Bagging Improves Reproducibility of Functional Parcellation of the Human Brain

Aki Nikolaidis (1), Anibal Solon Heinsfeld (1), Ting Xu (1), Pierre Bellec (2), Joshua Vogelstein (3), Michael Milham (1).

1- The Child Mind Institute, 101 East 56th Street. New York, NY 10022
2- John's Hopkins University 3400 N. Charles St Baltimore, MD 21218
3- University of Montreal, PO Box 6128 Downtown STN Montreal QC H3C 3J7 CANADA

Abstract:
Increasing the reproducibility of neuroimaging measurement addresses a central impediment to the clinical impact of human neuroscience. Recent efforts demonstrating variance in functional brain organization within and between individuals shows a need for improving reproducibility of functional parcellations without long scan times. We apply bootstrap aggregation, or bagging, to the problem of improving reproducibility in functional parcellation. We use two large datasets to demonstrate that compared to a standard clustering framework, bagging improves the reproducibility and test-retest reliability of both cortical and subcortical functional parcellations across a range of sites, scanners, samples, scan lengths, and clustering parameters. With as little as six minutes of scan time bagging creates more reproducible parcellations than standard approaches with twice as much data. This suggests bagging may be a key method for improving functional parcellation and bringing functional neuroimaging-based measurement closer to clinical impact.

Keywords: reproducibility, bagging, parcellation, cluster ensembles, reliability, discriminability, biomarkers, striatum, thalamus, subcortical, basal ganglia, fMRI, bootstrap

1. Introduction

Reproducibility is a prerequisite for building solid domain knowledge in the neurosciences, and recent efforts have focused on addressing problems that lead to lack of reproducibility (Millman, Brett, Barnowski, & Poline, 2018; Nichols et al., 2017; Zuo et al., 2014; Zuo, Biswal, & Poldrack, 2019). Given that functional parcellation remains a central goal for neuroscience, finding methods to enhance reproducibility of functional parcellation is important for improving robustness of findings. The past decade has yielded a high degree of success in the identification of large-scale functional networks that are highly reproducible across studies and laboratories, and can be measured at the individual level with moderate to high test-retest reliability (Arslan et al., 2018; Buckner, Krienen, Castellanos, Diaz, & Yeo, 2011; Noble et al., 2017; Yang et al., 2012; Yeo et al., 2011; X. N. Zuo et al., 2010). In contrast, efforts to delineate functionally-defined areal parcels have yielded more variable results across studies, with few focusing on the individual level (Glasser et al., 2016; Xu et al., 2016). Among those that have attempted to quantify either the reproducibility or reliability of parcellation results, several have pointed to the need for optimization of methodologies. Critically, the minimum data requirements (25+ minutes) needed to achieve desirable levels of reliability appear to supercede most clinical and cognitive neuroscience datasets already collected or currently underway (Laumann et al., 2015; Xu et al., 2016; Zuo, Xu, & Milham, 2019). As such, there is an increasing need for methods that can optimize the value of resting state acquisitions with insufficient data.

Here, we draw upon advances in the machine learning literature, bagging and ensemble clustering in particular, to assess whether we can bring down the minimum data needed to achieve reproducible and reliable functional parcellations. The motivation behind bagging, a resampling technique that aggregates bootstrap samples (Breiman, 1996), is to reduce variability in the estimation process through averaging (Dudoit & Fridlyand, 2003). While originally designed for prediction, bagging has more recently become an important technique for ensemble clustering

(Fischer & Buhmann, 2003; Hong, Kwong, Wang, & Ren, 2009; Jia, Xiao, Liu, & Jiao, 2011; Li & Ding, 2008; Zhou, 2012). The essential advantage of ensemble clustering derives from the added value of combining multiple cluster assignments into a single clustering (Strehl & Ghosh, 2002). In other words, the aggregation of cluster solutions themselves become the features for ensemble clustering. Numerous studies have found that the instability and sensitivity inherent to the cluster optimization procedure are significantly attenuated by cluster ensembles generated through bagging (Dudoit & Fridlyand, 2003; Hoyos-Idrobo, Schwartz, Varoquaux, & Thirion, 2015; Jia et al., 2011; Li & Ding, 2008; Strehl & Ghosh, 2002; Zhou, 2012), making this a promising area of research in machine learning.

Here, we examine the impact of bagging on the reproducibility and reliability of functionally-defined cortical and subcortical parcellations generated using resting state fMRI data. We leverage and extend the Bootstrap Analysis of Stable Clusters (BASC) framework. BASC applies multi-level bagging to create group and individual-level functional parcellations. We extend this framework to leverage the power of nilearn (https://nilearn.github.io/), Scikit learn (Pedregosa et al., 2011), and Nipype (Gorgolewski et al., 2011) by building an open source python package for BASC, PyBASC (Nikolaidis et al., 2019); GitHub: https://github.com/AkiNikolaidis/PyBASC; PyPI: https://pypi.org/project/PyBASC/) Though introduced to the imaging literature nearly a decade ago (Bellec, Rosa-neto, Lyttelton, Benali, & Evans, 2010), the full merits of bagging are yet to be tested or appreciated in functional parcellation.

Our strategy for establishing and assessing reproducibility is motivated by the COBIDAS report (Nichols et al., 2017) and broader outlines for the establishment of reproducibility in metrology and biomedical science (JCGM, 2008). These efforts have highlighted a range of challenges in reproducibility the field faces as we aim for more robust estimates of brain organization and biomarker discovery (Table 2; (Zuo, Biswal, et al., 2019; Zuo, Xu, et al., 2019). In brief, we focus on creating functional parcels that improve 1-Between Sample Reproducibility:

similarity in group-level parcellations obtained between independent samples; 2-Between Session Reproducibility: similarity of group and individual level parcellations obtained in the same sample at different time points; 3-Reliability: consistency of differences among individuals on repeat assessments at distinct time points (See Methods 4.5-4.6 for more detailed definitions of these terms).

Keeping in line with extensive efforts at both cortical (Bellec et al., 2010; Craddock, James, Holtzheimer, Hu, & Mayberg, 2012; Kelly et al., 2009; Kelly et al., 2012; Margulies et al., 2007) and subcortical parcellation (Barnes et al., 2010; Choi, Yeo, & Buckner, 2012; Garcia-Garcia et al., 2017; Janssen, Jylänki, Kessels, & van Gerven, 2015), we perform functional parcellation at multiple resolutions in both cortex (k = 7, 20, 197, 444) and subcortex (k = 10, 18) to assess the impact of bagging on reproducibility and reliability of these parcellations. While several cortical parcellations are well established (Glasser et al., 2016; Yeo et al., 2011), subcortex is an area where parcellations are less well-studied (Choi et al., 2012; Janssen et al., 2015). Subcortical regions are also in need of methods to improve reproducibility and reliability given that they show connectivity with notably lower reliability than the rest of the brain (Noble et al., 2017; O'Connor et al., 2017).

Functional parcellations often are created with strong spatial constraints to create more homogenous parcel shapes with smoother parcel boundaries (Craddock et al., 2012; Schaefer et al., 2018; Xu et al 2016). Such constraints can be considered a version of spatial regularization that enhances the generalizability of resulting parcellations. To provide the best test of the impact of bagging on parcellation reproducibility we use minimal spatial constraints in the current approach, enabling parcels that are discontinuous and even bilaterally homologous. Such an approach yields parcels with noisier boundaries than would be the case with stronger spatial constraints, such as Craddock et al. 2012, but our approach a allows the parcellation a better fit to a given sample and provides a stricted and more direct test of the impact of bagging.

The present study leverages openly available datasets that are well suited for the establishment of minimum data requirements for reproducibility (Table 1). We use the Hangzhou Normal University (HNU1) test retest dataset (Zuo et al., 2014), which is comprised of 30 participants each scanned 10 times for 10 minutes over the course of 30 days. We also use the Genomics Superstruct Project (GSP) dataset (Holmes et al., 2015), from which we select a subset of 300 participants with two 6 minute resting state scans, which we age and sex matched to the HNU1 dataset. Given the extensive computational burden of the current set of analysis, we run a full spectrum of parameters in the subcortical parcellation, and an extensive, but reduced set on the cortical parcellation. **We demonstrate that compared to a standard clustering framework, bagging improves the reproducibility and reliability of both cortical and subcortical functional parcellations across a range of sites, scanners, samples, scan lengths, and clustering parameters. These results suggest that bagging may be an important component for achieving more robust parcellations in functional neuroimaging.**

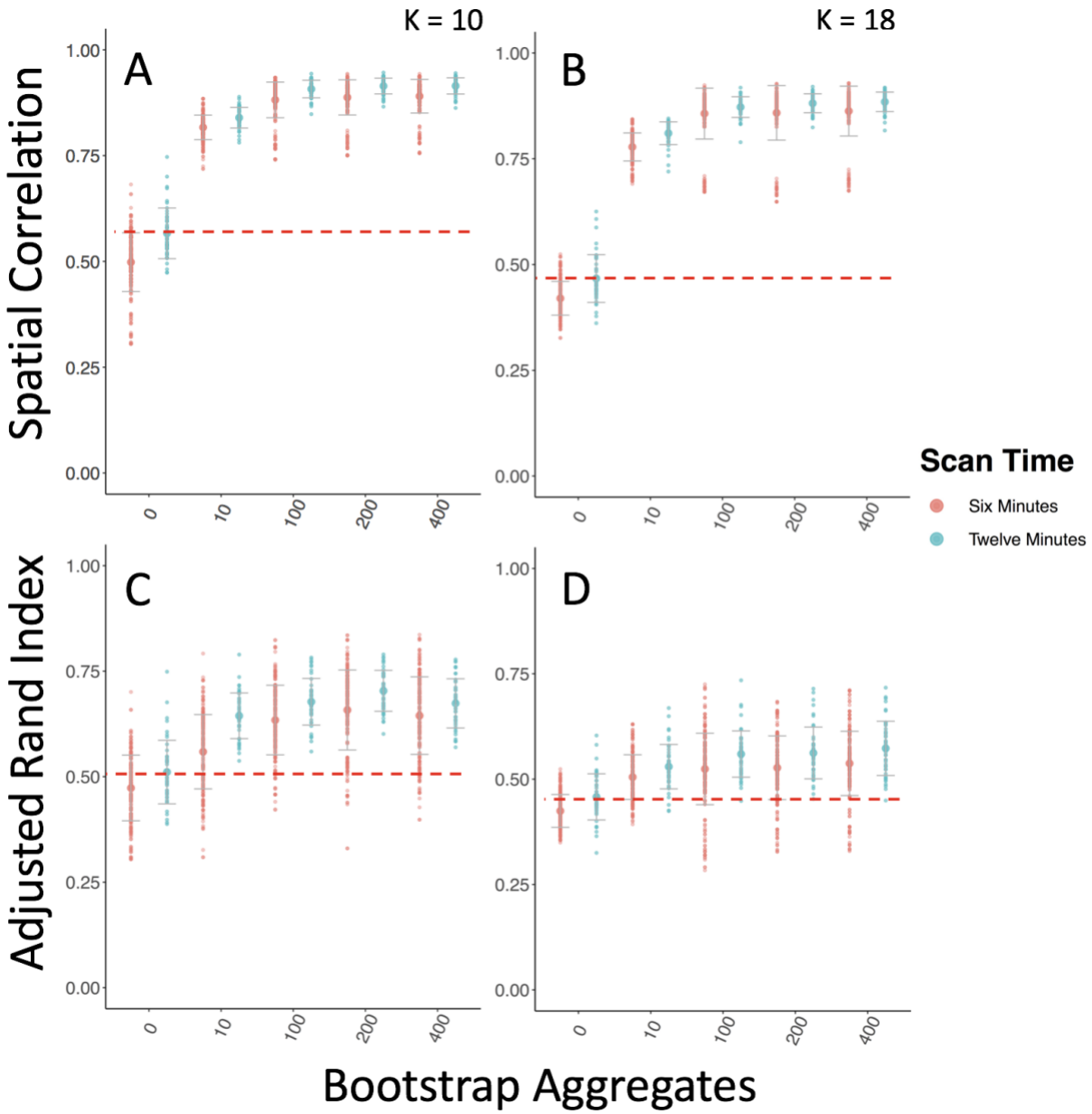


Figure 1- Bagging Improves Group Level Within Study/Between Sample Subcortical Reproducibility. Comparing the parcellations from 10 groups of 30 participants from the GSP sample to one another. The dotted red line shows the highest mean reproducibility of the standard parcellation approach. A-B Spatial correlation between mean group adjacency matrix of 10 GSP groups with 10 and 18 clusters. C-D Adjusted Rand Index between cluster labels of 10 GSP groups with 10 and 18 clusters.

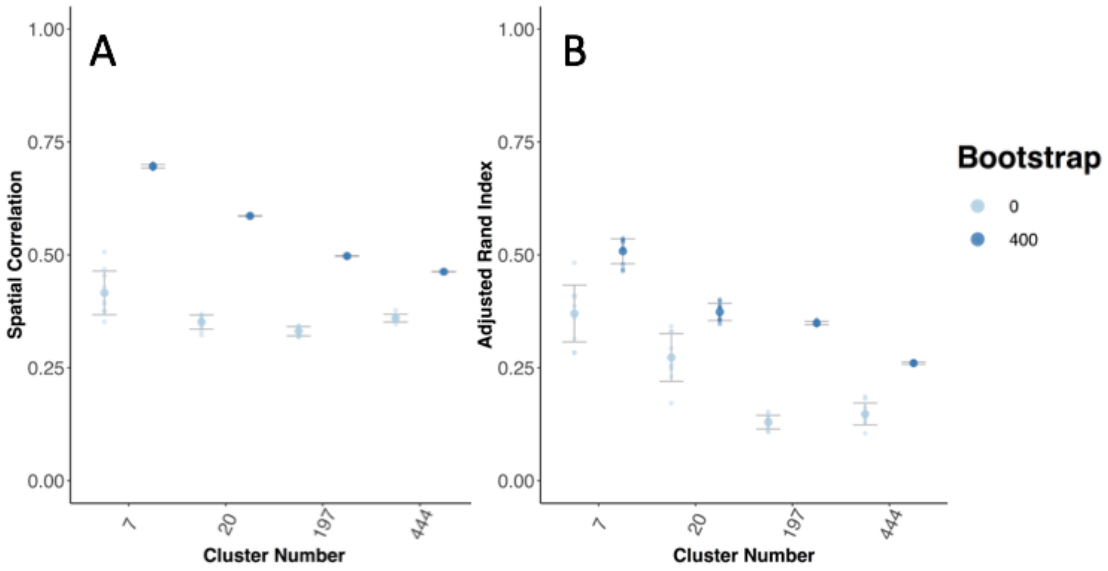


Figure 2- Bagging Improves Between Study/Between Sample Cortical Reproducibility. We compared a reference parcellation created from the 12 minute combined scan from 300 individuals from the GSP to 10 repeated parcellations of 10 minutes of HNU1 data. A) Spatial correlation between the group mean adjacency matrix for the GSP and HNU1 parcellations. B) Adjusted Rand Index between the cluster labels for the GSP and HNU1 parcellations.

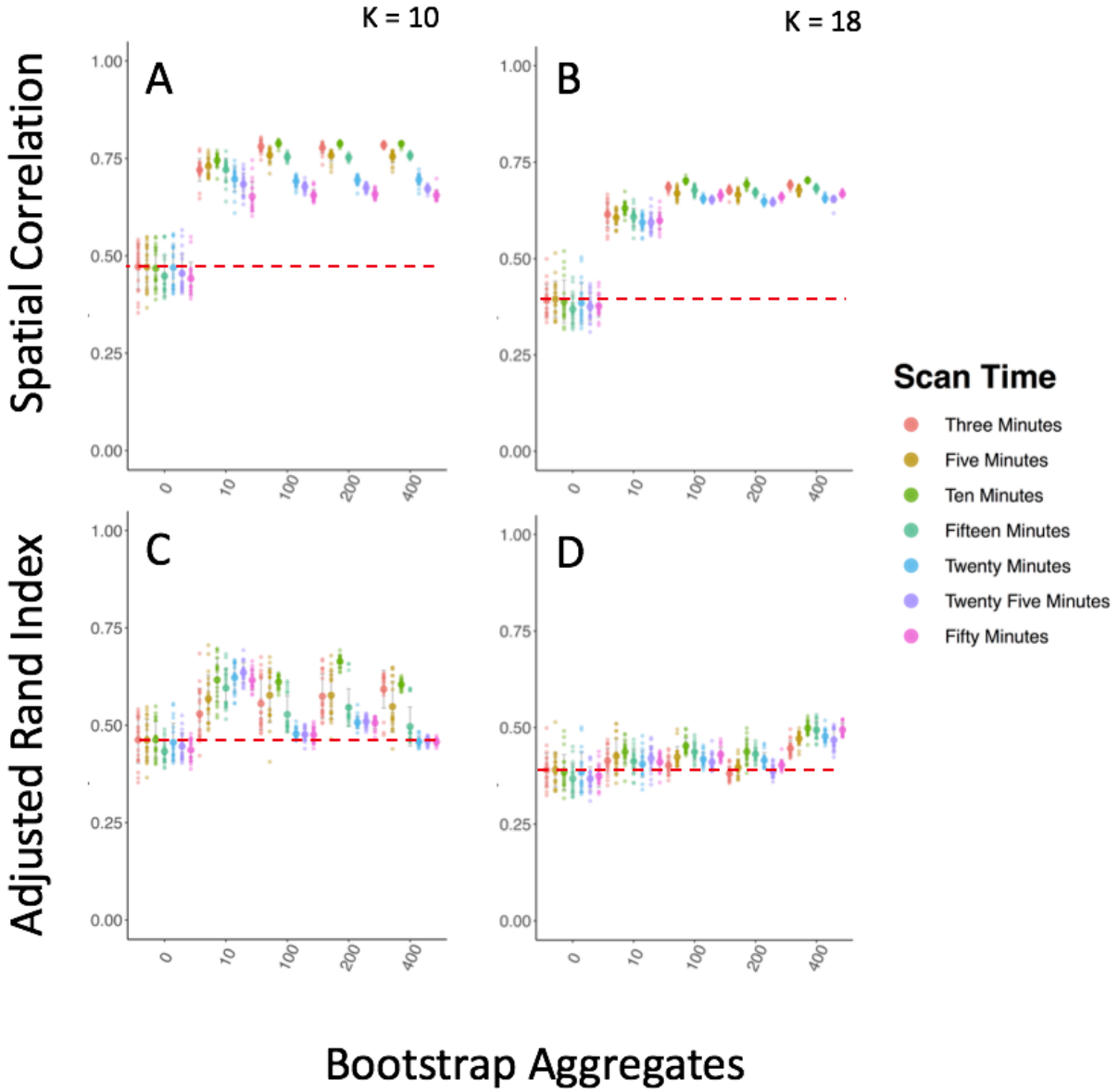


Figure 3- Bagging Improves Between Sample Subcortical Reproducibility. Bagging improves the reproducibility of subcortical brain parcellation between the GSP and HNU1 datasets, though results more mixed than the cortical results. We compared a reference parcellation created from the 12 minute combined scan from 300 individuals from the GSP to 10 repeated parcellations of 10 minutes of HNU1 data. The dotted red line shows the highest mean reproducibility of the standard parcellation approach. A) Spatial correlation between the group mean adjacency matrix for the GSP and HNU1 parcellations. B) Adjusted Rand Index between the cluster labels for the GSP and HNU1 parcellations.

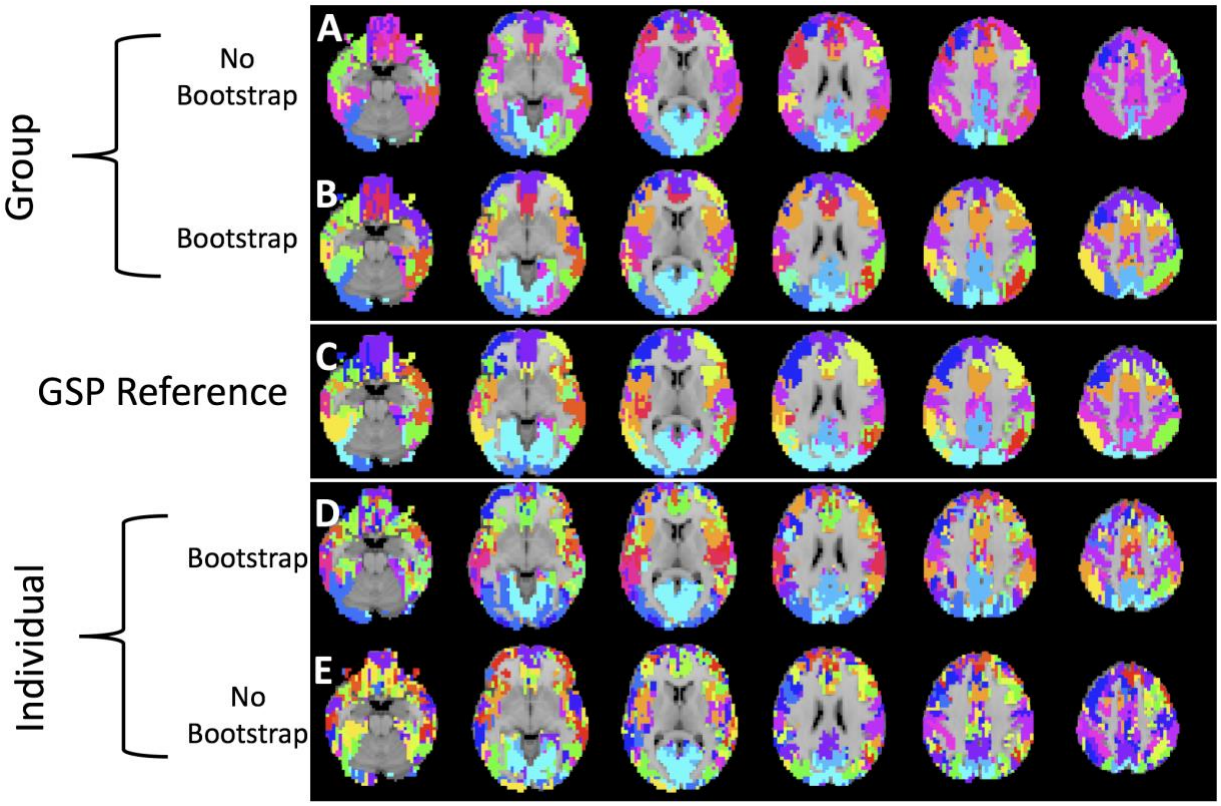


Figure 4 - Visualization of Between Study Cortical Parcellations. We show the similarity of the HNU1 parcellation to the GSP reference parcellation with 20 clusters as shown in Figure 2. Notably, both the group and individual level parcellations with 400 bootstraps are closer to the GSP reference parcellation than either of the standard parcellations. A) Group-level parcellation with standard approach, B) and bootstrapping. C) GSP Reference parcellation. D) Individual-level parcellation with Bootstrapping, E) and standard approach.

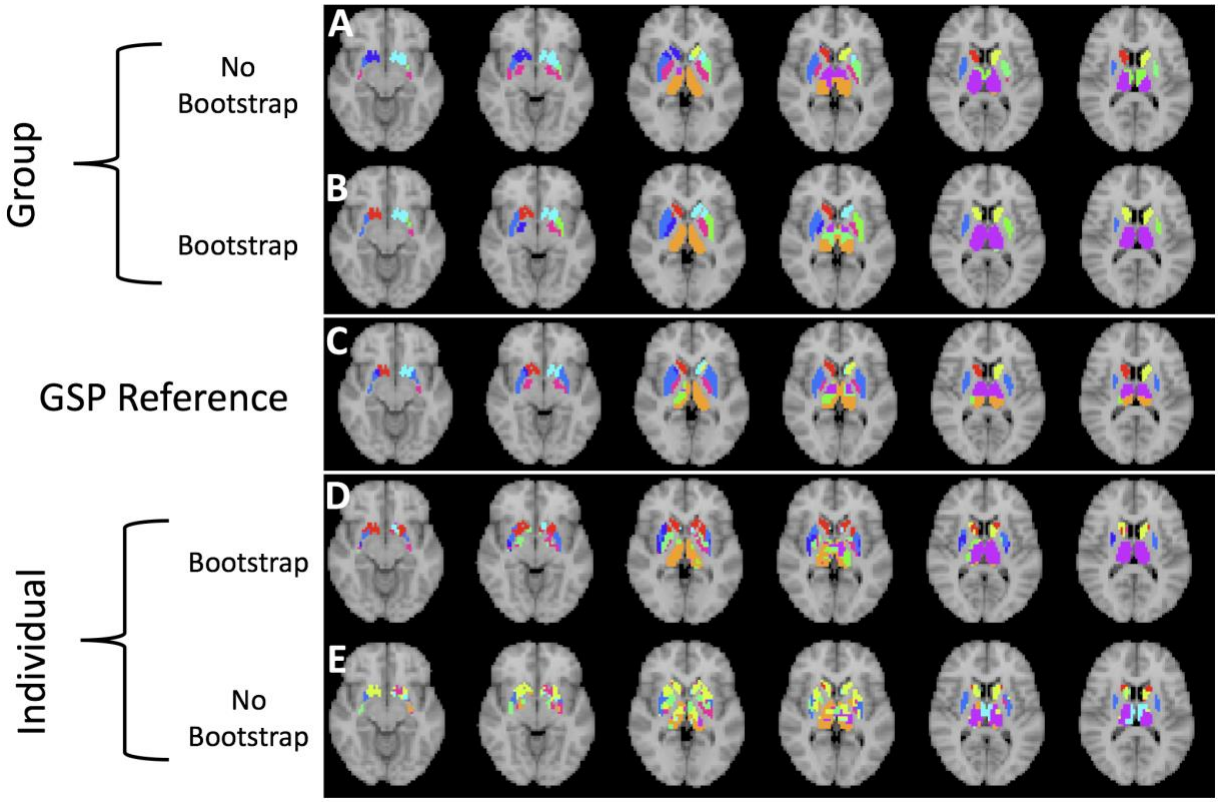


Figure 5 - Visualization of Between Study Subcortical Parcellations. We show the similarity of the HNU1 parcellation to the subcortical GSP reference parcellation with 10 clusters as shown in Figure 3. Both the group and individual level bootstrapped parcellations are notably more similar to the GSP reference than either of the standard parcellation counterparts. A) Group-level parcellation with standard approach, B) and bootstrapping. C) GSP Reference parcellation. D) Individual-level parcellation with Bootstrapping, E) and standard approach.

2 Results

2.1 Between Sample Reproducibility

Improving between sample reproducibility is key for creating generalizable scientific discovery. We assess two different levels of between sample reproducibility here: within study and between study, and we hypothesized that bagging should have a beneficial impact on both levels. As a first step, we investigated the effect of bagging on the within study/between sample reproducibility of our subcortical parcellations (Figure 1). We split the 300 GSP participants into 10 groups of 30, and computed the correlation of the group mean adjacency matrices (Figure 1 A, B), and the ARI of their cluster labels between groups (Figure 1 C, D). We found that while increasing the time of the scan from 6 to 12 minutes significantly improved reproducibility when measured by both correlation and ARI (K = 10; Correlation: Kruskal-Wallis chi-squared = 58.517, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 10.788, df = 1, p-value < 0.005), we found that the impact of bagging improved reproducibility more than increasing the scan length (K = 10; Correlation: Kruskal-Wallis chi-squared = 401.26, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 278.9, df = 1, p-value < 0.0001), indicating that the 6 minute scans could become more reproducible than scans twice as long with the use of bagging. We also found that bagging significantly decreased the variance in the reproducibility estimates for both 6 minute and 12 minute scans, but had a greater impact on the 12 minute scans (6 min ARI variance 0 vs 400 bootstraps: F = 0.716, num df = 267, denom df = 267, p-value < 0.01; 12 min ARI variance 400 vs 0 bootstraps F = 1.6712, num df = 89, denom df = 89, p-value < 0.05).

As a further test of generalizability we tested between study/between sample reproducibility of both cortical parcellation between the GSP and HNU1 studies (Figure 2; Figure 4). These studies differ on many criteria including scanner type and population, while the GSP data was collected in the USA with Siemens 3T Magnetom Tim Trio scanners, the HNU1 data was collected

in China with a GE 3T scanner. We hypothesized that bagging should not only improve the reproducibility of the parcellations, but also that it may have an impact on the consistency of the replication parcellations as well. We conducted a range of tests to assess these concepts. First, using the 300 individuals from the GSP study to create a reference cortical parcellation, we assessed the correlation of the group mean adjacency matrices (Figure 2 A), and the adjusted rand index (ARI) of the group cluster labels (Figure 2 B). We demonstrated that cortical parcellations were more reproducible between studies using bagging; for example, for K = 444, bagging improves ARI (Kruskal-Wallis test comparing 400 vs 0 bootstraps: chi-squared = 14.30; $p < 0.0005$). We also found that bagging improved the consistency of reproducibility, meaning that across runs bagging produced cluster solutions whose varied less between one another (Test of unequal variance of ARI 400 vs 0 bootstraps: $F = 94.8$, $p < 0.0001$).

Third, we found that bagging had a significant impact on the between study/between sample reproducibility of subcortical parcellation, comparing the GSP reference to the HNU1 dataset (Figure 3; Figure 5). Both correlation and ARI of the parcellations improves with scan time from 10 to 50 minutes as expected. However, we expected that bagging would improve the parcellation labels to improve equally for both the 10 and 50 minute replication samples. What we found instead, is that while both improved significantly from 0 to 400 bootstraps, the 10 minute HNU1 dataset improved much more than the 50 minute (K = 10, 10 minute: $t = -13.087$, $df = 21.389$, p-value < 0.0001; 50 minute: $t = -3.2407$, $df = 21.75$, p-value < 0.005). When visualized (Figure 5), we see this difference is largely due to the fact that both the GSP reference and the 10 minute HNU1 scans split the bilateral caudate into two clusters, whereas the 50 minute HNU1 parcellation did not. In fact, the 50 minute HNU1 replication and reference data yielded a seemingly better quality parcellation overall, with less variance on the edges between parcels, and more homotopic similarity, which can be expected given homotopic connectivity is known to be especially high (Zuo et al., 2010). We also visualized the voxelwise differences in parcellation

labels on average over all HNU bootstrapped conditions and found most differences could be found in the boundary regions between parcels (See Supplementary Figure 1). Of note here, is that the clusters created have minimal spatial constraints applied to them, implying that neither the cluster solutions themselves, nor the improvement in reproducibility come from the clusters being forced into a similar spatially or anatomically constrained configuration. In fact we see many clusters in the HNU1 50 minute Reference and Replication solution that are anatomically distinct but functionally united bilateral homologues (ie the left and right putamen cluster together, the left and right caudate cluster together, etc).

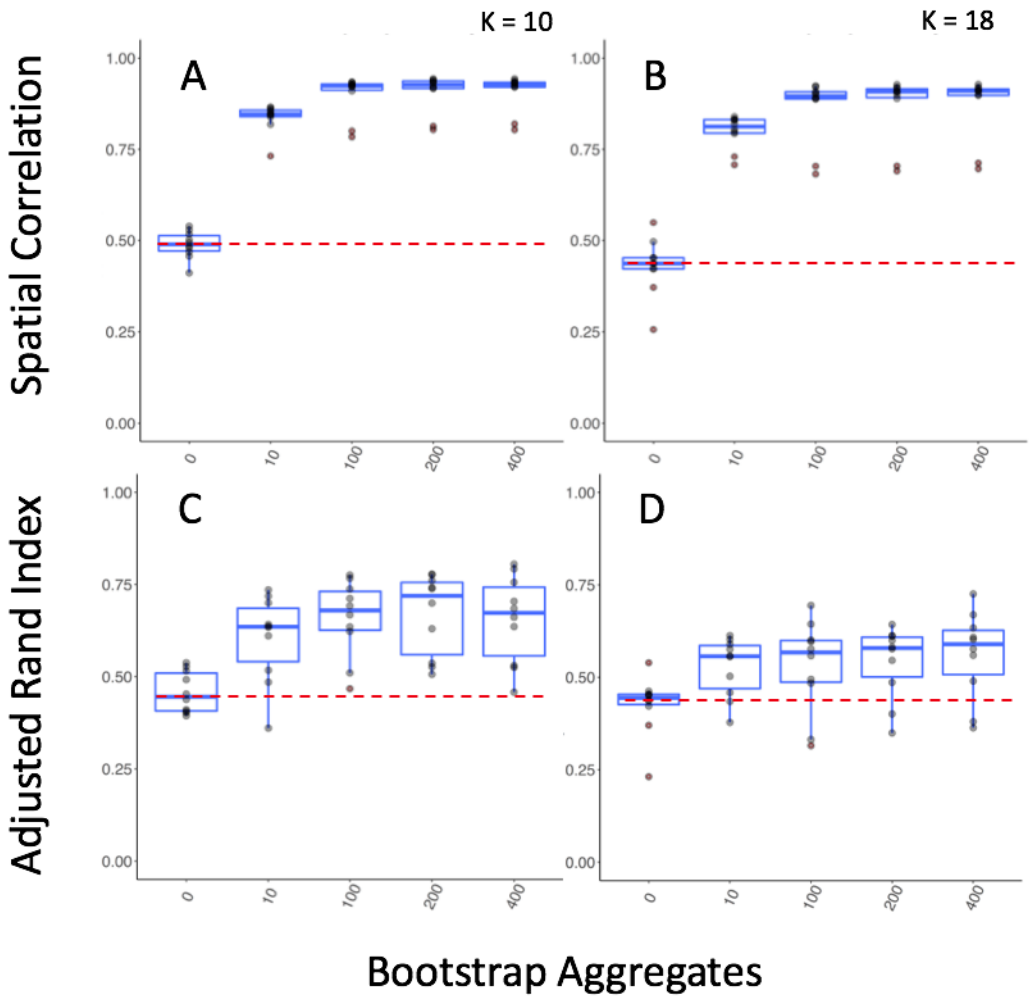


Figure 6- Bagging Improves Within Sample/Between Scan Subcortical Reproducibility. We split the 300 participants of the GSP dataset into 10 groups of 30, and compare the parcellations of their first and second functional scan to one another. The dotted red line shows the highest mean reproducibility of the standard parcellation approach. A-B show the improvement in bagged parcellations measured by spatial correlation of the group mean adjacency matrix for 10 and 18 clusters respectively. C-D show the improvement in bagged parcellations measured by the adjusted Rand index of the cluster labels for 10 and 18 clusters respectively.

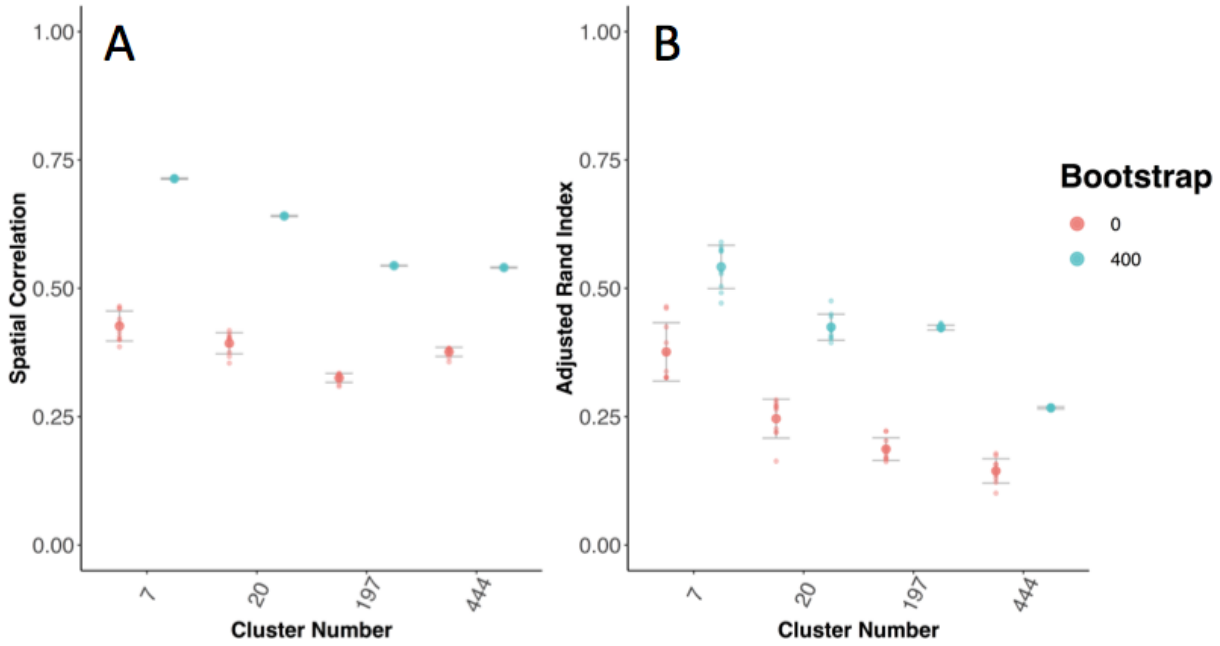


Figure 7- Bagging Improves Between Session Cortical Parcellation. We compared the HNU1 reference to the HNU1 replication cortical parcellation. A) Spatial correlation between the group mean adjacency matrix for the reference and replication parcellations. B) Adjusted Rand Index between the cluster labels for the reference and replication parcellations.

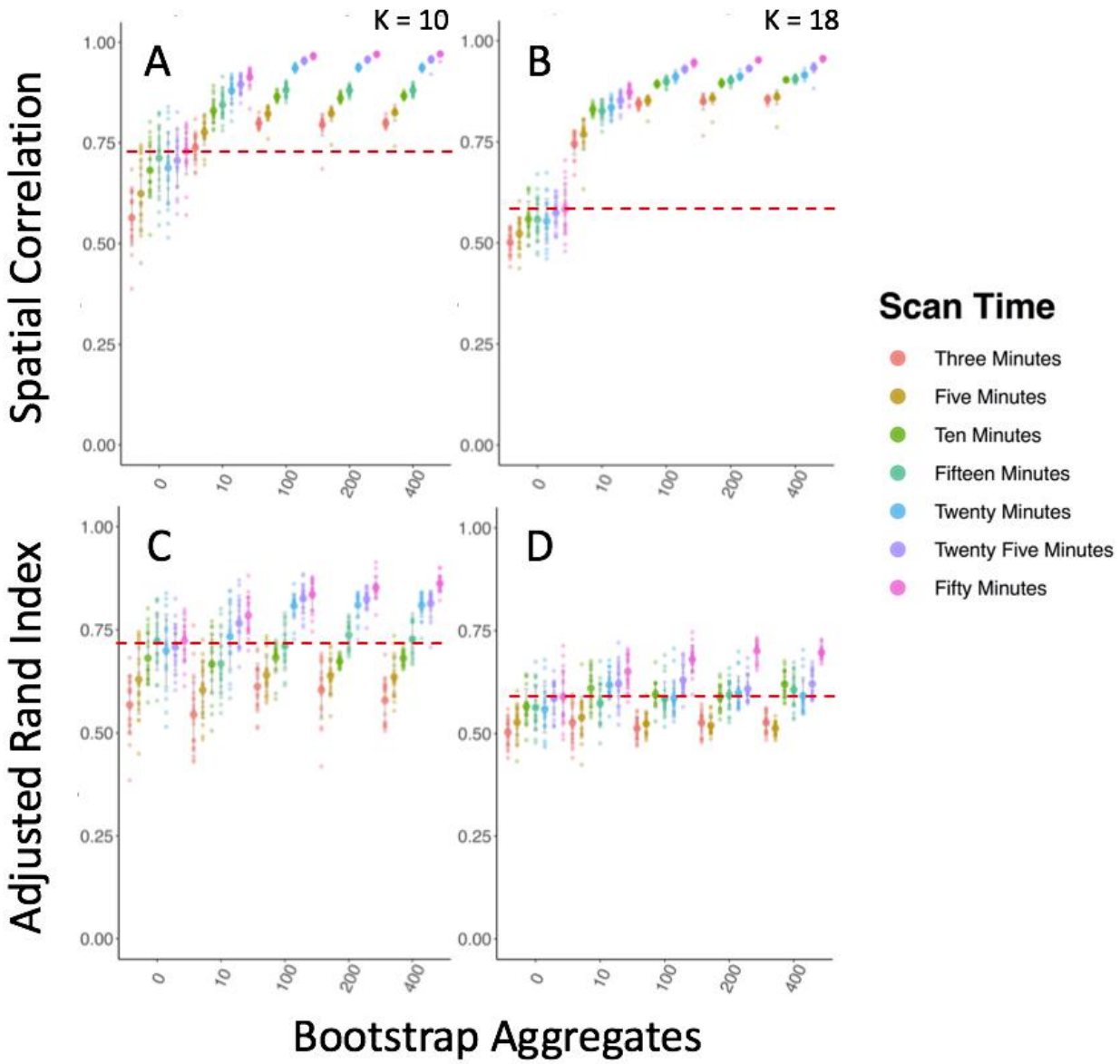


Figure 8-Bagging Improves Between Session Reproducibility of Subcortical Parcellations. We compared the HNU1 reference to the HNU1 replication subcortical parcellations across different lengths of time. The dotted red line shows the highest mean reproducibility of the standard parcellation approach. A) Spatial correlation between the group mean adjacency matrix for the reference and replication parcellations. B) Adjusted Rand Index between the cluster labels for the reference and replication parcellations.

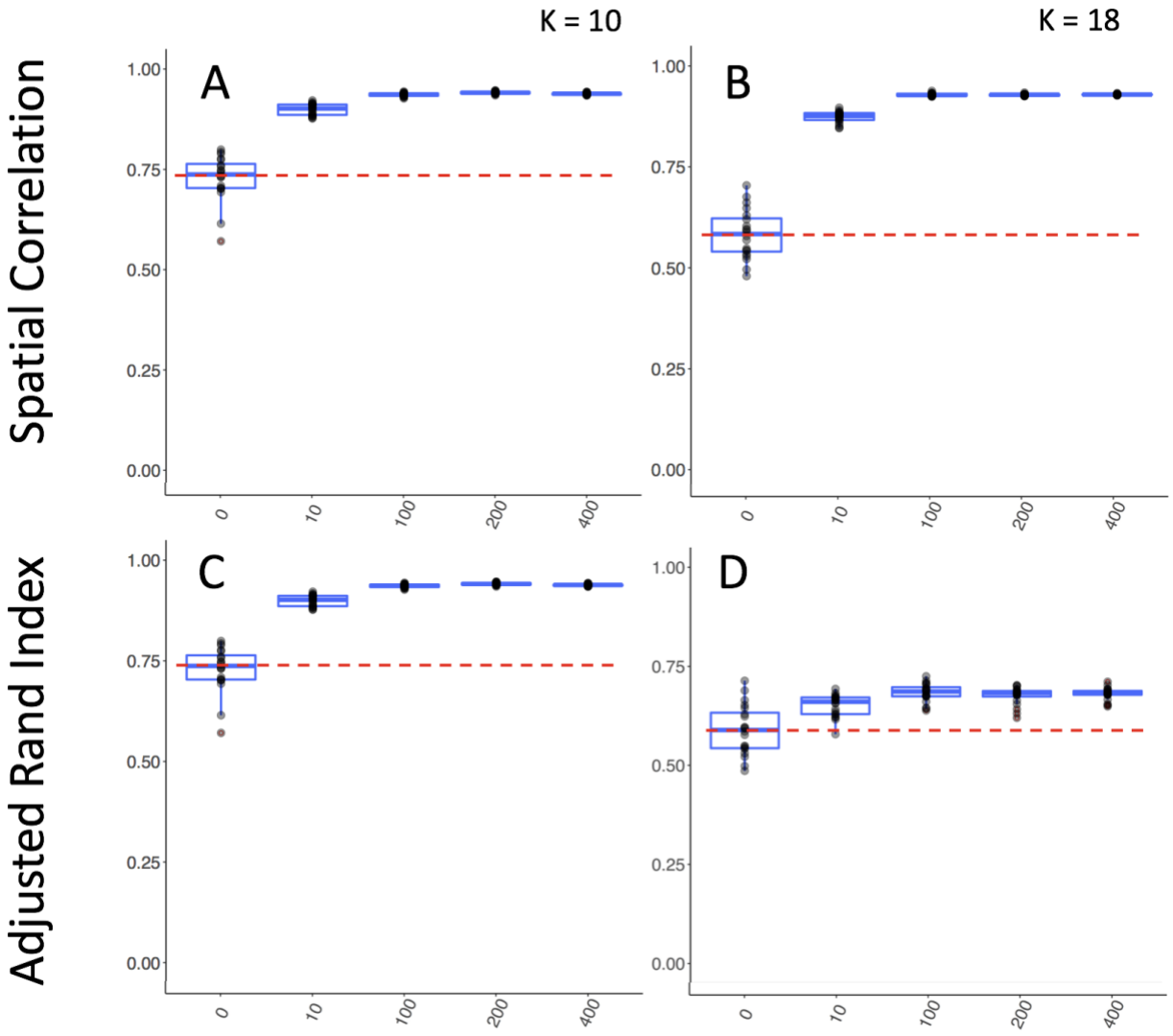


Figure 9- Group Bagging Alone Improves Between Session Reproducibility. We compare the subcortical parcellations of the 50 minute replication and reference datasets from HNU1. The dotted red line shows the highest mean reproducibility of the standard parcellation approach. A & B show the improvement in the spatial correlation of the group-level mean adjacency matrix for K = 10 and K = 18, while C & D show the improvement in the adjusted rand index of the group cluster labels for K = 10 and K = 18.

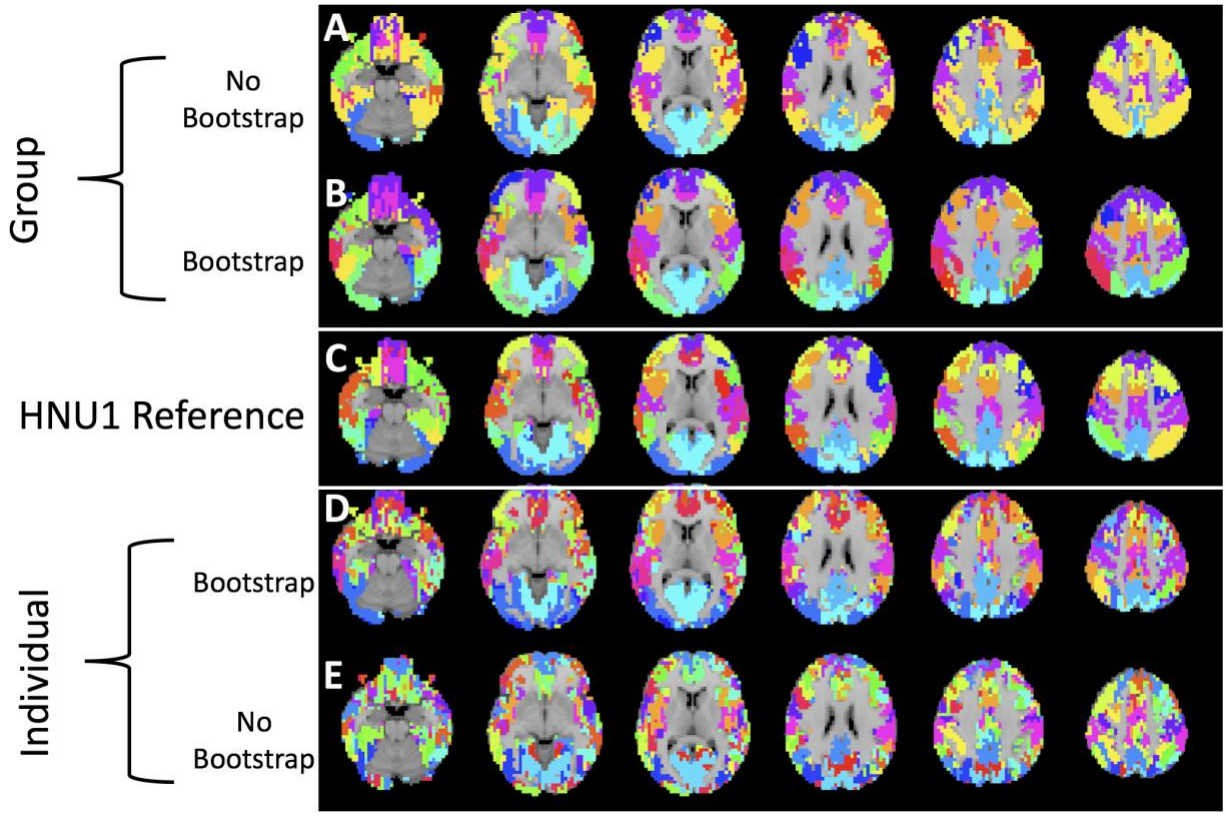


Figure 10 - Visualization of Between Session Cortical Parcellations. We show the similarity of the HNU1 reference parcellation to the subcortical HNU1 replication parcellation with 20 clusters as shown in Figure 7. Both the group and individual level bootstrapped parcellations are notably more similar to the HNU1 reference than either of the standard parcellation counterparts. We see notably more evidence for large-scale network similarities between the HNU1 reference and the bootstrapped parcellations at both the individual and group level, though the individual level is less obvious. A) Group-level parcellation with standard approach, B) and bootstrapping. C) HNU1 Reference parcellation. D) Individual-level parcellation with Bootstrapping, E) and standard approach.

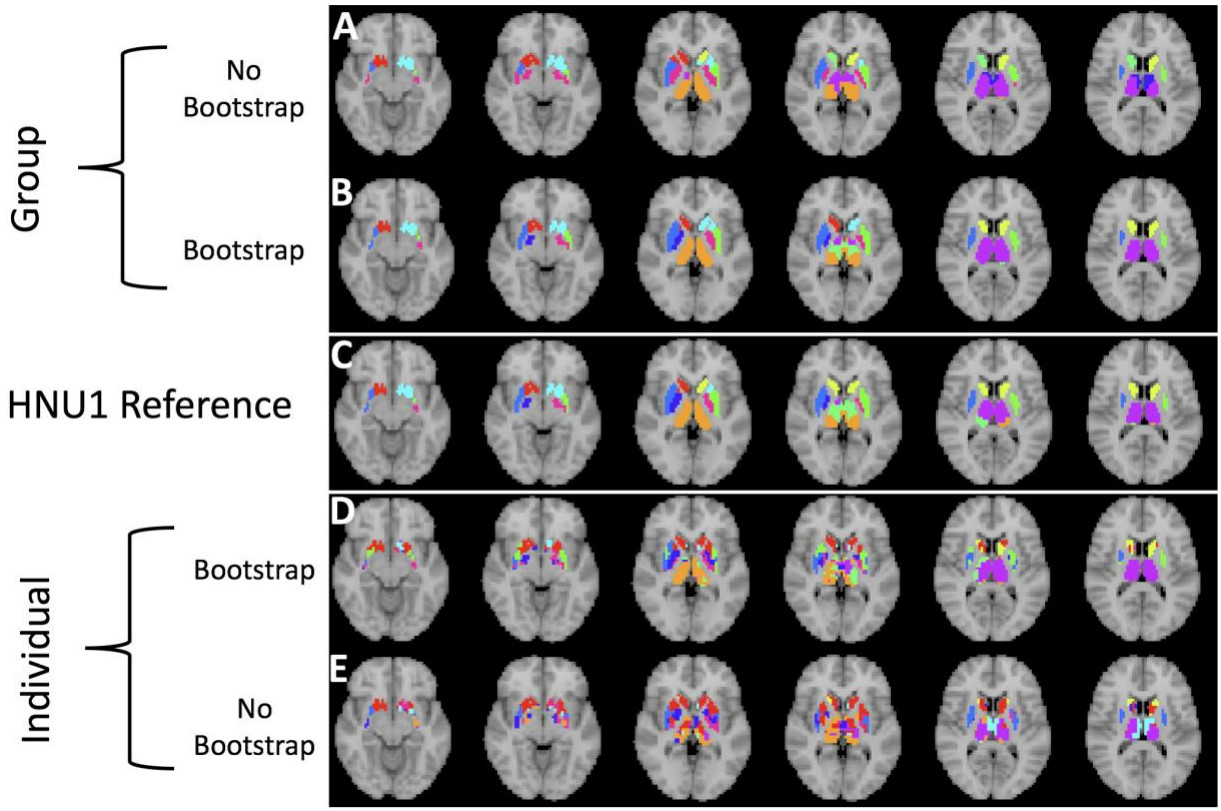


Figure 11 - Visualization of Between Session Subcortical Parcellations. We show the similarity of the HNU1 reference parcellation to the subcortical HNU1 replication parcellation with 10 clusters, as shown in Figure 8. Both the group and individual level bootstrapped parcellations are notably more similar to the HNU1 reference than either of the standard parcellation counterparts. Group-level parcellation with standard approach (A), and bootstrapping (B). HNU1 Reference parcellation (C). Individual-level parcellation with Bootstrapping (D), and standard approach (E).

2.2 Between Session and Between Scan Reproducibility

Whereas creating parcellations with high between sample reproducibility is key for generalizing scientific discovery, creating parcellations with higher between session and between scan reproducibility is key for decreasing uncertainty in measuring change over time. We tested between scan reproducibility using the subcortical parcellations of the GSP dataset. We tested between session reproducibility using both subcortical and cortical parcellations of the HNU1 dataset.

First we found that our subcortical parcellations of the GSP dataset, where we compared the effect of bagging on between scan reproducibility of 10 separate groups of 30 participants, had higher between scan reproducibility with bagging with as little as 6 minutes of data (Figure 6). This suggests that long scan times are not required for bagging to have a significantly beneficial impact on a parcellation's between scan reproducibility of the group mean adjacency matrix and cluster labels (K = 10; Correlation: Kruskal-Wallis chi-squared = 14.286, df = 1, p-value < 0.0005; ARI: Kruskal-Wallis chi-squared = 10.566, df = 1, p-value < 0.005).

We also found that our cortical parcellations had higher between session reproducibility with bagging than without in the HNU1 dataset (Figure 7; Figure 10). Comparing the K = 444 parcellation, we see significant improvements in both between session reproducibility of group mean adjacency matrices (t = -57.86, df = 9.2157, p-value = < 0.0001), and group cluster labels (t = -16.353, df = 9.1687, p-value < 0.0001). If parcellations have higher between session reproducibility on average, but have high variance across repeated parcellations we might not have a good grasp on the extent to which a particular parcellation will match across time. Therefore, it's also important for the variance in the between session reproducibility of our parcellations to decrease. We found that bagging also significantly decreases the variance in between session reproducibility of the group mean adjacency matrix (F = 83.456, num df = 9, denom df = 9, p-value < 0.0001), suggesting that the bagged parcellations are not only more reproducible, but that this reproducibility is more stable as well.

Finally, we tested the effect of bagging on the between session reproducibility of subcortical parcellations in the HNU1 dataset (Figure 8; Figure 11). We found that scan length had a highly significant impact on the between session reproducibility of parcellations. For instance, comparing 5 and 50 minute group mean adjacency matrix correlation and cluster label ARI with 0 bootstraps (K = 10; Correlation: Kruskal-Wallis chi-squared = 14.137, df = 1, p-value < 0.0005; ARI: Kruskal-Wallis chi-squared = 12.176, df = 1, p-value < 0.0005), and with 400

bootstraps (K = 10; Correlation: Kruskal-Wallis chi-squared = 29.268, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 29.271, df = 1, p-value < 0.0001). Notably, the impact of time was enhanced in the bagging condition, suggesting bagging and scan time can work synergistically to provide parcellations with high between session reproducibility. We also found that while between-session reproducibility was improved with bagging, these reproducibility estimates were not improved as was the case with longer datasets (K = 10; Correlation: F = 0.5374, num df = 9, denom df = 9, p-value > 0.05; ARI: F = 0.22301, num df = 9, denom df = 9, p-value < 0.05). We also visualized the voxelwise differences in parcellation labels on average over all HNU bootstrapped conditions and found most differences could be found in the boundary regions between parcels, along a dorsal-anterior ventral-posterior gradient (See Supplementary Figure 2), supporting the existence of small scale gradients in thalamic functional organization (Lambert, Simon, Colman, & Barrick, 2017).

We also see that the improvement in between session reproducibility from bagging is significantly greater than increasing scan length. For example, for the 0 bootstrap condition, increasing the scan time from 20 minutes to 50 minutes does not yield a significant improvement in between session reproducibility of group mean adjacency matrix correlation or ARI (K=10; Correlation: Kruskal-Wallis chi-squared = 2.9041, df = 1, p-value >0.05; ARI: Kruskal-Wallis chi-squared = 0.79683, df = 1, p-value > 0.05), but increasing the bagging from 0 to 400 has a significant effect on the between session reproducibility of the 20 minute data (K=10; Correlation: Kruskal-Wallis chi-squared = 29.268, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 17.579, df = 1, p-value < 0.0001). This demonstrates that bagging can improve reproducibility even more than doubling scan time, replicating our previous findings with the between-sample reproducibility in the GSP dataset. We also found that bagging had a significant impact on reducing the variance in between session reproducibility across multiple parcellations, replicating our results with the cortical parcellation (20 Minutes; K = 10; Correlation: F = 311.14, num df =

19, denom df = 19, p-value < 0.0001; ARI: F = 9.7003, num df = 19, denom df = 19, p-value < 0.0001). While the cortical parcellation demonstrated that only 10 minutes of data was needed for bagging to improve between session reproducibility in the HNU1 data, with the subcortical parcellation we saw that bagging did not improve the reproducibility for less than 20 minutes of data. This may be due to random elements of sampling those first 3, 5, 10, or 15 minute portions of the HNU1 dataset, or differences in the parcellation favored by lower amounts of data compared to higher amounts, as we saw in comparing the HNU1 50 minute parcellation to the GSP 12 minute reference parcellation in section 2.1.

Importantly, we also find that even bagging at the group level alone is enough to improve the between session reproducibility of the functional parcellations (Figure 9). This shows that individual level parcellations that are not bagging enhanced can be bagged to improve the group-level representation. This finding is important for other multi-level cluster estimation problems, such as diffusion weighted parcellation, which may benefit from the addition of group-level bagging (Glasser et al., 2016; Klein et al., 2007; Mars et al., 2011). Finally, we see that bagging at the individual level alone also improves individual between session reproducibility. This is key for many approaches aiming to improve individual-level parcellation (Gordon et al., 2017; Laumann et al., 2015; Xu et al., 2016).

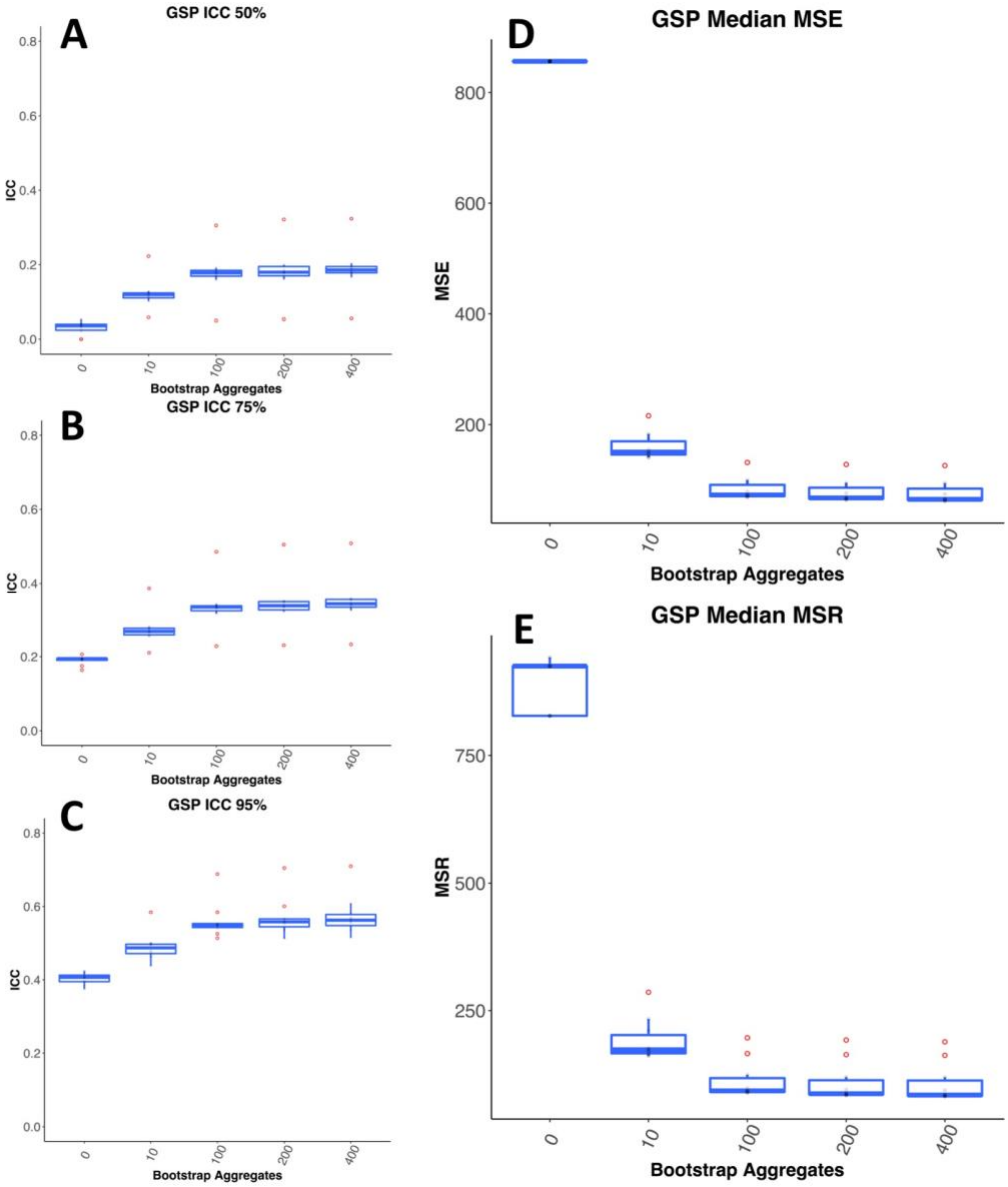


Figure 12 - Bagging Improves Univariate Subcortical ICC. A,B,C show the top 50%, 75%, and 95% ICC values for the individual-level mean adjacency matrices in the 10 GSP groups. D and E show the mean squared error within individuals and between individuals respectively. We see that bagging improves ICC. It also has a dramatic impact on both within subject variance (MSE) and between subject variance (MSR), with samples with the lowest drop in MSR showing the greatest pickup in ICC.

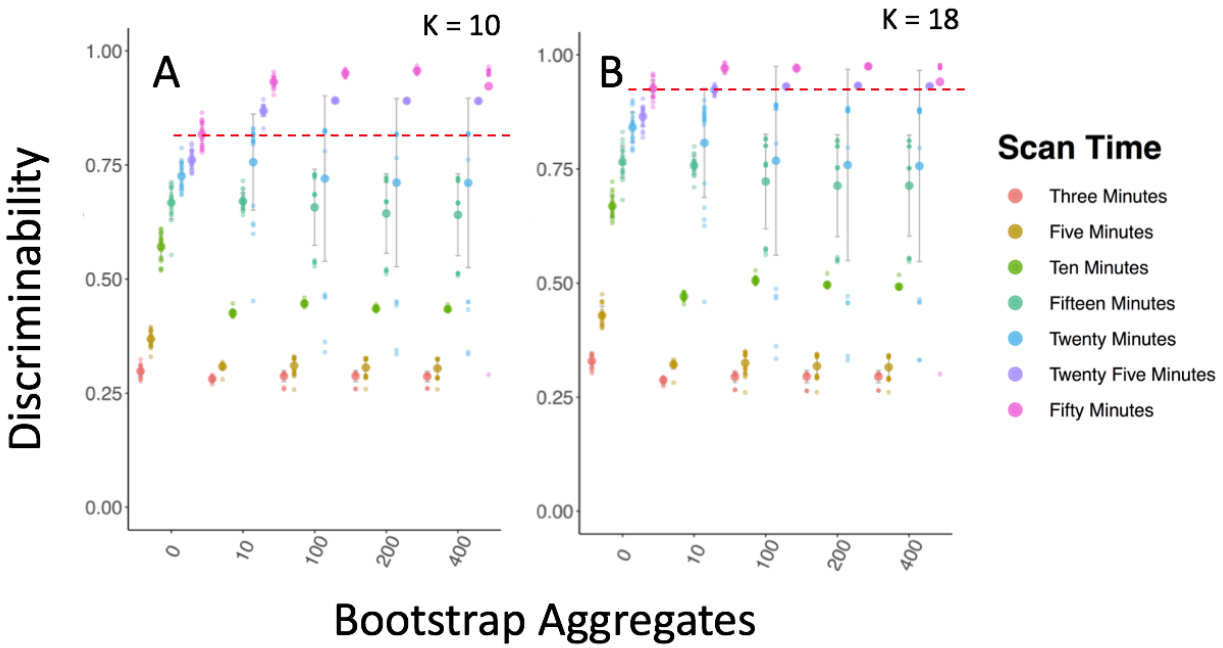


Figure 13-Bagging Improves Multivariate Reliability in Subcortical Parcellation. Figure A and B show the discriminability results for the subcortical parcellation for K = 10 and K = 18 respectively. The dashed red line shows the top mean performance for the non-bagged parcellations. Bagging leads to improved discriminability for longer scan times.

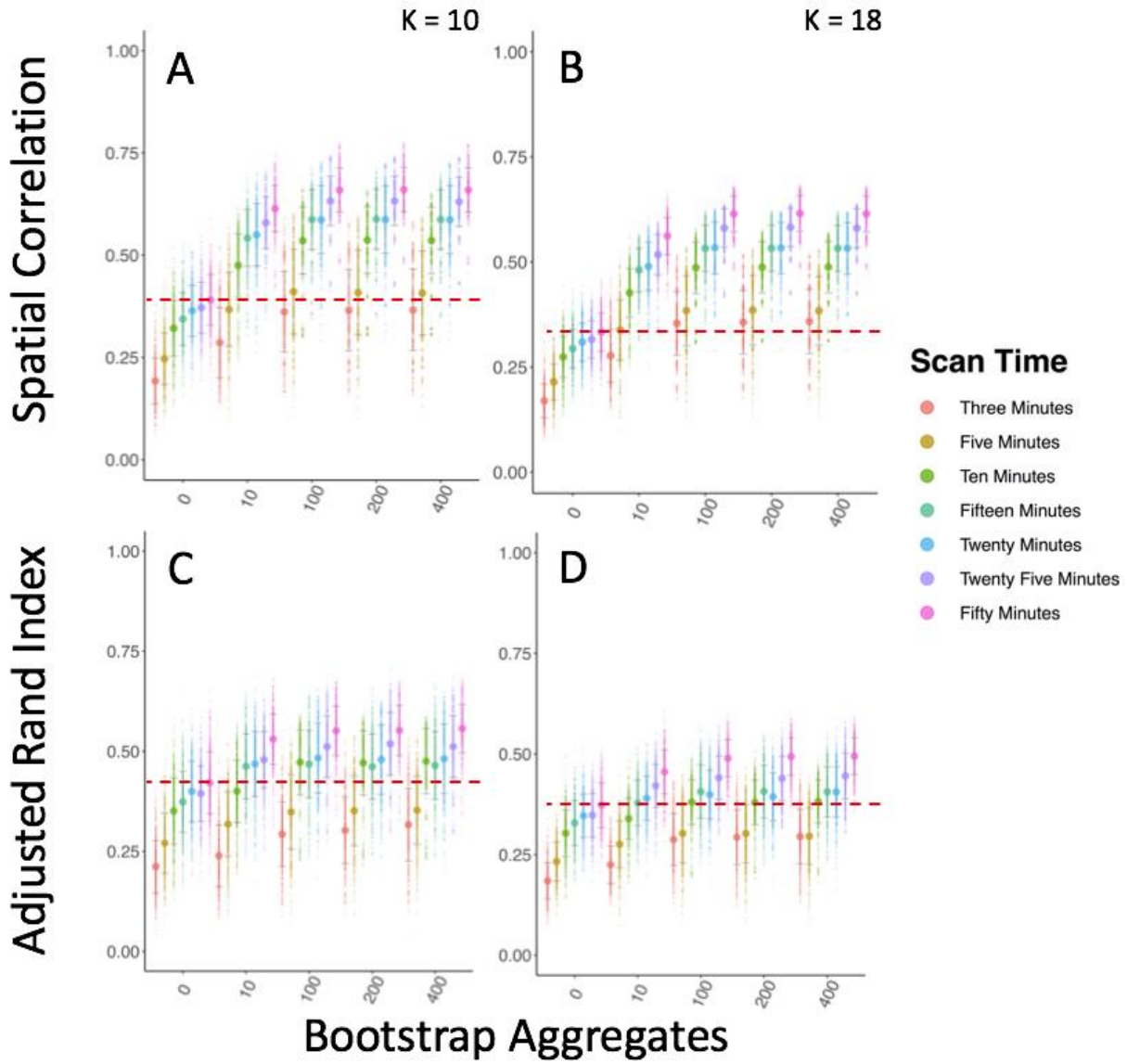


Figure 14 - Bagging Increases Group-Individual Parcellation Similarity. Overall this figure shows how bagging improves coherence between individual and group parcellations. The dotted red line shows the highest mean coherence of the standard parcellation approach. A and B show the improvement in correlation between the group-mean adjacency matrix and each individual's mean adjacency matrix for the K = 10 and K = 18 subcortical parcellation respectively. C & D show the improvement of the adjusted Rand Index between the clusters labels of the group and individual level parcellations.

2.3 Reliability

We found that bagging led to improvements in univariate estimates of reliability calculated with ICC of the voxel-wise mean adjacency matrices (Figure 12 A,B,C). Each of the entries in these matrices is the number of times across individual-level bootstraps that two voxels are put into the same cluster. As such, the individual mean adjacency matrix is a representation of the average clustering solution across bootstrap aggregations for a given individual. ICC is calculated based on the relative size of variance between subjects compared to variance within subjects. We see that bagging has a large impact on both between subject variance (MSR), and within subject variance (MSE). Notably, the decrease in MSE is slightly higher than the decrease in MSR, leading to a rise in ICC (Figure 12 D,E). In this case, ICC does not capture that bagging leads to a ten-fold decrease in variance, both within and between subjects' individual-level parcellation estimates.

Motion can be a source of highly reproducible signal in MRI data and can highly influence estimates of reliability (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012; Yan et al., 2013). We wanted to make sure the impact of bagging wasn't being driven by motion, and so we regressed out mean framewise displacement from each individual mean adjacency matrix and calculated ICC, MSE, and MSR. As was the case without motion regression, here we see that bagging improves the ICC, MSE, and MSR of parcellations (Supplementary Figure 4).

The individual voxel-voxel ICCs were moderate, reflecting that each of the voxel-voxel relationships that make up the individual parcellations are not highly statistically reliable; however, it is still possible that when taken as a whole, the parcellation patterns in these adjacency matrices would be reliably distinct between individuals. We tested this using the multivariate metric of reliability called discriminability (Figure 13). This method allows us to consider the extent to which individual parcellation patterns are unique to that individual compared to the rest of the group. We

found that discriminability of the parcellations for 0 bootstraps was heavily impacted by the amount of data acquired, and was quite high for 25 and 50 minutes of data (25 minute mean = 0.761; 50 minute mean = 0.817). We found that discriminability was significantly increased through bagging but not for all lengths of data. It seems that the individual-level signals were not strong enough in 3-15 minutes of data to improve the discriminability of parcellation through bagging in the HNU1 data (all $p > 0.05$); however, we did see significant improvements from 0 to 400 bootstraps for 20 minutes (Kruskal-Wallis chi-squared = 7.1762, df = 1, p-value < 0.01), 25 minutes (Kruskal-Wallis chi-squared = 29.44, df = 1, p-value < 0.0001); and 50 minutes (Kruskal-Wallis chi-squared = 24.584, df = 1, p-value < 0.0001), showing individual parcellation became even more highly discriminable for 20 (mean = 0.71), 25 (mean = 0.89), and 50 (mean = 0.92) minutes.

Finally, when making generalizations from group-level information to the individual, it's critically important that the group-level information is representative of the individual (Adolf & Fried, 2019; Fisher, Medaglia, & Jeronimus, 2018). Such group-to-individual generalizability is critical for deploying any group-defined models on individual level data, and as such is central for any kind of biomarker discovery process. Given that recent work has demonstrated that individual-level brain networks can differ substantially from group level estimates (Gordon, Laumann, Adeyemo, & Petersen, 2015; Laumann et al., 2015), we wanted to test whether bagging could improve the group-to-individual generalizability of parcellation, improving their potential for scientific and clinical applications (Figure 14). We found that bagging led to significantly more generalizable parcellations (Twenty minute parcellation; 0 vs 400 bootstraps; K = 10; Correlation: Kruskal-Wallis chi-squared = 800.81, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 248.43, df = 1, p-value < 0.0001), and increasing scan time did as well (K=10; Correlation: Kruskal-Wallis chi-squared = 50.985, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 16.593, df = 1, p-value < 0.0001). We also found that bagging led to an overall decrease in the

variance of the group-individual similarity (K = 10; Correlation: F = 0.56163, num df = 599, denom df = 599, p-value < 0.0001; ARI: F = 0.85283, num df = 1199, denom df = 1199, p-value < 0.01), but this decreased variance was less pronounced here than in other analyses.

3. DISCUSSION

3.1 Overview- We find that bagging improves the reproducibility of group-level functional parcellations with as little as six minutes of data, improving reproducibility more than by doubling scan time. We find that bagging may be a potentially valuable means of bringing down data requirements for reproducible study of functional brain organization. Bagging enhanced parcellations are also more reliable on the individual level, with an order of magnitude decrease in the MSE and MSR of individual-level parcellations. With twenty or more minutes we see increases in already highly discriminable individual-level parcellations, meaning that the highly detailed, individual-specific attributes of parcellations can be enhanced through bagging. We find that bagging enhanced parcellations also have higher between session and between scan reproducibility, meaning that they are better suited for within-sample repeated measures that are commonly used in intervention or other longitudinal studies in clinical and cognitive neuroscience. Finally, we found that bagging improved group-to-individual generalization of parcellations, which is key for using any common parcellation across participants. Overall, this indicates that bagging enhanced parcellations outperform the standard approach on a wide variety of reproducibility and reliability indices, and should be considered for further implementation in cutting-edge cluster and gradient-based parcellation and approaches for potential improvements in making more robust measurements of the connectome (Glasser et al., 2016; Laumann et al., 2015; Schaefer, Kong, & Yeo, 2016; Xu et al., 2016).

3.2 Implications Reaching clinical impact for cognitive and clinical neuroscience necessitates the development of robust domain knowledge that can be used by researchers in academia and industry to create the advances that will change people's lives. One of the largest impediments to the development of such robust domain knowledge is the widespread reproducibility problem in human neuroscience. It has been estimated that the United States spends about $28,000,000,000 on non-reproducible preclinical research (Freedman & Gibson, 2015), and some estimate that irreproducibility may reach as high as 90% (Begley & Ioannidis, 2015). According to a recent poll in Nature, 86% agree that there is a crisis of reproducibility in their field ("Checklists work to improve science," 2018). Therefore investigating, establishing, and disseminating methods that may improve the reproducibility of human neuroscience findings is an essential step towards moving human neuroscience closer to clinical impact.

Given that functional parcellation misalignment has been found to control most of the variance in functional connectivity weights (Bijsterbosch, Beckmann, Woolrich, Smith, & Harrison, 2019; Bijsterbosch et al., 2018), it's possible that some issues in reproducibility arise due to irreproducible mapping of functional networks from inaccurately applied parcellations. While the field has not established any 'best' technique for functional parcellation (Arslan et al., 2018), bagging does not rely on any particular clustering technique. In fact, bagging can be applied across a wide range of cluster and gradient based functional parcellation approaches. With continuing advances in functional and multi-modal parcellations, bagging can be employed in new methodological approaches to continually improve parcellation reproducibility, reliability, and group-individual coherence.

The cluster ensemble literature has shown that bagging is one of several methods that may improve clustering. Other approaches include combining cluster solutions across multiple numbers of clusters (Kuncheva & Hadjitodorov, 2004), varying random parameters of the clustering algorithm, or combining multiple clustering techniques (Hu & Yoo, 2004; Lancichinetti & Fortunato, 2012). This implies that there may be other avenues to pursue ways in which both

ensemble methods or bagging can be used to improve the generalizability of functional parcellations and other MRI approaches that use clustering (Blumensath et al., 2013; Craddock et al., 2012). For example, prior work has demonstrated the utility of using clustering for tissue class segmentation in lesion detection (Bosc et al., 2003; Juang & Wu, 2010), and these efforts may be furthered by the use of cluster ensemble methods for detecting and evaluating differences lesions more reproducibly across samples and timepoints.

The impact of bagging on reproducibility observed between completely independent samples is particularly promising, as it suggests a means of improving the generalizability of parcellation atlases across sites, scanners, and populations. The choice of a parcellation is an important step in the creation of a connectome, and it has a significant impact on the sensitivity of the resulting connectome to phenotypic differences (Abraham et al., 2017). Recent work has demonstrated that up to 60% of the variance in the edges of a connectome derive from the variance in the match of the reference parcellation to the data in question (Bijsterbosch et al., 2019); this is troublesome because when parcellations do not generalize well between samples, the resulting connectome edges will conflate differences between group network structure with differences in underlying parcellation fit (Zuo & Xing, 2014). In this way, enhancing the reproducibility of the parcellation fit between samples may also increase sensitivity of network edges to phenotypic differences between groups and individuals. This is an important step for creating the robust brain-behavior associations required for efforts aimed at using brain networks as biomarkers for various phenotypic variables.

It may be possible that with improved reliability of parcellations we may be able to also elucidate brain-behavior relationships with more robustness than previously possible. The strength of the association between two variables is limited to the reliability of the variables in question (Zuo & Xing, 2014). With improvements in between session reproducibility of measurement we also improve the ability to clinically assess changes over time that may occur in a clinical trial or behavioral intervention, or in the case of a longitudinal development study,

such as the ongoing ABCD study (Volkow et al., 2018). As measurement error decreases over time, it becomes more feasible for estimating the change over time that occurs as the result of a process of interest rather than uncontrolled variance.

An important recurrent finding was that bagging decreased the variance in the reproducibility estimates. While the importance of improving reproducibility of parcellations are key for improving the fit of a parcellation across samples or over time, if the improvements are highly variable then it's unclear when a procedure such as bagging is helping or hurting for a given case. Seeing that bagging significantly reduced this variance is not only indicative that it is more likely to produce reproducible parcellations, but also that each time it is applied to a given dataset these parcellations are likely to be more in line with one another. We show that cluster ensembles have a better grasp on the central tendency of the data for each participant and the group as a whole, and by demonstrating that bagging improves the consistency of parcellations we show that bagging also helps prevent small perturbations in the data from having outsized effects on the parcellation results. Given that an individual system may be unique in ways that share little relationship to group models (Medaglia 2017), and extensive individual-variation has been demonstrated in functional networks (Gordon 2015; Xu 2016), it's also critical that if group models are used, they are as representative of each individual in the group as possible. In other words, group-individual representativeness is critical to make accurate assessments of individual-level functional connectivity. Otherwise, differences between individuals in edge-wise connectivity will be the result of the parcellation fit, rather than actually the functional connection of interest (Bijsterbosch 2018; Bijsterbosch 2019). We saw bagging had a significant impact in not only reducing variance across repeated group level parcellations, but also in reducing variance across individuals, and between the individuals and the group level maps, without decreasing individualization of the parcellations. These results offer compelling evidence that bagging provides parcellations that better generalize from group to individual, and may give more reproducible individual level estimates of the underlying connectome as well.

3.3 Limitations- It is also worth noting that the functional parcellations created in the current work are not intended to be new reference atlases that are more reproducible than all others. Instead, we intend to demonstrate that bagging can help create functional parcellations generalize better between multiple sessions and samples. Most likely, the application of such cluster ensemble methods on multi-modal structural and functional data would lead to even greater improvements on the test-retest reliability of the parcellation, such as those applied in recent advanced efforts (Glasser et al., 2016; Xu et al., 2016). We also did not examine the impact of bagging on the reliability and reproducibility of the resulting edges of the parcellation, though this is important work that should be considered in the future. Another limitation of the current study is that we did not exhaustively test the space of cluster methods, even though PyBASC inherently is capable of such extensions (See Supplementary Methods). Bagging and cluster ensembles are notable for being agnostic to clustering technique and have been successfully demonstrated in a wide range of clustering techniques however, and we feel this limitation is notable but not a significant one.

3.4 Future Directions- The current effort aimed at establishing the role of bagging in improving reproducibility and reliability of functional parcellations, but the number of tests were far from exhaustive. There are more ways that bagging and cluster ensembles could be leveraged to improve measurements of the brain. For instance, varying and aggregating over a range of time series window lengths, cluster techniques, or resampling methods may lead to better individual-level and group-level estimates. Some research has demonstrated that selecting diverse sets of cluster solutions for aggregation can outperform simply aggregating all cluster solutions (Hadjitodorov, Kuncheva, & Todorova, 2006). Such a diversity selection approach could also be employed in a similar context and may likely improve reproducibility and reliability of the cluster solutions as well.

4. Materials and Methods

4.1 Overview

We aimed to approach a full-scale assessment of the impact of bagging on between sample reproducibility, between session reproducibility, and between session reliability. Towards this end, we apply PyBASC, a multi-level bagging approach for functional parcellation (See Methods Section 4.4 on PyBASC).

Data Source	Sample Size	Data Length	Age	Sex	Voxel Size	TR
HNU1	30	10x 10-Minute	24.37, +/- 2.37	50/50 % F/M	3.4, 3.4, 3.4 mm	2000ms
GSP	300	2x 6-Minute	23.67, +/- 2.15	50/50 % F/M	3.0, 3.0, 3.0 mm	3000ms

Table 1: Overview of datasets used in the current study. HNU1- Hanzhou Normal University. GSP- Genomics Superstruct Project.

4.2 HNU1 and GSP Data

We assessed the impact of bagging on reproducibility of functional parcellation using the HNU1 (Zuo et al., 2014), and GSP (Holmes et al., 2015). In the HNU1 dataset, 10 minute resting state scans were acquired in 30 young adults every three days for a month, for a total of 10 sessions

per person. To test between session reproducibility and reliability we split the even and odd numbered sessions into fifty minute reference and replication datasets. From the 300 participants in the GSP, we created 10 groups of 30 age and gender matched participants to assess within study/between sample reproducibility. Using all 300 GSP participants combined, we also created a reference parcellation and compared it to the HNU1 replication dataset to assess the impact of bagging on the between study/between sample reproducibility.

4.3 MRI Data

4.3.1 MRI Acquisition- *HNU1. Anatomical*: 3D SPGR images were acquired on a GE 3T scanner (General Electric, Milwaukee, WI) with an 8 Channel head coil. Flip angle: 8 Degrees. TI: 450ms, TE: 60ms, TR: 8.06s, 180 slices, slice thickness 1mm. Acquisition time: 5:01. *Rest*: T2* BOLD EPI sequence, GE 3T, 8 Channel head coil. Flip angle 90 Degree. TE 30ms, TR 2000ms, 43 slices, slice thickness: 3.4 mm. Acquisition time: 10:00.

GSP Anatomical: T1 MEMPRAGE images were acquired on Siemens 3T Magnetom Tim Trio scanner (Siemens Medical Solutions, Erlangen, Germany) with a 12 channel headcoil. Flip angle: 7 degrees. TI 1.1s, TE:1.5 / 3.4 / 5.2 / 7.0. TR 2.2s. 144 slices, slice thickness 1.2mm. Acquisition time: 2:12. *Rest*: T2* BOLD epfid2d1_72 sequence, Siemens 3T Magnetom Tim Trio, 12 channel headcoil. Flip angle 85 degrees. TE 30ms, TR 3.0s, 47 slices, 3mm slice thickness. Acquisition time: 6:12.

4.3.2 Preprocessing- *Anatomical:* 1) AFNI skull stripping. 2) MNI 152 2 mm template anatomical nonlinear registration with FNIRT. 3) Data transformed to 3mm. *Functional:* 1) Transformed to 3 mm. 2) Nuisance regression applied (white matter, CSF, global signal, motion, linear, and quadratic components of motion). 3) Friston 24-Parameter model used for volume realignment. 4) Nuisance band pass filtering between 0.01 and 0.1 Hz. 5) Spatial smoothing applied 6mm

FWHM. Preprocessing in GSP and HNU1 structural and rest data were identical with the addition of de-spiking to the GSP resting state data. (HNU1- CPAC 1.0; GSP- CPAC Version 1.2).

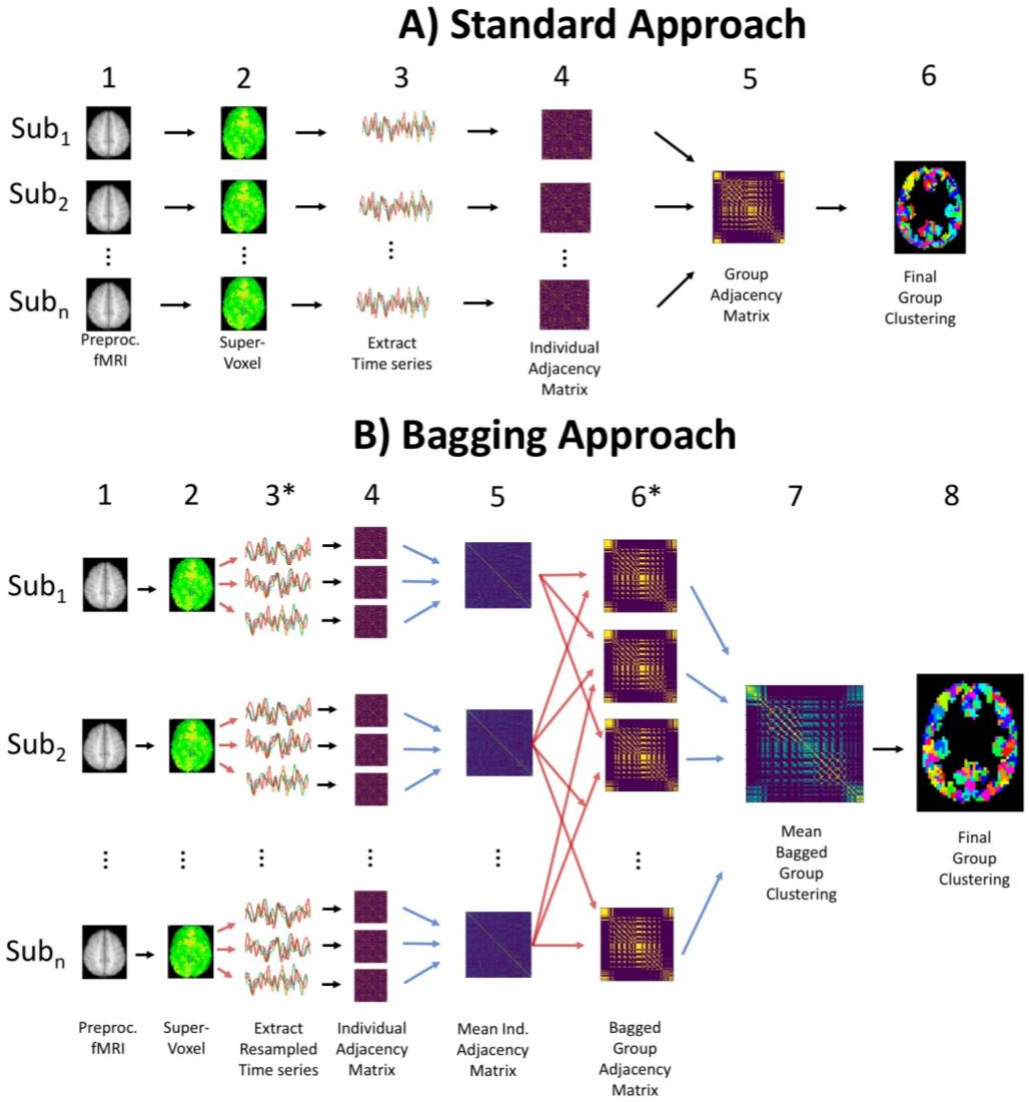


Figure 15 Parcellation Overview- A) Outline of different steps in standard clustering approach. B) Creating cluster ensembles through bagging follows the same overall structure as A, with the addition of individual level time-series resampling at step 3, and bagging across individuals in step 6. Red arrows show when bootstrap resampling occurs to create multiple resampled individual (3) and group-level data (6). Blue arrows show when bootstrap aggregation happens to create averaged individual (5) and group-level data (7).

4.4 PyBASC Parcellation

We first apply clustering in the same manner as is commonly applied in functional parcellation methods when not using bagging (Figure 1; Craddock et al., 2012; van den Heuvel, Mandl, & Pol, 2008). 1) We start with preprocessed functional MRI data; 2) this data is transformed from voxel representation to a supervoxel representation through feature reduction such as Ward's criterion based hierarchical agglomerative clustering 3) time series for each supervoxel is extracted, and a correlation matrix is calculated, 4) clustering is applied to extract a specific number of clusters (K) for each individual's correlation matrix and an adjacency matrix is created, and 5) individual level adjacency matrices are averaged together and 6) clustered again to reveal the group level clustering.

This approach can be extended by adding bagging to both within each individual's time series and across individuals to create resampled groups, as in the case in BASC (Figure 1B). We explain this process briefly here (See Supplementary Methods for more detail). At the individual level, bagging is applied using a circular block bootstrap (window size = sqrt(number of time steps)), which has been shown to outperform autoregressive process modeling and other methods (Adolf, Baecke, Kahle, Bernarding, & Kropf, 2011; Adolf et al., 2014; Bellec, Marrelec, & Benali, 2008). We used this method on the individual time series to create many resampled 4D functional scans for each individual, and clustering is applied to each individual and resulting in cluster solutions for each individual-level bootstrap. These cluster solutions are transformed to an adjacency matrix and are averaged together to create a mean adjacency matrix for each participant. We then bagged across individuals to create resampled groups (6), which are then averaged to create the group mean adjacency matrix (7) that is clustered to create the final parcellation (8). This multilevel bagging approach has been previously applied in the original BASC implementation in Octave and has been explained in greater detail elsewhere (Bellec et al., 2010; Garcia-Garcia et al., 2017); See supplementary materials for more details about PyBASC). Here we have created a python based implementation of this framework, PyBASC, to

conduct the following reproducibility and reliability assessments. Overall, PyBASC and BASC follow the same algorithmic structure, however there are a few differences worth noting for the application here: 1) Dimension reduction is conducted through a region growing algorithm in BASC, whereas in PyBASC Ward's criterion hierarchical clustering is applied. 2) Dimension reduction is applied at the group level in BASC, though can be at the individual or group level in PyBASC. Here we apply individual-level dimension reduction for subcortical parcellation and group-level dimension reduction for cortical parcellation to reduce computational load.

In the current work, we applied parcellation to both cortical and subcortical regions across a range of number of clusters (Subcortical K = 10, 18; Cortical K = 7, 20, 197, 444). We chose K = 10, and 18 for the subcortical parcellation given our work and other literature has repeatedly shown these parcellations fit best for the striatum and thalamus (Behrens et al., 2003; Choi et al., 2012; Garcia-Garcia et al., 2017; Johansen-Berg et al., 2005). We chose K = 7, 20, 197, and 444 following previous analysis exploring multi-scale parcellation with BASC (Bellec, 2013; Bellec et al., 2010). Also worth noting, many functional parcellations have strong spatial constraints placed on the clusters (Craddock et al., 2012; Schaefer et al., 2018), which plays a significant role in homogenizing parcel shape and smoothing parcel boundaries. In this way such constraints can be considered a spatial regularizer for the cluster optimization procedure, thus increasing the reproducibility of the clustering outcomes. To better appreciate the ability of bagging to enhance reproducibility of parcellation, we use minimal spatial constraints which are only implemented in our dimension reduction step.

Level of Generalization	Participants		MRI		Data Used	Figures
	Pop.	Sample	Scanner	Data		
Between Sample Reproducibility I	=	X	X	X	GSP 10 groups	1
Between Sample Reproducibility II	X	X	X	X	GSP 300 HNU Replication	2, 3, 4, 5
Between Scan Reproducibility	=	=	=	X	GSP 10 groups TRT	6
Between Session Reproducibility	=	=	=	X	HNU Reference & Replication	7, 8, 9,10, 11
Reliability	=	=	=	X	GSP 10 groups TRT HNU Reference & Replication	12, 13, 14

Table 2- Group-Level Analyses. Level of Generalization depicts which type of generalizability we tested with each analysis, see Methods sections 4.5-4.6. Columns show which kind of parameters would be kept constant (=) or changed (x) across analyses in the current study.

4.5 Reproducibility

4.5.1 Concept- The definitions of concepts of reproducibility and replicability are numerous and often conflicting (Barba, 2018). We choose to follow the Joint Committee for Guides in Metrology (JCGM). This group is comprised of eight international organizations specializing in measurement, including the International Standards Organization (ISO), which the COBIDAS report follows closely (Nichols et al., 2017). According to the JCGM, Reproducibility is the

similarity between the results of measurements of the same or similar measurands carried out under changed conditions of measurement (JCGM, 2008).

4.5.2 Data Requirements- The changed conditions may include the method of measurement, observer, measuring instrument, reference standard, location, conditions of use, and time. Notably, statements of reproducibility must include a description of the specific conditions changed. In other words, when discussing reproducibility, it's important to be specific about what is being changed and across which dimension(s) measurements are being reproduced. Reproducibility across sessions but not samples requires multiple assessments of the same sample, and is known as between session reproducibility. Similarly, reproducibility across investigators requires different investigators but can use the same method and data, which is often referred to as computational reproducibility (Millman et al., 2018). We extend this formal definition of reproducibility to include reproducibility across samples, which we call between sample reproducibility.

To support our tests of between sample, between session, and between scan reproducibility, we created several reference and replication datasets using the HNU1 and GSP datasets. In the HNU1 dataset, we created a test-retest dataset by combining the ten, 10-minute sessions into two five-session, 50-minute scans. Our reference and replication datasets were created through a concatenation of each participant's even and odd numbered session's respectively. To assess the interaction of bagging and length of acquired data on reproducibility, we performed parcellation across the first 3, 5, 10, 15, 20, and 25, and 50 minutes of the replication data. We split the 300 participants of the GSP data into 10 groups of 30 participants that were age and sex matched to the HNU1 dataset, and we used these 10 groups to assess both between sample and between scan reproducibility, given these two 6-minute scans were acquired in the same session. We also concatenated both 6 minute sessions and used all 300 participants of the GSP dataset to create a large reference parcellation, and we compared this reference to the replication sets of the HNU1 data to test between sample reproducibility.

4.5.3 Testing Methods- To measure between session reproducibility, we assessed the similarity between the reference and each of the replication datasets using spatial correlation of these group-level mean adjacency matrices, and the adjusted Rand Index of the group-level cluster labels. Given that bagged clustering is non-deterministic, we assessed the variance in our parcellations by repeating our parcellations 20 times for each replication dataset across each amount of scan time and bootstrap aggregation for the subcortical parcellation. For the cortical parcellation we repeated the parcellations 10 times, with the HNU1 ten minute replication data.. Since the standard, non-bagged, clustering approach clustering is deterministic, we needed a method to assess the variance in between session reproducibility of the parcellations. To overcome this issue, we created 20 new datasets for each amount of the scan lengths by resampling the 50 minute replication dataset (See supplementary methods for details on this resampling method). We applied PyBASC without bagging across each of these 20 datasets to estimate the variance in between session reproducibility for the standard parcellation condition.

We assessed the effect of bagging on within study/between sample reproducibility of functional parcellation by comparing the parcellations between 10 age and sex-matched groups of 30 participants from the GSP dataset. Since we had 10 groups, we did not need to resample and repeat our bagged or standard parcellations as above. We used each of the two 6 minute scans for each group, as well as a concatenated 12 minute scan. We compared the similarity of group-level parcellations between the 10 groups for each scan length, and we compared the parcellations of the ten groups for both 6 minute scans. Second, we assessed the effect of bagging on between study/between sample reproducibility. We used the same 12 minute scans from all 300 GSP participants together to create a reference parcellation. For assessing cortical parcellation reproducibility, we compared the GSP reference to the 10 minute HNU1 data. For assessing subcortical parcellation reproducibility we assessed the GSP parcellation against a range of HNU1 scan lengths.

4.6 Reliability & Discriminability

4.6.1 Concept- Reliability is a metric of the inter- and intra-individual stability of a measurement across multiple occasions (Zuo & Xing, 2014), and can be measured with many different indices. Intraclass correlation (ICC) is a descriptive statistic that relates the within-participant to between subject variance to get an indication of univariate reliability (Shrout & Fleiss, 1979). ICC is most commonly considered as the intra-individual variability relative to the inter-individual variability. Discriminability is multivariate metric of reliability that takes the full set of features across all observations into account to compute a multivariate index of reliability (https://github.com/neurodata/r-mgc).

4.6.2 Data Requirements- Calculating test retest reliability requires at least two observations of the same individuals in the same group. In the current study, we compute ICC using two measurements from each individual in both the HNU1 and GSP samples.

4.6.3 Testing Methods- We assessed whether bagging would improve both the univariate and multivariate reliability of our functional parcellations, and how length of scan acquisition would have an impact on these results. Using the same set of functional parcellations across multiple datasets calculated in Section 4.5 Reproducibility, namely the reference and multi-length replication datasets, we calculated the ICC for each cell in the individual level mean adjacency matrix data, and discriminability for the individual mean adjacency matrix as a whole.

Acknowledgements: We would like to thank Lei Ai and Yu Tong for their assistance with the GSP and HNU1 datasets.

Funding: AN is supported by R21MH118556-02 and a NARSAD Young Investigator Award from the Brain and Behavior Research Foundation.

References

Abraham, A., Milham, M. P., Di Martino, A., Cameron Craddock, R., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, Vol. 147, pp. 736–745. https://doi.org/10.1016/j.neuroimage.2016.10.045

Adolf, D., Baecke, S., Kahle, W., Bernarding, J., & Kropf, S. (2011). Applying multivariate techniques to high-dimensional temporally correlated fMRI data. *Journal of Statistical Planning and Inference*, *141*(12), 3760–3770.

Adolf, D., Weston, S., Baecke, S., Luchtmann, M., Bernarding, J., & Kropf, S. (2014). Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Frontiers in Neuroinformatics*, *8*, 72.

Adolf, J. K., & Fried, E. I. (2019). Ergodicity is sufficient but not necessary for group-to-individual generalizability. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(14), 6540–6541.

Arslan, S., Ktena, S. I., Makropoulos, A., Robinson, E. C., Rueckert, D., & Parisot, S. (2018). Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage*, *170*, 5–30.

Barba, L. A. (2018). Terminologies for Reproducible Research. Retrieved from http://arxiv.org/abs/1802.03311

Barnes, K. A., Cohen, A. L., Power, J. D., Nelson, S. M., Dosenbach, Y. B. L., Miezin, F. M., … Schlaggar, B. L. (2010). Identifying Basal Ganglia divisions in individuals using resting-state functional connectivity MRI. *Frontiers in Systems Neuroscience*, *4*(June), 18.

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research*, *116*(1), 116–126.

Behrens, T. E. J., Johansen-Berg, H., Woolrich, M. W., Smith, S. M., Wheeler-Kingshott, C. A. M., Boulby, P. A., … Matthews, P. M. (2003). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience*, Vol. 6, pp. 750–757. https://doi.org/10.1038/nn1075

Bellec, P. (2013). Mining the Hierarchy of Resting-State Brain Networks: Selection of Representative Clusters in a Multiscale Structure. *2013 International Workshop on Pattern Recognition in Neuroimaging*, 54–57. ieeexplore.ieee.org.

Bellec, P., Marrelec, G., & Benali, H. (2008). A bootstrap test to investigate changes in brain connectivity for functional MRI. *Statistica Sinica*, *18*, 1253–1268.

Bellec, P., Rosa-neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). NeuroImage Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, *51*(3), 1126–1139.

Bijsterbosch, J. D., Beckmann, C. F., Woolrich, M. W., Smith, S. M., & Harrison, S. J. (2019). The relationship between spatial configuration and functional connectivity of brain regions revisited. *eLife*, *8*. https://doi.org/10.7554/eLife.44890

Bijsterbosch, J. D., Woolrich, M. W., Glasser, M. F., Robinson, E. C., Beckmann, C. F., Van Essen, D. C., … Smith, S. M. (2018). The relationship between spatial configuration and functional connectivity of brain regions. *eLife*, Vol. 7. https://doi.org/10.7554/elife.32992

Blumensath, T., Jbabdi, S., Glasser, M. F., Van Essen, D. C., Ugurbil, K., Behrens, T. E. J., & Smith, S. M. (2013). Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *NeuroImage*, *76*, 313–324.

Bosc, M., Heitz, F., Armspach, J. P., Namer, I., Gounot, D., & Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage*, *20*(2), 643–656.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C., & Yeo, B. T. T. (2011). The

organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(July 2011), 2322–2345.

Checklists work to improve science. (2018). *Nature*, *556*(7701), 273–274.

Choi, E. Y., Yeo, B. T. T., & Buckner, R. L. (2012). The organization of the human striatum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *108*(8), 2242–2263.

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*(8), 1914–1928.

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* , *19*(9), 1090–1099.

Fischer, B., & Buhmann, J. M. (2003). Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(11), 1411–1415.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6106–E6115.

Freedman, L. P., & Gibson, M. C. (2015). The impact of preclinical irreproducibility on drug development. *Clinical Pharmacology and Therapeutics*, *97*(1), 16–18.

Garcia-Garcia, M., Nikolaidis, A., Bellec, P., Craddock, R. C., Cheung, B., Castellanos, F. X., & Milham, M. P. (2017). Detecting stable individual differences in the functional organization of the human basal ganglia. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.07.029

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Gordon, E. M., Laumann, T. O., Adeyemo, B., & Petersen, S. E. (2015). Individual Variability of the System-Level Organization of the Human Brain. *Cerebral Cortex* , bhv239.

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., … Dosenbach, N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron*, *95*(4), 791–807.e7.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Hadjitodorov, S. T., Kuncheva, L. I., & Todorova, L. P. (2006). Moderate diversity for better cluster ensembles. *An International Journal on Information Fusion*, *7*(3), 264–275.

Holmes, A. J., Hollinshead, M. O., O'Keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., … Buckner, R. L. (2015). Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Scientific Data*, *2*, 150031.

Hong, Y., Kwong, S., Wang, H., & Ren, Q. (2009). Resampling-based selective clustering ensembles. *Pattern Recognition Letters*, *30*(3), 298–305.

Hoyos-Idrobo, A., Schwartz, Y., Varoquaux, G., & Thirion, B. (2015). Improving sparse recovery on structured images with bagged clustering. *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*, 73–76. IEEE.

Hu, X., & Yoo, I. (2004). Cluster Ensemble and Its Applications in Gene Expression Analysis. *Proceedings of the Second Conference on Asia-Pacific Bioinformatics - Volume 29*, 297–302. Darlinghurst, Australia, Australia: Australian Computer Society, Inc.

Janssen, R. J., Jylänki, P., Kessels, R. P. C., & van Gerven, M. A. J. (2015). Probabilistic model-based functional parcellation reveals a robust, fine-grained subdivision of the striatum. *NeuroImage*, *119*, 398–405.

JCGM. (2008). Evaluation of measurement data—Guide to the expression of uncertainty in measurement. *Int. Organ. Stand. Geneva ISBN*, *50*, 134.

Jia, J., Xiao, X., Liu, B., & Jiao, L. (2011). Bagging-based spectral clustering ensemble

selection. *Pattern Recognition Letters*, *32*(10), 1456–1467.

Johansen-Berg, H., Behrens, T. E. J., Sillery, E., Ciccarelli, O., Thompson, A. J., Smith, S. M., & Matthews, P. M. (2005). Functional–Anatomical Validation and Individual Variation of Diffusion Tractography-based Segmentation of the Human Thalamus. *Cerebral Cortex* , *15*(1), 31–39.

Juang, L.-H., & Wu, M.-N. (2010). MRI brain lesion image detection based on color-converted K-means clustering segmentation. *Measurement*, *43*(7), 941–949.

Kelly, A. M. C., Di Martino, A., Uddin, L. Q., Shehzad, Z., Gee, D. G., Reiss, P. T., … Milham, M. P. (2009). Development of Anterior Cingulate Functional Connectivity from Late Childhood to Early Adulthood. *Cerebral Cortex* , *19*(3), 640–657.

Kelly, C., Toro, R., Di Martino, A., Cox, C. L., Bellec, P., Castellanos, F. X., & Milham, M. P. (2012). A convergent functional architecture of the insula emerges across imaging modalities. *NeuroImage*, *61*(4), 1129–1142.

Klein, J. C., Behrens, T. E. J., Robson, M. D., Mackay, C. E., Higham, D. J., & Johansen-Berg, H. (2007). Connectivity-based parcellation of human cortex using diffusion MRI: Establishing reproducibility, validity and observer independence in BA 44/45 and SMA/pre-SMA. *NeuroImage*, *34*(1), 204–211.

Kuncheva, L. I., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, *2*, 1214–1219 vol.2. ieeexplore.ieee.org.

Lambert, C., Simon, H., Colman, J., & Barrick, T. R. (2017). Defining thalamic nuclei and topographic connectivity gradients in vivo. *NeuroImage*, *158*, 466–479.

Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, *2*, 336.

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., … Petersen, S. E. (2015). Functional System and Areal Organization of a Highly Sampled

Individual Human Brain. *Neuron*, *87*(3), 657–670.

Li, T., & Ding, C. (2008). Weighted Consensus Clustering. In *Proceedings*. *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 798–809). Society for Industrial and Applied Mathematics.

Margulies, D. S., Kelly, A. M. C., Uddin, L. Q., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2007). Mapping the functional connectivity of anterior cingulate cortex. *NeuroImage*, *37*(2), 579–588.

Mars, R. B., Jbabdi, S., Sallet, J., O'Reilly, J. X., Croxson, P. L., Olivier, E., … Rushworth, M. F. S. (2011). Diffusion-Weighted Imaging Tractography-Based Parcellation of the Human Parietal Cortex and Comparison with Human and Macaque Resting-State Functional Connectivity. *Journal of Neuroscience*, Vol. 31, pp. 4087–4100. https://doi.org/10.1523/jneurosci.5102-10.2011

Millman, K. J., Brett, M., Barnowski, R., & Poline, J.-B. (2018). Teaching Computational Reproducibility for Neuroimaging. *Frontiers in Neuroscience*, *12*, 727.

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., … Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, *20*(3), 299–303.

Nikolaidis, A., Heinsfeld, A. S., Xu, T., Bellec, P., Vogelstein, J., & Milham, M. (2019). Bagging Improves Reproducibility of Functional Parcellation (p. 343392). https://doi.org/10.1101/343392

Noble, S., Spann, M. N., Tokoglu, F., Shen, X., Constable, R. T., & Scheinost, D. (2017). Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. *Cerebral Cortex* , *27*(11), 5415–5429.

O'Connor, D., Potler, N. V., Kovacs, M., Xu, T., Ai, L., Pellman, J., … Milham, M. P. (2017). 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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(Oct), 2825–2830.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., … Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex* , *28*(9), 3095–3114.

Schaefer, A., Kong, R., & Yeo, B. T. T. (2016). Chapter 1 - Functional connectivity parcellation of the human brain. In G. Wu, D. Shen, & M. R. Sabuncu (Eds.), *Machine Learning and Medical Imaging* (pp. 3–29). Academic Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations-uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Strehl, A., & Ghosh, J. (2002). Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research: JMLR*, *3*(Dec), 583–617.

van den Heuvel, M., Mandl, R., & Pol, H. H. (2008). Normalized cut group clustering of resting-state fMRI data. *PloS One*, *3*(4). https://doi.org/10.1371/journal.pone.0002001

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., … Weiss, S. R. B. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7.

Xu, T., Opitz, A., Craddock, R. C., Wright, M. J., Zuo, X.-N., & Milham, M. P. (2016). Assessing Variations in Areal Organization for the Intrinsic Brain: From Fingerprints to Reliability. *Cerebral Cortex* , *26*(11), 4192–4211.

Yan, C.-G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., … Milham, M. P. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *NeuroImage*, *76*, 183–201.

Yang, Z., Zuo, X.-N., Wang, P., Li, Z., LaConte, S. M., Bandettini, P. A., & Hu, X. P. (2012). Generalized RAICAR: discover homogeneous subject (sub)groups by reproducibility of their intrinsic connectivity networks. *NeuroImage*, *63*(1), 403–414.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). *The organization of the human cerebral cortex estimated by intrinsic functional connectivity*. 1125–1165.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., … Milham, M. P. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, *1*, 140049.

Zuo, X.-N., Biswal, B. B., & Poldrack, R. A. (2019). *Reliability and Reproducibility in Functional Connectomics*. Frontiers Media SA.

Zuo, X. N., Kelly, C., Adelstein, J. S., Klein, D. F., Castellanos, F. X., & Milham, M. P. (2010). Reliable intrinsic connectivity networks: Test-retest evaluation using ICA and dual regression approach. *NeuroImage*, *49*(3), 2163–2177.

Zuo, X.-N., Kelly, C., Di Martino, A., Mennes, M., Margulies, D. S., Bangaru, S., … Milham, M. P. (2010). Growing together and growing apart: regional and sex differences in the lifespan developmental trajectories of functional homotopy. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(45), 15034–15043.

Zuo, X.-N., & Xing, X.-X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neuroscience and Biobehavioral Reviews*, *45*, 100–118.

Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research.

Nature Human Behaviour. https://doi.org/10.1038/s41562-019-0655-x

5. Supplementary Materials.

5.1 Supplementary Methods

5.1.1 Standard Clustering Resampling Datasets- Without bagging, Ward's criterion hierarchical agglomerative cluster is deterministic and has no variance over multiple runs. To offer a better understanding of the sampling variance that could lead to differences in clustering when bagging is not applied, we apply resampling to the original replication datasets when comparing them to a given reference. To create the multiple 3 and 5 minute datasets for the non bootstrapped condition, we randomly sampled continuous 3 or 5 minute blocks of the full 50 minute replication dataset. To create the 10, 15, 20, 25, and 50 minute resampled datasets, we randomly sampled continuous 5 minute blocks of the full 50 minute replication dataset. We did not calculate resampling in the GSP datasets because we compared the 10 groups against one another.

5.1.2 PyBASC Originally, this method was implemented in Octave, and we have recently implemented this method in Python to take advantage of the broad range of powerful libraries available for machine learning and neuroimaging, including NiBabel, Nilearn, Scikit Learn, and NiPype. This enables PyBASC to be highly flexible and extensible in adopting new clustering algorithms and implementations, but more importantly allows for dramatic increases in computational power through the use of NiPype's easy-to-use and adjust multi-core processing capabilities. We have also made edits to the processing that make PyBASC run very efficiently with minimal memory loads. This makes PyBASC just as easy to run on a laptop as on a large server in the cloud. Here, we briefly address the PyBASC overall methodology. For more information, we have also explained these methods in more detail elsewhere (Garcia-Garcia et

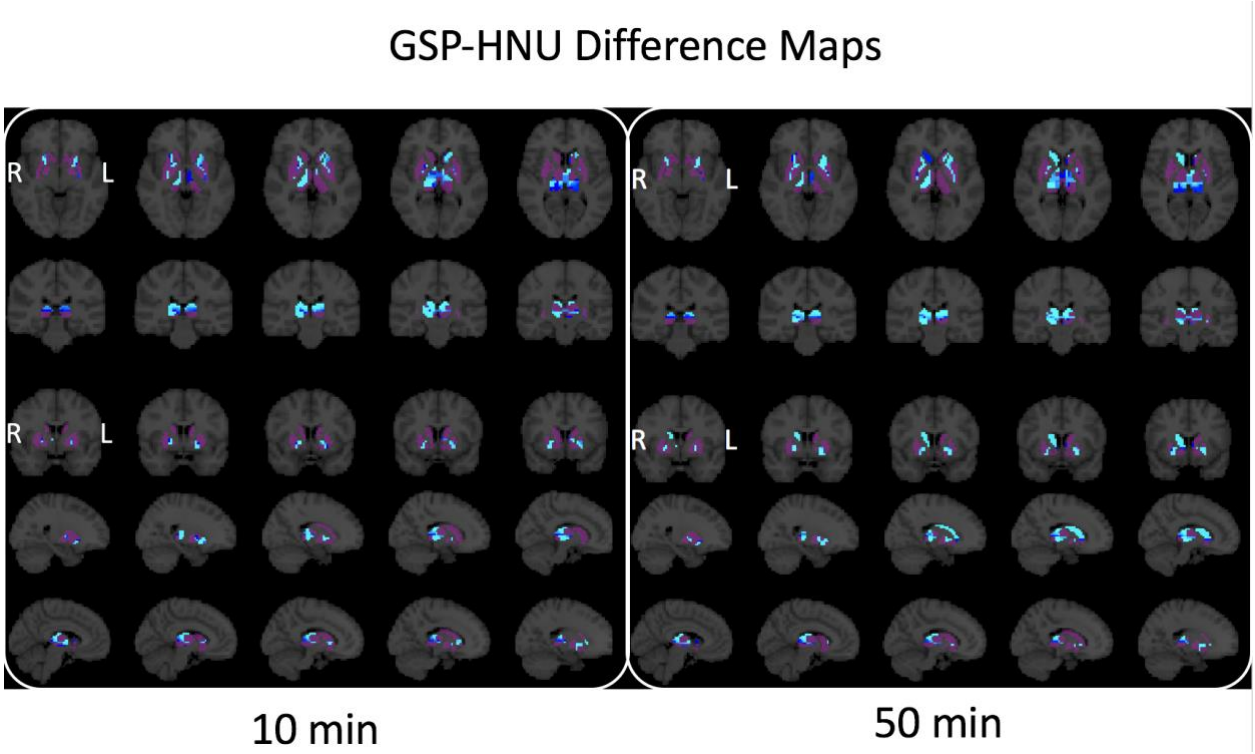
al., 2017). PyBASC has a wide range of functionalities that can be customized according to user preferences. Well- performing default values are provided where applicable, but users can choose 1) the number of initial supervoxels, 2) the number of individual and 3) group- level bootstraps, the 4) region to be parcellated and 5) the secondary region (optional) to use as the basis of cross-network clustering. Users can also define the affinity threshold and distance metric used in the clustering procedures.

5.1.3 Individual-Level- First to reduce the dimensionality of the voxel wise time series, and to allow for better computational efficiency, we apply a hierarchical feature agglomeration from Scikit Learn with a neighboring voxel spatial constraint to each individual's data (Subcortical Parcellation: 600 supervoxels; Cortical Parcellation: 1800 supervoxels). Then, a circular block bootstrap procedure is applied to create resampled time series data from the original functional image. We follow prior recommendations to set the block length to greater than 20 time steps (Adolf et al., 2014), and more specifically, equal to the square root of the number of points in a time series (Bellec et al., 2010; Bellec et al., 2008) Each of these resampled time series are then clustered using the clustering technique of choice. Currently k-means, Ward's criterion hierarchical agglomerative, spectral clustering, and Gaussian mixture models are implemented. With each application of clustering, the cluster labels are transformed into an adjacency matrix of 0s and 1s representing whether a given voxel belongs to the same cluster as another voxel. Each of these adjacency matrices are in supervoxel space, which differs between individuals. Therefore the dimension reduction procedure is reversed to create a voxel wise adjacency matrix that puts all individual-level data back into the same space. All adjacency matrices are averaged together to create an individual mean adjacency matrix (IMA), which represents a summary of all clustering assignments across all bootstraps. The IMA can be clustered to produce a final consensus cluster assignment for each individual, or it can used in group-level bagging.

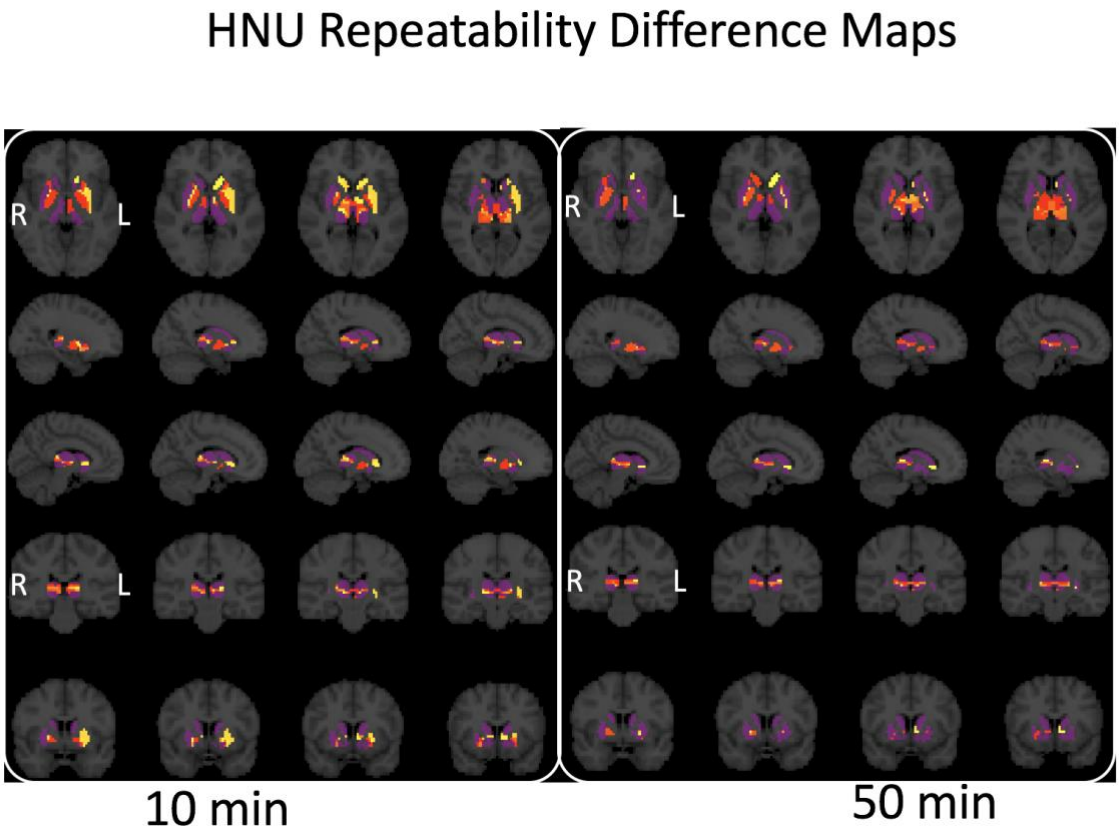
5.1.4 Group Level- PyBASC applies bagging to IMAs across individuals, recreating resampled

groups of IMAs from the original pool of participants. The IMAs for each group-level bootstrap are averaged and another group adjacency matrix is created through clustering. The adjacency matrix across all bootstraps are averaged to create a group mean adjacency matrix (GMA), and this GMA is clustered to create the final group-level clustering. The GMA is the final output of the group-level bagging, along with the cluster labels. While the IMA represents a summary of the clusterings across all bootstraps for each individual, the GMA represents a summary of the clustering across all bootstraps for the group. We compare each individual's IMA to the GMA as a detailed way of assessing similarities between the individual and group-level clusterings.

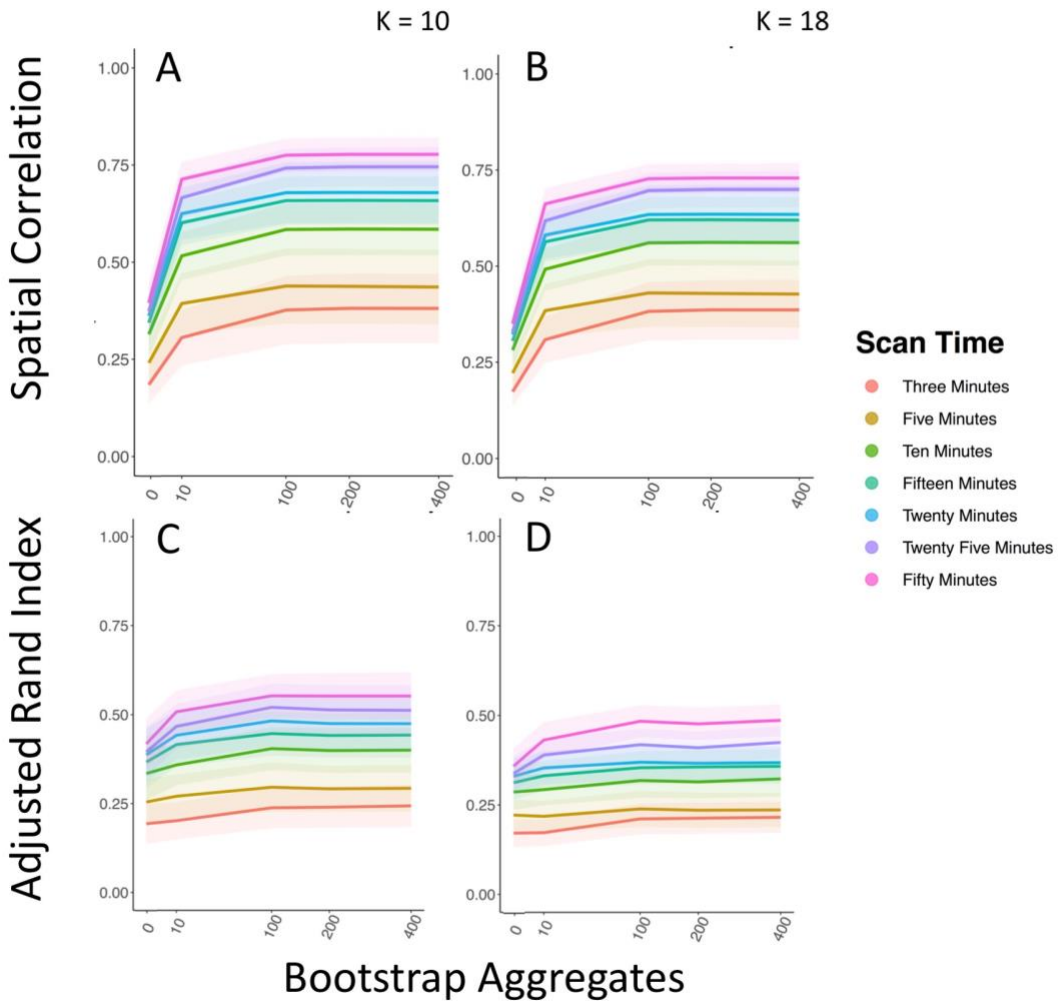
5.2 Supplementary Results



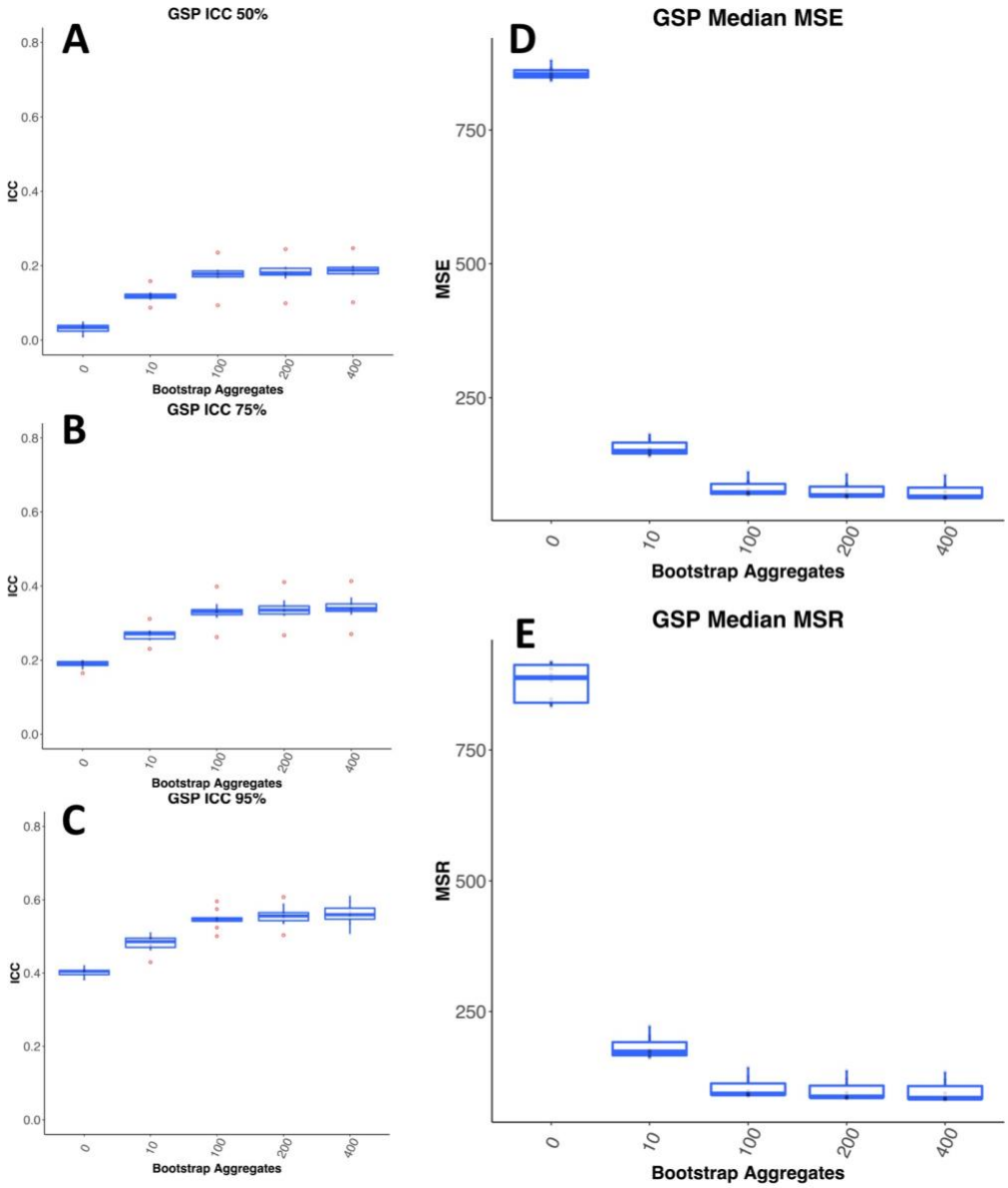
5.2.1- Here we we compare the voxelwise cluster assignment differences between the Reference GSP dataset created with 400 bootstraps and the 10 and 50 minute HNU parcellations. Bright blue corresponds to all comparisons being different between the two parcellations across all replication samples with 400 bootstraps, and dark purple shows where there are no differences between parcellations. We find that both 10 and 50 minute parcellations mostly differ from the GSP in transition zones between parcels. In particular we see these difference maps reveal a superior-inferior thalamic transition zone, and a putamen-pallidum transition zone. Notably, the HNU 50 minute parcellation differs from the GSP reference in differential classification of the caudate. Whereas the 50 minute HNU data clusters the bilateral caudate into the same cluster, both the GSP and the 10 minute HNU data cluster the caudate into separate, non-homologous clusters.



5.2.2- Here we compare differences the HNU Replication and Reference parcellations. Bright yellow corresponds to all comparisons being different between the two parcellations across all replication samples with 400 bootstraps, and dark purple shows where there are no differences between parcellations. Notably we see differentiation between reference and the 10 minute parcellation is higher than the 50 minute replication set. Furthermore, we see a reprise of the transition zones seen in S1, notably a clear superior-inferior transition, and a putamen-pallidum transition as well.



5.2.3- Individual-level bagging improves between session reproducibility of individual-level parcellations across a range of bagging levels and range of scan times.. We compare the individual mean stability matrix for each participant (A, B) and the individualized cluster labels (C, D).



5.2.4- Univariate Reliability in GSP with motion regressed out. Motion can be a source of highly reproducible signal in MRI data. We wanted to make sure the impact of bagging wasn't being driven by motion, and so we regressed out motion from each individual mean adjacency matrix before calculating ICC. As was the case without motion regression, here we see that bagging improves the ICC of parcellations. A,B,C show the top 50%, 75%, and 95% ICC values for the individual-level mean adjacency matrices. Each point corresponds to the 50th, 75th, and 95th percentile ICC values for the individual-level mean adjacency matrices for each of the 10 GSP groups. D and E show the mean squared error within individuals and between individuals respectively. We see that bagging improves ICC, it has a dramatic impact on both MSE and MSR, with samples with the lowest drop in MSR showing the greatest pickup in ICC.