note that this for-loop can be parallelized over T cores, as the loops are independent processes
for i in $1, \ldots, r$ **do** $\boldsymbol{s} = sample(n, replace = FALSE)$ ▷ *resample our sample indices
$d_i = Discriminability(\boldsymbol{X}, \boldsymbol{Y}[\boldsymbol{s}])$ ▷ *Compute discriminability with random order of sample ids
$p = \frac{1}{r+1}\left(\sum_{i=1}^{r} \mathbb{I}_{\{d_a > d_i\}} + 1\right)$ ▷ *p-value is fraction of times observed is more extreme than under null
return p

Figure A.1: **Discriminability One-Sample Test Overview**. Our implementation of the permutation test for the one-sample test of the hypothesis given in Equation (2.1) 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.

lower discriminability, and equal discriminability relative a second pipeline; we implement all three in the software implementation of the two-sample test. The Algorithm for the two-sample test is shown in Figure A.2, with the alternative hypothesis as specified in Equation (2.2). The computational complexity is then $\mathcal{O}\left(\frac{r}{T}N^2 \max(p, \max_i(s_i))\right)$. Note that for each permutation, the limiting step is the computation of the discriminability in $\mathcal{O}\left(N^2 \max(p, s)\right)$. This is then offset through parallelization over T cores in the implementation. We extend this two-sample test to both `ICCoPCA` and `I2C2` to provide a robust p-value associated with both reference statistics of interest, for similar reasons to the above.

Appendix B. Simulations.

B.1 Algorithms

B.2 Benchmark Settings

With $X_j^{(i)}$ the random variable for the jth sample of the ith class. Note that for all simulations, $n_0 = n_1$.

One Sample Testing and Bayes Error The following simulations were constructed, where $\sigma_{min}, \sigma_{max}$ as indicated, and settings were run at 15 intervals on $[\sigma_{min}, \sigma_{max}]$ for 500 repetitions per setting. Dimensionality was 2, and number of individuals as indicated by K on the respective setting. The data has increasingly substantial noise added. Typically, i indicates the individual identifier, and t the measurement index.

1. No Signal: $K = 2$
   - $\boldsymbol{z}_{it} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $i = 1, \ldots, 2$, $t = 1, \ldots, 64$. Note: $\boldsymbol{0} \in \mathbb{R}^2$ is $\boldsymbol{0}_2$, and likewise for $\boldsymbol{I}$
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$, $\sigma \in [0, 20]$
   - $\boldsymbol{x}_{it} = \boldsymbol{z}_{it} + \boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\left(\boldsymbol{0}, (1 + \sigma^2)\boldsymbol{I}\right)$
2. Cross: $K = 2$
   - $\pi_1 = \pi_2 = \lambda_1 = \lambda_2 = \frac{1}{2}$
   - $\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 0.1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 2 \end{bmatrix}$
   - $\boldsymbol{z}_{it} \overset{iid}{\sim} \lambda_1 \mathcal{N}(\boldsymbol{0}_2, \boldsymbol{\Sigma}_i)$, $i = 1, 2$
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\left(\boldsymbol{0}_2, \sigma^2 \boldsymbol{I}\right)$, $\sigma \in [0, 20]$
   - $\boldsymbol{x}_{it} = \boldsymbol{z}_{it} + \boldsymbol{\epsilon}_{it}$

Algorithm A.2 Discriminability Two-Sample Test

$\boldsymbol{X}^{(1)} \in \mathbb{R}^{n \times d}$ a data matrix with n samples in d dimensions.
$\boldsymbol{X}^{(2)} \in \mathbb{R}^{n \times p}$ a data matrix with n samples in p dimensions.
$\boldsymbol{Y} \in \{1, \ldots, K\}^n$ sample ids for each sample, matched between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$.
$r \in \mathbb{Z}_+$ the number of permutations.
$\delta : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ a distance function. $p \in [0, 1]$ the p-value associated with the test. $\hat{D}^{(1)} = Discriminability(\boldsymbol{X}^{(1)}, \boldsymbol{Y})$ ▷ *The discriminability of the first sample. $\hat{D}^{(2)} = Discriminability(\boldsymbol{X}^{(2)}, \boldsymbol{Y})$ ▷ *The discriminability of the second sample. $d_a = \hat{D}^{(1)} - \hat{D}^{(2)}$ ▷ *The observed difference in discriminability between samples 1 and 2. ▷ *The for-loop below can be parallelized over T cores, as each loop is an independent process
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** $\boldsymbol{q}^{(k)_1} = sample(n)$▷ *The first id for the sample. $\boldsymbol{q}_2^{(k)} = sample(n)$▷ *The second id for the sample. $\lambda_j \overset{iid}{\sim} \mathrm{Unif}(0, 1)$ ▷ *for $j = 1, \ldots, n$, where $\boldsymbol{\Lambda} = (\lambda_j)_{j=1}^n$ $\boldsymbol{X}_n^{(k)} = \boldsymbol{\Lambda}' \cdot \boldsymbol{X}^{(k)}[\boldsymbol{q}_1^{(k)},] + (\boldsymbol{1} - \boldsymbol{\Lambda})' \cdot \boldsymbol{X}^{(k)}[\boldsymbol{q}_2^{(k)},]$ ▷ *Convex combination of the elements of $\boldsymbol{X}^{(k)}$ $d_{n,i}^{(k)} = Discriminability(\boldsymbol{X}_n^{(k)}, \boldsymbol{Y})$ ▷ *Compute discriminability of the convexly combined elements ▷ *Compute all pairs differences in discriminability using the convexly-combined samples
  for i in $1, \ldots, r-1$ **do**
    for j in $i, \ldots, r$ **do** $d_n \leftarrow c\Big(d_n, d_{n,i}^{(1)} - d_{n,j}^{(2)}, d_{n,j}^{(2)} - d_{n,i}^{(1)}\Big)$ ▷ *Null distribution of the difference $p = \frac{2}{r(r-1)} \sum_{i=1}^{|d_n|} \mathbb{I}_{\{d_a < d_{n,i}\}}$ ▷ *p-value is fraction of times that observed discriminability is less than under the null **return** p

Figure A.2: **Discriminability Two-Sample Test Overview** Our implementation of the permutation test for the hypothesis given in Equation (2.2) requires $\mathcal{O}\big(\frac{r}{T} N^2 \max(p, s)\big)$ 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.

3. Gaussian: $K = 16$
   - $\pi_1 = \pi_2 = \frac{1}{2}$
   - $\boldsymbol{\mu}_i \overset{iid}{\sim} \pi_1 \mathcal{N}(\boldsymbol{0}_2, 4\boldsymbol{I})$, $i = 1, \ldots, 16$
   - $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$
   - $\boldsymbol{z}_{it} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\big(\boldsymbol{0}_2, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 20]$
   - $\boldsymbol{x}_{it} = \boldsymbol{z}_{it} + \boldsymbol{\epsilon}_{it}$
4. Annulus/Disc: $K = 2$
   - $\pi_1 = \pi_2 = \frac{1}{2}$
   - $\boldsymbol{z}_{1t} \overset{iid}{\sim} \mathbb{B}(r = 1) + \mathcal{N}(\boldsymbol{0}_2, 0.1\boldsymbol{I})$ samples uniformly on unit ball of radius 2 with gaussian error
   - $\boldsymbol{z}_{2t} \overset{iid}{\sim} \mathbb{S}(r = 1.5) + \mathcal{N}(\boldsymbol{0}_2, 0.1\boldsymbol{I})$ samples uniformly on unit sphere of radius 2 with gaussian error
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\big(\boldsymbol{0}_2, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 10]$
   - $\boldsymbol{x}_{it} = \boldsymbol{z}_{it} + \boldsymbol{\epsilon}_{it}$

The 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 was the 8 left most centers $\{\mu_i\}$, and the second class the right most 8 centers in $\{\mu_i\}$.

Two Sample Testing The following simulations were constructed, where $\sigma_{min}, \sigma_{max}$ as indicated, and settings were run at 15 intervals on $[\sigma_{min}, \sigma_{max}]$ for 500 repetitions per setting. Dimensionality was 2, and number of classes as indicated. Typically, j indicates the pipeline choice, i indicates the individual identifier, and t the measurement index. The second pipeline has added gaussian error compared to the first pipeline, and therefore, we would anticipate an observed discriminability of the first pipeline exceeding the second pipeline. Note that in the below simulations, noise is added for the second pipeline, so the natural hypothesis is:

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

1. No Signal: $K = 2$
   - $\boldsymbol{z}_{jit} \overset{d}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $j = 1, 2$ $i = 1, \ldots, 2$, $t = 1, \ldots, 64$
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\big(\mathbf{0}, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 10]$
   - $\boldsymbol{x}_{jit} = \begin{cases} \boldsymbol{z}_{1it} & j = 1 \\ \boldsymbol{z}_{2it} + \boldsymbol{\epsilon}_{it} & j = 2 \end{cases}$
2. Cross: $K = 2$
   - $\pi_1 = \pi_2 = \lambda_1 = \lambda_2 = \frac{1}{2}$
   - $\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 0.1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 2 \end{bmatrix}$
   - $\boldsymbol{z}_{jit} \overset{iid}{\sim} \lambda_1 \mathcal{N}(\mathbf{0}_2, \boldsymbol{\Sigma}_i)$, $i = 1, 2$
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\big(\mathbf{0}_2, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 2]$
   - $\boldsymbol{x}_{jit} = \begin{cases} \boldsymbol{z}_{1it} & j = 1 \\ \boldsymbol{z}_{2it} + \boldsymbol{\epsilon}_{it} & j = 2 \end{cases}$
3. Gaussian: $K = 16$
   - $\pi_1 = \pi_2 = \frac{1}{2}$
   - $\boldsymbol{\mu}_i \overset{iid}{\sim} \pi_1 \mathcal{N}(\mathbf{0}_2, 4\boldsymbol{I})$, $i = 1, \ldots, 16$
   - $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$
   - $\boldsymbol{z}_{jit} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\big(\mathbf{0}_2, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 2]$
   - $\boldsymbol{x}_{jit} = \begin{cases} \boldsymbol{z}_{1it} & j = 1 \\ \boldsymbol{z}_{2it} + \boldsymbol{\epsilon}_{it} & j = 2 \end{cases}$
4. Annulus/Disc: $K = 2$
   - $\pi_1 = \pi_2 = \frac{1}{2}$
   - $\boldsymbol{z}_{j1t} \overset{iid}{\sim} \mathbb{B}(r = 1) + \mathcal{N}(\mathbf{0}_2, 0.1\boldsymbol{I})$ samples uniformly on unit ball of radius 2 with gaussian error
   - $\boldsymbol{z}_{j2t} \overset{iid}{\sim} \mathbb{S}(r = 1.5) + \mathcal{N}(\mathbf{0}_2, 0.1\boldsymbol{I})$ samples uniformly on unit sphere of radius 2 with gaussian error, $i = 1, 2$, $j = 1, 2$
   - $\boldsymbol{\epsilon}_{it} \overset{iid}{\sim} \mathcal{N}\big(\mathbf{0}_2, \sigma^2 \boldsymbol{I}\big)$, $\sigma \in [0, 10]$
   - $\boldsymbol{x}_{jit} = \begin{cases} \boldsymbol{z}_{1it} & j = 1 \\ \boldsymbol{z}_{2it} + \boldsymbol{\epsilon}_{it} & j = 2 \end{cases}$

Appendix C. Connectomics Application.

C.1 Data Collection and Processing

fMRI Preprocessing 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 preprocessing 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. Specific data processing instructions for deployment in `AWS` can be found in the neurodata-arxiv/f2g repository. All data preprocessing was performed in the `AWS` cloud using `CPAC` version `3.9.2` [3].

dMRI Preprocessing Pipelines The dMRI connectomes were acquired as follows. The dMRI scans were pre-processed for eddy currents using FSL's `eddy-correct` [30]. FSL's "standard" linear registration pipeline was used to register the sMRI and dMRI images to the MNI152 atlas [11, 17, 30, 38]. A tensor model is fit using DiPy [9] to obtain an estimated tensor at each voxel. A deterministic tractography algorithm is applied using DiPy's EuDX [8, 9] 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 [18], anatomical [14, 16, 21, 33], or functional [4, 5, 12, 31] 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. $w(v_i, v_j) = \sum_{u \in P(v_i)} \sum_{w \in P(v_j)} \mathbb{I}\{F_{u,w}\}$ 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. [13], consisting of parcellations defined in the MNI152 space [4, 5, 12, 14, 16, 21, 31, 33].

C.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. Ideally, for a particular summary statistic, a high value will generally correlate with a positive effect size. For a dataset $i = 1, \ldots, D$ where D is the total number of datasets and a pipeline $j = 1, \ldots, 192$ for 192 total pipelines and $k = 1, \ldots, 3$ are our summary statistics of interest, we posit the model:

$$Y_{ij} = \beta_{ik} X_{ijk} + \epsilon_{ijk}$$

Where we model the effect Y_{ij} estimated by `MGC` [27] as a linear combination of a fixed effect X_{ijk}, the observed sample statistic for approach k (discriminability, `ICCoPCA`, or `I2C2`), and random noise ϵ_{ijk}, with coefficient β_{ijk}. Note that the interpretation of β_{ijk} is the expected change in the response Y_{ij} due to a single unit change in the observed sample statistic X_{ijk}. Both Y_{ij} and X_{ijk} are uniformly normalized across all pipelines within a single dataset to facilitate intuitive comparison across methods. We pose the following hypothesis:

$$H_0 : \beta_{ik} = 0; \quad H_A : \beta_{ik} > 0$$

Acceptance of the alternative hypothesis would have the interpretation that an increase in the observed sample statistic X_{ijk} would tend to correspond to an increase in the observed effect size Y_{ij}. The preceding hypothesis is tested using a 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 processed nor the sample statistics were computed. This provides evidence that the statistic is informative for reference selection within the context of this investigation. Model fitting for this investigation is conducted using the `lm` package in the `R` programming language [24].

Useful Data Links All relevant analysis scripts and data for figure reproduction in this manuscript made publicly available, and can be found at neurodata/r-mgc github link.